American Journal of Theoretical and Applied Statistics
2016; 5(4): 173-179
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20160504.12
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Evaluation of Error Rate Estimators in Discriminant Analysis with Multivariate Binary Variables

Egbo Ikechukwu

Department of Mathematics, Alvan Ikoku Federal College of Education, Owerri, Nigeria

Email address:
egboike@gmail.com

To cite this article:
Egbo Ikechukwu. Evaluation of Error Rate Estimators in Discriminant Analysis with Multivariate Binary Variables. *American Journal of Theoretical and Applied Statistics.* Vol. 5, No. 4, 2016, pp. 173-179. doi: 10.11648/j.ajtas.20160504.12

Received: April 1, 2016; **Accepted**: April 19, 2016; **Published**: June 4, 2016

Abstract: Classification problems often suffers from small samples in conjunction with large number of features, which makes error estimation problematic. When a sample is small, there is insufficient data to split the sample and the same data are used for both classifier design and error estimation. Error estimation can suffer from high variance, bias or both. The problem of choosing a suitable error estimator is exacerbated by the fact that estimation performance depends on the rule used to design the classifier, the feature-label distribution to which the classifier is to be applied and the sample size. This paper is concerned with evaluation of error rate estimators in two group discriminant analysis with multivariate binary variables. Behaviour of eight most commonly used estimators are compared and contrasted by mean of Monte Carlo Simulation. The criterion used for comparing those error rate estimators is sum squared error rate (SSE). Four experimental factors are considered for the simulation namely: the number of variables, the sample size relative to number of variables, the prior probability and the correlation between the variables in the populations. From the analysis carried out the estimators can be ranked as follows: DS, O, OS, U, R, JK, P and D.

Keywords: Discriminant Analysis, Error Rate, Monte Carlo Simulation, Error Rate Estimators

1. Introduction

It is common to use the estimated error rate to evaluate the performance of a classifier. In the nonparametric framework the leave-one-out method (also referred to as cross-validation or the U method) proposed by [16] has been shown to have a much smaller bias than the resubstitution method [17], and has become a popular nonparametric error estimator in small sample size situations. However, [18] has shown that the leave-one-out method can have a much larger variance than competing estimators. In some cases, this variance is sufficiently large that competitors with slightly larger bias but smaller variance will outperform the leave-one-out estimator. Error estimation is critical to classification because the validity of the resulting classifier model, composed of the classifier and its error estimate, is based on the accuracy of the error estimation procedure [19, 20, 21, and 22]. Given a large set of sample data, the data can be split between training and test data, with a classifier being designed on the training data and its error being estimated on the test data. The downside in splitting the data is that there are less data available for design, thereby hurting the design process. This negative impact is negligible when there is an abundance of data but can be significant when samples are small [22, 23, 24, and 25]. In this paper our focus is on using the same data for training and testing. Since it is impossible to know the accuracy of a particular error estimate for a specific sample, estimation quality is judged based on the properties of the estimation procedure. Performance can be judged in various ways. We consider error-estimation performance relative to accuracy, correlation with the true error, regression between the true and estimated errors, conditional bounds on the true error, the number of variables, the sample size relative to number of variables and the prior probability.

In this paper, the problem of estimating the error rate in two group discriminant analysis is considered. Given the existence of two groups of individuals, one want to find a classification rule for allocating new individuals or observations into one of the existing two groups. Corresponding to each classification rule, there is a probability of misclassifications if that classification rule is

used 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 [23, 24 and 25]. The error rate considered in this paper is the conditional error rate. Here the word conditional refers to the conditioning of the training samples from which the classification rule is constructed. One may also think of this as the probability that the given classification rule would inaccurately 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 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. The rest of the paper is organized as follows; the classification rule which is used in this study is described in section 2, error rates of the discriminant rules in section 3, simulation study plan is given in section 4while results and conclusion is given in section 5.

2. Classification Rule

The classification rule considered in the current study is the maximum likelihood rule, which can be described as follows;

Maximum Likelihood Rule (ML-Rule)

The maximum likelihood discriminant rule for allocating an observation x to one of the populations; $\pi_1,..\pi_n$, is to allocate x to the population which gives the largest likelihood to x. Classify in π_1 if $P(w_1 / x) > P(w_2 / x)$ or to π_2 if

$$P(w_1 / x) < P(w_2 / x) \tag{1}$$

where $P(w_1 / x)$ is the posterior probability which can be found by the Bayes Rule. But this is the same as: classify to π_1 if

$$\frac{P(x/w_1)\cdot P(w_2)}{p(x)} > \frac{P(x/w_2)\cdot P(w_2)}{p(x)} \tag{2}$$

where $P(x / w_i)$ is the class conditional probability density function and $P(w_i)$ is the prior probability. By denoting the classes as $\pi_1,\pi_2 \ldots \pi_n$, the maximum likelihood classifier is based on the assumed multivariate normal probability density function for each class given by

$$f(x / \pi_i) = \frac{1}{(2\pi)^{p/2}|\hat{\sum}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}((x-\hat{\mu}_i))^T \hat{\sum}_i^{-1}((x-\hat{\mu}_i))} \tag{3}$$

where $\hat{\mu}_i$ is the estimated mean vector for class i and $\hat{\sum}_i$ is the estimated variance covariance matrix for class π_i and p is the number of characteristics measured (ie the length of each vector x into one of the classes, recall that the density function $f(x / \pi_i)$ is evaluated for each of the k classes and the x is assigned to π_i if (assuming equal costs of misclassification and equal a prior probabilities) one has

$$f(x / \pi_i) > f(x / \pi_j) \text{ for all } j \neq i \tag{4}$$

We assumed that the data can be modeled adequately by a multi-normal distribution. If the class-conditional probability density function $P(x / w_i)$ is estimated by using the frequency of occurrence of the measurement vectors in the training data, the resulting classifier is non-parametric. An important advantage of the non-parametric classifier is that any pattern, however irregular it may be, can be characterized exactly. This advantage is generally outweighed by two difficulties with the non-parametric approach.

(i). It is difficult to obtain a large enough training sample to adequately characterize the probability distribution of a multi-band data set.

(ii). Specification of a meaningful n-dimensional probability density function requires a massive amount of memory or very clever programming.

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, only the case with equal prior probabilities and equal cost due to misclassifications is considered.

3. Type of Error Rate of the Discriminant Rules

One of the objectives of evaluating a discriminant function is to determine its performance in the classification of future observations. There are several types of error rates associated with discriminant rules.

3.1. The Optimum Error Rate

This is the error rate that would hold if we know the parameter of the distribution. Let $\alpha_i(\tau), i = 1,2$ be defined as the probability that a random member of the π_i is misallocated when the rule $\boldsymbol{\varsigma}$ is used.

$$\alpha_i(\boldsymbol{\varsigma}) = \text{pr}\{\text{l}_\theta(\text{x}) \leq \text{k}/\text{x} \in \pi_1\} \tag{5}$$

$$\alpha_i(\boldsymbol{\varsigma}) = \text{pr}\{\text{l}_\theta(\text{x}) > k/x \in \pi_2\} \tag{6}$$

These are known as the optimum error rates, they are the error rates that would occur if θ were known. Since π_1 and π_2 are labeled arbitrary, it is necessary only to consider $\alpha_1(\boldsymbol{\varsigma})$. To study $\alpha_2(\boldsymbol{\varsigma})$, the labels of the populations are simply interchanged. Therefore, subsequently, any unknown observation, X is assumed to come from π_1, the subscript on α is dropped and $\alpha(\boldsymbol{\varsigma}) = \alpha_j(c)$. The optimum error rate is now given by

$$\alpha(\boldsymbol{\varsigma}) = \text{pr}\{\text{l}_\theta(\text{x}) \leq \text{k}\} \tag{7}$$

3.2. The Conditional Actual Error Rate

The conditional actual error rate is defined as the probability that a random observation from π_1 is misallocated when the rule $\hat{\boldsymbol{\varsigma}}$ is used.

$$\alpha(\hat{\boldsymbol{\varsigma}}) = \text{pr}\{\text{l}_{\hat{\theta}}(\text{x}) \leq \text{k}\} \tag{8}$$

Note that this error rate is conditional on the estimated parameters which in turn are determined by the training samples.

3.3. Expected Actual Error Rate

This is the probability that randomly chosen training samples yield a decision rule which misclassifies a randomly chosen member of π_1. If the expected value operator is defined with respect to all possible training samples, then the expected actual error rate is written as

$$E[\alpha(\hat{\boldsymbol{\varsigma}})] = E[\mathrm{pr}\{\mathrm{l}_{\hat{\theta}}(\mathrm{x}) \leq \mathrm{k}\}] = \mathrm{pr}\{\mathrm{l}_{\hat{\theta}}(\mathrm{x}) \leq \mathrm{k}\} \quad (9)$$

Note the hierarchy associated with these error rates: the optimum error rate is a function only of the distributions of X for the two populations, the expected actual error rate is a function of the distributions of X and the training sample sizes, while the conditional actual error rate is a function of the distributions of X and particular training samples selected. In order to compare error rate estimators it is necessary to specify the error rate being estimated. Assuming θ is unknown; estimates of the optimum error rate and the expected actual error rate are valuable for deciding whether or not a discriminant analysis should be performed, for comparing possible discriminant rules and for determining the advantages of increasing the size of the training samples. However, an experimenter is most likely to be concerned with the performance of his or her discriminant rule after the training samples have been selected. Although the performance of the rule can vary greatly with the choice of the training samples, the optimum error rate and the expected actual error rate are independent of that choice. Therefore, once a discriminant rule $\hat{\boldsymbol{\varsigma}}$ has been determined, it is the conditional error rate, $\alpha(\hat{\boldsymbol{\varsigma}})$, which is of interest.

3.4. Expression for $\alpha(\boldsymbol{\varsigma})$, $\alpha(\hat{\boldsymbol{\varsigma}})$ and $E[\alpha(\hat{\boldsymbol{\varsigma}})]$ Under Normality

Throughout this work the costs of misclassification are assumed to be equal, this may be done without loss of generality since this assumption does not restrict the range of the constant k. Now consider the situation where π_1 and π_2 refer to r-variate normal parent distributions with unknown means, μ_1 and μ_2, respectively, a common covariance matrix, Σ, which may be known or unknown, and let

$$\Delta^2 = (\mu_1 - \mu_2)^1\Sigma^{-1}(\mu_1 - \mu_2) \quad (10)$$

be the mahalanobis distance between the populations. Also assume equal prior probabilities and therefore, k =1. Now let $\bar{X}_1 = \bar{x}_1, \bar{X}_2 = \bar{x}_2$ and $\hat{\Sigma} = \mathrm{S}$ be the minimum variance unbiased estimates of μ_1, μ_2 and Σ based on the training samples [1].

Note that $\hat{\Sigma}$ refers to a random variable and S to a realization of that random variable. In this situation, the linear discriminant function or Anderson's W statistic is defined as

$$\mathrm{W(X)} = \begin{cases} [\mathrm{X} - \frac{1}{2}(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2)]^1\Sigma^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2) \text{ if } \Sigma \text{ is known} \\ \mathrm{X} - \frac{1}{2}(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2)]^1\hat{\Sigma}^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2) \text{ if } \Sigma \text{ is unknown} \end{cases} \quad (11)$$

and the decision rule $\hat{\boldsymbol{\varsigma}}$ reduces to

$$\mathrm{IfW(x)} = \begin{cases} > 0 \; assign \; x \; to \; \pi_1 \\ \leq 0 \text{ assign x to } \pi_2 \end{cases} \quad (12)$$

The optimum error rate is simply

$$\alpha(\boldsymbol{\varsigma}) = \phi(-\Delta/2) \quad (13)$$

Where

$$\phi(\mathrm{t}) = \int_{-\infty}^{\mathrm{t}} (2\pi)^{-\frac{1}{2}} \mathrm{e}^{-\mathrm{x}^{2/2}} \, \mathrm{dx}[5] \quad (14)$$

Conditional on the training samples (and therefore on $\bar{\mathrm{x}}_1, \bar{\mathrm{x}}_2$ and S), W(X) has a univariate normal distribution:

$$\mathrm{W(X)} \sim \begin{cases} \mathrm{N}\{[\mu_1 - \frac{1}{2}(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2)]^1\Sigma^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2), (\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)]^1\Sigma^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)\} \text{ if } \Sigma \text{ is known} \\ \mathrm{N}\{[\mu_1 - \frac{1}{2}(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2)]^1\mathrm{S}^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2), (\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)^1\mathrm{S}^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)\} \text{ if } \Sigma \text{ is unknown} \end{cases} \quad (15)$$

The conditional actual error rate is the probability that W(X) is less than or equal to zero and hence can be given as

$$\alpha(\hat{\boldsymbol{\varsigma}}) = \begin{cases} \phi\left[-\frac{[\mu_1 - \frac{1}{2}(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2)]^1\Sigma^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)}{[(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)^1\Sigma^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)]^{1/2}}\right] \text{ if } \Sigma \text{ is known} \\ \phi\left[-\frac{[\mu_1 - \frac{1}{2}(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2)]^1\mathrm{S}^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)}{[(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)^1\mathrm{S}^{-1}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)]^{1/2}}\right] \text{ if } \Sigma \text{ is unknown} \end{cases} \quad [8] \quad (16)$$

The expected actual error rate is more complicated. For Σ unknown, an asymptotic distribution of $E[\alpha(\hat{\boldsymbol{\varsigma}})]$ was given by [8], [9] and [14] used numerical integration to tabulate values of $\mathrm{E}[\alpha(\hat{\boldsymbol{\varsigma}})]$ for r=1…4 and $\mathrm{n}_1 = \mathrm{n}_2 = 25, 50, 100$. These results were compared and were found to be in close agreement.

In the Univariate case, r =1, the situation simplifies considerably. Equation (13) involves only $\Delta^2 = (\mu_1 - \mu_2)^2/\sigma^2$. Equation (16) reduces to

$$\alpha(\hat{\boldsymbol{\varsigma}}) = \begin{cases} \phi\{-[\mu_1 - \frac{1}{2}(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2)]/\sigma\} \text{ if } \bar{\mathrm{x}}_1 > \bar{\mathrm{x}}_2 \\ 1 - \phi\{-[\mu_1 - \frac{1}{2}(\bar{\mathrm{x}}_1 + \bar{\mathrm{x}}_2)]/\sigma\} \text{ if } \bar{\mathrm{x}}_1 \leq \bar{\mathrm{x}}_2 \end{cases} \quad (17)$$

3.5. Criteria for Comparing Error Rate Estimators

Let $\hat{\alpha}$ represent an arbitrary estimate of the conditional actual error rate, $\alpha(\hat{\boldsymbol{\varsigma}})$, based on the training samples. The most reasonable criteria for comparing estimators is felt to be

$$E[\hat{\alpha} - \alpha(\hat{\boldsymbol{\varsigma}})]^2 \quad (18)$$

called the Unconditional mean square error (UMSE) by [15]. Two other possible criteria are the conditional mean square error

$$E\left[\{\hat{\alpha} - \alpha(\hat{\boldsymbol{\varsigma}})\}^2 \middle| \hat{\theta}\right] \quad (19)$$

and the mean absolute error

$$E|\hat{\alpha} - \alpha(\hat{\boldsymbol{\varsigma}})| \quad (20)$$

The results obtained using the criterion of conditional mean square are functions of $\hat{\theta}$, this criterion could be used if

it were desirable to have the choice of the error rate estimator depend on the training sample. However, the goal of this study is to compare estimators chosen independently of the training samples. Therefore, UMSE, which is the expected value of the conditional mean square error over the distribution of $\hat{\theta}$, is the preferred criterion. The mean absolute error is also felt to be a reasonable criterion, but it is not considered further because it is not as sensitive to the variability of the error as the unconditional mean square error.

4. Error Rate Estimators

In this paper, we considered nine major error rate estimators namely; Plug-in estimator (D-method), Resubstitution estimator or Apparent error rate (R-method) and the leave-one-out estimator, (U-method).

4.1. Plug-in Estimator

This is the earliest error rate estimator proposed by [3]

$$\text{Let } D^2 = \begin{cases} (\bar{x}_1 - \bar{x}_2)^1 \Sigma^{-1} (\bar{x}_1 - \bar{x}_2) \text{ if } \Sigma \text{ is known} \\ (\bar{x}_1 - \bar{x}_2)^1 \hat{\Sigma}^{-1} (\bar{x}_1 - \bar{x}_2) \text{ if } \Sigma \text{ is unknown} \end{cases} \quad (21)$$

The plug-in estimate is defined by $\hat{\alpha}^{\Delta} = \phi(-\Delta/2)$
The probability of misclassification P is given by

$$P_1 = P\left[\frac{t<-\mu_1^1 s^{-1}(\bar{x}_1-\bar{x}_2)+\frac{1}{2}(\bar{x}_1+\bar{x}_2)^1 s^{-1}(\bar{x}_1-\bar{x}_2)}{\sqrt{(\bar{x}_1-\bar{x}_2)^1 s^{-1}\Sigma^{-1}s^{-1}(\bar{x}_1-\bar{x}_2)}}\right] \quad (22)$$

where T is a standard normal deviate. If we replace μ_1 and Σ by $\bar{x}_1$ and s we have that for normally distributed variables, the estimate of P_1 is

$$P_1 = P\left[\frac{t<-\bar{x}^1 s^{-1}(\bar{x}_1-\bar{x}_2)+\frac{1}{2}(\bar{x}_1+\bar{x}_2)^1 s^{-1}(\bar{x}_1-\bar{x}_2)}{\sqrt{(\bar{x}_1-\bar{x}_2)s^{-1}ss^{-1}(\bar{x}_1-\bar{x}_2)}}\right] \quad (23)$$

$$= P\left[\frac{t<-s^{-1}(\bar{x}_1-\bar{x}_2)+\frac{1}{2}(\bar{x}_1+\bar{x}_2)^1 s^{-1}(\bar{x}_1-\bar{x}_2)}{\sqrt{(\bar{x}_1-\bar{x}_2)s^{-1}ss^{-1}(\bar{x}_1-\bar{x}_2)}}\right] \quad (24)$$

$$= P\left[\frac{t<-\frac{1}{2}(\bar{x}_1-\bar{x}_2)^1(\bar{x}_1-\bar{x}_2)}{\sqrt{(\bar{x}_1-\bar{x}_2)^1 s^{-1}(\bar{x}_1-\bar{x}_2)}}\right] \quad (25)$$

$$= \phi\left[\frac{-D^{2/2}}{D}\right] = \phi\left[\frac{-D}{2}\right] \quad (26)$$

Also if we replace μ_2 and Σ by $\bar{x}_2$ and s in the case of P_2 then the estimate of P_2 is

$$P_2 = \phi\left[\frac{-D}{2}\right] \quad (27)$$

Where $D^2 = (\bar{x}_1 - \bar{x}_2)^1 s^{-1} (\bar{x}_1 - \bar{x}_2)$ is the Mahalanobis' sample distance. These estimates are good if the degrees of freedom are large since D^2 is consistent for δ^2. If the degrees of freedom are not large, this may be badly biased and give much too favourable an impression of the probability of error. Another way to derive this estimate is that since $P_1 = \phi\left(\delta/2\right)$ when the parameters are known, by estimating the parameters μ_1, μ_2 and Σ by $\bar{x}_1, \bar{x}_2$ and s we should arrive at reasonable results.

4.2. Resubstitution Estimator

The other commonly used error rate estimator is called the Resubstitution estimator, apparent error rate or the R-method. This is the proportion of the observations in the training sample from π_1 which is misclassified by the discriminant rule. In this method, the sample used to compute the discriminant function is reused to estimate the error rate. This means that if n_1 and n_2 are samples from population π_1 and π_2 respectively, then we use n_1 and n_2 to compute the discriminant function. If the number of misclassification on π_1 and π_2 are $m_1$and m_2, then the estimates of the error rate P_1 and P_2 are $\frac{m_1}{n_1}$ and $\frac{m_2}{n_2}$ respectively. Hence the Resubstitution error rate estimator of the Apparent error rate estimator (APER) is given by

$$\text{APER} = \frac{m_1+m_2}{n_1+n_2} \quad (28)$$

4.3. Leave-One-Out Estimator

In the leave-one-out estimator or procedure, all but one observation is used to complete the classification rule, and this rule is then used to classify the omitted observation. We repeat this procedure for each observation, so that in a sample of size $N = \Sigma i n_1$, each observation is classified by a function based on the N-1 observations. When $g = 2$, that is, two-fold cross-validation, this is the rotation method. When $g = n$, that is the n-fold cross-validation error estimator, R(cv), attributed to [6], where in the case of two populations $R(cv) = \Sigma_{i=1}^{2}\Sigma_{j=1}^{2} n_{ij} / n_i$. This method is also known as the "leave-one-out" or U estimate. Studies undertaken by numerous authors including [2] have shown that n-fold cross-validation has large variance. Thus, although R(cv) may be an Unbiased estimate, the confidence with which the user can expect R(cv) for his/her sample to approach. R(T) is not great. The main advantage of this method is felt to be that it obtains an unbiased estimate of the expected actual error rate for a discrimination problem with training samples of size $n_1 - 1$ and n_2[6]. However, this does not mean that the leave-one-out estimator has small bias with respect to the conditional actual error rate, which is the error rate of interest here. One disadvantage of this estimator is that it requires more computation then the resubstitution estimator. However, ways have been found to reduce this problem. Another disadvantage of the leave-one-out estimate is its large variance. The main consideration of most investigators when comparing estimators has been the bias, but the variance is also important factor. [4] Performed a sampling experiment in order to demonstrate the importance of the variance. In the Univariate normal case, he found that the bias with respect to $E[\alpha(\hat{\varsigma})]$ is very small for the leave-one-out estimator, larger for the plug in estimator and largest for the resubstitution estimator, as expected. However, he also compared the variance of the estimators and found that the leave-one-out estimator had a much larger variance than the resubstitution estimator, which in turn had a larger variance than the plug-in-estimator. Unfortunately, Glick did not consider the mean square error and hence, left Unanswered whether the resubstitution estimator over performs better than leave-one-out.

4.4. Jackknife Error Rate Estimator

This method was due to [13]. The method involves omitting each observation in turn from the learning sample and to obtain the apparent error rate for the learning sample with the jth observation omitted, $R_j^*(A)$, so that

$$R_j^*(A) = \frac{1}{n}\sum_{j=1}^{n} R_j^*(A). \quad (29)$$

So that $\hat{w}_j$, the Jacknife estimate of the bias of R(A) is $\hat{w}_j = (n-1)\left[R_j^*(A) - R(A)\right]$ leading to the Jacknife estimate of the error rate

$$R(J) = nR(A) - (n-1)R_j^*(A). \quad (30)$$

4.5. The DS Method Estimator

This estimator DS method is based on the plug-in estimator which assumes multivariate normality and contains a bias correction. When Σ is unknown, D^2 is a biased estimator of Δ^2. [7] described a consistent estimator of Δ^2 which has less bias than D^2. This estimator of Δ^2 is

$$D^2 = (n_1 + n_2 - r - 3)D^2 / (n_1 + n_2 - 2) \quad (31)$$

and hence the estimator of $\alpha(\hat{\varsigma})$, called the DS method is

$$\alpha^{DS} = \phi\left\{-(DS)^{\frac{1}{2}}/_2\right\} \quad (32)$$

4.6. The O and OS Estimators

The distribution of Anderson's W statistic is very complicated and is not known exactly. [11] Provided an asymptotic expansion for $\Pr\left\{W(x) < \frac{1}{2}\Delta^2 + a\Delta\right\}$ where a is a real constant. Since $\alpha(\hat{\varsigma}) = \Pr\{W(x) < 0\}$, one could substitute an estimate of Δ^2 into Okamoto's expansion in order to estimate $\alpha(\hat{\varsigma})$. [7] Suggested two such estimators: the O method is obtained by replacing Δ^2 with D^2, and the OS method is obtained by replacing D^2 with DS. These estimators were explicitly obtained in the Univariate case with δ^2 known by [15]:

$$\hat{\alpha}^{O} = \phi\{-|\bar{x}_2 - \bar{x}_1|2\delta\} + \frac{1}{8}(n_1^{-1} + n_2^{-2})\phi^1\{|\bar{x}_2 - \bar{x}_1| / 2\delta\} \quad (33)$$

$$\hat{\alpha}^{OS} = \phi\left\{\frac{|\bar{x}_2 - \bar{x}_1|(n_1+n_2-4)^{\frac{1}{2}}}{2\delta(n_1+n_2-4)^{\frac{1}{2}}}\right\} + \frac{1}{8}(n_1^{-1} + n_2^{-2})\,\phi^1\left\{\frac{|\bar{x}_2 - \bar{x}_1|(n_1+n_2-4)^{\frac{1}{2}}}{2\delta(n_1+n_2-4)^{\frac{1}{2}}} - \right\} \quad (34)$$

Where

$$\phi^1(t) = -t(2\pi)^{-\frac{1}{2}}\exp\left(-t^2/_2\right) \quad (35)$$

4.7. Posterior Probability Estimator

This estimator was described by [10]. Assuming equal prior probabilities, if θ is known and the discriminant rule is **ς**, the posterior probability of misclassification is

$$[\min\{f(x,\theta_1), f(x,\theta_2)\}] / [f(x,\theta_1) + f(x,\theta_2)] \quad (36)$$

when θ is estimated, the posterior probability of misclassification by the rule $\hat{\varsigma}$, given x_i is estimated by

$$\left[\min\left\{f(x_i,\hat{\theta}_1), f(x_i,\hat{\theta}_2)\right\}\right] / \left[f(x_i,\hat{\theta}_1) + f(x_i,\hat{\theta}_2)\right] \quad (37)$$

This function is evaluated for each of the x_i and the mean is the estimator of $\alpha(\hat{\varsigma})$.

5. The Simulation Experiments and Results

In this comparative study, some existing estimators are compared using Monte Carlo Simulations. The usefulness of a Monte Carlo assessment is that the population parameters and the true distribution from which the training data are obtained are known. Thus, the true error rates can always be computed. Hence, the estimated error rates can be compared with the true error rate for choosing the best estimator.

The eight estimators' procedures are evaluated at each of the 118 configurations of n, r and d. The 118 configurations of n, r and d are all possible combinations of n=40, 60, 80, 100, 200, 300, 400, 600, 700, 800, 900, 1000, r=3, 4, 5 and d = 0.1, 0.2, 0.3, and 0.4. A simulation experiment which generates the data and evaluates the procedures is now described.

(i). A training data set of size n is generated via R-program where $n_1 = {}^{n}/_{2}$ observations are sampled from π_1 which has multivariateb Bernoulli distribution with input parameter p_1 and p_2 and $n_2 = {}^{n}/_{2}$ observations sampled from π_1, which is multivariate Bernoulli with input parameter $p_2, j = 1 \ldots r$. These samples are used to construct the various estimators.

(ii). The likelihood ratios are used to define classification rule. The estimators of error rates are determined for each of the methods.

(iii). Step (i) and (ii) are repeated 1000 times and the mean error rate and variances for the 1000 trials are recorded.

The following table contains a display of one of the results obtained.

Table 1. *Mean error rates for estimators under different parameter values, sample sizes and Replications.*

P_1 = (.5,.5,.5,.5,.5) P_2 = (.6,.6,.6,.6,.6)

Sample sizes	DS	R	U	P	JK	D	O	OS
40	0.365212	0.237006	0.254587	0.252975	0.251887	0.494812	0.382087	0.362220
60	0.376908	0.278807	0.287591	0.286358	0.287816	0.500316	0.393791	0.375385
100	0.389975	0.316222	0.323300	0.324600	0.323815	0.500990	0.401335	0.384240

Sample sizes	DS	R	U	P	JK	D	O	OS
140	0.393925	0.336808	0.342721	0.343307	0.343153	0.501775	0.406560	0.396101
200	0.4007250	0.355295	0.359190	0.359465	0.359842	0.499727	0.411195	0.398143
300	0.402866	0.370199	0.373166	0.373318	0.373128	0.499428	0.412693	0.402204
400	0.404201	0.379041	0.382187	0.381760	0.381406	0.500437	0.414523	0.402156
600	0.405495	0.386957	0.389576	0.389626	0.389663	0.500395	0.415382	0.403902
700	0.406001	0.390346	0.392677	0.392478	0.391590	0.499647	0.416030	0.403770
800	0.406843	0.392932	0.394805	0.394511	0.394905	0.500325	0.416420	0.405535
900	0.406832	0.394140	0.395937	0.396352	0.396006	0.500902	0.416912	0.404521
1000	0.407625	0.395220	0.397488	0.396858	0.397428	0.499799	0.417174	0.405044

p (mc) = 0.16308

Table 2. *Standard error for the estimator rules under different parameter values, sample sizes and replications.*

P_1 = (.5,.5,.5,.5,.5) P_2 = (.6,.6,.6,.6,.6)$|p(mc) - \hat{p}(mc)|$

Sample sizes	DS	R	U	P	JK	D	O	OS
40	0.047146	0.069485	0.046432	0.047218	0.045033	0.059599	0.0451534	0.074752
60	0.040174	0.059342	0.040172	0.041581	0.040481	0.047988	0.0392	0.060813
100	0.031479	0.049585	0.034399	0.033562	0.033471	0.039786	0.030936	0.047585
140	0.026298	0.042871	0.028142	0.029753	0.028542	0.037205	0.026595	0.040519
200	0.023616	0.038008	0.026386	0.027106	0.025865	0.031779	0.023459	0.03599
300	0.019186	0.031847	0.022209	0.022355	0.022309	0.029105	0.019309	0.028217
400	0.016343	0.029209	0.019176	0.019488	0.018798	0.026048	0.01636	0.023954
600	0.013147	0.023879	0.016622	0.01523	0.015892	0.02399	0.013488	0.019303
700	0.012653	0.02271	0.015258	0.015375	0.015703	0.02423	0.012725	0.019036
800	0.012157	0.021352	0.014518	0.014808	0.0145799	0.023763	0.012257	0.01706
900	0.010951	0.021304	0.014157	0.014231	0.013759	0.023136	0.011209	0.016578
1000	0.010528	0.019691	0.013182	0.012785	0.013094	0.023139	0.010844	0.015555

Tables 1 and 2 present the mean error rates and sum of square error rates for estimators under different parameter values. The mean error rates increases with the increase in sample sizes and sum of square error decreases with the increase in sample sizes. From the analysis, DS is ranked first, followed by O, OS, U, R, JK, P, and D came last.

Estimators	Position
DS	1
O	2
OS	3
U	4
R	5
JK	6
P	7
D	8

6. Conclusion

We obtained two major results from this study. Firstly, using the simulation experiments we ranked the estimators as follows: DS, O, OS, U, R, JK, P and D. The best method was the DS estimator. Secondly, we concluded that it is better to increase the number of variables because accuracy increases with increasing number of variables. Also, the general trend for the estimators was an increase in error rate as sample size decreases while decreasing the distance between populations generally increase the error rate. DS estimator was the most consistent and thus reliable over all combinations of probability pattern and sample sizes.

References

[1] Anderson, T. W. (1951), Classification by Multivariate analysis, *Psychometric, 16*, 631-650.

[2] Efron, B. (1983), Estimating the error rate of a prediction rule: improvement on cross validation. *Journal of the American Statistical Association, 78,* 316-331.

[3] Fisher, R. A. (1936). The use of multiple measurements in taxanomic problem. *Annals of Eugenics, 7,* 179-188.

[4] Glick, N. (1978), Additive estimators for probabilities of correct classification. *Pattern Recognition, 10,* 211-222.

[5] John, N. (1961) "Errors in discrimination" *Annals of Mathematical Statistics, 32*, 1125-1144

[6] Lachenbruch, P. A. (1967), an almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics, 23,* 639-645.

[7] Lachenbruch, P. A. & Michey, M. R. (1968), Estiamtion of error rates in discriminant analysis, *Technometrics, 10*, 1-11.

[8] McLachlan, G. J. (1972), An Asymptotic Unbiased Techniques.

[9] McLachlan, G. J. (1974)," The Asymptotic Unbiased distribution of the conditional error rate and risk in Discriminant Analysis", *Biometrics 61,* 239-249.

[10] Moore, D. H. (1973) "Evaluation of five Discriminant procedures for binary variables' *Journal of the American Statistical Association, 68,* 399-404.

[11] Okamoto, M. (1963), An Asymptotic Expansion for distribution of linear Discriminant function, *Ann Math Stat, 34,* 1286-1301.

[12] Okamoto, M. (1971) "Correction to the Asymptotic expansion for distribution of the linear Discriminant function" *Annals of Mathematical Statistics 39,* 1358-1359.

[13] Quenouille, M. (1949), Approximate tests of correlation in time series. *Journal of the Royal Statistical Society Series B, 11,* pp 18-84.

[14] Sayre, J. W. (1980) "The distributions of the actual error rates in linear Discriminant Analysis". *Journal of American Statistical Association, 75,* 201-205.

[15] Sedranski, N. &Okamoto, M. (1971) "Estimation of the probabilities of misclassification for a linear Discriminant function in the Univariate normal case. *Annals of the Institute of Statistical Mathematics, 23,* 419-435.

[16] Lachenbruch, P. & Mickey, M. (1968) "Estimation of error rates in discriminant analysis". *Technometrics, vol 10,* pp 167-178.

[17] Devijver, P. A. & Kittler, J. (1982). Pattern Recognition: A Statistical approach, Englenood cliffs, NJ: Prentice-Hall international.

[18] Efron, B. & Gong, G. (1983). Estimating the error rate of prediction rule, Improvement on Cross validation. *Journal of American Statistical Association, vol 78*, pp 316-331.

[19] Dongherty, E. R. & Braga-Neto, U. M. (2006). Epistemology of computational Biology: Mathematical models and Experimental prediction as the Basis of their validity. *Biological Systems, vol 14 no. 1,* pp 65-90.

[20] Vishwa Nath Maurya; Madaki, U. Y.; Vijay, V. S. 7 Babagana, M. (2015). Application of Discriminant Analysis onb Broncho-pulmonary Dysplasia among infants: A case study of UMTH and UDUS Hospitals in Maiduguri, Nigeria. *American Journal of Theoretical and Applied Statistics, 4 (2-1):* 44-51.

[21] Vishwa N. M.; Ram, B. M.; Chandra, K. J. & Avadhesh, K. M. (2015). Performance analysis of powers osskewness and kurtosis based multivariate normality tests and use of estended Monte Carlo Simulation for proposed novelty algorithm. *American Journal of Theoretical and Applied Statistics, 4 (2-1):* 11-18.

[22] Egbo, I.; Onyeagu, S. I.; Ekezie, D. D. & Uzoma, P. O. (2014). A comparison of the optimal classification Rule and maximum likelihood Rule for Binary Variables. *Journal of Mathematics Research, vol 6 No. 4*.

[23] Egbo, I.; Onyeagu, S. I. & Ekezie, D. D. (2014). A comparison of multinomial classification Rules for Binary variables. *International Journal of Maths. Sci. & Eng. Appls., vol 8 No V*.

[24] Egbo, I.; Egbo, M. &Onyeagu, S. I. (2015). Performance of Robust linear classifier with multivariate Binary variables. *Journal of Mathematics Research, vol 7 No 4.*

[25] Egbo, I. (2015). Discriminant analysis procedures under non-optimal conditions for Binary variables. *American Journal of Theoretical and Applied Statistics, 4 (6):* 602-609.