ESTIMATING THE PROBABILITY OF MISCLASSIFICATIONS IN TWO-GROUPS DISCRIMINANT ANALYSIS

I WAYAN MANGKU

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

Abstract. This paper is a survey study on estimation of the probability of misclassifications in two-groups discriminant analysis using the linear discriminant function as the classification rule. Here we consider two groups of estimators, namely parametric estimators and empirical estimators. The results of 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, parametric estimator, empirical estimator.

1. Introduction

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

One of the problems 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 misclassifications 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 misclassifications, which also called error rates.

There are three types error rates that have been frequently considered for study, 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 practice, 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 work associated with error rate 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 experimentor who has already determined the classification rule. This conditional error rate is also referred to as 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.

2. Classification rule

Now we defined the classification rule which is used in the current study. 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 real situations 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.

A variety of 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

$$L(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2). \tag{2.1}$$

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

$$W(\underline{\mathbf{x}}) = W(\underline{\mathbf{x}}, \underline{\mathbf{t}}) = \left(\underline{\mathbf{x}} - \frac{1}{2}\left(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2\right)\right)^T \mathbf{S}^{-1}\left(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2\right). \tag{2.2}$$

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 (2.2) 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 $\underline{\mathbf{x}}$.

3. The probability of misclassifications

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

$$\begin{aligned} P_1 &= \mathsf{P}(W(\underline{\mathbf{x}}, \underline{\mathbf{t}}) < 0 \text{ when } \underline{\mathbf{x}} \text{ is from } \Pi_1 | \underline{\mathbf{t}} \text{ fixed}), \\ P_2 &= \mathsf{P}(W(\underline{\mathbf{x}}, \underline{\mathbf{t}}) \geq 0 \text{ when } \underline{\mathbf{x}} \text{ is from } \Pi_2 | \underline{\mathbf{t}} \text{ fixed}). \end{aligned} \tag{3.1}$$

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

$$AC = \frac{n_1}{n_1 + n_2} P_1 + \frac{n_2}{n_1 + n_2} P_2. \tag{3.2}$$

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

$$P_1 = \Phi\left[\frac{-\left(\underline{\mu}_1 - \frac{1}{2}(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2)\right)^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)}{\left((\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)^T \mathbf{S}^{-1} \Sigma \mathbf{S}^{-1} (\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)\right)^{1/2}}\right] \tag{3.3}$$

and

$$P_2 = \Phi\left[\frac{\left(\underline{\mu}_2 - \frac{1}{2}(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2)\right)^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)}{\left((\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)^T \mathbf{S}^{-1} \Sigma \mathbf{S}^{-1} (\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)\right)^{1/2}}\right] \tag{3.4}$$

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

From the expressions above, we can see 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.

There is vast amount of literature available on estimation of error rates in discriminant analysis using LDF given by (2.2). Extensive bibliographies can be found in Toussaint (1974), and see also McLachlan (1986). However, this paper only deals with some of the error rate estimators which are either have been shown to be robust or have been implemented in computer software for estimating misclassification probabilities. 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.

In this paper we consider two different types of estimators, namely (1) parametric estimators, (2) empirical (non bootstrap) estimators. The bootstrap estimators will be discussed in a follow up paper (see also Mangku (1992)).

4. Parametric Estimators

Parametric estimators mainly depend on the assumptions of multivariate normality and common covariance structure for the parent populations. The oldest parametric estimator is the *D-estimator* or *"plug-in estimator"* proposed by Fisher (1936). Many investigators found that this estimator is optimistically biased for estimating the actual error rate, particularly when the training samples are small (see for example, Dunn and Varady (1966), and Hills (1966)). The other parametric estimators denoted by DS, O, OS, L, and M are modifications (improvements) of the D-estimator with various bias reduction terms, proposed and compared by Okamoto (1963), Lachenbruch (1968), Lachenbruch and Mickey (1968), McLachlan (1974a), Page (1985), Ganeshanandam and Krzanowski (1990), and some others. Some authors also proposed smoothed estimators as bias corrections to D and DS estimators (see Glick (1978), Snapinn (1983), Snapinn and Knoke (1985)). The above studies suggest OS (Okamoto, 1963), M (McLachlan, 1974a), and NS (Snapinn and Knoke, 1985), to be robust mainly for normal-variables data. Hence, in this paper, we present and discuss these three estimators, which are described as below:

(a) Okamoto's estimator (OS): Okamoto (1963) considered an asymptotic approach to the distribution of $W(\underline{\mathbf{x}})$ and proposed asymptotic expansions in terms of n_1 , n_2 , p, and D_s^2 for estimating P_1 and P_2. The OS estimator for P_1 is given by

$$\hat{P}_1(OS) = \Phi\left(-\frac{D_s}{2}\right) + \phi\left(-\frac{D_s}{2}\right)\left[\frac{D_s^3/4 + 3D_s(p-1)}{4n_1D_s^2}\right]$$
$$+\phi\left(-\frac{D_s}{2}\right)\left[\frac{D_s^3/4 - D_s(p-1)}{4n_2D_s^2} + \frac{D_s(p-1)}{4(n-2)}\right], \tag{4.1}$$

where $D_s^2 = (n-p-3)D^2/(n-2)$, $D^2 = (\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)^T\mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)$, ϕ and Φ are respectively the density and distribution functions of a standard normal variate, and p is the number of variables in the training data. The expression for $\hat{P}_2(OS)$ can be obtained by interchanging n_1 and

n_2 in (4.1). Note here that D^2 is the estimated (sample) *Mahalanobis squared distance* between the populations Π_1 and Π_2.

(b) McLachlan's estimator (M): McLachlan (1974a) gave an asymptotic unbiased estimator for the actual error rate P_1 in the form

$$\begin{aligned}\hat{P}_1(M) &= \Phi\left(-\frac{D}{2}\right)+\phi\left(\frac{D}{2}\right)\left[\frac{p-1}{Dn_1}+\frac{4D(4p-1)-D^3}{32n-64}\right.\\ &+\frac{D(p-1)(p-2)}{4n_1^2}+\frac{(p-1)(-D^3+8D(2p+1)+16/D)}{64n_1(n-2)}\\ &+\frac{3D^7-4(24p+7)D^5+16(48p^2-48p-53)D^3}{12288(n-2)^2}\\ &\left.+\frac{192D(15-8p)}{12288(n-2)^2}\right]. \qquad (4.2)\end{aligned}$$

Similarly, $\hat{P}_2(M)$ is obtained by interchanging n_1 and n_2 in (4.2).

(c) The Smoothed estimator (NS): Glick (1978) proposed a new class of estimators called the smoothed estimators. Glick's study was later generalized by Snapinn (1983), and Snapinn and Knoke (1985) who suggested an improved smoothed estimator called the NS estimator. The NS estimates for P_1 and P_2 are given by

$$\hat{P}_1(NS)=\Phi\left[-\frac{D}{2}\left(\frac{n_1}{c^2n_1+n_1-1}\right)^2\right] \text{ and}$$

$$\hat{P}_2(NS)=\Phi\left[-\frac{D}{2}\left(\frac{n_2}{c^2n_2+n_2-1}\right)^2\right], \qquad (4.3)$$

where

$$c=\left[\frac{(p+2)(n_1-1)+(n_2-1)}{n_1(n_1+n_2-p-3)}\right]^{1/2}.$$

5. Empirical Estimators (Non Bootstrap)

The error rate estimators which are free from the assumption of multivariate normality are generally called the empirical estimators. These empirical estimators include the *resubstitution* estimator as well as the estimators using resampling techniques such as *cross-validation, jackknifing,* and *bootstrapping*. We will discuss the bootstrap estimators separately in a follow up paper. In this paper, we consider the resubstitution or the R estimator (smith, 1947), the U estimator (Lachenbruch, 1967), the $\bar{U}$ estimator (Lachenbruch and Mickey, 1968), and the jackknife estimator.

(a) The Resubstitution (R) estimator: This estimator was proposed by Smith (1947). The basic idea is to reallocate each individual in the training sample $\underline{\mathbf{t}}$ using the rule $W(\underline{\mathbf{x}},\underline{\mathbf{t}})$ to asses the performance of this rule. Estimate of the error rate is then given by the proportion of those individuals which are misclassified by the rule $W(\underline{\mathbf{x}},\underline{\mathbf{t}})$. Let the

"counting criterion" $Q(i,j) = 0$ if $i = j$, and $Q(i,j) = 1$ if $i \neq j$, for any i and j. Then the R estimator can be defined as

$$\hat{P}_1(R) = \frac{1}{n_1}\sum_{j=1}^{n_1} Q[1, W(\underline{\mathbf{x}}_{1j}, \underline{\mathbf{t}})] \text{ and } \hat{P}_2(R) = \frac{1}{n_2}\sum_{j=1}^{n_2} Q[2, W(\underline{\mathbf{x}}_{2j}, \underline{\mathbf{t}})]. \quad (5.1)$$

These are obvious nonparametric estimators of the actual error rates, and the overall estimator is often referred to as the *Apparent Error Rate*. This overall estimator provides a highly overoptimistic assessment of the actual error rate, since it is obtained by applying the classification rule $W(\underline{\mathbf{x}}, \underline{\mathbf{t}})$ to the same data used in its construction.

(b) The U estimator: Lachenbruch (1967) introduced this empirical estimator which depends on the well-known *leave-one-out* technique (a particular choice of the general *cross-validation* procedure). The basic idea is to estimate the actual error rates by deleting an individual from the training sample $\underline{\mathbf{t}}$ each time, and then construct a classification rule using the remainder to allocate the deleted individual using this classification rule. This process is repeated until each individual has been deleted once from $\underline{\mathbf{t}}$. The estimate of the error rate is then given by the proportion of the deleted individuals which are misclassified by the corresponding classification rules. Let $\underline{\mathbf{t}}_{[ij]}$ denotes the original training sample $\underline{\mathbf{t}}$ with observation $\underline{\mathbf{x}}_{ij}$ omitted ($i = 1, 2$ and $j = 1, 2, \ldots, n_i$). The U estimators are then given by

$$\hat{P}_1(U) = \frac{1}{n_1}\sum_{j=1}^{n_1} Q[1, W(\underline{\mathbf{x}}_{1j}, \underline{\mathbf{t}}_{[1j]})] \text{ and}$$
$$\hat{P}_2(U) = \frac{1}{n_2}\sum_{j=1}^{n_2} Q[2, W(\underline{\mathbf{x}}_{2j}, \underline{\mathbf{t}}_{[2j]})]. \quad (5.2)$$

Furthermore, Lachenbruch and Mickey (1968) suggested a procedure to avoid inversions of several sample covariance matrices associated with various $\underline{\mathbf{t}}_{[ij]}$, and hence to accelerate the computations. This procedure can be summarized as below. Using the notations before, let $\bar{\underline{\mathbf{x}}}_1$ and $\bar{\underline{\mathbf{x}}}_2$ be the sample mean vectors and $\mathbf{S}$ be the pooled sample covariance matrix computed from the original training sample $\underline{\mathbf{t}}$. For $j = 1, 2, \ldots, n_1$, delete $\underline{\mathbf{x}}_{1j}$ from $\underline{\mathbf{t}}$, and let $\underline{\mathbf{t}}_{[1j]}$ denote the training data $\underline{\mathbf{t}}$ without $\underline{\mathbf{x}}_{1j}$. Let $\underline{\mathbf{u}}_{1j} = \underline{\mathbf{x}}_{1j} - \bar{\underline{\mathbf{x}}}_1$, $c_1 = n_1/((n_1 - 1)(n - 2))$, $a_{1j} = \underline{\mathbf{u}}_{1j}^T \mathbf{S}^{-1} \underline{\mathbf{u}}_{1j}$, and $\mathbf{S}_{1j}$ be the pooled sample covariance matrix from $\underline{\mathbf{t}}_{[1j]}$. Then $\mathbf{S}_{1j}^{-1}$ can be obtained as

$$\mathbf{S}_{1j}^{-1} = \mathbf{S}^{-1} + \left[\frac{c_1 \mathbf{S}^{-1} \underline{\mathbf{u}}_{1j} \underline{\mathbf{u}}_{1j}^T \mathbf{S}^{-1}}{1 - c_1 a_{1j}}\right].$$

Note that the only inversion required here is $\mathbf{S}^{-1}$ which is computed only once for the data under study. This $\mathbf{S}_{1j}^{-1}$ is used in the rule $W(\underline{\mathbf{x}})$

re-written as

$$W_{1j}(\underline{\mathbf{x}}_{1j}) = \left(\frac{n-3}{n-2}\right)\left\{\underline{\mathbf{x}}_{1j} - \frac{1}{2}\left(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2\right) + \frac{\underline{\mathbf{u}}_{1j}}{2(n_1-1)}\right\}^T \mathbf{S}_{1j}^{-1} \left\{\left(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2\right) - \frac{\underline{\mathbf{u}}_{1j}}{(n_1-1)}\right\} \quad (5.3)$$

to be used as $W(\underline{\mathbf{x}}_{1j}, \underline{\mathbf{t}}_{[1j]})$ in $\hat{P}_1(U)$ of (5.2). Similar construction can be performed to obtain $\hat{P}_2(U)$. Here for $j = 1, 2, \ldots, n_2$, delete $\underline{\mathbf{x}}_{2j}$ from $\underline{\mathbf{t}}$, and let $\underline{\mathbf{t}}_{[2j]}$ denote $\underline{\mathbf{t}}$ without $\underline{\mathbf{x}}_{1j}$. Also let $\underline{\mathbf{u}}_{2j} = \underline{\mathbf{x}}_{2j} - \bar{\underline{\mathbf{x}}}_2$, $c_2 = n_2/((n_2-1)(n-2))$, $a_{2j} = \underline{\mathbf{u}}_{2j}^T \mathbf{S}^{-1} \underline{\mathbf{u}}_{2j}$, and $\mathbf{S}_{2j}$ be the pooled sample covariance matrix from $\underline{\mathbf{t}}_{[2j]}$. Then

$$\mathbf{S}_{2j}^{-1} = \mathbf{S}^{-1} + \left[\frac{c_2 \mathbf{S}^{-1} \underline{\mathbf{u}}_{2j} \underline{\mathbf{u}}_{2j}^T \mathbf{S}^{-1}}{1 - c_2 a_{2j}}\right],$$

and the classification rule for allocating $\underline{\mathbf{x}}_{2j}$ is

$$W_{2j}(\underline{\mathbf{x}}_{2j}) = \left(\frac{n-3}{n-2}\right)\left\{\underline{\mathbf{x}}_{2j} - \frac{1}{2}\left(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2\right) + \frac{\underline{\mathbf{u}}_{2j}}{2(n_2-1)}\right\}^T \mathbf{S}_{2j}^{-1} \left\{\left(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2\right) - \frac{\underline{\mathbf{u}}_{2j}}{(n_2-1)}\right\}. \quad (5.4)$$

(c) The $\bar{U}$ estimator: This estimator was also proposed by Lachenbruch and Mickey (1968) by blending the empiricism of the U estimator with the application of normal distribution theory to the classification rules. Let $\bar{w}_1$ and s_{w1} be the means and standard deviations of values $W_{11}(\underline{\mathbf{x}}_{11}), W_{12}(\underline{\mathbf{x}}_{12}), \ldots, W_{1n_1}(\underline{\mathbf{x}}_{1n_1})$. Also let $\bar{w}_2$ and s_{w2} be the means and standard deviations of values $W_{21}(\underline{\mathbf{x}}_{21}), W_{22}(\underline{\mathbf{x}}_{22}), \ldots, W_{2n_2}(\underline{\mathbf{x}}_{2n_2})$. Then normality is assumed on the LDF's $W_{1j}(\underline{\mathbf{x}}_{1j})'s$ and $W_{2j}(\underline{\mathbf{x}}_{2j})'s$ separately ($j = 1, 2, \ldots, n_i, i = 1, 2$), and the error rates P_1 and P_2 are estimated by

$$\hat{P}_1(\bar{U}) = \Phi\left(-\frac{\bar{w}_1}{s_{w1}}\right) \text{ and } \hat{P}_2(\bar{U}) = \Phi\left(-\frac{\bar{w}_2}{s_{w2}}\right). \quad (5.5)$$

(d) The Jackknife estimator (JK): The jackknife technique was firstly introduced by Quenouille as a nonparametric method for estimating bias (Efron, 1982). Then this method was adapted and widely applied in various statistical estimations, for examples see Efron (1981, 1982), Efron and Gong (1983), Efron and Stein (1981), Parr (1983), Beran (1984), Frangos and Stone (1984), Hinkley and Wei (1984), Abel and Berger (1986), Gong (1986), McLachlan (1986), Simonoff (1986), Wu (1986), Kunsch (1989), and Schucany and Sheater (1989). However, in this section we only focus on the jackknife method for estimating the probability of misclassification in discriminant analysis.

As opposed to the leave-one-out estimator, the jackknife procedure first computes the resubstitution error rate each time an observation is omitted from the training sample $\underline{\mathbf{t}}$. Then the standard jackknife technique is used to correct the overall bias from the resubstitution

error. Following Efron (1982, chapter 7) and McLachlan (1986), the jackknife estimate of the actual error rates P_1 and P_2 can be written as

$$\hat{P}_1(JK) = \hat{P}_1(R) + (n-1)(R_1^+ - R_{1(.)}), \text{ and}$$
$$\hat{P}_2(JK) = \hat{P}_2(R) + (n-1)(R_2^+ - R_{2(.)}). \tag{5.6}$$

Here, $\hat{P}_1(R)$ and $\hat{P}_2(R)$ are the resubstitution estimators given by (5.1), R_1^+ and R_2^+ are given respectively by

$$R_1^+ = \left(\frac{1}{n_1}\right)^2 \sum_{j=1}^{n_1}\sum_{k=1}^{n_1} Q\left[1, W(\underline{\mathbf{x}}_{1k}, \underline{\mathbf{t}}_{[1j]})\right] \text{ and}$$
$$R_2^+ = \left(\frac{1}{n_2}\right)^2 \sum_{j=1}^{n_2}\sum_{k=1}^{n_2} Q\left[2, W(\underline{\mathbf{x}}_{2k}, \underline{\mathbf{t}}_{[2j]})\right], \tag{5.7}$$

while $R_{1(.)}$ and $R_{2(.)}$ are given by

$$R_{1(.)} = \frac{1}{n_1}\sum_{j=1}^{n_1}\left\{\frac{1}{n_1-1}\sum_{k\neq j}^{n_1} Q\left[1, W(\underline{\mathbf{x}}_{1k}, \underline{\mathbf{t}}_{[1j]})\right]\right\} \text{ and}$$
$$R_{2(.)} = \frac{1}{n_2}\sum_{j=1}^{n_2}\left\{\frac{1}{n_2-1}\sum_{k\neq j}^{n_2} Q\left[2, W(\underline{\mathbf{x}}_{2k}, \underline{\mathbf{t}}_{[2j]})\right]\right\}. \tag{5.8}$$

Note that $R_{1(.)}$ and $R_{2(.)}$ represent the average apparent error rates associated with $W(\underline{\mathbf{x}}, \underline{\mathbf{t}}_{[1j]})'s$ and $W(\underline{\mathbf{x}}, \underline{\mathbf{t}}_{[1j]})'s$, averaged over n_1 and n_2 respectively.

In exhibiting the close relationship between the jackknife and the cross-validation methods, Efron (1982) showed that the right-hand side of equations (5.6) can be rearranged to give

$$\hat{P}_1(JK) = \hat{P}_1(R) + \hat{P}_1(U) - R_1^+ \text{ and } \hat{P}_2(JK) = \hat{P}_2(R) + \hat{P}_2(U) - R_2^+. \tag{5.9}$$

Here $\hat{P}_1(R)$ and $\hat{P}_2(R)$ are given by (5.1), $\hat{P}_1(U)$ and $\hat{P}_2(U)$ are given by (5.2), and R_1^+ and R_2^+ are given by (5.7). The jackknife estimators are often used as alternatives to the leave-one-out estimators.

6. Discussion

In the previous section we have given descriptions of some error rates estimators, namely the OS, M, NS, R, U, $\bar{U}$, and JK estimators. The natural question is, which one is the best among those methods. To answer this question, it is important to conduct some comparative studies, to compare the performances of those existing estimators. Some comparative studies, which comparing the performances of parametric and empirical (non-bootstrap) estimators, have been done by Lachenbruch and Mickey (1968), McLachlan (1974a, 1974b, 1974c), Snapinn and Knoke (1984), and Page (1985). The results of their studies can be summarized as follows.

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, 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 and jackknife. The best estimators among these empirical techniques, as reported by some comparative studies mention above, are the U, $\bar{U}$, and JK 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 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. Estimation of the actual error rates using the bootstrap techniques will be discussed in a follow up paper (see also Mangku (1992)).

References

[1] Abel, U., and Berger, J. (1986). "Comparison of Resubstitution, Data Splitting, the Bootstrap, and the Jackknife as Methods for Estimating Validity Indices of New Marker Test: A Monte Carlo Study," *Biometrical Journal*, **28**, 899-908.

[2] Anderson, T.W. (1951). "Classification by Multivariate Analysis," *Psychometric*, **16**, 631-650.

[3] Beran, R. (1984). "Jackknife Approximation to Bootstrap Estimates," *Annals of Statistics*, **12**, 101-118.

[4] Dunn, O.J., and Varady, P.D. (1966). "Probabilities of Correct Classification in Discriminant Analysis," *Biometrics*, **22**, 908-924.

[5] Efron, B. (1981). "Nonparametric Estimates of Standard Error: the Jackknife, the Bootstrap and Other Methods," *Biometrika*, **68**, 589-599.

[6] Efron, B. (1982). *The Jackknife, The Bootstrap and Other Resampling Plans*, SIAM-CBMS Monograph 38. Philadelphia: S.I.A.M.

[7] Efron, B., and Gong, G. (1983). "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, **37**, 36-48.

[8] Efron, B., and Stein, C. (1981). "The Jackknife Estimate of Variance," *Annals of Statistics*, **9**, 586-596.

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

[10] Frangos, C.C., and Stone, M. (1984). "On Jackknife, Cross-Validatory, and Classical Methods of Estimating a Proportion With Batches of Different Sizes," *Biometrika*, **71**, 361-366.

[11] 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.

[12] Glick, N. (1978). "Additive Estimators for Probabilities of Correct Classification," *Pattern Recognition*, **10**, 211-222.

[13] Gong, G. (1986). "Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression," *Journal of the American Statistical Association*, **81**, 108-113.
[14] Hills, M. (1966). "Allocation Rules and Their Error Rates," *Journal of The Royal Statistical Society*, Ser.B, **28**, 1-20.
[15] Hinkley, D., and Wei, B. (1984). "Improvements of Jackknife Confidence Limit Methods," *Biometrika*, **71**, 331-339.
[16] Kunsch, H.R. (1989). "The Jackknife and the Bootstrap for General Stationary Observations," *Annals of Statistics*, **17**, 1217-1241.
[17] Lachenbruch, P.A. (1967). "An Almost Unbiased Method of Obtaining Confidence Intervals for The Probability of Misclassification in Discriminant Analysis," *Biometrics*, **23**, 639-645.
[18] Lachenbruch, P.A. (1968). "On Expected Probabilities of Misclassification in Discriminant Analysis, Necessary Sample size, and a Relation With Multiple Correlation Coefficient," *Biometrics*, **24**, 823-834.
[19] Lachenbruch, P.A., and Mickey, M.R. (1968). "Estimation of Error Rates in Discriminant Analysis," *Technometrics*, **10**, 1-11.
[20] Mangku, I W. (1992). *Error Rate Estimation in Discriminan Analysis: Another Look at Bootstrap and Other Empirical Techniques.* Unpublish Master Thesis, Curtin University of Technology, Perth, Australia.
[21] McLachlan, G.J. (1974a). "An Asymptotic Unbiased Techniques for Estimating The Error Rate in Discriminant Analysis," *Biometrics*, **30**, 239-249.
[22] McLachlan, G.J. (1974b). "Estimation of The Error of Misclassification on the Criterion of Asymptotic Mean Square Error," *Technometrics*, **16**, 255-260.
[23] McLachlan, G.J. (1974c). "The Relationship in Term of Asymptotic Mean Square Error Between The Separate Problems of Estimating Each of The Three Types of Error Rate of The Linear Discriminant Function," *Technometrics*, **16**, 569-574.
[24] 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.
[25] Okamoto, M. (1963). "An Asymptotic Expansion for The Distribution of The Linear Discriminant Function," *Ann. Math. Stat.*, **34**, 1286-1301.
[26] Page, J.T. (1985). "Error-Rate Estimation in Discriminant Analysis," *Technometrics*, **27**, 189-198.
[27] Parr, W. C. (1983). "A Note of the Jackknife, the Bootstrap, and the Delta Method Estimators of Bias and Variance," *Biometrika*, **70**, 719-722.
[28] Schucany, W.R., and Sheater, S.J. (1989). "Jackknifing R-Estimators," *Biometrika*, **76**, 393-398.
[29] Simonoff, J.S. (1986). "Jackknifing and Bootstrapping Goodness-of-Fit Statistics in Sparse Multinomials," *Journal of the American Statistical Association*, **81**, 1005-1011.
[30] Smith, C.A.B. (1947). "Some Examples of Discrimination," *Annals of Eugenics*, **13**, 272-282.
[31] Snapinn, S.M. (1983). "An Evaluation of Smoothed Error Rate Estimators in Discriminant Analysis," Inst. Stat. and Mines Ser. 1438, Univ. North Carolina at Chapel Hill.
[32] Snapinn, S.M., and Knoke, J.D. (1984). "Classification Error Rate Estimators Evaluated by Unconditional Mean Square Error," *Technometrics*, **26**, 371-378.
[33] Snapinn, S.M., and Knoke, J.D. (1985). "An Evaluation of Smoothed Classification Error-Rate Estimators," *Technometrics*, **27**, 199-206.
[34] Toussaint, G.T. (1974). "Bibliography on Estimation of Misclassification," *IEEE Transactions on Information Theory*, **20**, 472-479.
[35] 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.
[36] Wu, C.F.J. (1986). "Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis," *Annals of Statistics*, **14**, 1261-1350.