J. Ind. Soc. Agril. Statist.
61(3), 2007 : 336-340

Probability of Misclassification for Sample Linear Discriminant Function when Training Samples have Misallocated Observations

B. Singh
IVRI, Izatnagar (U.P.)

(Received : September 2002)

SUMMARY

Approximate expressions for probability of misclassification (PMC) are derived for sample linear discriminant function (SLDF) when training samples have misallocated observations. The PMC of SLDF using misallocated training samples is also obtained through simulated samples from two multivariate ncrnal populations for examining the validity of derived expressions for practical applications. The numerical results reveal that the misallocation in training samples increases PMC. The effect is small for low level of misallocation and moderate correlation and serious for high level of misallocation and high correlation among component variables. Further, the derived expressions for PMC provide numerical results close to the simulated values for small and moderate values of Δ^2 (Mahalanobis distance).

Key words : Discriminant function, Misallocated observations, PMC, SLDF, Training samples.

1. INTRODUCTION

Fisher's linear discriminant function is the popular technique in the field of discriminant analysis. This yields optimal results in the sense of smallest probability of misclassification (PMC) when parameters are known. The performance of Fisher's linear discriminant function has been studied by Singh (2001) when parameters are unknown. The assumptions involved in the construction of LDF include that the training samples are correctly classified. Sometimes the training samples may have misallocated observations. Lachenbruch (1966) has given large sample expression for PMC when training samples have misallocated observations. McLachlan (1972) obtained the asymptotic results for the same model. In this paper, we derive approximate expressions for PMC of sample linear discriminant function (SLDF) when training samples have misallocated observations and examine its validity for practical applications by using the simulated samples from multivariate normal populations.

2. SAMPLE LINEAR DISCRIMINANT FUNCTION

Consider two p-variate normal populations π_1 and π_2. Suppose that we have a sample $x^{(1)}_{\alpha}$, ($\alpha = 1, 2, \ldots, N_1$) from population π_1 with distribution $N(\mu_1, \Sigma)$ and a sample $x^{(2)}_{\alpha}$, ($\alpha = 1, 2, \ldots, N_2$) from population π_2 with distribution $N(\mu_2, \Sigma)$. These are taken as training samples to obtain estimates of μ_1, μ_2 and Σ. The estimates are defined as

$$\bar{X}_1 = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} x^{(1)}_{\alpha}, \quad \bar{X}_2 = \frac{1}{N_2}\sum_{\alpha=1}^{N_2} X^{(2)}_{\alpha} \text{ for } \mu_1 \text{ and } \mu_2,$$

respectively and

$$S = \frac{1}{n}\sum_{\alpha=1}^{N_1}(x^{(1)}_{\alpha} - \bar{X}_1)(x^{(1)}_{\alpha} - \bar{X}_1)']$$

$$+ \sum_{\alpha=1}^{N_2}(x^{(2)}_{\alpha} - \bar{X}_2)(x^{(2)}_{\alpha} - \bar{X}_2)'] \text{ for } \Sigma$$

where $\quad n = (N_1 + N_2 - 2)$

The classification statistic is defined as

$$W = X'S^{-1}(\overline{X}_1 - \overline{X}_2) - \frac{1}{2}(\overline{X}_1 + \overline{X}_2)' S^{-1}(\overline{X}_1 - \overline{X}_2) \qquad (2.1)$$

Suppose that some of the observations in two training samples are misallocated. It is assumed that α_1 is the proportion of the N_1 observations in training sample one that really belong to population π_2 and α_2 is the proportion of the N_2 observations in training sample two that really belong to population π_1.

Let $\overline{X}_{iT}$ is the mean of those members correctly classified as coming from population π_i and $\overline{X}_{iM}$ is the mean of those members from π_i incorrectly classified as member of π_{3-i}, i = 1, 2. The sum and difference of the sample means can be expressed as

$$\overline{X}_1 + \overline{X}_2 = (1-\alpha_1)\overline{X}_{1T} + \alpha_1 \overline{X}_{2M} + \alpha_2 \overline{X}_{1M} + (1-\alpha_2)\overline{X}_{2T}$$

$$\overline{X}_1 - \overline{X}_2 = (1-\alpha_1)\overline{X}_{1T} + \alpha_1 \overline{X}_{2M} - \alpha_2 \overline{X}_{1M} - (1-\alpha_2)\overline{X}_{2T}$$

The sample covariance matrix is expressed as

$$S = S_T + \frac{c_1}{n}(\overline{X}_{1T} - \overline{X}_{2M})(\overline{X}_{1T} - \overline{X}_{2M})'$$
$$+ \frac{c_2}{n}(\overline{X}_{1M} - \overline{X}_{2T})(\overline{X}_{1M} - \overline{X}_{2T})' \qquad (2.2)$$

where $c_i = \alpha_i(1-\alpha_i)N_i$, i = 1, 2 and S_T is the true sample covariance matrix.

Lachenbruch (1966), observing the expressions (2.1 and 2.2) to be quite difficult to analyse, performed a series of sampling experiments to study the effects of misallocation in training samples. His results are not useful in computing PMC from numerical data. Here, we develop approximate theoretical expressions for PMC of SLDF using misallocated training samples in the following section.

3. PROBABILITY OF MISCLASSIFICATION

The two probabilities of misclassification are defined as

$$P(2 \mid 1) = P(W \le 0 \mid X \in \pi_1) \text{ and}$$
$$P(1 \mid 2) = P(W > 0 \mid X \in \pi_2) \qquad (3.1)$$

We write W in (2.1) approximately (see, McLachlan 1972) as

$$W = u'S_T^{-1}v/t$$

where

$$u = (\overline{X}_1 - \overline{X}_2)$$
$$= (1-\alpha_1)\overline{X}_{1T} + \alpha_1 \overline{X}_{2M} - \alpha_2 \overline{X}_{1M} - (1-\alpha_2)\overline{X}_{2T}$$

$$v = X - \frac{1}{2}(\overline{X}_1 + \overline{X}_2)$$
$$= X - \frac{1}{2}[(1-\alpha_1)\overline{X}_{1T} + \alpha_1 \overline{X}_{2M} + \alpha_2 \overline{X}_{1M}$$
$$+ (1-\alpha_2)X_{2T}]$$

$$t = 1 + \frac{c_1}{n}(\overline{X}_{1T} - \overline{X}_{2M})'S_T^{-1}(\overline{X}_{1T} - \overline{X}_{2M})$$
$$+ \frac{c_2}{n}(\overline{X}_{1M} - \overline{X}_{2T})'S_T^{-1}(\overline{X}_{1M} - \overline{X}_{2T})$$

and verify the validity of this approximation through simulation technique.

Now $t > 0$ and S_T^{-1} is positive definite. Hence, for obtaining PMC, we express W equivalent to

$$W = u'S_T^{-1}V \qquad (3.2)$$

Suppose $X \in \pi_1$, then

$$u \sim N[(\mu_1 - \mu_2)(1-\alpha_1-\alpha_2), (N_1^{-1} + N_2^{-1})\Sigma] \text{ and}$$

$$v \sim N[(\mu_1 - \mu_2)(1+\alpha_1-\alpha_2)/2,$$
$$(1 + (4N_1)^{-1} + (4N_2)^{-1})\Sigma]$$

Let $u_1 = \sqrt{}[N_1N_2/(N_1+N_2)]\,u$ and

$v_1 = \sqrt{}[4N_1N_2/(N_1+N_2+4N_1N_2)]\,v$

then

$$u_1 \sim N[(\mu_1 - \mu_2)(1-\alpha_1-\alpha_2)\sqrt{\{N_1}N_2/(N_1+N_2)\}, \Sigma]$$

$$v_1 \sim N[(\mu_1 - \mu_2)(1+\alpha_1-\alpha_2)$$
$$\sqrt{\{N_1}N_2/(N_1+N_2+4N_1N_2)\}\Sigma]$$

$$W = k[(u_1 + v_1)'S_T^{-1}(u_1 + v_1) - (u_1 - v_1)'S_T^{-1}(u_1 - v_1)] \qquad (3.3)$$

where $k = (1/8N_1N_2)\sqrt{[(N_1+N_2)(N_1+N_2+4N_1N_2)]}$

Note that (u_1+v_1) and (u_1-v_1) are independently normally distributed (see Moran 1975) as

$$(u_1+v_1) \sim N(\delta_1, k_1\Sigma) \text{ and } (u_1-v_1) \sim N(\delta_2, k_2\Sigma)$$

where

$$\delta_1 = (\mu_1-\mu_2)[\{(1-\alpha_1-\alpha_2)/(N_1+N_2)^{-1/2}\}$$
$$+\{(1+\alpha_1-\alpha_2)/(N_1+N_2+4N_1N_2)^{-1/2}\}]\sqrt{(N_1N_2)}$$
$$k_1 = 2\,[1+(N_1-N_2)/\{(N_1+N_2)\,(N_1+N_2+4N_1N_2)\}^{1/2}]$$
$$\delta_2 = (\mu_1-\mu_2)[\{(1-\alpha_1-\alpha_2)/(N_1+N_2)^{-1/2}\}$$
$$-\{(1+\alpha_1-\alpha_2)/(N_1+N_2+4N_1N_2)^{-1/2}\}]\sqrt{(N_1N_2)}$$

and

$$k_2 = 2\,[1-(N_1-N_2)/\{(N_1+N_2)(N_1+N_2+4N_1N_2)\}^{1/2}]$$

Now let

$$t_1 = (u_1+v_1)\,k_1^{-1/2} \text{ and } t_2 = (u_1-v_1)\,k_2^{-1/2}$$

Then, one writes

$$W = k[k_1\,t_1'\,S_T^{-1}t_1 - k_2\,t_2'\,S_T^{-1}t_2] \qquad (3.4)$$

where t_1 and t_2 are independently distributed as

$$t_1 \sim N(\delta_1/\sqrt{k_1}, \Sigma) \text{ and } t_2 \sim N(\delta_2/\sqrt{k_2}, \Sigma)$$

Now, by using Theorem (5.2.2) of Anderson (1984, p-163), we write the classification statistic W as

$$W = (U_1/V_1) - (U_2/V_2) \qquad (3.5)$$

where U_1 and U_2 are independent to V_1 and V_2, $U_1 \sim g_1\chi^2_p(\Delta_1^2)$, $U_2 \sim g_2\chi^2_p(\Delta_2^2)$ and V_1 and V_2 are identically distributed as chi-square on $(n-p+1)$ degrees of freedom. The constants $g_i = (nkk_i)$, and Δ_i^2, $i = 1, 2$ are defined as

$$g_1 = (n/4N_1N_2)[N_1-N_2+\{(N_1+N_2)(N_1+N_2+4N_1N_2)\}^{1/2}]$$

and

$$g_2 = (n/4N_1N_2)[N_2-N_1+\{(N_1+N_2)(N_1+N_2+4N_1N_2)\}^{1/2}]$$
$$\Delta_i^2 = (1/k_i)^{-1}\,\delta_i'\,\Sigma^{-1}\,\delta_i, \; i=1, 2, \text{ that is}$$
$$\Delta_1^2 = (N_1N_2/k_1)[(1-\alpha_1-\alpha_2)\,(N_1+N_2)^{-1/2}$$
$$+(1+\alpha_1-\alpha_2)\,(N_1+N_2+4N_1N_2)^{-1/2}]^2\,\Delta^2$$
$$\Delta_2^2 = (N_1N_2/k_2)[(1-\alpha_1-\alpha_2)\,(N_1+N_2)^{-1/2}$$
$$-(1+\alpha_1-\alpha_2)(N_1+N_2+4N_1N_2)^{-1/2}]^2\,\Delta^2$$

and Δ^2, the Mahalanobis distance between two multivariate populations, is defined as

$$\Delta^2 = (\mu_1-\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2)$$

The exact distribution of W (3.5) is difficult to obtain since V_1 and V_2 are not necessarily same except that they are identically distributed. Thus V_1 and V_2 are asymptotically same. So one can assume V_1 and V_2 as approximately equal (say, V) for all values of p and derive the approximate expressions for PMC of SLDF from misallocated training samples. We examine the validity of this approximation by comparing wit:: corresponding results based on simulated samples. These simulated results may not be correct but provide quite good results for practical purposes.

So, with the assumption of same denominator in (3.5) we write W as

$$W = (U_1-U_2)/V \qquad (3.6)$$

Since V, a chi-square variate, is positive so for obtaining PMC, W (3.6) is equivalently expressed as

$$W = U_1-U_2 \qquad (3.7)$$

The exact distribution of U_1 and U_2 can be expressed as linear combination of chi-square probabilities (see Johnson and Kotz 1970). These expressions would be computationally tedious for practical applications. For simplicity, we assume that U_1 and U_2 are approximately distributed as $a\chi^2_b$ and $c\chi^2_d$ respectively, where the constants a, b, c and d are obtained by using the Patnaik's two moments approximation (Patnaik 1949) as under

$$a = Var(U_1)/2E(U_1), \; b = 2E^2(U_1)/Var(U_1)$$
$$c = Var(U_2)/2E(U_2) \text{ and } d = 2E^2(U_2)/Var(U_2)$$

By using the expression for r-th raw moment of a non-central chi-square variate (Johnson and Kotz 1970) we obtain that

$$E(U_1) = g_1(p+\Delta_1^2), \; Var(U_1) = 2g_1^2(p+2\Delta_1^2)$$
$$E(U_2) = g_2(p+\Delta_2^2), \; Var(U_2) = 2g_2^2(p+2\Delta_2^2)$$

The probability of misclassifying X to Δ_2, when it actually belongs to π_1, is given by

$$\begin{aligned} P(2\mid 1) &= P(W \le 0 \mid \pi_1) \\ &= P(U_1 \le U_2 \mid \pi_1) \\ &= P(a\chi^2_b \le c\chi^2_d \mid \pi_1) \\ &= I_{wo}(b/2, d/2) \qquad (3.8) \end{aligned}$$

where, I_x (a, b) is the value of incomplete beta and $wo = c/(a+c)$.

Similarly, the expression for $X \in \pi_2$ can be obtained as

$$\begin{aligned} P(1 \mid 2) &= P(W > 0 \mid \pi_2) \\ &= P(U_1 > U_2 \mid \pi_2) \\ &= 1 - P(U_1 \le U_2 \mid \pi_2) \\ &= 1 - I_{w1}(b^*/2, d^*/2\ ) \\ &= I_w(d^*/2, b^*/2\ ) \qquad (3.9) \end{aligned}$$

By interchanging Δ^2_i by Δ^2_{i*}, $i = 1,2$ as

$$\Delta_{1*}{}^2 = (N_1N_2/k_1)[(1-\alpha_1-\alpha_2)\,(N_1+N_2)^{-1/2} + (1+\alpha_1-\alpha_2)\,(N_1+N_2+4N_1N_2)^{-1/2}]^2\ \Delta^2$$

$$\Delta_{2*}{}^2 = (N_1N_2/k_2)[(1-\alpha_1-\alpha_2)\,(N_1+N_2)^{-1/2} - (1+\alpha_1-\alpha_2)\,(N_1+N_2+4N_1N_2)^{-1/2}]^2\ \Delta^2$$

where

$$\begin{aligned} w &= 1 - w_1 \\ &= a^* / (a^* + c^*) \end{aligned}$$

and a^*, b^*, c^*, d^* are constants corresponding to the case when $X \in \pi_2$.

4. NUMERICAL RESULTS

Here, we generate $N_1 + N_2 + 2$ observations from two p-variate normal populations, $N_1 + 1$ from π_1 and N_2+1 from π_2, with certain apriori values of parameters. The first $N_1 + N_2$, p-variate observations are used to obtain SLDF. The remaining two observations, one from each population were used to get numerical value for SLDF for each group, separately. This process was repeated 1000 times to get PMC for SLDF for each group, separately, for one fixed set of parameters p, N_1 and N_2. These values for PMC are for training samples with no misallocation. The results corresponding to misallocation are obtained by interchanging some observations (say, C) in two training samples, randomly. The corresponding theoretical values are computed from the formulae (3.7, 3.8) for necessary comparison with simulated results.

Since the PMCs are invariant under linear transformations, so the numerical results presented in Tables 1 and 2 are obtained for following apriori values

$\Sigma = I$, $\mu_1 = (0, 0, \ldots, 0)$, $\mu_2 = (\Delta, 0, 0, \ldots, 0)$, $p = 3, 5$ $\Delta^2 = 1(2)7$, $N_1 = N_2 = 20$, $N_1 = 25$, $N_2 = 15$, $\alpha_1 = C/N_1$, $\alpha_2 = C/N_2$, $C = 0, 2, 4, 6$

The numerical results reveal that the misallocation in training samples increases PMC. The effect is small for low level of misallocation (10%) and serious for high level of misallocation (30%) and high correlation. Further, the simulated (S) values of PMC for SLDF agree with the corresponding theoretical (T) values for small and moderate values of Mahalanobis distance (Δ^2) between the two multivariate normal populations. This implies that the derived expressions for PMC give good results and hence may be used for practical applications when training samples have misallocated observations.

REFERENCES

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis* (second edition). John Wiley & Sons, New York.

Johnson, N.L. and Kotz, C. (1970). *Continuous Univariate Distributions* John Wiley & Sons, New York.

Lachenbruch, P.A. (1966). Discriminant analysis when initial samples are misclassified. *Technometrics,* **8**, 657-662.

McLachlan, G.J.(1972). Asymptotic results for discriminant analysis when the initial samples are misclassified. *Technometrics,* **14**, 415-422.

Moran, M.A. (1975). On the expectation of errors of allocation associated with linear discriminant function. *Biometrika,* **62**, 141-148.

Patnaik, P.B. (1949). The non-central χ^2 and F distributions and their applications. *Biometrika,* **36**, 202-232.

Singh, B. (2001). On the performance of two sample linear discriminant function. *J. Ind. Soc. Agril. Statist.,* **54**, 209-220.

Table 1. Probability of misclassification under equal misallocation

p	Δ^2	Popula-tion	$\alpha_1 = \alpha_2 = .00$		$\alpha_1 = \alpha_2 = .10$		$\alpha_1 = \alpha_2 = .20$		$\alpha_1 = \alpha_2 = .30$	
			T	S	T	S	T	S	T	S
3	1	π_1	.328	.333	.343	.347	.352	.367	.393	.386
		π_2	.328	.347	.343	.352	.352	.392	.393	.406
	3	π_1	.212	.195	.210	.207	.220	.230	.254	.282
		π_2	.212	.222	.210	.242	.220	.274	.254	.300
	5	π_1	.140	.138	.141	.151	.150	.182	.168	.232
		π_2	.140	.154	.141	.173	.150	.202	.168	.242
	7	π_1	.099	.100	.099	.119	.103	.149	.117	.199
		π_2	.099	.114	.099	.124	.103	.143	.117	.213
5	1	π_1	.344	.335	.356	.366	.368	.394	.411	.410
		π_2	.344	.354	.356	.378	.368	.402	.411	.421
	3	π_1	.210	.215	.222	.247	.238	.285	.281	.314
		π_2	.210	.215	.222	.263	.238	.306	.281	.343
	5	π_1	.144	.153	.151	.185	.163	.232	.193	.282
		π_2	.144	.149	.151	.189	.163	.252	.193	.296
	7	π_1	.102	.114	.105	.141	.112	.193	.140	.255
		π_2	.102	.116	.105	.144	.112	.214	.140	.264

Table 2. Probability of misclassification under unequal misallocation

p	Δ^2	Popula-tion	$\alpha_1 = \alpha_2 = .00$		$\alpha_1 = .13, \alpha_2 = .08$		$\alpha_1 = .27, \alpha_2 = .16$		$\alpha_1 = .40, \alpha_2 = .24$	
			T	S	T	S	T	S	T	S
3	1	π_1	.323	.319	.340	.339	.361	.377	.405	.421
		π_2	.340	.371	.350	.361	.362	.365	.408	.422
	3	π_1	.201	.218	.215	.220	.237	.271	.281	.349
		π_2	.207	.229	.206	.232	.209	.251	.251	.312
	5	π_1	.137	.138	.152	.155	.170	.209	.218	.304
		π_2	.142	.166	.134	.167	.132	.190	.152	.245
	7	π_1	.097	.097	.111	.118	.129	.176	.161	.259
		π_2	.102	.121	.091	.120	.086	.144	.102	.208
5	1	π_1	.331	.334	.344	.371	.378	.389	.405	.412
		π_2	.362	.375	.372	.387	.396	.415	.436	.475
	3	π_1	.205	.217	.222	.271	.254	.305	.300	.360
		π_2	.221	.253	.224	.276	.236	.293	.288	.383
	5	π_1	.140	.148	.157	.198	.182	.247	.236	.324
		π_2	.149	.183	.146	.193	.150	.228	.193	.313
	7	π_1	.099	.110	.114	.152	.135	.211	.178	.301
		π_2	.107	.127	.098	.153	.101	.189	.130	.292