ISSN - 2277 - 0593

Journal of Natural Sciences, Engineering and Technology

ERROR RATES STABILITY OF THE HOMOSCEDASTIC DISCRIMINANT FUNCTION

A. ADEBANJI[1], S. NOKOE[2] AND O. IYANIWURA[3]

[1]*Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.
[2]Department of Applied Mathematics and Computer Science, University for Development Studies, Navrongo, Ghana. E-mail: nokoe_biomaths@yahoo.co.uk.
[3]Department of Statistics, University of Ibadan, Ibadan, Nigeria.
Email: jo_iyaniwura@yahoo.com
***Corresponding author:** tinuadebanji@yahoo.com

ABSTRACT

In this study the stability of the observed error rates of the homoscedastic discriminant function relative to the number of parameters in the model using simulated data from multivariate normal populations was investigated. Three models were considered, the four, six and eight variables models, each having four values of the separator function (δ). Equal and unequal prior probabilities were considered for the different number of parameter and separator function configurations. The asymptotic performance of the models was considered using the cross validation error rate estimation procedure. Results indicate the six variable models as being more stable (displaying less variability in the estimated error rates) than the other models under consideration. Less deterioration was observed for the six-variable model specification as was evident in the other models and this was more pronounced for smaller values of δ .

Keywords: Homoscedastic, Discriminant function, prior probabilities, asymptotic.

2000 Mathematics Subject Classification: 62H30, 65C10

INTRODUCTION

Given two or more groups of populations and a set of associated variables, one wants to locate a subset of the variables and associated functions of the subset that leads to maximum separation among the centroids of the groups. The exploratory multivariate procedure of determining variables and a reduced set of functions called discriminants or discriminant functions is called discriminant analysis. Discriminants that are linear functions of the variables are called Linear Discriminant Functions (LDF) or homoscedastic discriminant function (derived from the assumption of homoscedasticity of the variance-covariance matrix).

In this study, estimation of the stability (determined as a measure of within sample variability) in the error rates of the different models under consideration is of interest. This is an overall indicator of the performance of the discriminant function.

Observations are drawn from two multivariate normal populations (groups) ($R_i, i = 1,2$). The mean vectors of R_1 and R_2 are given as $\mu_1 = (0,....,0) \in \Re^p$ and $\mu_2 = (\partial, 0,...0) \in \Re^p$ respectively and both matrices have identity variance covariance structures $\Sigma_1 = \Sigma_2 = \Sigma$ and $\Sigma \in \Re^{pxp}$. The performance of the function is not dependent on the location of δ in the mean vector. Under a homoscedastic normal model for the group conditional distributions of the feature vector X on an entity, it is assumed that

$$X \square MNV(\mu_i, \Sigma_i) \quad (1)$$

The ith group-conditional density $f_i(X_i, \theta_i)$ is given as

$$f_i(X_i, \theta_i) = \phi(X; \mu_i, \Sigma)$$

$$= (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(X-\mu_i)'\Sigma^{-1}(X-\mu_i)\right\} \quad (2)$$

where θ_i consists of the elements of μ_i and the $\frac{1}{2}P(P+1)$ distinct elements of Σ. The square root of the Mahalanobis distance is predetermined as 1,3,5 and 7 using

$$\delta = \{(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\}^{1/2} \quad (3)$$

MATERIALS AND METHODS

The Model

In the classical homoscedastic model, the likelihood ratio classification function for an observed vector x is derived as

$$\frac{f_1(x)}{f_2(x)} = \exp\left\{x'\Sigma^{-1}(\mu_1 - \mu_2) - \tfrac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\right\} \quad (4)$$

We assign a new p-variable observation x to R_1 if

$(x - \frac{1}{2}(\mu_1 + \mu_2))'\Sigma^{-1}(\mu_1 - \mu_2) \geq k$ (otherwise assign to R_2) (5)

$k = \log(\pi_2 C(1|2))/(\pi_1 C(2|1))$. π_i ($i =$ 1, 2) is the prior probability of an observed vector coming from population i and $C(i|j)$ C is the cost incurred when an observation from R_j is classified as having come from R_i ($i \neq j$), $i, j = 1,2$. Thus $C(i|i) = C(j|j) = 0$. In this study, we assume equal cost of misclassification and k reduces to the log of the ratio of the prior probabilities.

It may be remarked that a Bayes rule may result in a large probability of misclassification and several attempts have been made to overcome this difficulty. When prior probabilities are known, the Bayes rule is optimum in the sense that it minimises average expected cost (Giri - 2004).

For an observed vector x, if the plug-in rule

is given as $r_0(x;\hat{F})$, this provides a good approximation to the Bayes rule $r_0(x;F)$ if $\hat{F}$ is a good estimate of F. $r_0(.)$ is the classification rule obtained when sample estimates $\bar{X}_1, \bar{X}_2$ of μ_1, μ_2 and S of Σ are plugged into (5).

The probability that a randomly chosen entity from R_i is allocated to R_j on the basis of $r_0(x;F)$, has an error rate specific to the i[th] group as

$$eo_i(F) = \sum_{j \neq i}^{g} eo_{ij}(F) \quad (i, j = 1,2) \tag{6}$$

And the overall error rate is

$$eo(F) = \sum_{i=1}^{g} \pi_i eo_i(F) \tag{7}$$

where π_i is as earlier defined.

In using error rates to measure the performance of a sample-based allocation rule, it is the conditional error rates that are of primary concern once the rule has been formed from the training data **(**Johnson and Wichern-1998**).**

The overall error rate for equal priors, is given by

$$e(F) \cong eo(F) + n^{-1}\{\phi(\tfrac{1}{2}\delta)/4\}\{p\delta + 4(p-1)\delta^{-1} + O^{1}\} \tag{8}$$

where

$$eo(F) = \Phi(-\tfrac{1}{2}\delta) \tag{9}$$

and δ is as earlier defined.

If the dimension p is small, the sample sizes n_1, n_2 occurring in practice will probably be large enough to apply this result. However, if p is not small, extremely large sample sizes will probably be required to make this results relevant (Giri 2004).

The leave-one-out (cross validation) error rate estimation procedure of Lachenbruch and Mickey (1968) is used as the performance evaluator for the models under consideration.

The simulation study

From (7), we can determine approximately how large n must be for a specified δ and p in order for the unconditional error rate not to exceed too far the best obtainable as given by the optimal error rate. Indeed, if n is small, then for p>1, the error rate is not far short of ½, which is the error rate for a randomized rule that ignores the feature vector and makes a choice of groups according to the toss of a coin (McLachlan – 1992).

The sample sizes have to be specified, here we set the 21 values of n_1 at 25, 50, 75, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1,300, 1400, 1500, 2000, 2500, 3000 respectively. The size of n_2 is decided by the predetermined sample size ratios n_1:n_2. The ratios are 1:1 and 1:2 thus determining the prior probabilities to be considered.

Numerical values were assigned to δ, the group centroid separator factor and Mahalanobis' distance determinant. These are 1, 3, 5 and 7 respectively.

The number of variables in the multi-normal distributions to be generated is predetermined as 4, 6 and 8 following Murray (1977) that considered the selection of variables in

Discriminant Analysis. By making 100 replicates we aim at attaining an accurate estimate of the misclassification rate by reducing the between sample variability.

Hence 100 samples of random variates of the required specification are generated, and the analysis is carried out on the 100 samples. Thus resulting in 2,100 samples for each sample ratio consideration and four values of δ . This gives a total of 4200x4 (16,800) samples of various sizes. The error rate estimates are then averaged over the number of replicates.

The SAS V8 (1996) package was used for generating the matrices from a N (0,1) distribution for the predetermined models. Independent series of normal deviates of required length are drawn and then transformed (standardized) to have unit variance and zero covariance.
Similar simulation experiments had been constructed by Marks and Dunn (1976) and He and Fung (2000). These did not consider variable effects and number of replicates and sample sizes not this many.

Results of simulation

The total probability of misclassification, standard deviation and coefficient of variation are presented as decimals. The results of the simulation are presented in a series of figures The first set (1.1 to 1.6) present the results for the four variable model with equal prior probability scheme presented in figs 1.1 to 1.3 for different values of δ . Figures 1.1 to 1.3 are the total error rates, the standard deviation (SD) and coefficient of variation (CV) respectively.

Figures 1.4 to 1.6 present the same results for the unequal prior probability (n_1:n_2 = 1:2). The second series of figures are for the six (6) variable model and are presented in the same format (that is equal and unequal prior probabilities) as figures 2.1 to 2.3 and 2.4 to 2.6 respectively. The last series of figures are for the eight (8) variable model and are presented in the same sequence as figures 3.1 to 3.6 respectively.

In the equal prior scenario, for the four variable model, error rates are less stable for the smaller sample sizes than for the larger sample sizes as evidenced in the coefficients of variation recorded. Stability in the error rates deteriorates remarkably asymptotically as δ increased from 1 to 7 and worsens when δ =7 when the samples are so far apart that it does not justify the use of a classification function.

Changing the number of variables in the model to six resulted in some observable difference in the performance of the LDF. A reduction in error rate for the respective δ values was recorded. When δ =1, the values recorded for the standard deviation were consistently lower than those for the mean total error rate. The reduction of the CV is rapid. The highest value of 35.8% error rate was observed for sample size 50 (ratio 1:1) and the smallest recorded was 18.59% for sample size 6000 (ratio 1:4). Reduction in error rates was more rapid for ratio 1:1 than for 1:2. When compared with δ =1 (p=4 variables), the minimum and maximum recorded values are quite close. The four variable model (p=4) had values 33.76 and 18.6% for maximum and minimum misclassifications respectively.

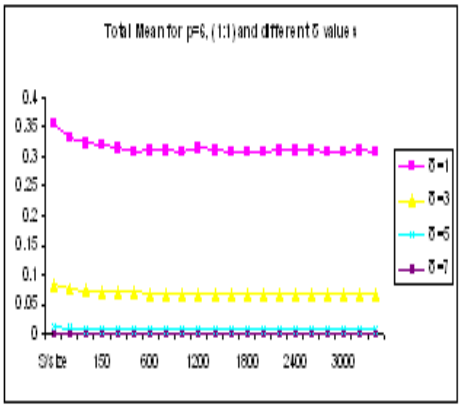

fig 2.1

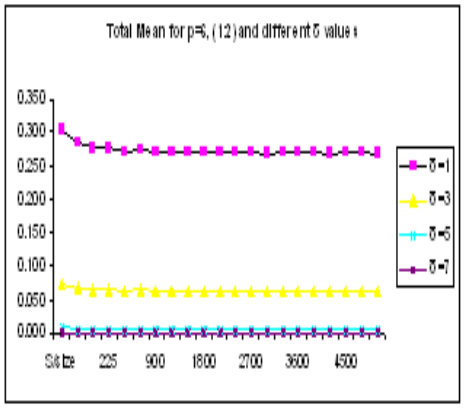

fig 2.4

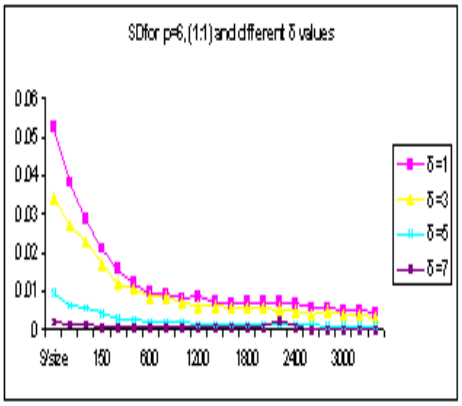

fig 2.2

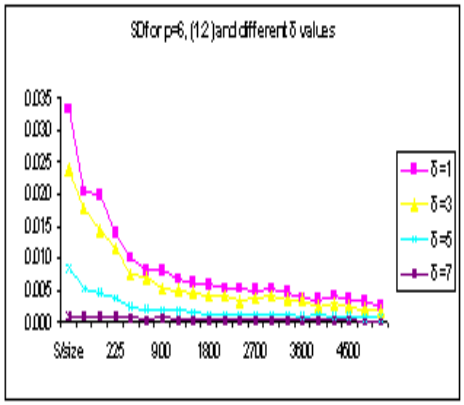

fig 2.5

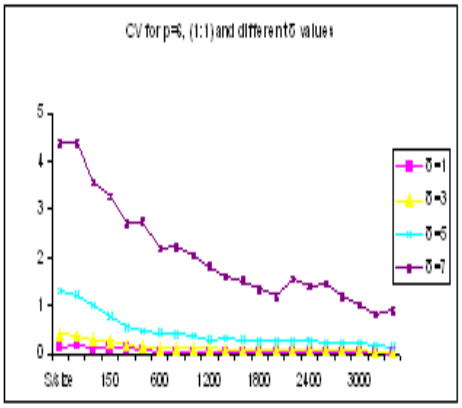

fig 2.3

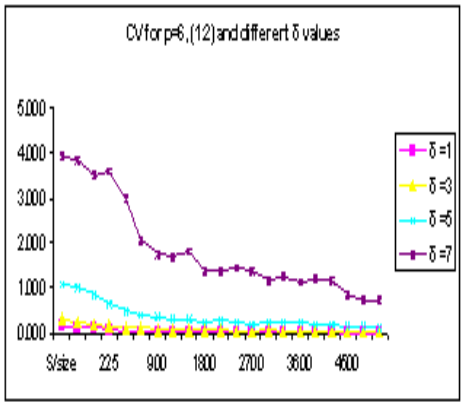

fig 2.6

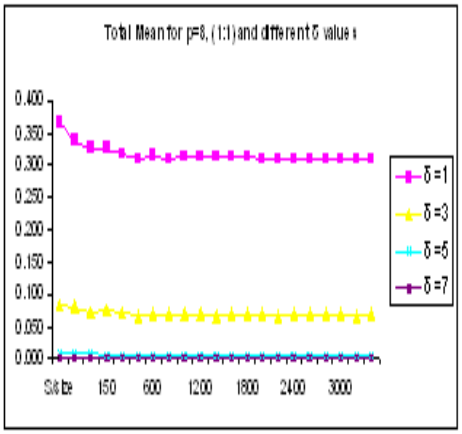

fig 3.1

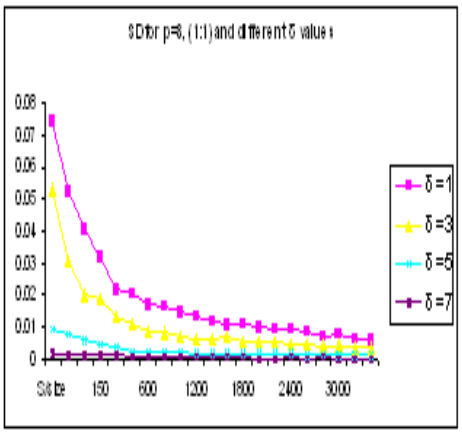

fig 3.2

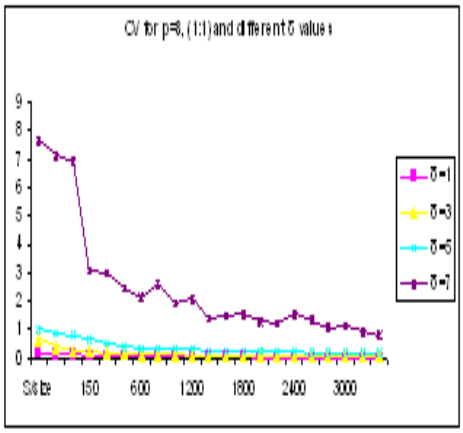

fig 3.3

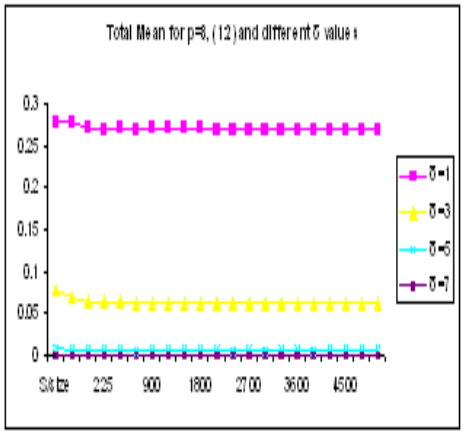

fig 3.4

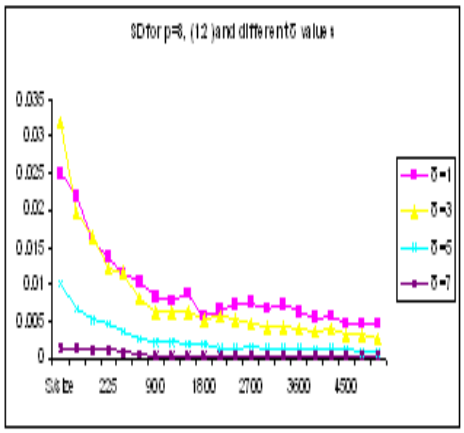

fig 3.5

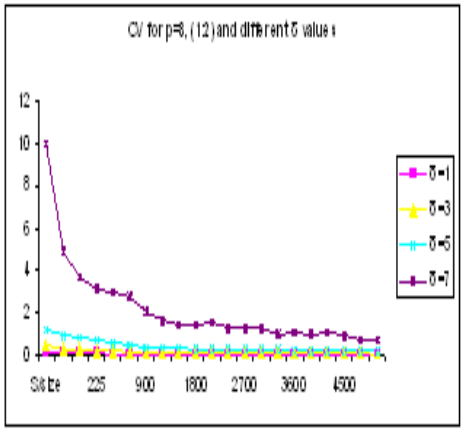

fig 3.6

When δ =3, the relationship between the mean error rates and CV is reversed (CV records higher values). The maximum error rate of 8.26% was recorded for sample size 50 (ratio 1:1) and the CV results recorded a maximum of 101.29% for sample size 1800 (ratio 1:2).The recorded error rates are however much lower than was earlier recorded for lower values of δ (especially when viewed in comparison with the four variable linear models). When δ =7, the values of CV are much higher than earlier recorded as a result of the high reduction in mean error rate; although little improvement was observed across sample sizes, the function recorded near zero misclassification rates.

When the number of variables is increased to eight (p=8), the observed pattern is similar to the pattern observed for P=6 with the mean error rates recording higher values than the CV for lower values of δ .The maximum value of mean error rate observed is 36.67% for sample size 50 (ratio 1:1) and reduction in CV is not as rapid as earlier recorded and the reduction in mean error rate is much more gradual than earlier observed.

At δ =3, the values of CV are larger than the mean error rate except at the tail end of the curves. A maximum error rate of 8.2% is recorded for sample size 50 (ratio 1:1) However, the asymptotic reduction in CV is no longer observed for δ =5. Here, an initial reduction is observed after which the values stabilizes. The mean error rates are close to zero.

Improvement in the performance of the function is more pronounced for the eight variable model than the four variable case. Maximum error rate of 0.256% was observed for sample size 100 (ratio 1:3) and minimum value of 0.102% for sample size 50 (ratio 1:1) was observed for δ =6 for δ =7, the values were 0.044% for sample size 50 (ratio 1:4) and minimum value of 0.014% for sample size 50 (ratio 1:1).

DISCUSSION

There appears a turning point in the reduction in the mean error (or misclassification) rates as well as improvement in their stability beyond p = 6. This is suggesting that not more than six variables should be included in discriminant analysis even when the sample size is as large as 15000. A reasonable corollary to this finding is the plausible conclusion that the smaller your sample sizes the fewer should be the number of variables. This decline in stability of observed error rates is observable for higher values of δ with δ =7 recording much higher instability than δ =5.

Also, increasing the sample size will cease to result in an improvement in the performance of the function once a thresh-hold is reached, beyond which, there is nothing to be gained by any further increase even to values that give sample estimates that would equal the population parameters.

CONCLUSION

These inconsistencies in the behaviour of average error rates are the observed deterioration in the stability of the error rates as p changes from 6 to 8 suggest that there is a turning point between p = 6 and p = 8 in the relationship between the number of variables and the magnitude of error rates and their variation. This is most plausible at p = 7. This pattern was observed for the different values of δ that we considered.

ACKNOWLEDGEMENT

This study was supported by the Third World Organization for Women in Sciences (TWOWS) fellowship.

REFERENCES

Adebanji, A.O., Nokoe, S. 2004. Evaluating the Quadratic Classifier; *Proceedings of the Third International Workshop on contemporary problems in Mathematical Physics,* P. 369-394.

Giri N.C. 2004. *Multivariate Statistical Analysis.* DEKKER Series © 2004 P. 435-477.

He, X.M., Fung, W.K. 2000. High breakdown estimation for multiple populations with applications to discriminant analysis; *Journal of Multivariate Analysis,* 72: 51-162.

Johnson, R.A., Wichern, D.W. 1998. *Applied Multivariate Statistical Analysis*, 4th Edition. Prentice Hall Inc. USA. P. 629 – 723.

Joossens, K. 2006. Robust Discriminant Analysis; Ph.D. Thesis of Katholieke Universiteit, Leuven.

Lachenbruch, P.A., Mickey, M.R. 1968. Estimation of error rates in discriminant analysis. *Technometrics,* 10: 1-11.

Marks, S., Dunn, O.J. 1974. Discriminant Functions when covariance matrices are unequal; *Journal of the American Statistic Association,* 69: 555-559.

McFarlan, R.H., Richards, D. 2001. Exact Misclassification Problems for Plug-in Normal Discriminant Functions. Equal Mean Case. *Journal of Multivariate Analysis,* 77: 21-53

McLachlan, G. 1992 *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Mathematical Statistics. P. 4-64.

Murray, G.D. 1977. A Cautionary Note on Selection of Variables in Discriminant Analysis. *Applied Statistics,* 26(3): 246-250.

Okomato, M. 1963. An Asymptotic Expansion for the Distribution of the Linear Discriminant Function. *Annals of Math Stat.,* 34: 1286-301.

(Manuscript received: 6th January, 2010; accepted: 24th June, 2010).