American Journal of Theoretical and Applied Statistics
2015; 4(6): 602-609
Published online December 9, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150406.32
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Discriminant Analysis Procedures Under Non-optimal Conditions for Binary Variables

I. Egbo

Department of Mathematics, Alvan Ikoku University of Education, Owerri, Nigeria

Email address:
Egboike@gmail.com, Egboikechukwuj@yahoo.com

To cite this article:
I. Egbo. Discriminant Analysis Procedures Under Non-optimal Conditions for Binary Variables. *American Journal of Theoretical and Applied Statistics*. Vol. 4, No. 6, 2015, pp. 602-609. doi: 10.11648/j.ajtas.20150406.32

Abstract: The performance of four discriminant analysis procedures for the classification of observations from unknown populations was examined by Monte Carlo methods. The procedures examined were the Fisher Linear discriminant function, the quadratic discriminant function, a polynomial discriminant function and A-B linear procedure designed for use in situations where covariance matrices are equal. Each procedure was observed under conditions of equal sample sizes, equal covariance matrices, and in conditions where the sample was drawn from populations that have a multivariate normal distribution. When the population covariance matrices were equal, or not greatly different, the quadratic discriminant function performed similarly or marginally the same like Linear procedures. In all cases the polynomial discriminate function demonstrated the poorest, linear discriminant function performed much better than the other procedures. All of the procedures were greatly affected by non-normality and tended to make many more errors in the classification of one group than the other, suggesting that data be standardized when non-normality is suspected.

Keywords: Apparent Error Rates, Fisher's Linear Discriminant, Quadratic Discriminant Function, A-B Discriminant Function, Polynomial Discriminant Function

1. Introduction

Many practical problems can be reduced to the assignment of various objects to different classes. For example in the case of the medical diagnosis, it is a question of recognizing the pathology of a given patient, the purposes correspond to the patients and the classes with various pathologies. In the economy field, a bank wants to know if a customer applying for a loan is a good or bad customer while being based on several variables like the age, the profession, former fidelity, the required credit. A review of these appears in [16]. In assignment problems in biomedical research, one or more of these techniques is often used. The assumptions underlying these techniques are not always evident to the user, nor are the consequences of their violation. The assumptions include multivariate normality, common covariance matrices and correct assignment of the initial groups [17], [18] and [19]. While a good deal is known in the two group situation, the robustness of these procedures under non-optimal conditions for binary variable is essentially unknown. The purpose of this paper is to compare and delineate these problems systematically and to suggest useful areas of research.

The problem of classifying an individual into one of two concerned groups (called populations), arises in many areas, typically in anthropology, education, psychology, medical diagnosis, biology, engineering, etc. An anthropometrician may wish to identify ancient human remains in two different racial groups or in two different time periods by measuring certain skull characters [2]. A plant breeder discriminates a desired from an undesirable species by observing some heritable characters [14]. A company hires or rejects an applicant frequently based on a certain measurement. Similarly a college accepts or denies a prospective student usually based on his entrance examination scores. In a hospital, a patient maybe diagnosed and classified into a certain potential disease group by a battery of tests, usually it is assumed that there are two populations, say π_1 and π_2, the individual to be classified comes from either π_1 or π_2; furthermore, it is assumed that from previous experiments or records we have in our possession the characteristic measurements of n_1 individuals who were known to belong to π_1, and of n_2 individuals who were known to belong to π_2. Based on the available data obtained from previous $n_1 +$

n_2 individuals and the corresponding characteristic measurements of a new individual, we would like to classify the new individual into either π_1 or π_2 by using certain criterion. The case of more than two populations will not be considered in this paper.

In this inferential setting, the researcher can commit one of the following errors. An object from π_1 may be misclassified into π_2 likewise, an object from π_2 may be misclassified into π_1. If misclassification occurs a loss would be suffered. Let $C(i/j)$ be the cost of misclassifying an object π_j, into π_i. For the two population setting, we have that $C(2/1)$ means cost of misclassifying an object into π_2 given that it is from π_1.

$C(1/2)$ is the cost of misclassifying an object into π_1 given that it is from π_2. The relative magnitude of the loss $L(j,i) = C(i/j)$ depends on the case in question: for example failure to detect an early cancer in a patient is costlier than stating that a patient has cancer and discovering otherwise.

2. Classification Procedures

2.1. The Fisher's Linear Discriminat Function (FLDF Rules)

The linear discriminant function for discrete variables is given by

$$\hat{L}(x) = \sum_i \sum_k (\hat{p}_{2i} - \hat{p}_{1i}) s^{kj} x_k - \frac{1}{2} \Sigma\Sigma (\hat{p}_{2i} - \hat{p}_{1i}) s^{kj} (\hat{p}_{2k} + \hat{p}_{1k}) \quad (1)$$

Where s^{kj} are the element of the inverse of the pooled sample covariance matrix p_{1i} and $\hat{p}_{2j}$ are the elements of the sample means in π_1 and π_2 respectively. The classification rule obtained using this estimation is classify an item with response pattern X into π

If $\sum_i \sum_k (\hat{p}_{2i} - \hat{p}_{1i}) s^{kj} x_k - \frac{1}{2} \Sigma\Sigma (\hat{p}_{2i} - \hat{p}_{1i}) s^{kj} (\hat{p}_{2k} + \hat{p}_{1k}) > 0$ and to π_2 or otherwise (2)

2.2. The Quadratic Discriminant Function

When an observation vector $\underline{x}$, is drawn from a MVN distribution with mean vector μ_I and covariance matrix $\underline{\Sigma}_I$, the MVN density function f ($\underline{x}$), can be expressed as:

$$f_i(\underline{x}) = (2\pi)^{-p/2} |\Sigma_i|^{-1/2} \exp\left[-\tfrac{1}{2}(\underline{x} - \underline{\mu}_i)' \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i)\right] \quad (3)$$

In the case of two groups an individual is classified as belonging to population 1 if $p_1 f_1(\underline{x}) / p_2 f_2(\underline{x}) > 1$, that is, if $f_1(\underline{x}) / f_2(\underline{x}) > p_2/p_1$. Alternatively, an individual is assigned to population 2 if $p_1 f_1(\underline{x}) / p_2 f_2(\underline{x}) \leq 1$, that is, if $f_1(\underline{x})/f_2(\underline{x}) \leq p_2/p_1$. Where p_1 and p_2 are the proportions of individuals from the two groups in the populations [7].When the two groups have a common covariance matrix, $\underline{\Sigma}$, and mean vectors $\underline{\mu}_1$ and $\underline{\mu}_2$ the above rule becomes

$$f_1(\underline{x})/f_2(\underline{x}) = \exp\left[-\tfrac{1}{2}(\underline{x} - \underline{u}_1)' \Sigma^{-1} (\underline{x} - \underline{u}_1) + \tfrac{1}{2}(\underline{x} - \underline{u}_2)' \Sigma^{-1} (\underline{x} - \underline{u}_2)\right] \quad (4)$$

$$= \exp\left[x' \Sigma^{-1} (\underline{u}_1 - \underline{u}_2) - \tfrac{1}{2}(\underline{u}_1 + \underline{u}_2)' \Sigma^{-1} (\underline{u}_1 - \underline{u}_2)\right]$$

And taking logarithms, the rule is to assign an individual to population 1 if

$$D_t(\underline{x}) = In[f_1(\underline{x})/f_2(\underline{x})] = \left[(\underline{x}) - \tfrac{1}{2}(\underline{u}_1 + \underline{u}_2)\right]' \Sigma^{-1} (\underline{u}_1 - \underline{u}_2) > In(p_2/p_1) \quad (5)$$

And to the group 2 otherwise. The sample analogue of the above equation is

$$D_s(\underline{x}) = [\underline{x} - \tfrac{1}{2}(\overline{\underline{x}}_1 + \overline{\underline{x}}_2)]' s^{-1} (\overline{\underline{x}}_1 - \overline{\underline{x}}_2) \quad (6)$$

And the coefficients $s^{-1}\begin{pmatrix} - & - \\ x_1 - x_2 \end{pmatrix}$ are seen to be identical to Fisher's result for the LDF.

When covariance matrices are unequal and cannot be pooled, but the population distributions are multivariate normal, the classification rule has the form

$$Q_t(\underline{x}) = in\,[f_1(\underline{x})/f_2(\underline{x})] > In\,[p_2/p_1] \quad (7)$$

$$= \tfrac{1}{2} In\left|\frac{\Sigma_2}{\Sigma_1}\right| - \tfrac{1}{2}(\underline{x}-\underline{u}_1)'\Sigma^{-1}(\underline{x}-\underline{u}_1) + \tfrac{1}{2}(\underline{x}-\underline{u}_2)'\Sigma_2^{-1}(\underline{x}-\underline{u}_2)\, s^{-1}\left(\bar{x}_1 - \bar{x}_2\right)$$

$$= \tfrac{1}{2}\left[In\left|\frac{\Sigma_2}{\Sigma_1}\right| - \underline{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\underline{x} + 2\underline{x}'(\Sigma_1^{-1}\underline{u}_1 - \Sigma_2^{-1}\underline{u}_2) - \underline{u}[\,\Sigma_1^{-1}\underline{u}_1 + \underline{u}_2'\Sigma_2^{-1}\underline{u}_2\right] \quad (8)$$

In these cases, the discriminant function is quadratic, since the term $\Sigma_1^{-1} - \Sigma_2^{-2}$ is still present [7]. From the above with $\underline{u}_1$, $\underline{u}_2, \Sigma\Sigma_1$ and Σ_2 estimated bytheir respective mean vectors and covariance matrices $\bar{\underline{x}}_1$, $\bar{\underline{x}}_2$, s_1 and s_2 the sample analogue of $Q_t(\underline{x})$ is

$$Q_s(\underline{x}) = In\left(\left|s_1/s_2\right|\right) - (\underline{x}-\bar{\underline{x}}_1)'s_1^{-1}(\underline{x}-\bar{\underline{x}}_1) + (\underline{x}-\bar{\underline{x}}_2)'S_2^{-1}(\underline{x}-\bar{\underline{x}}_2) > 2In(p_2/p_1) \quad (9)$$

In each of the conditions of the present study the proportions of each group is the population were assumed to be equal to each other and not proportional to sample size since the true proportion are not usually know in most areas of psychological research. When population proportions are equal the quadratic decision rule is the to classify an individual into population 1 if $Q_s(\underline{x}) > 0$ or into population 2 if $Q_S(\underline{x}) \le 0$ since $In(p_2/p_1) = 0$

2.3. The A-B Discriminant Function

[1] proposed a Linear discriminant function of the form $\underline{b}'\underline{x}$, with $\underline{b}' = [b_1, \ldots, b_p]$ chosen so that $\underline{x}$ is classified as from population 1 if $\underline{b}'\underline{x}, >$ c and from population 2 if $\underline{b}'\underline{x}, \le c$, where c is also suitably determined. With this procedure, the misclassification probabilities are:

$$P_1 = \Pr ob.(\underline{b}'\underline{x} \le c \,|\, \underline{x} \;\varepsilon\; pop\ 1) = 1 - \Phi(y_1), \quad \text{And}$$

$$\boldsymbol{P}_2 = \Pr \boldsymbol{ob}.(\underline{\boldsymbol{b}}'\underline{\boldsymbol{x}} > \boldsymbol{c} \,|\, \underline{\boldsymbol{x}} \;\boldsymbol{\varepsilon}\; \boldsymbol{pop}\ 2) = 1 - \boldsymbol{\Phi}(\boldsymbol{y}_2), \quad (10)$$

Where Φ is the cumulative distribution function of a standard normal variable. The y_1 and y_2 are determined by

$$y_1 = \frac{c - \underline{b}'\underline{u}_1}{(\underline{b}'\Sigma_1\underline{b})^{1/2}} \quad \text{and} \quad y_2 = \frac{c - \underline{b}'\underline{u}_2}{(\underline{b}'\Sigma_2\underline{b})^{1/2}} \quad (11)$$

Where $\underline{u}_1$ and $\underline{u}_2$ are the means of population 1 and population 2, Now y_1 can be expressed as"

$$y_1 = \frac{-\underline{b}'(\underline{u}_1 - \underline{u}_2) - y_2(\underline{b}'\Sigma_2\underline{b})^{1/2}}{(\underline{b}'\Sigma_1\underline{b})^{1/2}}, \quad (12)$$

The $\underline{b}$ is then chosen which maximizes y_1 for a given y_2. By differentiating y_1 withrespectto $\underline{b}$. It can be show that the solution consists of solving the following equations in $\underline{b}$ and a scalar t:

$$[\,t\Sigma_1 + (1-t)\Sigma_2]\underline{b} = (\underline{u}_1 - \underline{u}_2), \text{ and}$$

$$y_2 = (1-t)(\underline{b}'\Sigma_2\underline{b})^{1/2} \quad (13)$$

The solution to these equations is obtained by a trial – and-error procedure and c is then obtained by:

$$c = \underline{b}'\underline{u}_1 + t\underline{b}'\Sigma_1\underline{b}' = \underline{b}'\underline{u}_2 - (1-t)\underline{b}'\Sigma_2\underline{b} \quad (14)$$

Now y_1 can be obtained from

$$y_1 = t(\underline{b}'\Sigma_1\underline{b})^{1/2} \quad (15)$$

[1] also considered an alternative method when the two misclassification probability are equal, i.e. $y_1 = y_2$. In this case, $\underline{b}$ and *t*are found from:

$$0 = y_1^2 - y_2^2 = t^2\underline{b}'\Sigma_1\underline{b} - (1-t)^2\underline{b}'\Sigma_2$$

$$= \underline{b}'[t^2\Sigma_1 - (1-t)^2\Sigma_2]. \quad (16)$$

The determination of the value of t was accomplished by using the result due to [3], in which Σ_1 *and* Σ_2 were expressed as:

$$\Sigma_1 = N'\Lambda N, \text{ and}$$

$$\Sigma_2 = N'N, \quad (17)$$

Where $\underline{\Lambda} = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_p$ *and* $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the roots of the determinantalequation

$\left|\Sigma_1 - \lambda \Sigma_2\right| = 0$. *if* $\vee = (1-t^2)/t^2$, then $\vee$ must lie between the minimum and maximum roots of the above characteristic equation.

In the present study the optimal value of t was approximated by evaluating t for $\vee$ equal to the minimum and maximum characteristic roots and computing the vector $\underline{b}$ in each case from the equation (13). The value of c was then calculated from Equation (14). And the observation population 1 if $\underline{b}' \underline{x} > c$ or population 2 if $\underline{b}' \underline{x} < c$. In this manner, the interval $[\min,\lambda_1, \max \lambda_1]$ was successively bisected, and for reach value of t, the proportion of correct classification calculated. The interval was bisected a maximum of five times or until classification did not improve. The resultant discriminant function was then applied to the cross validation sample, and the proportion of correct classifications was calculated.

2.4. The Polynomial Discriminant Function

In this case, the discriminant function was constructed by estimating the probability density function for each sample directly from the observed data, as described in [15]. This was accomplished by expanding the estimate $\hat{p}(\underline{x})$, *of* $p(\underline{x})$ in in a series which represents the probability density function of the *ith* population, $p(\underline{x})/pop_1)$. Tou and Gonzalez show that if it is required that the estimate of the probability density function minimize a mean –square error function defined as:

$$R = \int_{\underline{x}} w(\underline{x}) \, [p(\underline{x}) - \hat{\mathrm{P}}(\underline{x})]^2 \, d\underline{x} \, , \qquad (18)$$

Where w $(\underline{x})$ is a waiting function, then $p(\underline{x})$ may be expanded in the series

$$\hat{\mathrm{P}}(\underline{x}) = \sum_{j=1}^{m} c_j \, \Phi_j \, (\underline{x}) \qquad (19)$$

Where the c_j are coefficient to be determined, and the $\{\Phi_j \, (\underline{x})\} m_1$ are a set of specified basis functions.

A set of univariate basis functions associated with the normal distribution from which multivariate basis functions can be obtained, are Hermite polynomials, $\mathrm{H}_n(x)$, generated by the recursive relation

$$\mathrm{H}_{\mathrm{K}+1}(x) - 2x\,\mathrm{H}_k(x) + 2\mathrm{K}\mathrm{H}_{\mathrm{K}-1}(x) = 0, \mathrm{K} \geq 1 \qquad (20)$$

Where $\mathrm{H}_0(x) = 1$ *and* $\mathrm{H}_1(x) = 2x$. The first few Hermite polynomials are: $\mathrm{H}_0(x) = 1$:

$$\begin{aligned} \mathrm{H}_1(x) &= 2x; \\ \mathrm{H}_2(x) &= 4x^2 - 2; \\ \mathrm{H}_3(x) &= 8x^3 - 12x; \\ \mathrm{H}_4(x) &= 16x^4 - 48x^2 + 12. \end{aligned} \qquad (21)$$

Substituting the expansion of $\mathrm{P}(x)$ into the mean-square error function yields

$$R = \int_{\underline{X}} w(\underline{x}) \, [\mathrm{P}(\underline{x}) - \sum_{j=1}^{m} c_j \, \Phi_j \, (\underline{x})]^2 \, d\underline{x}, \qquad (22)$$

And minimizing R with respect to the coefficient, $\partial R/\partial C_{\mathrm{K}} = 0$, K $= 1, \ldots, m$ yields.

$$\sum_{j=1}^{m} c_j \int_{\underline{x}} w(\underline{x}) \, \Phi_j \, (\underline{x}) \Phi_{\mathrm{K}} \, (\underline{x}) \, d\underline{x} = \int_{\underline{x}} w(\underline{x}) \Phi_k \, (\underline{x}) \mathrm{P} \, (\underline{x}) \, d\underline{x}. \qquad (23)$$

The right side of this equation is the definition of the expected value of the function $w(\underline{x}) \, \Phi_{\mathrm{K}} \, (\underline{x})$ and may be approximated from the sample average

$$\sum_{j=1}^{m} c_j \int_{\underline{x}} w(\underline{x}) \, \Phi_j \, (\underline{x}) \Phi_{\mathrm{K}} \, (\underline{x}) \, d\underline{x} = \frac{1}{\mathrm{N}} \sum_{i=1}^{n} w(\underline{x})_1 \, \Phi_{\mathrm{K}} \, (\underline{x}_i) \qquad (24)$$

Since the basic functions $\{\Phi_j \, (\underline{x})\}$ are orthonormal and are chosen orthogonal with respect to the weighting function $w(\underline{x})$, the coefficients may be determined from

$$c_{\mathrm{K}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{n} \Phi_{\mathrm{K}} \, (\underline{x}_i) \, , \; \mathrm{K} = 1 \, , \, 2 \, , \ldots , \, m \qquad (25)$$

And the resultant density may be obtained from

$$\hat{\mathrm{P}} \, (\underline{x}) = \sum_{j=1}^{m} c_j \, \Phi_j \, (\underline{x}) \, , \qquad (26)$$

By using Bayes' formula

$$\mathrm{P} \, (Pop_1 | \underline{x}) = \frac{\mathrm{P} \, (Pop_1) \; \mathrm{P} \, (\underline{x} | Pop_1)}{\mathrm{P} \, (\underline{x})} \qquad (27)$$

Where $\mathrm{P}(Pop_i)$ is the probability of the *ith* population, the discriminant function for this problem are then given by:

$$d_1 \, (\underline{x}) = \hat{\mathrm{P}}(\underline{x} | Pop_1) \; \mathrm{P}(Pop_1), \text{and} \qquad (28)$$

$$d_2 \, (\underline{x}) = \hat{\mathrm{P}}(\underline{x} | Pop_2) \; \mathrm{P}(Pop_2), \qquad (29)$$

And if $\mid \mathrm{P}(Pop_1), = \mathrm{P}(Pop_2)$, the decision boundary is given by $d_1 \, (\underline{x}) - d_2 \, (\underline{x}) = 0$.

In the present study a two-dimensional set of orthogonal

function was obtained by forming pairwise combinations of the one-dimensional functions. Six terms were used to appreciate the density function and were constructed as follows:

$$\Phi_1(\underline{x}) = \Phi_1(x_1, x_2) = H_0(x_1)\, H_0(x_2) = 1;$$

$$\Phi_2(\underline{x}) = \Phi_2(x_1, x_2) = H_1(x_1)\, H_0(x_2) = 2x_1;$$

$$\Phi_3(\underline{x}) = \Phi_3(x_1, x_2) = H_0(x_1)\, H_1(x_2) = 2x_2;$$

$$\Phi_4(\underline{x}) = \Phi_4(x_1, x_2) = H_1(x_1)\, H_1(x_2) = 4x_1 x_2 \quad (30)$$

$$\Phi_5(\underline{x}) = \Phi_5(x_1, x_2) = H_2(x_1)\, H_0(x_2) = 4x_1^2 - 2;$$

$$\Phi_6(\underline{x}) = \Phi_6(x_1, x_2) = H_0(x_1)\, H_2(x_2) = 4x_2^2 - 2;$$

The set of original functions for the six-variable case was constructed in the same manner as for the bivariate case by forming the product of one dimensional Hermite polynomials. In order for the estimates of the density functions to be polynomials of degree two for all the variables, 28 terms were constructed as follow:

$$\Phi_1(x_1, \ldots, x_6) = H_0(x_1)\, H_0(x_2) \ldots H_0(x_6) = 1;$$

$$\Phi_2(x_1, \ldots, x_6) = H_1(x_1)\, H_0(x_2) \ldots H_0(x_6) = 2x_1;$$

$$\vdots$$

$$\Phi_7(x_1, \ldots, x_6) = H_0(x_1)\, H_0(x_2) \ldots H_1(x_6) = 2x_6;$$

$$\Phi_8(x_1, \ldots, x_6) = H_1(x_1)\, H_1(x_2) H_0(x_3 \ldots H_0(x_6) = 4x_1x_2;$$

$$\vdots$$

$$\Phi_{22}(x_1, \ldots, x_6) = H_0(x_1) \ldots H_0(x_4) H_1(x_5) H_1(x_6) = 4x_5x_6;$$

$$\Phi_{23}(x_1, \ldots, x_6) = H_2(x_1)\, H_0(x_2) \ldots H_0(x_6) = 4x_1^2 - 2;$$

$$\vdots$$

$$\Phi_{28}(x_1, \ldots, x_6) = H_0(x_1)\, H_0(x_2) \ldots H_0(x_5) H_2(x_6) = 4x_6^2 - 2; \quad (31)$$

The vector of coefficients c, was then computes for each sample from equation (25), and the polynomial estimates of the density functions were constructed as in Equation (26). The two estimates of the density functions were the subtracted to form the polynomial discriminant function, which was then applied to the observations in each of the original and cross-validation samples. Finally, the proportion of correct classification was calculated.

2.5. Testing Adequacy of Discriminant Coefficient

Consider the discriminant problems between two multinomial populations with mean μ_1, μ_2 and common matrixΣ. The coefficient of the MLD discriminant function ax are given by $\alpha = \Sigma^{-1}\delta$ *where* $\delta = \mu_1 - \mu_2$ in practice of course the parameters are estimated by

$$x_1, x_2 \text{ and } S = m^{-1}\{(n_1 - 1)s_1 + (n_2\ 1)s_2\}, . \quad (32)$$
$$\text{where } m = n_1 + n_2 - 2$$

Letting, the coefficient of sample MLDF given by $a = MW^{-1}d$

A test of hypothesis H0: $\alpha_1 = 0$ using the sample Mahalanobis distance $D_p^2 = Md^1W^{-1}d$ *and* $D_t^2 = Md_1^1W_{11}d_1$ has been proposed by [12] this test statistics uses the statistic:

$$\left\{\frac{m\,p+1}{p-k}\right\} C\left\{D_p^2 - D_t^2 \Big/ (m + c^2\, D_p^2\right\} \quad (33)$$

Where $c^2 = \dfrac{n_1\, n_2}{n}$, under the null hypothesis has $F_{p-k}, m-p+1$ distribution and we reject H_0 for large value of this statistics.

2.6. Evaluation of Classification Functions

One important way of judging the performance of any classification procedures is to calculate the errors rates or misclassification probability [13]. When the forms of parent populations are known completely, misclassification probabilities can be calculated with relative ease. Because parent populations are rarely know, we shall concentrate on the error rates associated with the sample classification functions. Once this classification function is constructed a measure of its performance in future sample is of interest. The total probability of misclassification (TPM) is given as:

$$TPM = P_1 \int_{R_1}, f_1 dx + P_2 \int_{R_2} f_2 dx \quad (34)$$

The smallest value of this quantity by a judicious choice of R_1 *and* R_2 is calculated the optimum error rate (OFR)

OFR = Minimum TPM

2.7. Probability of Misclassification

In constructing a procedure of classification, it is desires to minimize on the average the bad effects of misclassification [10], [13] and [11]. Suppose we have an item with response pattern x from either π_1 *or* π_2. We think of an item as a point in a r-dimensional space. We partition the space R into regions R_1 *and* R_2 which are mutually exclusive. If the

item falls in R_1 , we classify it as coming from π_1 and if it falls in R_2 we classify it as coming from π_2 .In following a given classification procedure, the researcher can make two kinds of errors in classification. If the item is actually from π_1 the researcher can classify it as coming from π_2 .Also the researcher can classify an item from π_2 as coming from π_1. We need to know the relative undesirability of these two kinds of errors in classification. Let the prior probability that an observation comes from π_1 be q_1, and from π_1 be q_2 .Let the probability mass function of π_1 be $f_1(x)$ and that of π_2 be $f_1(x)$. Let the regions of classifying into π_1 be R_1 .Then the probability of correctly classifying an observation that is actually from π_1 into π_1 is;

$$p(1/1) = \sum_{R_2} f_1(x) \tag{35}$$

Similarly, the probability of correctly classifying an observation from π_1 is

$$p(2/1) = \sum_{R_2} f_1(x) \tag{36}$$

Similarly, the probability of correctly classifying an observation from π_2 into π_2 is $p(2/2) = \sum_{R_2} f_1(x)$ and the probability is misclassifying an item from π_1 into π_2 is

$$p(1/2) = \sum_{R_2} f_1(x) \tag{37}$$

The total probability of misclassification using the rule is

$$TPMC\ (R) = q_1 \sum_{R_2} f_1(x) + q_2 \sum_{R_1} f_1(x) \tag{38}$$

In order to determine the performance of a classification rule R in the classification of future items, we compute the total probability of misclassification know as the error rate. [7] defined the following types of error rates.

i. Error rate for the optimum classification rule $R_{opt.}$. When the parameter of the distributions are known the errors is $TPMC\ (R) = q_1 \sum_{R_2} f_1(x) + q_2 \sum_{R_1} f_1(x)$ which is optimum for this distribution.

ii. Actual error rate: The error rate for the classification rule as it will perform in future samples

iii. Expected actual error rate: The expected error for classification rules based on sample size c from π_1 and π_2 from π_2 .

iv. The plug-in estimate of error rate obtained by using the estimated parameters for π_1 and π_2 .

v. The apparent error rate: This is defined as the fraction of items in the initials sample which is misclassified by the classification rule.

Table 1. *Confusion matrix of Apparent error rate.*

	π_1	π_2	
π_1	n_{11}	n_{12}	n_1
π_2	n_{11}	n_{22}	n_{11}
	π_1	π_2	n

The table above is called the confusion matrix and the apparent error rate is given by

$$\hat{P}(mc) = \frac{n_{12} + n_{21}}{n} \tag{39}$$

[6] called the second error rate the actual error rate and the third expected actual error rate. Hills showed that the actual error rate is greater than the optimum error rate and it in turns, is greater than the expectation of the plug –in estimate of the error rate. [9] proved a similar inequality. An algebraic expression for the extract bias of the apparent error rate of the sample multinomial discriminant rule was obtained by [5], who tabulated it under various combinations of the sample size n_1 and n_2 the number of multinomial cells and the cell probabilities. Their result demonstrated that the bound described above is generally loose.

3. The Simulation Experiments and Results

The four classification 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, 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 multivariate Bernoulli distribution with input parameter p_1 and $n_2 = n/2$ observations sampled from π_1, which is multivariate Bernoulli with input parameter $p_2, j = 1...r$. These samples are used to construct the rule for each procedure and estimate the probability of misclassification for each procedure is obtained by the plug-in rule or the confusion matrix in the sense of the full multinomial.

(ii) The likelihood ratios are used to define classification rules. The plug-in estimates of error rates are determined for

each of the classification rules.

(iii) Step (i) and (ii) are repeated 1000 times and the mean plug-in error and variances for the 1000 trials are recorded. The method of estimation used here is called the resubstitution method.

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

Table 2(a). *Mean apparent error rates.*

Sample sizes	A-B	Polynomial	LDA	Quadratic
40	0.157125	0.110074	0.110787	0.204512
60	0.161900	0.127855	0.127958	0.207491
100	0.163290	0.143526	0.143680	0.209940
140	0.162967	0.149837	0.150407	0.209826
200	0.162565	0.156384	0.155280	0.211542

Table 2(b). *Actual Error rates.*

Sample sizes	A-B	Polynomial	LDA	Quadratic
40	0.040271	0.052706	0.037112	0.041686
60	0.032751	0.042691	0.031487	0.033007
100	0.027786	0.037015	0.026152	0.027125
140	0.022462	0.031623	0.022112	0.024082
200	0.017981	0.026657	0.018218	0.019071

Tables 2(a) and (b) present the mean apparent error rates and standard deviation (actual error rates) for classification rules under different parameter values. The mean apparent error rates increases with the increase in sample sizes and actual error rate decreases with the increase in sample sizes. From the analysis, linear discriminant function is ranked first, followed by A-B Discriminant, Quadratic function and Polynomial discriminant function came last.

Table 3. *Performance of classification rules by rank.*

Classification Rule	Performance/rank
Linear Discriminant	1
A-B Discriminant	2
Quadratic function	3
Polynomial Discriminant function	4

4. Discussion and Conclusion

The results in table 3.1b indicate that, in general, with samples drawn from MVN populations with equal covariance matrices, the fisher LDF, the A-B procedure, the Quadratic Discriminant function (QDF) and Polynomial discriminant function (PDF) performed similarly, but as the degree of heterogeneity increases (not shown in the table), the QDF outperformed the other procedures. These results are consistent with those of [8] and [4], since it can be observed that the fisher LDF performed well, with respect to the QDF, for mild departures from homogeneity of covariance matrices, but as the degree of heterogeneity increased, the QDF outperformed the fisher LDF, A-B procedure and Polynomial discriminant function.

However, we obtained two major results from this study. Firstly, using the simulation experiments we ranked the procedures as follows: Linear Discriminant Function, A-B Discriminant function Quadratic and Polynomial Discriminant function. The best method was the linear discriminant procedure. Secondly, we concluded that it is better to increase the number of variables because accuracy increases with increasing number of variables. Moreover, our study showed that the linear discriminant function is more flexible in such a way to allow the analyst to incorporate some priori information in the models. Nevertheless, this does not exclude the use of other statistical techniques once the required hypotheses are satisfied.

References

[1] Anderson, T.W. and Bahadur, R.R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *Annals of Mathematics Statistics, 33*.420-431.

[2] Barnard, M.M. (1935). The Secular variations of skull characteristics in four series of Egyptian skulls. *Annals Eugenics, v. 6,* 352-371.

[3] Benerjee, K.S. & Marcus, L.F. (1965).In a Minimax Classification Procedure. *Biometrics, 52*, 654-654

[4] Gilbert, S.E. (1968). "On Discrimination using Qualitative Variables" *Journal of the American Statistical Association 1399-1418.*

[5] Gold Stein M. &Wolf (1977). On the problem of Bias multinomial classification. *Biometrics 33*, 325-331.

[6] Hills, M. (1967). "Discriminantion and allocation with discrete data", *Applied Statistics*. *16* 237-250.

[7] Lachenbruch, P.A. (1975) Discriminant Analysis. Hafner Press New York.

[8] Marks, S. & Dunn, O.J. (1974). Discriminant functions when Covariance matrices are unequal. *Journal of the American Statistical Association, 69,* 555-559.

[9] Martins, D.C., & Bradley, R.R. (1972). Probability Models, Estimation and Classification for Multivariate Dichotomous Populations, *Biometrics, 23*, 203-221.

[10] Onyeagu S.I. (2003). Derivation of an optimal classification rule for discrete variables *Journal of Nigerian Statistical Association, 73,* 724-745.

[11] Oluadare. S. (2011). Rubust Linear classifier for equal Cost Ratios of misclassification. CBN *Journal of Applied Statistics*.(2) (1)

[12] Rao, C.R. (1965). Linear Statistical Inference and Its Applications: John Willey New York.

[13] Richard A.J. & Dean W.W. (1988). Applied Multivariate Statistical Analysis.4th edition Prentice Hall. Inc. New Jessey.

[14] Smith, H.F. (1936). A discriminant function for plant selection. *Ann. Eugn. 7,* 240 – 250.

[15] Tou, J.T. & Gonzalez, R. C. (1974). Pattern Recognition Principles. Reading, Mass.; Addison –Wesley.

[16] Slah, B.Y. & Abdelwaheb Rebai (2007). Comparison between Statistical Approaches and linear programming for resolving classification problem. *International Mathematics Forum, 2,(63),* 3125-3141.

[17] Egbo, I., Onyeagu, S.I. & Ekezie, D.D. (2014). A comparison of multinomial classification rules for binary variables. *International Journal of Mathematical Science and Engineering Applications (IJMSEA), 8,* 141-157.

[18] Ekezie, D.D. (2012). Comparison of seven Asymptotic Error Rate Expansion for the sample linear Discriminant function. Unpublished Ph.D thesis submitted to Department of Statistics, Imo State University, Owerri, Nigeria.

[19] Egbo, I., Onyeagu, S.I. & Ekezie, D.D. (2014). A Comparison of Multivariate Discrimination of Binary Data. *International Journal of Mathematics and Statistics Studies*, *2(4),* 40-61.