Aligarh Journal of Statistics
Vol. 27 (2007), 1-9

PERFORMANCE OF FISHER DISCRIMINANT FUNCTION IN CLASSIFICATION PROBLEM– *A* SIMULATION STUDY

A.V. Dattatreya Rao and Nunna Srinivasa Rao

ABSTRACT

The criterion of error of misclassification of Fisher linear classifier has not been fully investigated in the context of intra class correlations in classifying two Multivariate Normal populations [Johnson and Wichern (2001)]. Here an attempt is made to study the performance of Fisher classifier in classifying a new observation X_0 into one of the two multivariate normal populations $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$ in the above said context using simulation study.

1. INTRODUCTION

The purpose of pattern recognition is to determine to which category or class (population) a given sample belongs to. Through an observation or measurement process, we obtain an observation vector. The observation vector serves as the input to a decision rule (statistic) by which we assign the sample to one of the given classes. Let us assume that the observation is a random vector whose conditional density function depends on its class. If the conditional density function for each class is known, then the pattern recognition problem becomes a problem in statistical hypothesis testing [Fukunaga (1990)]. From this angle, the classification problem can be viewed as a Pattern Recognition problem. Here, it is aimed to study the performance of Fisher classifier in the said context.

As it is proposed to study the performance of Fisher classifier in case of two Multivariate Normal populations, the decision rule with usual notations, to test

$$H_0: \ X_0 \in \pi_1 \cong N_p(\mu_1, \Sigma) \qquad \text{Vs} \quad H_1: \ X_0 \in \pi_2 \cong N_p(\mu_2, \Sigma)$$

is presented here for ready reference:

Fisher Rule given by

Allocate X_0 to π_1 if

$$(\mu_1 - \mu_2)'\Sigma^{-1}X \geq \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \tag{1.1}$$

Otherwise allocate X_0 to π_2.

When the parameters are unknown, then they are estimated based upon two samples of sizes n_1 and n_2 respectively from π_1 and π_2, before the classification is done, and the decision rule is obtained by replacing μ_1, μ_2, Σ respectively by $\overline{X}_1$, $\overline{X}_2$ and S_{pooled} in (1.1).

2. COMPUTATION OF TOTAL PROBABILITY OF MISCLASSIFICATION (TPM)

Computing probability of error of misclassification is laborious as it involves evaluation of multiple integrals [Fukunaga (1990)]. However TPM can be estimated by means of the confusion matrix using simulation and the process is as under:

Total Probability of Misclassification (TPM) is defined as,

$TPM = \Pr($ misclassifying a π_1 observation or misclassifying a π_2 observation)

$= \Pr($ observation comes from π_1 and is misclassified as $\pi_2)$

$+ \Pr($ observation comes from π_2 and is misclassified as $\pi_1)$.

$= \Pr(X \in \pi_2 \mid \pi_1)\Pr(\pi_1) + \Pr(X \in \pi_1 \mid \pi_2)\Pr(\pi_2)$.

Therefore the confusion matrix is of the form

		Predicted	
		π_1	π_2
Actual	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$
	π_2	$n_{2M} = n_1 - n_{2C}$	n_{2C}

n_{1C} = Number of π_1 items correctly classified as π_1 items,

n_{2C} = Number of π_2 items correctly classified as π_2 items,

n_{1M} = Number of π_1 items misclassified as π_2 items,

n_{2M} = Number of π_2 items misclassified as π_1 items

and $$T\hat{P}M = \frac{n_{1M} + n_{2M}}{n_1 + n_2}. \qquad (2.1)$$

3. METHODOLOGY

Let $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$ be the two normal populations to be discriminated for the specified parameters using above said method. Without loss of generality μ_1 is taken as null vector, μ_2, the competitive vector is taken as $k\,1'$, k varying from $0.5(0.5)3$.

Therefore to test the hypothesis that

H_0 : The sample is drawn from $N_p(\mu_1, \Sigma)$

Vs

H_1 : The sample is drawn from $N_p(\mu_2, \Sigma)$

utilizing the relationship $V^{1/2}\rho V^{1/2} = \Sigma$, where V and ρ are respectively the diagonal matrix of variances and intra class correlations, and setting $\Sigma = \rho$. The dimensionality (p) ranging from $3(1)10$ and the correlations in the correlation matrix ranging from 0.2 to 0.8 spreading with equidistant along the rows are taken in ascending fashion. In another case it is taken in the descending order. The priori probabilities are $p(\pi_1) = p_1$ and $p(\pi_2) = p_2 = 1 - p_1$. Here we have taken $p_1 = 0.1(0.1)0.9$.

a) Generation of multivariate normal data

The vectors of multivariate normal data with the specified parameters can be obtained starting from univariate standard normal data and the operational procedure is as follows:

Let U_1 and U_2 be uniform random variates between $[0,1]$. According to Box and Muller (1958)

$$Z_1 = \sqrt{-2\ln U_1}\cos(2\pi U_2),\; Z_2 = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

will follow $N(0, 1)$.

Let $Y_1 = Z_1 + \mu_1$ and $Y_2 = Z_2 + \mu_2$ will follow $N(\mu_i, 1)$, $i = 1, 2$.

Let Λ and Φ be respectively the matrices of eigen values and the corresponding eigen vectors of the Σ.

Let X be $\Phi\Lambda^{1/2}Y$ where $Y = (Y_1\; Y_2\; \cdots\; Y_p)'$. It can be seen that X follows bivariate normal with

$\mu = (\mu_1\; \mu_2\; \cdots\; \mu_p)'$ and Σ as the parameters. Using this methodology we can generate multivariate normal data for any specified set of parameters.

When the parameters are not specified then the simulation study involves two phases - *Training* (estimation) and *Validation* (classification). In training phase, based on samples of size n_1 and n_2 drawn respectively from the specified populations, estimation of parameters is done, then the classifier is constructed, while in validation phase, another set of pseudo random vectors are drawn from multivariate normal and is used to study the performance of classifiers under consideration.

b) Orthogonal transformation

Classification using orthogonal transformation is worth attempting for study of relative performance.

In the more general case where $\Sigma \neq I$, the observation noise is correlated and is often called coloured noise. In such a case, one can introduce Whitening transformation $Y = A X$, where $A' \Sigma A = I$.

Then the procedure discussed so far applies to Y if we replace μ_i by $A \mu_i$. This transformation is reversible, and the observation Y can be classified as effectively as X_0. Construction of A is as explained in section 3 a).

c) Jackknife procedure

i) Start with the n_1 group of observations. Omit one observation from this group and develop a classification rule based on the remaining $(n_1 - 1)$, n_2 observations.

ii) Classify the observation using the function constructed in step i).

iii) Repeat steps i) and ii) until all the n_1 observations are classified. Let n_{1M} be the number of observations misclassified in this group.

iv) Repeat steps i) through iii) for the π_2 observations. Let n_{2M} be the number of observations misclassified in this group. Once n_{1M} and n_{2M} are obtained, TPM is estimated using (2.1).

d) Choice of parameters for the construction of tables

If π_1 and π_2 denote populations with $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$ respectively, then

$\mu_1 = (0\ 0 \cdots 0)'_{p\times 1}$

$\mu_2 = k\,(1\ 1 \cdots 1)'_{p\times 1}$

$k = 0.5\,(0.5)\,3$ and Σ the variance-covariance matrix is given by

$\Sigma = (\Sigma_{ij})_{p\times p}$ with $\Sigma_{ij} = 1$, if $i = j$.

and

Σ_{ij}, if $i \neq j$.

Case I:

$$\Sigma_{12} = 0.8$$

$$\Sigma_{1j+1} = \Sigma_{1j} - [0.6/(p-2)] \text{ for } j = 2, \cdots, p.$$

$$\Sigma_{ij} = \begin{cases} \Sigma_{1j} & \text{if } i < j \text{ for } j > 1 \\ 1 & \text{if } i = j \end{cases}$$

Case II:

$$\Sigma_{12} = 0.2$$

$$\Sigma_{1j+1} = \Sigma_{1j} + [0.6/(p-2)] \text{ for } j = 2, \cdots, p.$$

$$\Sigma_{ij} = \begin{cases} \Sigma_{1j} & \text{if } i < j \text{ for } j > 1 \\ 1 & \text{if } i = j \end{cases}.$$

4. SALIENT OBSERVATIONS ON THE SIMULATION RESULTS

The tolerance limit for TPM is taken as 10% in the present study. The exercise involved construction of 14 tables [two for known cases (one in decreasing and the other is in increasing order of correlations) and 12 for unknown cases of which each of six possible sample cases $(10, 10)$, $(10, 15)$, $(10, 20)$, $(15, 15)$, $(15, 20)$, $(20, 20)$ for decreasing and increasing order of correlations]. The conclusions are drawn consolidating all 14 tables are as follows. Of which only 4 tables bearing numbers 4.1 to 4.4 Tables (4.1 and 4.2 pertaining to known cases; 4.3 and 4.4 are for $(10, 20)$ in increasing and decreasing order of correlations respectively) are as appended in the annexure to save space and the remaining tables are available with the authors.

a) When parameters are specified

i) Fisher classifier performs better under orthogonal transformation for the decreasing pattern of correlations in Σ matrix.

ii) When the tolerance condition is relaxed Fisher classifier performs better under *given vector* (a new vector to be classified) for the increasing pattern of correlations in Σ matrix.

b) When parameters are not specified

i) Fisher classifier under *given vector* is performing better when compared to Orthogonal transformation and Jackknifing.

ii) Under *Jackknifing* TPM is within tolerance limit with the gradual increase in sample sizes.

ANNEXURE

Percentage TPM values of Fisher classifier in classifying two populations
$\pi_1 : N_p(0,\Sigma)$ and $\pi_2 : N_p(\mu_2,\Sigma)$

Table 4.1: Decreasing order of correlations in Σ matrix when the parameters are specified

μ_2	$(3.0)1'$		$(2.5)1'$		$(2.0)1'$		$(1.5)1'$		$(1.0)1'$		$(0.5)1'$	
P	❶	❷	❶	❷	❶	❷	❶	❷	❶	❷	❶	❷
3	2.0	6.0	5.0	1.6	8.8	4.0	15.3	9.2	23.9	15.3	9.2	23.9
4	3.3	0.3	6.2	1.2	11.6	1.9	19.2	7.9	30.1	19.2	7.9	30.1
5	3.2	0.1	7.8	0.5	10.6	1.2	17.7	4.8	27.5	17.7	4.8	27.5
6	2.0	2.0	2.7	1.0	8.2	1.6	14.9	5.5	26.5	14.9	5.5	26.5
7	3.8	0.0	5.2	0.1	11.1	1.4	15.0	4.3	24.4	15.0	4.3	24.4
8	2.7	0.0	5.3	0.0	10.9	0.5	12.9	4.0	25.9	12.9	4.0	25.9
9	2.0	0.0	4.2	0.0	8.0	0.5	14.8	2.3	25.1	14.8	2.3	25.1
10	4.1	0.0	6.2	0.0	10.1	0.6	16.4	2.4	25.3	16.4	2.4	25.3

❶ For a given vector
❷ Using orthogonal transformation.

Table 4.2: Increasing order of correlations in Σ matrix when the parameters are specified

μ_2	$(3.0)1'$		$(2.5)1'$		$(2.0)1'$		$(1.5)1'$		$(1.0)1'$		$(0.5)1'$	
P	❶	❷	❶	❷	❶	❷	❶	❷	❶	❷	❶	❷
3	24.2	77.9	26.0	73.9	26.6	75.0	31.6	74.7	35.5	31.6	74.7	35.5
4	20.1	71.2	23.2	69.9	26.3	73.5	30.9	72.5	36.1	30.9	72.5	36.1
5	17.0	61.8	19.4	64.6	28.2	67.3	30.3	68.5	37.4	30.3	68.5	37.4
6	16.4	59.7	21.1	60.6	27.3	62.0	31.0	61.1	36.2	31.0	61.1	36.2
7	18.2	54.4	21.7	56.3	26.7	57.3	32.0	57.4	38.1	32.0	57.4	38.1
8	18.1	50.2	20.9	51.7	25.7	52.7	30.6	53.4	40.0	30.6	53.4	40.0
9	18.3	44.6	21.9	43.5	28.5	42.2	30.1	44.9	37.4	30.1	44.9	37.4
10	16.4	33.1	21.6	34.0	25.2	33.4	31.0	36.8	36.3	31.0	36.8	36.3

❶ For a given vector
❷ Using orthogonal transformation.

Table 4.3: Decreasing order of correlations in Σ matrix when the parameters are not specified

Size of first sample: 10 Size of second sample: 20

μ_2	$(3.0)1'$			$(2.5)1'$			$(2.0)1'$		
P	❶	❷	❸	❶	❷	❸	❶	❷	❸
3	0.7	18.0	7.7	2.5	28.0	9.2	3.0	25.5	13.5
4	0.2	33.5	5.5	0.5	35.5	10.0	3.5	36.5	12.2
5	0.0	37.7	4.7	0.0	37.5	7.0	1.7	33.2	9.7
6	0.0	40.5	3.2	0.2	36.2	8.0	0.7	36.2	7.7
7	0.0	45.5	4.5	0.0	41.7	6.7	0.5	42.2	8.2
8	0.0	47.2	3.0	0.0	43.7	3.0	0.0	42.2	7.0
9	0.0	46.0	2.5	0.0	46.7	5.0	0.0	42.5	8.0
10	0.0	47.5	3.2	0.0	47.7	4.0	0.0	45.5	6.5
μ_2	$(1.5)1'$			$(1.0)1'$			$(0.5)1'$		
P	❶	❷	❸	❶	❷	❸	❶	❷	❸
3	10.2	34.5	17.0	21.2	37.5	22.7	34.0	44.0	34.5
4	6.5	34.7	13.7	8.7	13.7	9.2	32.0	44.7	31.0
5	4.2	39.7	12.5	11.7	39.5	17.5	22.5	42.0	26.0
6	4.0	36.7	10.0	12.0	39.0	16.2	22.5	44.0	25.7
7	2.0	41.0	12.2	8.7	41.5	15.2	20.7	46.0	25.5
8	1.5	38.2	10.0	6.2	42.5	16.2	22.0	43.0	25.5
9	1.0	44.0	9.7	7.5	45.7	14.0	25.7	43.5	28.7
10	2.2	44.5	7.7	5.0	41.5	14.2	22.0	47.7	25.5

❶ For a given vector

❷ Using orthogonal transformation

❸ Jackknife method.

Table 4.4: Increasing order of correlations in Σ matrix when the parameters are not specified

Size of first sample: 10 Size of second sample: 20

μ_2	$(3.0)1'$			$(2.5)1'$			$(2.0)1'$		
P	❶	❷	❸	❶	❷	❸	❶	❷	❸
3	0.0	19.7	4.5	0.7	22.0	9.5	3.7	25.5	13.7
4	0.5	15.7	6.5	0.7	15.5	8.5	1.5	18.0	13.2
5	0.0	7.5	5.5	0.0	5.5	7.7	1.7	9.5	9.2
6	0.0	0.5	3.7	0.2	2.5	4.7	1.7	6.0	8.2
7	0.0	0.2	3.2	0.5	2.2	6.2	0.2	4.5	9.7
8	0.0	6.5	3.5	0.0	7.0	4.2	0.5	9.5	6.5
9	0.0	12.0	1.7	0.0	13.5	4.5	0.2	17.5	6.0
10	0.0	34.5	2.5	0.5	33.7	3.0	0.2	28.5	6.0
μ_2	$(1.5)1'$			$(1.0)1'$			$(0.5)1'$		
P	❶	❷	❸	❶	❷	❸	❶	❷	❸
3	5.5	23.2	13.7	28.2	27.2	21.0	32.7	36.7	31.7
4	6.5	23.0	16.7	13.7	24.2	19.0	27.5	33.5	30.0
5	5.0	16.0	13.5	14.2	22.0	20.2	29.7	32.2	32.7
6	3.2	11.5	10.0	10.2	20.7	19.0	27.7	29.0	30.0
7	2.0	6.5	10.2	13.5	25.0	21.2	24.5	31.0	24.2
8	2.5	11.2	10.2	8.5	26.0	16.7	24.5	33.2	25.2
9	0.5	26.0	8.2	7.7	27.5	17.5	24.2	40.7	26.5
10	1.7	28.5	10.2	6.2	34.5	13.5	24.5	44.0	26.0

❶ For a given vector
❷ Using orthogonal transformation
❸ Jackknife method.

Acknowledgement

The authors thank University Grants Commission (UGC) New Delhi, India for extending financial support in the form of Minor Research Project to the Second author for carrying out this project.

REFERENCES

Box, G.E.P. and Muller, M.E. (1958): A note on the generation of normal variates. *Ann. Math. Statist*., **29**, 610-611.

Duda, R.O., Hart, P.E. and David, G. Stork (2002): *Pattern Classification*. Jonh Wiley, New York.

Fukunaga, K. (1990): *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

Johnson Richard, A., and Wichern Dean. W. (2001): *Applied Multivariate Statistical Analysis*. Prentice Hall International Inc., New Jersey.

Srinivasa Rao Nunna (2003): On some statistical and related approaches to pattern recognition, an unpublished Ph.D., Thesis, Nagarjuna University, India.

Received : 15-07-2004

A.V. Dattatreya Rao
Department of Statistics,
Nagarjuna University,
Nagarjuna Nagar-522 510, India.

Nunna Srinivasa Rao
Department of Statistics,
Andhra Loyola College, Vijayawada-520 008, India.
e-mail: nunna68@yahoo.co.in