Scientia Africana, Vol. 13 (No.1), June 2014. Pp 24-35
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

COMPARISON BETWEEN FISHERIAN AND BAYESIAN APPROACH TO CLASSIFICATION USING TWO GROUPS.

G. M. Oyeyemi [1], L. A. Oyebanji [2], I. S. Salawu [3]and A. I. Folorunsho[4]

[1, 2, 3]*Department of Statistics*
University of Ilorin
Ilorin, Nigeria
[1]*gmoyeyemi@gmail.com*
[2]*lateefabolarinwa@yahoo.com*
[3]*Salahus74@yahoo.com*

[4]*Department of Maths & Statistics*
Osun State Polytechnic Iree
Iree, Nigeria
[4]*folorunsoidowu@gmail.com*

Received:02-12-13
Accepted:03-02-14

ABSTRACT

Two approaches to discriminant analysis procedure are examined and compared based on their misclassification error rate. The Fisher's approach tends to find a linear combination of the variables which maximize the ratio of the between group sum of squares to that of the within group sum of squares in achieving a good separation. On the other hand, the Bayesian approach assigns an observed unit to a group with the greatest posterior probability. Fisher's linear discriminant analysis though is the most widely used method of classification because of its simplicity and optimality properties is normally used for two group cases. However, Bayesian approach is found to be better than Fisher's approach because of its low misclassification error rate.

Keywords: variance-covariance matrices, centroids, prior probability, mahalanobis distance, probability of misclassification.

INTRODUCTION

Classification is the allocation of an individual or object to a group or category on the basis of its own observed characteristics in the vector of observations for individual unit in which the linear or quadratic functions of the variables are employed to assign an individual unit to one of the groups (Anderson, 1958). The combination of measurements used is known as classification function and it is used to find the group or category to which the individual most likely belongs (Knoke, 1982).

Basically, the idea of classification analysis runs as follows; Suppose that we have samples from J populations of size n_j, j =1,2,…,J, with p measures on each of the N(=$\sum n_j$) units. Using the N x p data matrix, we want to determine from which of the J populations an (N +1)st unit is mostlikely to have been randomly sampled.

To accomplish this task we use the information in the given N x p data matrix to set up a rule for making the assignment. This situation is similar to that in multiple regression studies, where one is typically predicting a score on a continuous variable instead of predicting group membership. We could see that in both situations a rule based on a given data matrix is derived and may be used with "new" units (Rublik, 2008).

The Assumptions of Fishers Linear DiscriminantFunction

- The covariance matrix in first population is the same as the covariance matrix in second population i.e. $\Sigma_1 = \Sigma_2 = \Sigma$.
- The means μ_1, μ_2 and covariance matrix Σ are known.
- The two populations are multivariate normal i.e. $f_1(x, \theta_1)$ and $f_2(x, \theta_2)$ are multivariate normal density functions (Haerdle & Simar, 2003).

Fishers Linear Discriminant Function

Fisher's linear discriminant procedure works like the Principal Component Analysis which finds the most accurate data representation in lower dimensional space that projects data in the direction of maximum variance (Tipping, 2001). The main idea of Fisher's linear discriminant is finding projection to a line such that samples from different groups (populations) are well separated. This involves finding linear combination that maximizes the ratio of the between-group sum of squares and the within-group sum of squares such that a good separation is achieved (Fisher, 1938). It is also used for maximizing the sample mahalanobis distance between the two sets of data. Maximizing difference betweengroups may lead to reducing probability of misclassification.

Fisher suggested that using a linear combination of observations and choosing coefficients so that the ratio of the difference of means of linear combinations in the two groups to its variances is maximized. Fisher's linear discriminant function is known to be optimal for two multivariate normal populations with equal covariance matrices (Gorenstein et al, 1996).

Derivation of Fishers Linear Discriminant Function

Suppose we have a population consisting of two groups $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and a d- dimensional samples $x_1, x_2,\ldots,x_n$ such that n_1 samples comes from first group ($\boldsymbol{\beta}_1$) and n_2 samples comes from second group ($\boldsymbol{\beta}_2$) and must assign the individuals whose measurement are given by the d-dimensional samples into $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. We need a rule to do this, specifically, if the parameters of the distributions are known it can be used directly in construction of an assignment rule. However, if the parameters are unknown we use the n_1 samples from $\boldsymbol{\beta}_1$ and n_2 samples from $\boldsymbol{\beta}_2$ to estimate the parameters.

Consider a projection on a plane with v being the direction of projection; we can say that the projection of samples x_i onto a line in the direction v is given by $v'x_i$.Now, we want to measure separation between projections of different groups (populations).Suppose θ_1 and θ_2 are the means of projections of groups 1 and 2, respectively. Let μ_1 and μ_2 be means of groups 1 and 2, respectively.

Now, using $|\theta_1-\theta_2|$ as the measure of separation

Where;

$$\theta_1 = \frac{1}{n_1}\sum_{x_i \varepsilon C_1}^{n_1} V' X_i = V'\left(\frac{1}{n_1}\sum X_i\right) = V'\mu_1 \quad 1.4.1$$

and $\theta_2 = \boldsymbol{V'\mu_2}$ 1.4.2

It is important to note that the larger the $|\theta_1-\theta_2|$, the better the expected separation

Scientia Africana, Vol. 13 (No.1), June 2014. Pp 24-35
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

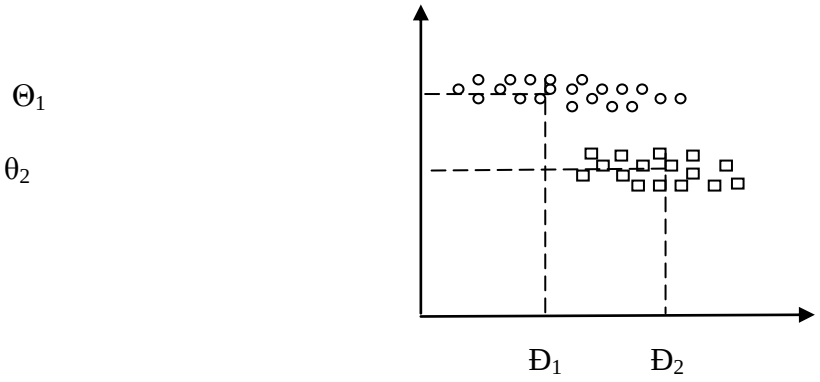

Figure 1: Projection of observations on both vertical and horizontal axis.

From Figure 1, the vertical axis is a better line of projection for separability than the horizontal axis but $|Ð_1-Ð_2|>|\theta_1-\theta_2|$. This is shows the problem with $|\theta_1-\theta_2|$ because it does not consider the variances of the classes. Normalizing $|\theta_1-\theta_2|$ by a factor which is proportional to variance, let the factor be S (scatter) and define as

$$S = \sum_{i=1}^{n} \left(Y_i - \mu_y\right)^2 \qquad 1.4.3$$

If we define $Y_i = V' X_i$ 1.4.4

$Y_i^{'s}$ are the projected samples, then S_1 is the scatter matrix for projected sample of group 1

$$\overset{*}{S}_1^{\,2} = \sum_{Y_i \varepsilon C_1} \left(Y_i - \mu_1\right)^2 \qquad 1.4.5$$

and S_2 is the scatter matrix for projected samples of group 2

$$\overset{*}{S}_2^{\,2} = \sum_{Y_i \varepsilon C_2} \left(Y_i - \mu_2\right)^2 \qquad 1.4.6$$

We need to normalize the function by both scatter matrices (Gilbert, 1969). Thus Fisher's linear discriminant is to project on line in the direction v which maximizes

$$\gamma(v) = \frac{\left(\theta_1 - \theta_2\right)^2}{\overset{*}{S}_1^{\,2} + \overset{*}{S}_2^{\,2}} \qquad 1.4.7$$

If we find v which makes $\gamma(v)$ large, we are guaranteed that the classes are well separated.

Expressing $\gamma(v)$ explicitly as a function v and maximizing it.

Defining the separate classes scatter matrices S_1 and S_2 for classes 1 and 2 respectively, these measure the scatter matrices of the original samples (x_{ij}) before the projection.

$$S_1 = \sum_{X_i \varepsilon C_1} \left(X_i - \mu_1\right)\left(X_i - \mu_1\right)' \qquad 1.4.8$$

$$S_2 = \sum_{X_i \varepsilon C_2} \left(X_i - \mu_2\right)\left(X_i - \mu_2\right)' \qquad 1.4.9$$

We now define the within group scatter matrix as

$$S_W = S_1 + S_2 \qquad 1.4.10$$

Recall from equation 1.4.5 that

$$S_1^{*2} = \sum_{Y_i \varepsilon C_1} (Y_i - \theta_1)^2$$

Using $Y_i = V'X_i$ and $\theta = V'\mu_1$

$$S_1^{*2} = \sum_{Y_i \varepsilon C_1} (V'X_i - V'\mu_1)^2$$

$$= \sum_{Y_i \varepsilon C_1} (V'(X_i - \mu_1))'(V'(X_i - \mu_1))$$

$$= \sum_{Y_i \varepsilon C_1} ((X_i - \mu_1)'V)'((X_i - \mu_1)'V)$$

$$= \sum_{Y_i \varepsilon C_1} V'(X_i - \mu_1)(X_i - \mu_1)'V$$

$$= V'S_1V \qquad 1.4.11$$

Similarly,

$$S_2^{*2} = V'S_2V \qquad 1.4.12$$

Hence,

$$S_1^{*2} + S_2^{*2} = V'S_1V + V'S_2V$$

$$= V'S_WV \qquad 1.4.13$$

Now, defining the between the group scatter matrix,

$$(\theta_1 - \theta_2) = (V'\mu_1 - V'\mu_2)^2$$

$$= V'(\mu_1 - \mu_2)(\mu_1 - \mu_2)'V$$

$$= V'S_BV \qquad 1.4.14$$

Thus $\gamma(v)$ can be rewritten as

$$\gamma(v) = \frac{(\theta_1 - \theta_2)^2}{S_1^* + S_2^*}$$

$$= \frac{V'S_BV}{V'S_WV}$$

Minimizing $\gamma(v)$ and equating to zero

$$\frac{d\gamma(v)}{dv} = \frac{\left(\frac{d}{dv}V'S_BV\right)V'S_WV - \left(\frac{d}{dv}V'S_WV\right)V'S_BV}{(V'S_WV)^2} = 0$$

$$= \frac{(2S_BV)V'S_WV - (2S_WV)V'S_BV}{(V'S_WV)^2} = 0$$

$$= V'S_WV(S_BV) - V'S_BV(S_WV) = 0$$

$$= S_BV - \frac{V'S_BV(S_WV)}{V'S_WV} = 0$$

$$= S_BV - \lambda S_WV = 0$$

$$S_BV = \lambda S_WV \text{ (General Eigen value problem).} \quad 1.4.15$$

If S_W has a full rank, we can convert this to a standard Eigen value problem

$$S_W^{-1}S_BV = \lambda V$$

But S_BV for any vector V points in same direction as $(\mu_1 - \mu_2)$.

$$S_BV = (\mu_1 - \mu_2)(\mu_1 - \mu_2)'V = \alpha(\mu_1 - \mu_2)$$

Thus the Eigen value problem can be solved immediately

$$S_W^{-1}S_BV = S_W^{-1}\{\alpha(\mu_1 - \mu_2)\}$$

$$= \alpha\{S_W^{-1}(\mu_1 - \mu_2)\}$$

$\alpha = \lambda$ And $V = S_W^{-1}(\mu_1 - \mu_2)$.

Therefore the linear discriminant function

$Y = V'X_i$, i=1…p and X=($x_1, x_2, \ldots, x_p$)

$$= S_W^{-1}(\mu_1 - \mu_2)X_i$$

Expected value of Y with respect to group 1 is given by

$$\bar{Y}_1 = S_W^{-1}(\mu_1 - \mu_2)\mu_1 = \theta_1,$$

And with respect to group 2 is given by

$$\bar{Y}_2 = S_W^{-1}(\mu_1 - \mu_2)\mu_2 = \theta_2$$

Scientia Africana, Vol. 13 (No.1), June 2014. Pp 24-35
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

Where, μ_i is the mean of measurement in the i[th]class.

Since the discriminant function

$$Y = V'X_i$$
$$= V_1X_1 + V_2X_2 + , \quad . \quad . \quad . + V_pX_p$$

V being the direction of projection.
The discriminant score for ith individual is defined by
$Y_i = V_1X_{1i} + V_2X_{2i} + ... + V_pX_{pi}$; i=1, 2, … ,$n_i$ and j=1…J .
Where X_{ij} represents the j[th] group with measurement X_i .

Assignment Rule

For $i = 1,2,...,n_i$ and $j = 1,2$; assign i[th] individual to group 1 if $V'\{X_i - \frac{1}{2}(\mu_1 - \mu_2)\} > 0$ and to group 2 if $V'\{X_i - \frac{1}{2}(\mu_1 - \mu_2)\} \leq 0$.
or
Assign i[th] individual to group 1 if $\bar{Y}_1 > \frac{\bar{Y}_1 + \bar{Y}_2}{2}$, and to group 2 if otherwise.

Bayesian Approach to Classification

The Bayesian approach towards classification when all parameters are known (estimated from the data) and misclassification cost are equal would begin with evaluation of the posterior probability that $X \in \Pi_j$ given $\underline{X}$, for each. $j = 1,2,...,J$ (Dunsmore,1966, Baldwin,1988). Then the posterior odds or ratios are computed for each pair of populations. Alternatively for $J > 2$, the population with the greatest posterior probability can be selected (Birnbaum & Maxwell, 1960).

When the costs of misclassification are unequal, the Bayesian would select the population thatproduces a minimum cost when average with respect to the posterior probability. This result alsoholds for all $J \geq 2$ when all parameters are known. Bayesian approach to classification is moregenerally applicable even when the covariance matrices are not equal and it requires nocomplicated distribution theory, though it is much more difficult to apply.

Construction of Classification Rule

For easy understanding of the classification rule, we present some definition of basic concepts and terms;

Typicality Probability $\{P(X/j)\}$

It is the probability that a randomly selected observation has a profile close to X, given that the unit is a member of population j. $j = 1,2,...,K$.

Posterior Probability $\{P(j/X)\}$.

It is the probability of an observation belonging to population j given that it has a particular observational vector X. Posterior in the sense that it is a probability conditioned on knowing $\underline{X}$.

It is reasonable to say that an observation be assigned to that population for which $P(j/\underline{X})$ is greatest.

$$P(j/\underline{X}) = \frac{P(\underline{X}/j)}{\sum_{j=1}^{K} P(\underline{X}/j)} \qquad 1.7.1$$

is the probability that an observation belong to population j given an observed score vector. It is equal to the ratio of the probability of its score vector in population j to the sum of the probabilities associated with its score vector in all j[’s].

Prior Probability (τ)

Examining the equation 1.7.1 above closely it is clear that the adequacy of $P\left(j/\underset{-i}{X}\right)$ will depend on goodness of the estimates of $P\left(\underset{-i}{X}/j\right)$, and that goodness in turn depends on size (representativeness) of the original samples on which these estimates are based.

This is one of the major problems of the Bayesian Approach (Rubin, et al, 2003).

If we let τ_j denote the proportion of observations in total K populations that is present in population j, then τ_j is the probability that the observation will be from population j.

This probability is known as the prior probability of membership in population j. Prior in the sense that it is a probability of population membership before $\underset{-i}{X}$ is known.

Hence τ_j should be taken into consideration when estimating $P\left(j/\underset{-i}{X}\right)$.

Now, letting $\tau_j.P\left(\underset{-i}{X}/j\right)$ denotes the probability that a randomly selected observational unit belongs to population j and at the same time has a score vector close to $\underset{-}{X}$. This can be used to arrive at values of $P\left(j/\underset{-i}{X}\right)$ by employing rules of probability due to T. Bayes (1701-1761).

So with respect to τ_j,

$$P\left(j/\underline{X_i}\right) = \frac{\tau_j P\left(\underline{X_i}/j\right)}{\sum_{j=1}^{k}\tau_j P\left(\underline{X_i}/j\right)}$$

$$= \frac{\tau_j f\left(\underline{X_i}/j\right)}{\sum_{j=1}^{k}\tau_j f\left(\underline{X_i}/j\right)} \qquad 1.7.2$$

Since $P\left(\underline{X_i}/j\right) \alpha\, f\left(\underline{X_i}/j\right)$, hence, the Bayes rule can be stated as;

Assign observation Xi to population j if;

$P\left(j/\underline{X_i}\right) > P\left(j'/\underline{X_i}\right)$; j ≠ j1, where $P\left(j/\underline{X_i}\right)$ is as earlier defined.

Or, Assign observation Xi to population j if;

$$\frac{P\left(j/\underline{X_i}\right)}{P\left(j^1/\underline{X_i}\right)} > 1$$

And to population j' if otherwise.

The rules above can be applied only if the probability density function $f\left(\underline{X_i}/j\right)$ is known. Most times the distribution parameters $\Sigma^{'s}$ and $\mu^{'s}$ are usually not known but can be estimated from the sample data.

Using $f\left(\underset{-}{X_i}/j\right)$ to be a multivariate normal density function define as

$$f\left(\underset{-}{X_i}/j\right) = (2\pi)^{-p/2}\left|\Sigma_j\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\underset{-}{X_i}-\underset{-}{\mu_j}\right)'\Sigma_j^{-1}\left(\underset{-}{X_i}-\underset{-}{\mu_j}\right)\right],$$

estimates of Σ_j and $\underset{-}{\mu_j}$ are respectively

$$\underset{-}{\bar{X}} = \underset{-}{\hat{\mu}_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} \text{ and}$$

Scientia Africana, Vol. 13 (No.1), June 2014. Pp 24-35
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

$$S_j = \hat{\Sigma}_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_{.j}\right)^2$$

Hence

$$\hat{f}\left(X_i / j\right) = (2\pi)^{-p/2}\left|S_j\right|^{-1/2}\exp\left[-\frac{1}{2}\left(X_i - \bar{X}_j\right)' S_j^{-1}\left(X_i - \bar{X}_j\right)\right] \qquad 1.7.3$$

Setting $\Delta^2_{ij} = \left[\left(X_i - \bar{X}_j\right)' S_j^{-1}\left(X_i - \bar{X}_j\right)\right]$ 1.7.4

We have

$$\hat{f}\left(X_i / j\right) = (2\pi)^{-p/2}\left|S_j\right|^{-1/2}\exp\left(-\frac{1}{2}\Delta^2_{ij}\right) \qquad 1.7.5$$

By substitution

$$\hat{P}\left(j / X_i\right) = \frac{\tau_j . \left|S_j\right|^{-1/2}\exp\left(-\frac{1}{2}\Delta^2_{ij}\right)}{\sum_{j=1}^{K}\tau_j . \left|S_j\right|^{-1/2}\exp\left(-\frac{1}{2}\Delta^2_{ij}\right)} \qquad 1.7.6$$

So for a p variate case the rule will be to assign an observation to population j if $R_{ij} = \frac{\hat{P}\left(j/X\right)}{\hat{P}\left(j'/X\right)} > 1$, $j \neq j'$ and assign to population j^1 if otherwise.

where $R_{ij}$is the predictive odd ratio for classifying X into any of the two populations.

Now, consider a case were $\Sigma_1 = \Sigma_2 = ... = \Sigma_K = \Sigma$, i.e. a case where we assume the equality of covariance matrices. Here an estimator for Σ is the pooled sample covariance matrix S_p where S_p is defined as

$$S_p = \frac{(n_1 - 1)S_1 + (n_2)S_2 + ... + (n_K - 1)S_K}{N - K}, N = \sum n_j$$

(Anderson & Bahadur, 1962, Chernoff, 1973).

Then

$$\Delta^2_{ij} = \left(X_i - \bar{X}_j\right) S_p \left(X_i - \bar{X}_j\right)$$

$$\hat{f}\left(X_i / j\right) = (2\pi)^{-p/2}\left|S_p\right|^{-1/2}\exp\left(-\frac{1}{2}\Delta^2_{ij}\right)$$

Note that Δ is a generalized distance estimator that is often attributed to an Indian statistician

P.C Mahalanobis.

Analysis and Comparisson

Analysis Using Fisher's Linear DiscriminantFunction

Assuming we have the following estimates of $\underline{\mu_1}$, $\underline{\mu_2}$, $\Sigma_1$and Σ_2 respectively

$$\bar{X}_1 = \begin{bmatrix} 67.87 \\ 61.00 \\ 61.50 \end{bmatrix} \quad \bar{X}_2 = \begin{bmatrix} 56.90 \\ 59.79 \\ 63.09 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 2673.84 & 485 & 621.5 \\ 485 & 6506 & 75 \\ 621.5 & 75 & 4599.5 \end{bmatrix} \quad S_2 = \begin{bmatrix} 6418.3 & 401.5 & 650.6 \\ 401.5 & 5083.79 & 153.29 \\ 650.6 & 153.29 & 4327.49 \end{bmatrix}$$

$$S_W = S_1 + S_2$$

The inverse matrix

$$S^{-1}{}_W = \begin{bmatrix} 0.00035319 & -0.00002604 & -0.00004966 \\ -0.00002604 & 0.00008825 & 0.00000145 \\ -0.00004966 & 0.00000145 & 0.00011906 \end{bmatrix}$$

The Fisher's linear discriminant function is given by

$$Y = \left(\bar{X}_1 - \bar{X}_2\right) S^{-1}{}_W \underline{X} \quad ,$$

where $\underline{X}$ is a 3x1 vector of observations.

$$\mathrm{Y} = \begin{pmatrix} 10.97 & 1.21 & -1.59 \end{pmatrix} \begin{bmatrix} 0.00035319 & -0.00002604 & -0.00004966 \\ -0.00002604 & 0.00008825 & 0.00000145 \\ -0.00004966 & 0.00000145 & 0.00011906 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$$= \begin{pmatrix} 0.00392199 & -0.00018116 & -0.00073236 \end{pmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$$= 0.0039219\,X_1 - 0.0001812\,X_2 - 0.0007323\,X_3$$

The discriminant scores for the two faculties are computed using the formulae

$$Y_i = 0.00392199\,X_{1i} - 0.00018116\,X_{2i} - 0.00073236\,X_{3i}$$

$$i = 1,2,\ldots n_j,\ j = 1,2$$

$$n_1 = n_2 = 70$$

From the above

$$\bar{Y}_1 = V'\bar{X}_1 = 0.21009456 \text{ and}$$

$$\bar{Y}_2 = V'\bar{X}_2 = 0.16612508$$

The discriminant function cut off is therefore computed as;

$$\frac{\left(\bar{Y}_1 + \bar{Y}_2\right)}{2} = 0.18810982$$

Scientia Africana, Vol. 13 (No.1), June 2014. Pp 24-35
 ISSN 1118 - 1931

So we assign i^{th} observation to population 1 if Y> 0.18810982, and to population 2 otherwise.

ANALYSIS USING BAYESIAN APPROACH

Using the estimates in section 2.1, the pooled covariance matrix, S_p, is defined as

$$S_P = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

$$S_P = \frac{S_W}{N - K}$$

is computed as

$$S^{-1}{}_P = \begin{bmatrix} 0.04872581 & -0.00359043 & -0.00684558 \\ -0.00359043 & 0.01217813 & 0.00020056 \\ -0.00684558 & 0.00020056 & 0.01642784 \end{bmatrix}$$

The prior probabilities (π) are

$\pi_1 = 0.4521$

$\pi_2 = 0.5479$

The distance estimates for each of the observed score vectors are obtained using

$$\Delta^2{}_{ij} = \left(\underset{\sim}{X}_i - \overline{X}_j\right) S_p{}^{-1} \left(\underset{\sim}{X}_i - \overline{X}_j\right)$$

$i = 1,2,...,n,\ j = 1,2$

Note that the distances to be computed here are of two types, namely,

(i) The distance of each of the observed vectors from their centroids,

(ii) The distance of the observed vectors from the centroid of the other population.

For group 1, the distance of the observed vectors from its means is,

$$\Delta^2{}_{1.} = \left(\underset{\sim}{X}_{1i} - \overline{X}_1\right) S_P{}^{-1} \left(\underset{\sim}{X}_{1i} - \overline{X}_1\right)$$

For group 2, the distance of the observed vectors from its means is

$$\Delta^2{}_{2.} = \left(\underset{\sim}{X}_{2i} - \overline{X}_2\right) S_P{}^{-1} \left(\underset{\sim}{X}_{2i} - \overline{X}_2\right)$$

Also for group 1, the distance of the observed vectors from the other means is

$$\Delta^2{}_{1.'} = \left(\underset{\sim}{X}_{1i} - \overline{X}_2\right) S_P{}^{-1} \left(\underset{\sim}{X}_{1i} - \overline{X}_2\right)$$

and for group 2, the distance of the observed vectors from the other means is

$$\Delta^2{}_{2.'} = \left(\underset{\sim}{X}_{2i} - \overline{X}_1\right) S_P{}^{-1} \left(\underset{\sim}{X}_{2i} - \overline{X}_1\right)$$

Hence,

$$P\left(1 / \underset{\sim}{X}_i\right) = \frac{0.5479 * \exp\left(-\frac{1}{2}\Delta^2{}_{1.}\right)}{\left\{0.5479 * \exp\left(-\frac{1}{2}\Delta^2{}_{1.}\right)\right\} + \left\{0.4521 * \exp\left(-\frac{1}{2}\Delta^2{}_{1.'}\right)\right\}}$$

$$P\left(2 / \underset{\sim}{X}_i\right) = \frac{0.4521 * \exp\left(-\frac{1}{2}\Delta^2{}_{2.}\right)}{\left\{0.4521 * \exp\left(-\frac{1}{2}\Delta^2{}_{2.}\right)\right\} + \left\{0.5479 * \exp\left(-\frac{1}{2}\Delta^2{}_{2'.}\right)\right\}}$$

So the predictive odd ratio for group 1 is;

$$P'\left(1/\underset{\sim}{X}_a\right) = \frac{0.4521 * \exp\left(-\frac{1}{2}\Delta^2_{1'.}\right)}{\left\{0.5479 * \exp\left(-\frac{1}{2}\Delta^2_{1}.\right)\right\} + \left\{0.4521 * \exp\left(-\frac{1}{2}\Delta_{1'}.^2\right)\right\}}$$

$$R_{ij} = \frac{P\left(1/\underset{\sim}{X}_i\right)}{P'\left(1/\underset{\sim}{X}_i\right)}$$

$$P'\left(2/\underset{\sim}{X}_a\right) = \frac{0.5479 * \exp\left(-\frac{1}{2}\Delta^2_{2'.}\right)}{\left\{0.4521 * \exp\left(-\frac{1}{2}\Delta^2_{2}.\right)\right\} + \left\{0.5479 * \exp\left(-\frac{1}{2}\Delta_{2'}.^2\right)\right\}}$$

And for group 2 is

$$R_{ij} = \frac{P\left(2/\underset{\sim}{X}_i\right)}{P'\left(2/\underset{\sim}{X}_i\right)}$$

Hence, the decision rule will be to assign individual with measurement $\underset{\sim}{X}_{ij}$ to any of the two groups whose R_j value is greater than 1, and to the other group otherwise, i = 1, 2, , , n_j; j = 1, 2

Table 1. Number of observations wrongly and correctly classified using Fisherian method

Faculty	Group 1	Group 2	Total
Group 1	60	19	79
Group 2	10	51	61
Total	70	70	140

The probability of misclassification of observations into group 1 is denoted by P_1 and calculated as;

$$p_1 = \frac{q_1}{n_1} = \frac{10}{70} = 0.1429$$

where q_1 is the number of misclassified observations into group 1 and n_1 is the number of samples of group 1.

Similarly, p_2 is the probability of misclassification of observations into group 2 and calculated as;

$$p_2 = \frac{q_2}{2} = \frac{19}{70} = 0.2714$$

` where q_2 is the number of misclassified observations into group 2 and n_2 is the number of samples in group 2.

Hence, the total probability of misclassification by the Fisherian approach is

$$\hat{P} = \frac{q_1 + q_2}{N + N} = \frac{10 + 19}{140 + 140} = 0.1036$$

Scientia Africana, Vol. 13 (No.1), June 2014. Pp 24-35
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

Table 2. Number of observations wrongly and correctly classified using Bayesian method

Faculty	Group 1	Group 2	Total
Group 1	60	17	77
Group 2	10	53	63
Total	70	70	140

The probability of misclassification of observations into group 1 is denoted by P_1 and calculated as;

$$p_1 = \frac{q_1}{n_1} = \frac{10}{70} = 0.1429$$

where q_1 is the number of misclassified observations into group 1 and n_1 is the number of samples of group 1.

Similarly, p_2 is the probability of misclassification of observations into group 2 and calculated as;

$$p_2 = \frac{q_2}{n2} = \frac{17}{70} = 0.2429$$

where q_2 is the number of misclassified observations into group 2 and n_2 is the number of samples in group 2.

Hence, the total probability of misclassification by the Bayesian approach is;

$$\widehat{P} = \frac{q_1 + q_2}{N + N} = \frac{10+17}{140+140} = 0.0964$$

DISCUSSION

From the comparison above, we can see that the probability of misclassification for the Fisher's approach is 0.1036 while that of the Bayesian approach is 0.0964, which makes the error rate of the Fisher's approach higher by 0.0072.

While the Bayesian approach provided only a predictive distribution for placing a vector of observations into the second population (in addition to the predictive odds) it falls short in defining prior probability for the groups. The Fisherian approach while providing only a decision that attempts to cope with the problem of risk associated with the classification decision falls short in requiring that sample size be large and that covariance matrices be equal, assumptions that are not always true.

Since we know that a rule is better than its alternative if its probability of misclassification is less than that of the alternative, we can conclude that the Baye's rule is better than the Fisher's rule, though this assertion may not always be true in all situations.

REFERENCES

Anderson, T. W. (1958). "An Introduction of Multivariate Analysis". John Wiley, New York.

Anderson, T. W. & Bahadur, R. R. (1962)."Classification into two Multivariate Distributions with Different Covariance Matrices", The Annals of Mathematical Statistics, 33, 420-431.

Baldwin, J. T. (ed.) (1988). "Classification Theory".J.T. Baldwin, "Classification theory: (Chicago 1985). Springer. pp. 1–23

Birnbaum, A.& Maxwell, A. E. (1960). "Classification Procedures Based on Baye's formula". Journal of the Royal Statistical Society.Series C (Applied Statistics), 9(3), pp. 152-169.

Chernoff, H. (1973). "Some Measures for Discriminating between Normal MultivariateDistributions with Unequal Covariance Matrices". Multivariate Analysis III, Ed., P. R. Krishinaiah, New York: Academic Press, pp. 337-344.

Dunsmore, I. R. (1966)."A Bayesian Approach to Classification", Royal Statistical Society.Series B (Methodological), Vol. 30, No. 2.

Fisher, R. A. (1938). *The Statistical Utilization of Multiple Measurements.*Ann. Eng. Lond. 7, 179-88.

Gilbert, E. S. (1969). "The Effect of Unequal Variance- Covariance Matrices on Fishers Linear Discriminant Functions", Biometrics. Ethel S. Gilbert *Biometrics* Vol. 25, No. 3 (Sep., 1969), pp. 505-515

Gorenstein, G., Lyons, R. & Solomon, R. (1996)."Classification of Finite Groups".Mathematical Surveys and Monographs, American Mathematical Society, vol. 40. Pp. 1994-2005.

Haerdle, W & Simar, L (2003)."Applied Multivariate Statistical Analysis". Springer Verlag, Berlin-Heidelberg-New York 2003, pp. XVIII/486, ISBN 3-540

Hazewinkel, M. (2001). "Bayesian Formula", Encyclopaedia of Mathematics, Springer. ISBN: 978-1-55608-010-4.

Knoke, J. D (1982) Discriminant Analysis with Discrete and Continuous Variables Biometrics, 38, (March, 1982) PP 191-200

Rublik, F. (2008). "On the Discriminant Analysis in the 2-population Case". Measurement Science Review, 8(3), 50 – 52

Rubin, D. B., Gelman, A., John, B. C. and Stern, H. (2003)."Bayesian Data Analysis (2nd. Ed.): Chapman & Hall/CRC. ISBN: 1-58488-388-1

Tipping, M. E. (2001). "Sparse Kernel Principal Component Analysis," Advances in Neural Information Processing Systems, 13 , pp. 633-639.