Indian Journal of Agricultural Sciences **82** (5): 426–9, May 2012

Linear discriminant function under multivariate non-normal rice (*Oryza sativa*) and maize (*Zea mays*) data

R K RAMAN[1], S D WAHI[2] and A K PAUL[3]

Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110 012

Received: 10 March 2011; Revised accepted: 28 February 2012

ABSTRACT

The performance of linear discriminant function was studied under multivariate non-normal situations. The different multivariate non-normal populations were simulated by using the mean vectors and dispersion matrices of rice (*Oryza sativa* L.) and maize (*Zea mays* L.) data sets. Further 50 different independent samples were simulated for different dimensions and sample sizes for maize and rice data to obtain empirical probabilities of misclassification. On fitting linear discriminant function to non-normal data the empirical probabilities of misclassification were higher as compared to misclassifying probabilities obtained by using normal approximation. In large sample sizes and in higher dimensions the differences between empirical and normal approximation of probabilities of misclassification were found almost negligible.

Key words: Discriminant analysis, Multivariate non-normal distribution, Probability of misclassification

Discriminant analysis deals with the problem of classification. We make a number of measurements on an individual and on the basis of these measurements we wish to classify it into one of several well-defined categories. The pioneer work done by Fisher (1936) have been used by biologists to solve the classificatory problems involving multiple measurements in different contexts. Ito and Schull (1964) discussed the robustness of T_o^2 statistics, when the condition of equality of covariance matrices is not satisfied. Minhajuddin *et al*. (2004) proposed a method to simulate the joint distributions which have equal positive pair-wise correlations and the method was illustrated for the p-dimensional families of beta and gamma distributions. Sever *et al.* (2005) compared Fisher's discriminant analysis under normal and skewed curved normal distribution based on the apparent error rates, which were used as a measure of classification performance and found that Fisher's linear discriminant analysis to be highly robust under skewed curved normal distribution. Rausch and Kelley (2009) compared different methods for discriminant analysis with respect to classification accuracy under non normality through Monte Carlo simulation. With this background the present study was taken to see the performance of Fisher's linear discriminant function under multivariate non-normal situation considering different dimensions of dispersion matrices and sample sizes.

[1] Ph D Student (e mail: Rohan4741@gmail.com), Sukhatme Hostel, IASRI; [2] Principal Scientist (e mail: sdwahi@iasri.res.in), [3] Senior Scientist (e mail: pal@iasri.res.in), Biometrics and Statistical Modelling Division

MATERIALS AND METHODS

The secondary data on 77 maize (*Zea mays* L.) genotypes grown at seven different locations reported in the Annual progress report for the year 2005–06 of All India Coordinated Maize Improvement Project, Directorate of Maize Research, New Delhi, on 10 morphological characters such as grain yield (kg/ha) at 15% moisture, days to 50% pollen shed, days to 50% silking, days to 50% dry husk, moisture percentage at harvest, plant aspect in kg/plant, Ear aspect in kg/plant, husk cover in kg/plant, plant height (cm), Ear height (cm), and 75 genotypes consist of 25 tall, 25 medium and 25 dwarf rice genotypes procured from Genetics Division of IARI, New Delhi, on nine morphological traits like tiller number/plot, plant height (cm), panicle length (cm), panicle weight (gm), 1000-grain weight (g), biomass/plant (g), harvest index, straw yield/plant and grain yield/plant (g) were used in the present investigation. The normality of both the data sets is tested by Mardia's (1980) skewness and kurtosis test. It was found that the probability for testing the kurtosis is 0.043 and 0.049 in maize and rice data, respectively, which shows that both the data sets are non-normal.

The Dirichlet distribution is a multivariate generalization of the beta distribution (Kotz *et al*. 2000). The RANDDIRICHLET (N, Shape) Function in SAS 9.2 was used for simulation of multivariate non-normal data. The inputs are as follows:

N is the number of desired observations sampled from the distribution.

Shape is a 1 x (p+1) vector of shape parameters for the distribution, shape[i] > 0.

The RANDDIRICHLET function returns an N × p matrix containing N random draws from the Dirichlet distribution.

If $X = (X_1\ X_2 \ldots X_p)$ with $\sum_{i=1}^{p} X_i < 1$ and $X_i > 0$ follows a Dirichlet distribution with shape parameter $\alpha = (\alpha_1\alpha_2 \ldots \alpha_{p+1})$, then

The probability density function for x is

$$f(x,\alpha) = \frac{\Gamma(\sum_{i=1}^{p+1}\alpha_i)}{\Pi_{i=1}^{p+1}\Gamma(\alpha_i)}\Pi_{i=1}^{p+1}x_i^{\alpha_i-1}(1 - x_1 - x_2 - \ldots - x_p)^{\alpha_{p+1}-1}$$

If p=1, the probability distribution is a beta distribution.

If $\alpha_0 = \sum_{i=1}^{p+1}\alpha_i$, then

The expected value of X_i is $\frac{\alpha_i}{\alpha_0}$, the variance of X_i is $\frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2\ (\alpha_0+1)}$, the covariance of X_i and X_j is $-\frac{\alpha_i\alpha_j}{\alpha_0^2\ (\alpha_0+1)}$

The classical Fisher's linear discriminant function between two populations have

$\Sigma_1 = \Sigma_2$ and the function is of the form be

$$Y = \sum_{i=1}^{p} b_i X_i \qquad \ldots (1)$$

where b_i's are coefficients of linear discriminant function and are obtained by maximizing the ratio of between and within class variances, i.e. $\frac{D_p^2}{S}$ ... (2)

where $D_p^2 = b_1d_1 + b_2d_2 + \ldots + b_pd_p$ and

$$S = \sum_{i}^{P}\sum_{j}^{P} b_i b_j W_{ij}$$

where W_{ij}'s are the elements of pooled dispersion matrix and d_i's are the elements of the vector of difference between the mean vector of the two populations. The solution for b_i's are obtained by differentiating (2) w.r.t., b_i's

$$\frac{D}{S^2}\left[2S\frac{\partial D}{\partial b} - \frac{D\partial S}{\partial b}\right] = 0 \qquad \Rightarrow \frac{1}{2}\frac{\partial S}{\partial b} = \frac{S}{D}\frac{\partial D}{\partial b}$$

Since S/D is a constant factor, the unknown b_i's are proportional to the solutions of the equations.

$b_1W_{11} + b_2W_{12} + \ldots + b_pW_{1p} = d_1$

$b_1W_{21} + b_2W_{22} + \ldots + b_pW_{2p} = d_2$

… …

$b_1W_{p1} + b_2W_{p2} + \ldots + b_pW_{pp} = d_p$

So, $\hat{b} = d'W^{-1}$ and $D_P^2 = d'W^{-1}d$

The D_p^2 is the estimate of variance of Y and the root of D_p^2 gives the discriminatory power of the linear discriminant function.

$$\hat{V}(Y) = \Sigma b_i b_j W_{ij} = \Sigma b_i d_i$$

The significance of D_p^2 can be tested for its significance by an F-test given by

$$F = \frac{(N_1 + N_2 - p - 1)N_1N_2D_p^2}{p(N_1 + N_2)(N_1 + N_2 - 2)}$$

with p and $(N_1 + N_2 - p - 1)$ degrees of freedom. The approximate probabilities of mis- classification for Fisher's linear discriminant function is given by $\Phi(-\frac{1}{2}D_p)$, where Φ is cumulative normal distribution and D_p is the square root of D_p^2.

Three distinct populations are simulated by using different mean vectors and dispersion matrices for both rice and maize data. Differences of mean vectors of these populations are tested by Hotelling's T^2 test and are found highly significant. A pooled dispersion matrix is formed using these three populations in both the data sets. Three distinct multivariate non-normal data sets for both maize and rice (*Oryza sativa* L.) of different dimensions (i e, four, six and nine characters) and different sample sizes (i e, 50, 100 and 150) were simulated by using these three mean vectors and pooled dispersion matrix with the help of RANDDIRICHLET function in SAS package (SAS 9.2, 2009). Linear discriminant function is fitted and D_p^2, discriminating power and probability of misclassification using normal approximation are obtained. Fifty samples each for different dimensions and sample sizes populations were simulated for obtaining empirical probability of misclassifications and their coefficient of variation.

RESULTS AND DISCUSSION

D_p^2, discriminating power, probability of misclassification and its coefficient of variation obtained by fitting linear discriminant function based on three different dimensions (i.e. four, six and nine) and three different sample sizes (i.e., 50, 100 and 150) for both multivariate non-normal maize and rice are given in Table 1 and Table 2 respectively. In case of four characters in maize data, it can be seen that D^2 values for different pairs of populations are found to be significant. Empirical probability of misclassifications decreased with increase in sample size. Coefficient of variation of probability of misclassification also decreases with increase in sample size. Similar trend was seen for D^2 values and empirical probabilities of misclassification in case of six and nine characters of maize data. In general, it was found that empirical probability of misclassification and

Table 1 D^2, discriminating power (DP), probability of misclassification (POM) and its coefficient of variation (CV) based on four, six and nine characters between maize populations of non-normal data

Population	Nine characters (maize)				Six characters (maize)				Four characters (maize)			
	D_9^2	DP	POM	CV	D_6^2	DP	POM	CV	D_4^2	DP	POM	CV
Sample size 50												
1–2	0.823*	0.907	0.379 (0.325)	12.573	0.712*	0.844	0.383 (0.337)	5.907	0.682*	0.826	0.402 (0.339)	7.847
2–3	0.907*	0.952	0.376 (0.317)	6.794	0.717*	0.847	0.391 (0.336)	8.458	0.421*	0.649	0.412 (0.372)	7.435
2–3	0.791*	0.889	0.421 (0.328)	6.158	0.731*	0.855	0.427 (0.335)	6.453	0.707*	0.841	0.431 (0.337)	6.495
Sample size 100												
1–2	0.544*	0.738	0.371 (0.356)	7.473	0.523*	0.723	0.373 (0.359)	7.514	0.652*	0.807	0.391 (0.343)	3.152
1–3	0.928*	0.963	0.362 (0.315)	3.148	0.733*	0.856	0.387 (0.334)	5.599	0.413*	0.643	0.402 (0.373)	4.672
2–3	0.417*	0.646	0.413 (0.373)	3.916	0.387*	0.622	0.419 (0.378)	2.972	0.692*	0.832	0.417 (0.338)	2.277
Sample size 150												
1–2	0.522*	0.722	0.367 (0.359)	5.453	0.573*	0.757	0.369 (0.353)	3.806	0.623*	0.789	0.387 (0.346)	1.541
1–3	0.876*	0.936	0.355 (0.320)	2.159	0.941*	0.970	0.355 (0.314)	4.151	0.393*	0.627	0.396 (0.376)	2.169
2–3	0.315*	0.561	0.406 (0.389)	3.163	0.291*	0.539	0.411 (0.394)	2.929	0.653*	0.808	0.412 (0.343)	1.324

Values in parentheses are probability of misclassification based on normal approximation, *represents the significance at 5% level

its coefficient of variation decreased with increase in sample size.

In case of four characters in rice data Table 2, it can be seen that D^2 values for different pairs of populations were found to be significant. Empirically probabilities of misclassification also decreased with increase in sample size. Similarly, coefficient of variation of probability of misclassification also decreased with increase in sample size as in maize data. Similar trend was seen for D^2 values and empirical probability of misclassifications in case of six and nine characters of maize data.

Considering the results obtained for rice and maize data sets it was observed that empirical probability of misclassification showed the decreasing trend with increasing dimension irrespective of sample size. These results are valid for given correlation structures of the data sets. The increase in dimension from 4 to 9 was done on the basis of the correlation among the characters and only those characters which are mostly having low to moderate positive correlation barring one or two low negative correlations were included. Empirical probabilities of misclassification also decreased as sample size increased irrespective of dimensionality of the data. It is also noticed that empirical probabilities of misclassification is always higher than the misclassification probabilities obtained by normal approximation in the both data sets. Hence it can be said that the probabilities of misclassification obtained by using normal approximation are always underestimated in case of non-normal data. In general, the difference between probabilities of misclassification obtained by empirical and normal approximation decreased with increase in sample size.

REFERENCES

Fisher R A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**: 179–88.

Ito K and Schull W J. 1964. On robustness of T_o^2 in multivariate analysis of variance when covariance matrices are not equal, *Biometrica* **51**: 71–82.

Kotz S, Balakrishnan N and Jhonson N L. 2000. *Continuous Multivariate Distributions,* pp 485–41. John Wiley & Sons, Inc., New York.

Mardia K V. (1980). (*in*) *Handbook of Statistics*, **1**: 279–320, P R Krishnaiah (Ed.) Test of univariate and multivariate normality. North Holland.

Table 2 D^2, discriminating power (DP), probability of misclassification (POM) and its coefficient of variation (CV) based on four, six and nine characters between rice populations of non-normal data

	Nine characters (maize)				Six characters (maize)				Four characters (maize)			
Population	D_9^2	DP	POM	CV	D_6^2	DP	POM	CV	D_4^2	DP	POM	CV
					Sample size 50							
1–2	2.077*	1.44 1	0.365 (0.236)	7.412	1.323*	1.150	0.387 (0.283)	10.13 1	0.626*	0.791	0.390 (0.346)	8.847
1–3	0.856*	0.925	0.398 (0.322)	13.164	0.471*	0.686	0.405 (0.333)	9.268	0.356*	0.597	0.412 (0.382)	8.435
2–3	1.562*	1.250	0.288 (0.266)	15.411	1.456*	1.207	0.318 (0.273)	8.055	0.683*	0.826	0.384 (0.339)	7.495
					Sample size 100							
1–2	1.658*	1.288	0.334 (0.260)	4.552	1.149*	1.072	0.346 (0.296)	8.527	0.586*	0.766	0.352 (0.350)	4.152
1–3	0.710*	0.843	0.357 (0.337)	10.367	0.462*	0.680	0.371 (0.367)	8.347	0.311*	0.558	0.406 (0.390)	5.672
2–3	2.280*	1.510	0.268 (0.225)	4.367	1.521*	1.233	0.331 (0.269)	7.798	0.577*	0.760	0.376 (0.352)	3.277
					Sample size 150							
1–2	1.129*	1.063	0.323 (0.298)	2.853	1.181*	1.087	0.331 (0.293)	7.076	0.544*	0.738	0.345 (0.356)	2.541
1–3	0.550*	0.742	0.345 (0.355)	5.586	0.428*	0.654	0.378 (0.372)	5.819	0.273*	0.522	0.392 (0.396)	3.169
2–3	2.070*	1.439	0.265 (0.236)	4.027	1.847*	1.359	0.290 (0.248)	6.504	0.493*	0.702	0.362 (0.362)	2.324

Values in parentheses are probability of misclassification based on normal approximation. (*)represents the significance at 5% level

Minhajuddin A M, Harris I R and Schucany W R. 2004. Simulating multivariate distribution with specific correlation. *Journal of Statistical Computation and Simulation* **74**(8): 599–607.

Rausch R J and Kelley. 2009. A comparison of linear and nonlinear models for discriminant analysis under non-normality. *Behaviour Research Methods* **41**(1): 85–98.

SAS. 2009. *SAS User Guide, Version* 9.2, SAS Institute Incorporates, USA.

Sever M, Lajovic J and Rajer B. 2005. Robustness of Fisher's discriminant function to skewed normal distribution. *Metodoliski Zvezki* **2**(2): 231–42.