APPLICATION OF BOOTSTRAP METHOD ON ESTIMATION OF THE ERROR RATES IN DISCRIMINANT ANALYSIS

I WAYAN MANGKU

Department of Mathematics,
Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University
Jl. Raya Pajajaran, Kampus IPB Baranang Siang, Bogor, Indonesia

Abstract. This paper is a survey study on applications of bootstrap methods for estimating the probability of misclassifications in two-groups discriminant analysis. Here we use the linear discriminant function as classification rule. Some comparative studies on the performances of the considered estimators are also discussed.

Key words: Discriminant analysis, classification rule, probability of misclassification, actual error rate, empirical estimator, bootstrap estimator.

1. Introduction

This paper is a survey study on applications of bootstrap methods for estimating the probability of misclassifications in two-groups discriminant analysis when the Linear Discriminant Function (LDF) is used as the classification rule.

1.1. Formulation of the problems.

A typical problem in two-groups discriminant analysis is as follows. Given the existence of two groups of individuals, we want to find a classification rule for allocating new individuals (observations) into one of the existing two groups. Corresponding to each classification rule, there is a probability of misclassification if we use that classification rule to classify new individuals (observations) into one of the two groups. The best classification rule is the one that leads to the smallest probability of misclassification, which also called error rates.

In this paper we consider three types of error rates, namely: (i) the *optimum error rate,* which describes the performance of a classification rule based on known parameters, (ii) the *conditional error rate,* which describes the performance of a classification rule based on parameters estimated by the statistics computed from the training samples, and (iii) the *expected error rate,* which describes the expected performance

of a classification rule based on parameters estimated by a randomly chosen training sample.

In real situation, the parameters are rarely known, and the expected (or unconditional) error rates depend heavily on the distribution of the discriminant function, which is very complicated (see for example, Wald (1944), Anderson (1951), Okamoto (1963) and Hills (1966)). Consequently most works associated with error rates have assumed that the samples, which are used to construct the estimated classification rule, are fixed. This leads to the exploration of the *conditional error rate.* Here the word 'conditional' refers to the conditioning of the training samples from which the classification rule is constructed. We may also think of this as the probability that the given classification rule would incorrectly classify a future observation. It should also be noted that the conditional error rate is the error rate that is important to an experimenter who has already determined the classification rule. This conditional error rate is also referred to the *actual error rate* or the *true error rate* by many authors. Hence, in this paper we concentrate only on the actual error rate and its estimation.

1.2. **Classification rule.** The classification rule which is used in the current study can be described as follows. Recall that we restrict our study to discriminant analysis problems involving only two groups or populations. These groups are denoted by Π_1 and Π_2. Suppose $\underline{\mathbf{X}} = (X_1, X_2, \ldots, X_p)^T$ is a p-dimensional vector of random variables associated with any individual. We assume that $\underline{\mathbf{X}}$ has different probability distributions in Π_1 and Π_2. Let $\underline{\mathbf{x}}$ be the observed value of $\underline{\mathbf{X}}$ (for an arbitrary individual), $f_1(\underline{\mathbf{x}})$ be the probability density of $\underline{\mathbf{X}}$ in Π_1, and $f_2(\underline{\mathbf{x}})$ be the probability density of $\underline{\mathbf{X}}$ in Π_2. Then the simplest intuitive classification decision is: classify $\underline{\mathbf{x}}$ into Π_1 if it has greater probability of coming from Π_1, that is if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) > 1$; or classify $\underline{\mathbf{x}}$ into Π_2 if it has greater probability of coming from Π_2, that is if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) < 1$; or classify $\underline{\mathbf{x}}$ arbitrarily into Π_1 or Π_2 if these probabilities are equal or if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) = 1$.

In practice it is reasonable to consider some important factors such as prior probabilities of observing individuals from the two populations and the cost due to misclassifications. However, in this paper, we only consider the case with equal prior probabilities and equal cost due to misclassifications.

Various classification rules has been established in the literature. The earliest and most well-known rule is Fisher's (1936) Linear Discriminant Function (LDF). Let $\underline{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip})^T$, be the means and Σ_i be the covariance matrices of $\underline{\mathbf{X}}$ in Π_i $(i = 1, 2)$. It is often assumed that $\Sigma_1 = \Sigma_2 = \Sigma$. Let $\bar{\underline{\mathbf{x}}}_1, \bar{\underline{\mathbf{x}}}_2, \mathbf{S}_1, \mathbf{S}_2$, and $\mathbf{S}$ be the sample estimates of $\underline{\mu}_1, \underline{\mu}_2, \Sigma_1, \Sigma_2$ and Σ respectively, using independent random samples of size n_1 and n_2 from Π_1 and Π_2. Denote these random samples (also called training samples) by $\underline{\mathbf{t}}_1$ and $\underline{\mathbf{t}}_2$ respectively, and let $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$

be the entire set of training data of $n = n_1 + n_2$ observations. Also let $N_p(\underline{\mu}, \Sigma)$ denotes the p-variate normal distribution with mean $\underline{\mu}$ and covariance matrix Σ. The estimated Fisher's LDF is then given by

$$L(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2). \tag{1.1}$$

This LDF was adopted later by Anderson (1951) to obtain a classification statistics $W(\underline{\mathbf{x}})$, given by

$$W(\underline{\mathbf{x}}) = W(\underline{\mathbf{x}}, \underline{\mathbf{t}}) = \left(\underline{\mathbf{x}} - \frac{1}{2}\left(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2\right)\right)^T \mathbf{S}^{-1}\left(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2\right). \tag{1.2}$$

Using this rule, a new individual $\underline{\mathbf{x}}$ will be allocated into Π_1 if $W(\underline{\mathbf{x}}) \geq 0$, otherwise into Π_2. In this paper we consider (1.2) as our classification rule, and sometime we will use the notation $W(\underline{\mathbf{x}}, \underline{\mathbf{t}})$, to give an emphasize that this classification rule is constructed using the training sample $\underline{\mathbf{t}}$, to classify the new individual $\underline{\mathbf{x}}$.

1.3. **The probability of misclassifications.** What we mean by the probability of misclassifications in this paper is the *actual error rates* of the linear discriminant function $W(\underline{\mathbf{x}}, \underline{\mathbf{t}})$. The actual error rates are given by

$$\begin{aligned} P_1 &= \mathsf{P}(W(\underline{\mathbf{x}}, \underline{\mathbf{t}}) < 0 \text{ when } \underline{\mathbf{x}} \text{ is from } \Pi_1 | \underline{\mathbf{t}} \text{ fixed}), \\ P_2 &= \mathsf{P}(W(\underline{\mathbf{x}}, \underline{\mathbf{t}}) \geq 0 \text{ when } \underline{\mathbf{x}} \text{ is from } \Pi_2 | \underline{\mathbf{t}} \text{ fixed}). \end{aligned} \tag{1.3}$$

Here, P_1 represents the probability of classifying the new individual $\underline{\mathbf{x}}$ in to Π_2 when it is actually belong to Pi_1 and P_2 represents the probability of classifying the new individual $\underline{\mathbf{x}}$ in to Π_1 when it is actually belong to Pi_2. The overall actual error rate is then defined by

$$AC = \frac{n_1}{n_1 + n_2} P_1 + \frac{n_2}{n_1 + n_2} P_2. \tag{1.4}$$

Under the assumptions that $\underline{\mathbf{X}} \sim N_p(\underline{\mu}_1, \Sigma)$ on population Π_1 and $\underline{\mathbf{X}} \sim N_p(\underline{\mu}_2, \Sigma)$ on population Π_2, it can easily be shown that

$$P_1 = \Phi\left[\frac{-\left(\underline{\mu}_1 - \frac{1}{2}(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2)\right)^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)}{\left((\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)^T \mathbf{S}^{-1} \Sigma \mathbf{S}^{-1} (\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)\right)^{1/2}}\right] \tag{1.5}$$

and

$$P_2 = \Phi\left[\frac{\left(\underline{\mu}_2 - \frac{1}{2}(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2)\right)^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)}{\left((\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)^T \mathbf{S}^{-1} \Sigma \mathbf{S}^{-1} (\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)\right)^{1/2}}\right] \tag{1.6}$$

where Φ is the distribution function of a standard normal variate.

We can see from the expressions above that the arguments are still functions of unknown parameters, so these error rates can not be computed directly from the given training data alone. Consequently a procedure for estimating these error rates is needed.

The literature on estimation of error rates in discriminant analysis using LDF given by (1.2) is enormous. Extensive bibliographies can be found in Toussaint (1974), and see also McLachlan (1986). However, this paper only deals with bootstrap error rate estimators. The *estimates* of the actual error rates P_1 and P_2 are denoted respectively

by $\hat{P}_1(e)$ and $\hat{P}_2(e)$, and the estimate of the overall actual error rate is given by $\hat{P}(e) = (n_1\hat{P}_1(e) + n_2\hat{P}_2(e))/n$, where e refers to the corresponding estimators.

2. The bootstrap error rates estimators

Bootstrap technique is a computer-based methodology introduced by Efron (1979) for assessing the variability of an estimator on the basis of the data at hand. The underlying idea of bootstrapping is that the sampling behaviour can often be studied by randomly resampling a given set of data and calculating the statistics of interest. Here, by resampling the original observations in a way such that the stochastic nature of the data is preserved, pseudodata (bootstrap samples) are generated on which the estimator of interest is assessed.

Because the use of computators have become very cheap and easy, the bootstrap methodology has become very popular in such a sort time. This technique has been widely modified and applied to solve various statistical estimation problems, especially when the parametric procedures are complicated or when the parametric assumptions are questioned. Plenty of works on bootstrap methods and their applications have been reported in literature.

In this section, we focus only on the applications of bootstrap techniques for estimating the probability of misclassification in discriminant analysis. Efron (1979) suggested an application of the bootstrap method in estimating the actual error rate as a bias correction procedure to the resubstitution error rate. Later, he developed more sophisticated estimators, in particular the *0.632 estimator* which was shown to outperform some other estimators based on cross-validation and jackknife techniques in his simulation studies. So we chose to focus more on this 0.632 estimator, in our present study. Since bootstrap samples can be obtained in two different ways namely *mixture sampling* and *separate sampling,* we shall consider both 'mixture and separate sampling versions' of the 0.632 estimator.

The procedures to compute Efron's 0.632 estimators can be summarized as follows:

Step 1: First, consider the case of mixture sampling. Let $\hat{F}$ be the empirical or sample probability distribution with mass $1/n$ for each $\underline{\mathbf{x}}_j$ in the original training data $\underline{\mathbf{t}}$, $(j = 1, 2, \ldots, n)$. A new training sample $\underline{\mathbf{t}}^* = \{\underline{\mathbf{x}}_1^*, \underline{\mathbf{x}}_2^*, \ldots, \underline{\mathbf{x}}_n^*\}$, called the bootstrap training sample, is generated by resampling the original training observations with replacement. This resampling is performed according to the above empirical probability distribution $\hat{F}$. Notice here that, we may relabel $\underline{\mathbf{x}}_1^*, \underline{\mathbf{x}}_2^*, \ldots, \underline{\mathbf{x}}_n^*$ so that $\underline{\mathbf{x}}_{1j}^*$ $(j = 1, 2, \ldots, n_1^*)$ denote those $\underline{\mathbf{x}}_j^*$ observations, n_1^* in number, which have been selected into the bootstrap sample from the training set $\underline{\mathbf{t}}_1$, and $\underline{\mathbf{x}}_{2j}^*$ $(j = 1, 2, \ldots, n_2^*)$ denote those $\underline{\mathbf{x}}_j^*$ observations, n_2 in number, which have been selected into the bootstrap sample from the training set $\underline{\mathbf{t}}_2$. Note also that n_i^* may not be equal to n_i $(i = 1, 2)$, but $n_1^* + n_2^* = n$.

Now, consider the case of separate samplings. Let $\hat{F}_1$ be the empirical or sample probability distribution with mass $1/n_1$ for each observation $\underline{\mathbf{x}}_{1j}$ in the first original training data $\underline{\mathbf{t}}_1$ $(j = 1, 2, \ldots, n_1)$. A new bootstrap training sample $\underline{\mathbf{t}}_1^* = \{\underline{\mathbf{x}}_{11}^*, \underline{\mathbf{x}}_{12}^*, \ldots, \underline{\mathbf{x}}_{1n_1}^*\}$, is generated by resampling $\underline{\mathbf{t}}_1$ with replacement. This resampling is performed according to the empirical probability distribution $\hat{F}_1$. Also let $\hat{F}_2$ be the empirical or sample probability distribution with mass $1/n_2$ for each observation $\underline{\mathbf{x}}_{2j}$ in the second original training data $\underline{\mathbf{t}}_2$ $(j = 1, 2, \ldots, n_2)$. Another bootstrap training sample $\underline{\mathbf{t}}_2^* = \{\underline{\mathbf{x}}_{21}^*, \underline{\mathbf{x}}_{22}^*, \ldots, \underline{\mathbf{x}}_{2n_2}^*\}$, is also generated by resampling the second original training sample $\underline{\mathbf{t}}_2$ with replacement. This resampling is performed according to the empirical probability distribution $\hat{F}_2$. Then the entire bootstrap sample obtained is $\underline{\mathbf{t}}^* = \{\underline{\mathbf{t}}_1^*, \underline{\mathbf{t}}_2^*\}$.

Step 2: Based on the bootstrap sample $\underline{\mathbf{t}}^*$, a new classification rule $W(\underline{\mathbf{x}}, \underline{\mathbf{t}}^*)$ is constructed in precisely the same manner as $W(\underline{\mathbf{x}}, \underline{\mathbf{t}}^*)$ is from the original $\underline{\mathbf{t}}$.

Step 3: Compute α and β; where β is the number of individuals in $\underline{\mathbf{t}}$ that are not drawn into $\underline{\mathbf{t}}^*$, and α is the number of individuals in $\underline{\mathbf{t}}$ that are not drawn into $\underline{\mathbf{t}}^*$ and are misclassified by the rule $W(\underline{\mathbf{x}}, \underline{\mathbf{t}}^*)$.

Suppose that steps 1, 2, and 3 are repeated b times so that we have b bootstrap training samples $\underline{\mathbf{t}}_1^*, \underline{\mathbf{t}}_2^*, \ldots, \underline{\mathbf{t}}_b^*$ and b corresponding classification rules

$$W_1(\underline{\mathbf{x}}, \underline{\mathbf{t}}_1^*), W_2(\underline{\mathbf{x}}, \underline{\mathbf{t}}_2^*), \ldots, W_b(\underline{\mathbf{x}}, \underline{\mathbf{t}}_b^*).$$

The bootstrap estimator of the overall actual error rate is then given by

$$\hat{P}(bst) = (1 - \nu)\hat{P}(R) + \nu\xi, \tag{2.1}$$

where, $\hat{P}(R)$ is the overall *apparent (resubstitution) error rate*, and

$$\begin{aligned} \xi &= \left(\sum_{m=1}^{b} \alpha_m\right) / \left(\sum_{m=1}^{b} \beta_m\right) \text{ or} \\ \xi &= \sum_{m=1}^{b}\sum_{j=1}^{n} \delta_{mj} Q[y_j, W_m(\underline{\mathbf{x}}_{ij}^*, \underline{\mathbf{t}}_m^*)] / \sum_{m=1}^{b}\sum_{j=1}^{n} \delta_{mj}, \end{aligned} \tag{2.2}$$

where $y_j = 1$ or 2 according as $\underline{\mathbf{x}}_{ij}^* \in \Pi_1$ or Π_2, and $\delta_{mj} = 1$ if $\underline{\mathbf{x}}_{ij}^* \notin \underline{\mathbf{t}}_m^*$ and 0 otherwise. The coefficient ν is a constant defined as below.

Efron (1983) developed this estimator by considering the probability distribution of the *(Mahalanobis type) distance* between the observation at which the classification rule is applied and the nearest observation in the training data. This probability distribution is equivalent to the probability that the observation at which the rule is applied is included in the bootstrap samples. In the mixture sampling situation, this probability denoted by ν can be expressed as

$$\left(1 - \left(1 - n^{-1}\right)^n\right) \tag{2.3}$$

which tends to 0.632 as $n \to \infty$. For separate sampling case, this probability is given by

$$\left\{1 - \left[\frac{n_1}{n_1+n_2}\left(1-n_1^{-1}\right)^{n_1} + \frac{n_2}{n_1+n_2}\left(1-n_2^{-1}\right)^{n_2}\right]\right\} \tag{2.4}$$

which also approaches 0.632 as both n_1 and $n_2 \to \infty$. Efron (1983) considered the limiting cases and defined the *0.632 estimator* as in (2.1) with $\nu = 0.632$.

Chatterjee and Chatterjee (1983) introduced a slightly different procedure to compute the '0.632 estimator' using separate sampling bootstrapping. Instead of using ξ as in (2.2), they used

$$\xi = \left(\frac{1}{b}\right)\sum_{m=1}^{b}\frac{\alpha_m}{\beta_m}, \tag{2.5}$$

in (2.1).

3. Comparative studies on error rates estimations

As mentioned previously, the literature on estimation of the error rates in discrimination and classification problems is enormous. However, in this section we focus only on some of those studies which deal with extensive comparisons of bootstrap estimators. The comparative studies considered here are by Efron (1983), Snapinn and Knoke (1985), and Ganeshanandam and Krzanowski (1990).

3.1. **The study by Efron (1983).**

Efron (1983), proposed three new bootstrap-based estimators, namely, the *randomized bootstrap*, the *double bootstrap* and the 0.632 estimator, as improvements to the *ordinary bootstrap* estimator of Efron (1979). Furthermore, he compared these estimators with the ordinary bootstrap estimator and the U estimator (which he called the cross-validation estimator), using a criterion called *mean squared error* (MSE).

These comparisons were carried out on *multivariate normal* situations with fixed separation between the two populations. The experimental factors which he considered were the number of variables in the data and the sample sizes. He considered two levels for the number of variables: $p = 2$ and $p = 5$, and also two levels for the sample sizes: $n = 14$ and $n = 20$.

He found the cross-validation technique (U estimator) to give a nearly unbiased estimate for the actual error rate, but often to have high variability, especially when n was small. The ordinary bootstrap gave an estimate of the actual error rate with low variability, but with a possibly large *downward bias.* The double bootstrap automatically corrected the bias of the ordinary bootstrap without increasing the MSE of estimation. The randomized bootstrap, on average, performed second best in the sampling experiments. The best method for estimating the actual error rate was the 0.632 estimator.

3.2. **The study by Snapinn and Knoke (1985).** Snapinn and Knoke (1985) proposed the NS estimator and compared it with R, U, and IB (ideal bootstrap) estimators in a variety of situations. Here the IB estimator uses a resampling plan which reduces the bias of the R estimator and adds no variance. This estimator (for $i = 1, 2$) is defined as

$$\hat{P}_i(IB) = \hat{P}_i(R) - [\mathsf{E}(\hat{P}_i(R)) - \mathsf{E}(P_i)],$$

where $\hat{P}_i(R)$ is the R estimator, P_i is the actual error rate, and the expectation is taken overall sampling experiments. In this study also they used the criterion $UMSE$, calculated by numerical integration and Monte Carlo samplings.

They concluded that the IB and U methods have lower $UMSE$ than NS method when Δ is small, but the NS method has the lowest UMSE when Δ is large. As usual, here Δ denotes the *true Mahalanobis distance* between the two populations. They also found the R, U, IB and NS methods to be robust to departures from normality. Finally, they suggested that the NS method should be chosen over the U and IB estimators if the ratio of the sample size to the number of variables is large (ie. when $n/p > 5$, $i = 1, 2$).

3.3. **The study by Ganeshanandam and Krzanowski (1990).** The most recent and wide spread comparative study on estimation of the actual error rate in discriminant analysis is due to Ganeshanandam and Krzanowski (1990). In this study, they compared the performance of the R, D, OS, L, M, NS, U, $\bar{U}$, JK, 0.632 and FK estimators. Here JK refers to the *jackknife estimator* and FK refers to the *leave-one-out* based estimator when the technique is applied to the *Quadratic Discriminant Function* (see Fukunaga and Kessel (1971)).

Regarding the parent populations, they considered both ideal and non-ideal conditions. The ideal condition refers to the case where both parent populations are *multivariate normal* with *common covariance matrix.* For non-ideal conditions, they considered the case where both populations consist of *multivariate binary variables*, a case which is far from normality assumption. Besides the *mean square error* (MSE) criterion, they also used another criterion called *optimism* (OPT). This OPT criterion indicates whether an estimator is overoptimistic or underoptimistic in estimating the actual error rate.

In their study, they worked with standardized variables, so that the covariance matrix Σ was in fact a correlation matrix. The mean vectors of the populations were $\underline{\mu}_1 = 0$ and the elements of $\underline{\mu}_2$ were determined by some of the experimental factors given below. For the multivariate normal situation, five experimental factors were considered. Two levels for the number of variables ($p = 10, 20$), two levels for sample sizes relative to the number of variables (small and large), and two levels for the Mahalanobis distances Δ, $\Delta = 1.01$ (close populations) and $\Delta = 2.53$ (well-separated populations) were chosen for their study. The fourth factor was ν which represents the interdependency of the variables and the fifth one denoted by d represents the variation in the elements of

$\underline{\mu}_2$. They considered two levels for factor ν as $\nu = 0.4$ (highly dependent variables) and $\nu = 0.8$ (almost independent variables), and also two levels for factor d: $d = 0.5$ (large differences among the elements of $\underline{\mu}_2$) and $d = 0.75$ (small differences among the elements of $\underline{\mu}_2$).

For multivariate binary case, they considered 5 and 10 variables with three levels of sample sizes relative to the number of variables (small, medium, and large). The other factors were $p_{i,j}$ which represents the probability that the jth variable takes value 1 in the ith population ($j = 1, 2, ..., p$ and $i = 1, 2$), and $r_i(jk)$ which represents the correlation between the jth and kth variables in the ith population. The values of $p_{i,j}$'s considered were as below,

Table 1: The values of p_{ij}'s considered.

Level	Π_1					Π_2				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
1	0.20	0.20	0.20	0.20	0.20	0.80	0.80	0.80	0.80	0.80
2	0.25	0.30	0.35	0.40	0.45	0.75	0.70	0.65	0.60	0.55
3	0.40	0.45	0.50	0.55	0.60	0.60	0.55	0.50	0.45	0.40
4	0.25	0.30	0.35	0.40	0.45	0.45	0.40	0.35	0.30	0.25
5	0.30	0.40	0.50	0.60	0.70	0.30	0.40	0.50	0.60	0.70

Here, the levels 1 to 4 represent progressively decreasing difference between Π_1 and Π_2, with level 5 being the limiting case of identical populations. For the case $p = 10$, they worked with two independent blocks of five binary variables in each block. When $p = 5$, they considered two situations for $r_i(jk)$: all $r_i(jk)$ equal to zero and all $r_i(jk)$ equal to 0.25. When $p = 10$, they set all $r_i(jk)$ equal to zero in the first block of 5 binaries and all $r_i(jk)$ equal to 0.25 in the second block. See Ganeshanandam and Krzanowski (1990) for exact details of the generation of simulated data.

Their conclusions were as follows. On average, the U (leave-one-out), JK (jackknife), L (Lachenbruch's), M (McLachlan's), and $\bar{U}$ (Lachenbruch and Mickey's) estimators form the best cluster of estimators for estimating the actual error rate, for both multivariate normal and multivariate binary parent populations. The OS estimator can also be recommended with a caution that it may produce very large overoptimistic bias. The NS (Smoothed), R (resubstitution), and D (plug-in) estimators perform poorly in general. They also reported that the 0.632 estimator is highly sensitive to the changes of the Mahalanobis distance Δ (for the multivariate normal case), and performs the best for small Δ and the worst for large Δ. It was also noted that the 0.632 estimator always estimated the actual error rate in the vicinity of $(0.3 - 0.4)$. We

refer to Mangku (1992) for the details explanations of the estimators considered in this comparative study.

4. Discussion

As mentioned in the previous section, the major problem in estimating the error rate in discriminant analysis arises when there are no data available beyond those used to define and estimate the classification rule. The resubstitution method has been reported optimistically biased because of using the same data to construct as well as to evaluate the classification rule.

One other alternative is using parametric estimators. When the parent populations are multivariate normal, on average, the OS, L, and M estimators are the best for estimating the actual error rate. However, the performance of these estimators deteriorate when the parent populations are not normal. So, when the normality assumption is questioned, we still need a better estimator.

Another alternative is using the empirical estimators such as the ones based on cross-validation, jackknife, and bootstrap. The best estimators among these empirical techniques, as reported by some comparative studies discussed in the previous section, are the U, $\bar{U}$, JK, and 0.632 estimators. These U, $\bar{U}$, and JK estimators are basically based on the leave-one-out technique. Since this procedure holds out one observation at a time, in turn, until each observation has been held once, the maximum number of pseudo data created here is the same as the original sample size. Because of this fact, the performance of these estimators deteriorate when the sample sizes become small. In other words, when the sample sizes are small, we still need a better estimator.

In the case of small samples, we expect the bootstrap based technique (0.632 estimator) to behave better, since the number of pseudo data that can be generated here is almost independent of the sample sizes. The number of bootstrap samples that can be re-sampled (with replacement) from a sample of size n is n^n . Here, we can notice that the number of pseudo data sets, namely n^n, is much larger than the size of the original sample, n, even for small values of n.

References

[1] Anderson, T.W. (1951). "Classification by Multivariate Analysis," *Psychometrika*, **16**, 31-50.

[2] Chatterjee, S. and Chatterjee, S. (1983). "Estimation of Missclassification Probabilities", *Commun. Statist-Simula. Computa.*, **12**, 645-656.

[3] Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, **7**, 1-26.

[4] Efron, B. (1983). "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, **78**, 316-331.

[5] Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problem," *Annals of Eugenics*, **7**, 179-188.

[6] Fukunaga, K. and Kessel, D.L. (1971). "Estimation of Classification Error," *IEEE Transactions on Computers*, **20**, 1521-1527.

[7] Ganeshanandam, S., and Krzanowski, W.J. (1990). "Error-rate Estimation in Two-Group Discriminant Analysis Using The Linear Discriminant Function," *J. Statist. Comput. Simul.*, **36**, 157-175.
[8] Hills, M. (1966). "Allocation Rules and Their Error Rates," *Journal of The Royal Statistical Society, Ser.B*, **28**, 1-20.
[9] Mangku, I W. (1992). *Error Rate Estimation in Discriminan Analysis: Another Look at Bootstrap and Other Empirical Techniques.* Unpublished Master Thesis, Curtin University of Technology, Perth, Australia.
[10] McLachlan, G.J. (1986). "Error Rate Estimation in Discriminant Analysis: Recent advances," *In Advances in Multivariate Statistical Analysis, ed. A. K. Gupta, Dordrecht: D. Reidel*, 233-252.
[11] Okamoto, M. (1963). "An Asymptotic Expansion for The Distribution of The Linear Discriminant Function," *Ann. Math. Stat.*, **34**, 1286-1301.
[12] Snapinn, S.M., and Knoke, J.D. (1985). "An Evaluation of Smoothed Classification Error-Rate Estimators," *Technometrics*, **27**, 199-206.
[13] Toussaint, G.T. (1974). "Bibliography on Estimation of Misclassification," *IEEE Transactions on Information Theory*, **20**, 472-479.
[14] Wald, A. (1944). "On a Statistical Problem Arising in the Classification of an Individual Into One of Two Group," *Annals of Mathematical Statistics*, **15**, 145-169.