Jour. Ind. Soc. Ag. Statistics
47(1), 1995 : 12–20

Use of Bootstrap Method in Comparing the Performance of Linear Discriminant Functions

S.D. Wahi and V.K. Bhatia
Indian Agricultural Statistics Research Institute,
New Delhi-110012

(Received : January, 1992)

SUMMARY

The linear discriminant function based on minimax linear procedure without assuming the equality of variance-covariance matrices among the two populations, together with corresponding D^2-values, were obtained. They were compared with Fisher's linear discriminant function which assumed the equality of such matrices. The genetic divergence (D^2-values) obtained by minimax linear procedure were higher than those for the linear discriminant function procedure where the covariance matrices were strikingly different. The bootstrap technique was used to further investigate the efficiency of the two procedures. The results of the bootstrap method have confirmed the superiority of minimax linear procedure method for different covariance matrices. The bootstrap technique yields higher values of D^2- statistics for both the minimax and Fisher's linear discriminant functions indicating the bias of the D^2-statistics.

Key Words : Fisher's Discriminant function, Minimax Linear Function and Bootstrap Procedure.

Introduction

The development of the theory if discriminant function has originally arisen from the classificatory analysis in taxonomic problems dealing many characteristics at a time. With the publication of the paper by Fisher [4], a large number of researchers contributed both in theoretical and applied aspects of classificatory problems concerning different situations. The concept of multiple measurements introduced by Mahalanobis [5] is also widely used in classificatory problems. The two basic assumptions required to be satisfied by the data for using discriminant function and/or D^2 are (i) the conformation to multivariate-normal distribution of the variates of the populations under study and (ii) the equality of variance- covariance matrices of the populations. In most of the applications of discriminant function/D^2-statistics in biological and social sciences, the multivariate normality and the equality of variance-covariance matrices are assumed, under the condition of large sample robustness. However, in practice it has been found in majority of crossbreeding data that the within group variance-covariance matrices are not equal due to genic segregation as well as other factors. In such situations the usual procedure of Fisher's linear discriminant function is not applicable and therefore needs modifications. To deal with this situation, a miniinax linear procedure given by Anderson and Bahadur [2] which does not assume the equality of variance-covariance-matrices, can be used with advantage.

Narain et al [6] attempted the use of minimax linear proceudre for comparing the performance of large number of genetic grades of sheep obtained under a crossbreeding programme and compared it with the Fisher's linear discriminant function. They have shown that the minimax linear procedure performs better or atleast same as Fisher's linear discriminant function in about three- fourths of comparisons. However, their conclusions were based on a single set of data and may vary with change in the data set and thus no conclusion could be drawn in a general sense. To deal with this situation an attempt has been made to utilise many samples of data with the bootstrap techniques for comparing the performance of two linear discriminant functions using data on different genetic grades of sheep. Findings based on procedure will certainly be more trustworthy and point out the superiorty of one procedure over the other.

2. *Material and Methods*

The basic data on sheep from the central sheep and wool Research Institute, Avikanagar consisting of single exotic breed Rambouillet (R), 3 indigenous breeds, Chokla (C), Malpura (M) and Jaisalmeri (J) and various genetic grades evolved by successive crossing under a crossbreeding project, in operation since 1964 were utilised. Upto 1970, the major breeding programme was to create halfbreds and backcross them with Rambouillet to produce 5/8ths and 3/4ths crossbreds. The data on the performance of crossbreds and exotic purebreds were very scanty till 1970 and were not therefore considered suitable for present analysis. The data for 1972-73 were only found suitable and hence used in the present investigation. Data on 3 wool quality characters viz staple length (cm), fiber diameter (μ) and medullation percentage and 6 months wool yield (kg) were available for sheep of both sexes, and were considered for the present study.

Initially for comparison purpose the equality of within grade variance-covariance matrices were tested by using the appropriate test statistic given by Anderson [1]. Further the standard linear discriminant function given by Fisher [4] and the D^2-statistics given by Mahalanobis [5] were used. Using minimax linear procedure given by Anderson and Bahadur [2] which did not assume the equality of variance-covariance matrices, D^2-statistics were also worked out. Although for comparison of these two discriminant functions, error rates which essentially are function of D^2-statistics, could also be obtained but still in the present situation only D^2-statistics were examined. In contrast to carrying out comparison based on one sample, the bootstrap technique following Efron [3] was used. The bootstrap technique is essentially a technique of resampling and in discriminant function problems, we have independent random samples from two unknown continuous probability distributions F and G on some k- dimensional space H^k

$$X_i = x_i \qquad X_i \sim \text{ind}\, F \qquad i = 1, 2, \ldots, m$$

$$Y_i = y_i \qquad Y_i \sim \text{ind}\, G \qquad j = 1, 2, \ldots, n$$

On the basis of the observed data $\underline{X} = \underline{x}$, $\underline{Y} = \underline{y}$ we used the two procedures of linear discriminant functions separately and obtained the corresponding square of discriminatory power of the two i.e. D^2-statistics to partition H^k into two complementary regions A and B with the intention of allocating a future observation Z, to the F distribution if Z belongs to A or to the G distribution if Z belongs to B. The obvious estimate of the D^2- statistics, associated with the partition (A, B) is $\hat{D}^2$ based on basic data. In order to study the performance of the two procedures the bootstrap estimate of D^2 can also be obtained. Let it be denoted by $\hat{D}^2_b$. We treat $\hat{D}^2$ as the population parameter and $\hat{D}^2_b$ obtained from bootstrap samples as an estimate of $\hat{D}^2$. In other words for comparison purposes we will be either interested in studying some statistical properties of $\hat{D}^2_b$ or in examining the distribution of the difference

$$\gamma\ [(\underline{X}, \underline{Y}), (F, G)] = \hat{D}^2_b - \hat{D}^2$$

Although one can directly consider the distribution of $\hat{D}^2_b$, but studying the distribution on the difference i.e. $\gamma[(X, Y), (F , G)]$ is much more efficient for comparing different linear discrimination procedures.

Given $\underline{X}$ and $\underline{Y}$, the estimate of D^2 can be obtained from both the Fisher's linear discriminant function and Minimax procedure as D^2 from Fisher's linear disriminant function as

$$\hat{D}^2_{(\text{FISHER})} = [(\bar{y} - \bar{x}), S^{-1} (\bar{y} - \bar{x})]$$

when $\bar{x} = \Sigma x_i/m$, $\bar{y} = \Sigma y_j/n$ and

$$S = \frac{1}{(m+n)} \left[\sum_i (x_i - \bar{x})(x_i - \bar{x})' + \sum_j (y_j - \bar{y})(y_j - \bar{y})' \right]$$

Under the assumption that there is no difference in variance-covariance matrices of the two population i.e. $\Sigma_1 = \Sigma_2 = \Sigma$. S being the estimate of Σ. In the case of different Σ_1 & Σ_2 the minimax procedure of Anderson & Bahadur [2] is followed and estimate of D^2 is obtained as

$$\hat{D}^2_{(\text{Minimax})} = \left\{ \frac{2b'd}{(b'S_ib)^{1/2} + (b'S_2b)^{1/2}} \right\}^2$$

where $d = (\bar{y} - \bar{x})$ and S_1 and S_2 are Sample estimates of Σ_1 and Σ_2 and b is to be obtained from

$$[t\,S_1 + (1-t)\,S_2]\ \ b = d$$

where $t\ (0 < t < 1)$ is obtained from

$$b'\,[t^2\,S_1 - (1-t)^2\,S_2]\,b = 0$$

There is one and only one value of t which will satisfy the above equation and ultimately give the co-efficients of minimax linear function. The procedure involves iteration on values of t and b.

Further in order to obtain the bootstrap estimate, it is implemented as follows, given the data **x, y**, bootstrap random samples

$$X_i^* = x_i^* \qquad X_i^* \sim \text{ind } \hat{F} \qquad i = 1, 2, \ldots, m$$

$$Y_j^* = y_j^* \qquad Y_j^* \sim \text{ind } \hat{G} \qquad j = 1, 2, \ldots, n$$

are generated, $\hat{F}$ and $\hat{G}$ being the sample probability distribution corresponding to F and G. For each of the procedure of Fisher's discriminant function and Minimax linear function, these yield $\hat{D}^2_{b(\text{Fisher})}$, $\hat{D}^2_{b(\text{Minimax})}$, $\gamma_{(\text{Fisher})}$ and $\gamma_{(\text{Minimax})}$. Further on repeated independent generations of (X^*, Y^*) which yield a sequence of independent realizations of

$$\hat{D}^2_{b(\text{Fisher})} \quad \text{as} \quad \hat{D}^{2*}_{b(F1)}, \hat{D}^{2*}_{b(F2)}, \ldots, \hat{D}^{2*}_{b(FN)}$$

$$\hat{D}^{2*}_{b(\text{Minimax})} \quad \text{as} \quad \hat{D}^{2*}_{b(M1)}, \hat{D}^{2*}_{b(M2)}, \ldots, \hat{D}^{2*}_{b(MN)}$$

$$\gamma^*_{(\text{Fisher})} \quad \text{as} \quad \gamma^*_{(F1)}, \gamma^*_{(F2)}, \ldots, \gamma^*_{(FN)}$$

and

$$\gamma^*_{(\text{Minimax})} \quad \text{as} \quad \gamma^*_{(M1)}, \gamma^*_{(M2)}, \ldots, \gamma^*_{(MN)}$$

which are then used to approximate the actual bootstrap distribution of $D^{2*}_{b(\text{Fisher})}$, $D^{2*}_{b(\text{Minimax})}$, $\gamma^*_{(\text{Fisher})}$ and $\gamma^*_{(\text{Minimax})}$ being the reasonable estimate of the unknown distribution of D_b^2 and γ. Further in addition to studies confined to γ, the first two moments of bootstrap estimates of D^2-statistics are also worked out. The co-efficient of variation as well as relative bias is used to compare the performance of the two procedures.

3. *Results and Discussions*

Classical Comparison of Fisher's and Minimax linear discriminant functions.

The equality of variance-covariance matrices is tested for all the possible pairs of covariance matrices among 7 grades and values of the test statistics so obtained are given in Table 1. From the results given in Table 1, based on Chi-square test statistics, it has been observed that all the covariance matrices are found to be significantly different among themselves. These results clearly show that the assumption of equality of variance-covariance matrices of the data under study does not hold good and in turn will effect the efficiency of the Fisher's linear discriminant function. On further examining the χ^2 values, it is observed that particularly the values, corresponding to comparison of Rambouillet with others, are on very higher side which clearly indicates the lower p-values. The values of D^2-statistics based on four characters for all the possible pairs among the grades by both Fisher's and Minimax discriminant procedures are given in Table 2. Of the 21 pairs of comparison, only the comparison of different grades with Rambouillet (R) has shown significantly higher D^2-values corresponding to the Minimax linear procedure as compared to Fisher's linear procedure. This may be because of high significant difference among the variance-covariance matrices. The remaining comparisons have however yielded either almost equal or lower D^2 values by the Minimax linear procedure as compared to the Fisher's linear procedure. It is concluded from these results that for the comparison of Rambouillet group with others, Minimax linear discriminant function is more efficient and in the other situations too, it is also equally efficient in relation to Fisher's linear discriminant procedure.

Table 1. Chi-square values of test criterion for testing the equality of variance-covariance matrices among the 7 grades of sheep.

Grades	R	C	M	RC (F_1)	RC (F_2)	RM (F_1)
C	253.42					
M	362.53	65.18				
RC (F_1)	187.46	104.95	38.85			
RC (F_2)	139.02	78.53	106.84	40.89		
RM (F_1)	247.14	56.54	58.08	53.17	58.92	
RM (F_2)	309.20	136.51	98.36	63.82	116.76	51.99

Note : The values in the table are significant at 1% level.

Table 2. D^2-values of Fisher's Minimax linear functions and their difference among 7 genetic groups.

Pairs	$D^2_{(\text{Fisher})}$	$D^2_{(\text{Minimax})}$	Difference
R, C	12.3904	17.7627	5.3723
R, M	38.3013	66.4996	28.1983
R, RC(F_1)	6.0815	8.2138	2.1323
R, RC(F_2)	3.5608	5.7751	2.2143
R, RM(F_1)	12.5674	17.7185	5.1511
R, RM(F_2)	7.0391	10.1260	3.0869
C, M	4.6311	4.5133	–0.1178
C, RC(F_1)	4.6777	4.3242	–0.3535
C, RC(F_2)	1.2414	1.2517	0.0103
C, RM(F_1)	1.5946	1.6095	0.0149
C, RM(F_2)	0.1925^{NS}	0.2043^{NS}	0.0118
M, RC(F_1)	17.1526	16.4229	–0.7297
M, RC(F_2)	11.1221	11.2432	0.1211
M, RM(F_1)	7.9491	7.7005	–0.2486
M, RM(F_2)	6.0899	5.7927	–0.2972
RC(F_1), RC(F_2)	1.6694	1.6215	–0.0479
RC(F_1), RM(F_1)	1.5462	1.4948	–0.0514
RC(F_1), RM(F_2)	1.1470	1.1282	–0.0188
RC(F_2), RM(F_1)	1.2327	1.2528	0.0201
RC(F_2), RM(F_2)	0.1763^{NS}	0.1875^{NS}	0.0112
RM(F_1), RM(F_2)	0.4774	0.5000	0.0226

Note : All values are significant excepting those marked with NS.

Table 3. Mean relative bias and percent co-efficient of variation of $\hat{D}_b^2$ values and their difference

Pairs	$D^2_{(Fisher)}$	$D^2_{(Minimax)}$	Difference
R, C	14.3762 (0.159,21)	20.1934 (0.139,19)	5.8172
R, M	43.3998 (0.133,28)	71.3332 (0.073,15)	27.9334
R, RC(F_1)	6.8454 (0.126,26)	9.1412 (0.113,24)	2.2958
R, RC(F_2)	3.8637 (0.085,18)	6.1666 (0.068,17)	2.3029
R, RM(F_1)	13.8439 (0.102,17)	19.4343 (0.097,15)	5.5904
R, RM(F_2)	7.6639 (0.089,23)	10.9359 (0.080,19)	3.2720
C, M	5.6207 (0.214,23)	5.5111 (0.221,23)	–0.1096
C, RC(F_1)	5.2679 (0.126,29)	4.9166 (0.137,28)	–0.3513
C, RC(F_2)	1.4173 (0.142,30)	1.4432 (0.153,30)	0.0259
C, RM(F_1)	1.7810 (0.117,37)	1.8043 (0.121,37)	0.0233
C, RM(F_2)	0.2620 (0.361,46)	0.2791 (0.366,46)	0.0171
M, RC(F_1)	17.9518 (0.047,20)	17.1107 (0.042,21)	–0.8411
M, RC(F_2)	12.0532 (0.084,18)	12.1680 (0.082,16)	0.1148
M, RM(F_1)	8.8025 (0.107,21)	8.4555 (0.098,21)	–0.3470
M, RM(F_2)	6.1879 (0.016,20)	5.9514 (0.027,20)	–0.2365
RC(F_1), RC(F_2)	1.8925 (0.134,40)	1.8431 (0.137,40)	–0.0449
RC(F_1), RM(F_1)	1.8932 (0.224,37)	1.8858 (0.262,38)	–0.0074
RC(F_1), RM(F_2)	1.7020 (0.484,44)	1.6635 (0.474,42)	–0.0385
RC(F_2), RM(F_1)	1.5388 (0.248,32)	1.5708 (0.254,32)	0.0320
RC(F_2), RM(F_2)	0.3396 (0.926,58)	0.3586 (0.912,57)	0.0190
RM(F_1), RM(F_2)	0.7876 (0.650,60)	0.8129 (0.626,60)	0.0253

Note : Figures in paranthesis are relative bias and percent co- efficient of variation respectively.

Thus it is inferred from the higher D^2 values which discriminate more efficiently as square root of it, is the discriminatory power, that Minimax linear procedure is having an advantage for almost all the situations. In the case of highly significant variance-covariance matrices, this power further gets enhanced and advocates the use of Minimax procedure.

Bootstrap comparison of Fisher's and Minimax linear disriminant functions.

For drawing the conclusions about the efficiency of one procedure in relation to other in a more general sense, the technique of bootstrap is employed by taking 100 samples from the basic data collected on 7 grades involving four characters. For each of generated sample, values of Fisher's D^2 and Minimax-D^2 statistics are obtained. The mean, relative bias and coefficient of variation of $(\hat{D}_b^2)$ statistics based no 100 samples are presented in Table-3. On examining the results from Table-3, it is again clearly seen that the values of D^2 statistics obtained by Minimax procedure are on the higher side in relation to Fisher's procedure for the comparison of Rambouillet genetic group with others. For rest of the comparisons, the differences between the D^2-statistics obtained by Minimax and Fisher's procedures are negligible. In other words, the findings of the bootstrap technique for comparison of two discriminant functions is in agreement with the finding based on one sample (Basic data) whose results are given in Table-2. The trend in the values for various comparisons are also exactly similar for the results based on Table-2 (Basic data) and Table-3 (Bootstrap technique).

On further examining the Table-3 for relative bias, it is clearly seen that for the comparison of Rambouillet with other, the values of relative bias are on the considerable lower side for Minimax procedure in comparison to values of Fisher's linear procedure. For the other comparisons, the relative bias of the two procedures are almost of the same order. Thus it is again concluded that for the situation where the dispersion matrices are widely different, it is advisable to use Minimax linear procedure for discriminatory problems. The co-efficient of variation of $\hat{D}_b^2$ given in Table-3 depicts its consistency. The lower values of co-efficient of variation in case of Minimax procedure in comparison to Fisher's linear procedure clearly indicate its superiority and consistency. The higher values of co-efficient of variation of $\hat{D}_b^2$ in some of the comparisons are probably due to inherent variability present in the basic data. Further it is also very interesting to observe the results of Table-3 that mean of the values of $\hat{D}_b^2$-statistics for both procedures are slightly greater in magnitude in relation to $\hat{D}^2$-statistics values given in Table-2 based on basic data. This clearly implies that although the trend of the comparisons remain same between two procedures but the power of discriminatory analysis of the procedure is more pronounced in the Bootstrap technique. In addition to this, the distribution aspect of the $\hat{D}_b^2$-values obtained from different samples has also been looked into and found that in most of the situations the form of the distribution is nearly normal with some positive skewness and because of this reason the mean $\hat{D}_b^2$-statistics are higher in relation of $\hat{D}^2$-statistics based on basic data.

ACKNOWLEDGEMENT

Authors are thankful to the referee for his valuable suggestion in the improvement of the paper.

REFERENCES

[1] Anderson, T.W., 1984. An introduction to multivariate statistical analysis. John Wiley & Sons, New York.

[2] Anderson, T.W. and Bahadur, R.R., 1962. Classification into two normal distributions with different co-variance matrices. *Annals of Mathematical Statistics*, 33, 420-31.

[3] Efron, B., 1982. The Jacknife, the Bootstrap and other resampling plans. STAM, *Philadelphia* 38, 1-91.

[4] Fisher, R.A., 1935. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179-88.

[5] Mahalanobis, P.C., 1936. On the generalised distance in statistics. *Proceedings of National Institute of Science, India*, 2, 49-55.

[6] Narain, P., Wahi, S.D., Malhotra, J.C. and Garg, L.K., 1991. Minimax linear procedure for comparing different grades of sheep in crossbreeding programmes. *Ind. J. Ani. Sci.*, 61(3), 305-10.