Austral. J. Statist. **38**(1), 1996, 75–82

CHANCE OF MISCLASSIFICATION IN THE CASE OF SEVERAL NORMAL POPULATIONS

PIL S. PARK[1] AND ANANT M. KSHIRSAGAR[1]

University of Michigan

Summary

When a new observation is to be classified into one of several multivariate normal populations with different means and the same covariance matrix, by Rao's method of scoring, the chance of misclassification is expressed as a multiple integral. This paper gives a practical method of obtaining reasonable approximations to this integral by using tables prepared by Gibbons, Olkin & Sobel (1977) for a different task.

Key words: Chance of misclassification; several normal populations.

1. Introduction

Let Π_i $(i = 1, \ldots, g)$ be g d-variate normal populations with mean vectors μ_i and a common covariance matrix Σ. Rao (1973) gives a method of classifying a new observation to one of these populations. The method consists of assigning scores C_i to Π_i for this new observation and assigning the observation to the population with the maximum score. In practice, we need training samples of sizes n_i $(i = 1, \ldots, g)$ from these g populations to estimate μ_i and Σ for use in the scores. The chance of misclassification can then be expressed as a multiple integral.

In this paper we give a useful practical approximation to this integral for estimating the chance of misclassification. Schervish (1981) uses the method of asymptotic expansion of this integral, but the result appears to be difficult for a practitioner to use. Our approximation is more user-friendly. We relate our method to the problem of selecting and ordering populations, considered by Gibbons, Olkin & Sobel (1977), and the tables they prepared for that purpose are useful in estimating this chance of misclassification. There is a considerable literature on different types of error rates in classification; McLachlan (1992) has given an excellent up-to-date account of various aspects of this problem.

Here, we first assume that the population parameters, such as the means and covariance matrix of the populations, are known. The chance of misclassification p is calculated in terms of these parameters. Then, for practitioners, we suggest the unknown parameters in the final expression be replaced by the corresponding sample estimates. Thus in both the approximations we are not estimating the

Received February 1995; revised November 1995; accepted December 1995.

[1] Dept of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029.

actual error rate but rather the optimal error rate. Consequently, our estimate is likely to underestimate the actual error rate, due to sampling fluctuations.

2. Chance of Misclassification

Denote by $\mathbf{x}_{ir}$ the d-vector of observations for the rth member of Π_i in the training sample ($i = 1, \ldots, g$; $r = 1, \ldots, n_i$). Let

$$S_i = \sum_{r=1}^{n_i} (\mathbf{x}_{ir} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ir} - \bar{\mathbf{x}}_i)', \qquad \text{where} \quad \bar{\mathbf{x}}_i = n_i^{-1} \textstyle\sum_{r=1}^{n_i} \mathbf{x}_{ir}, \tag{2.1}$$

be the matrix of the corrected sum of squares and products (SS & SP) for the ith sample. We estimate Σ by the pooled matrix

$$S = \sum_{r=1}^{g} S_i \tag{2.2}$$

whose degrees of freedom (df) are $f = N - g$ where $N = \sum_{i=1}^{g} n_i$.

For a new observation $\mathbf{x}_o$ to be classified, Rao (1973) defines the true score by

$$C_i = \boldsymbol{\mu}_i' \Sigma^{-1} \mathbf{x}_o - \tfrac{1}{2} \boldsymbol{\mu}_i' \Sigma^{-1} \boldsymbol{\mu}_i \qquad (i = 1, \ldots, g); \tag{2.3}$$

its sample estimate is

$$\widehat{C}_i = f \bar{\mathbf{x}}_i' S^{-1} \mathbf{x}_o - \tfrac{1}{2} f \bar{\mathbf{x}}_i' S^{-1} \bar{\mathbf{x}}_i. \tag{2.4}$$

The observation $\mathbf{x}_o$ is then assigned to Π_j if

$$C_j = \max(C_1, \ldots, C_g). \tag{2.5}$$

The chance of misclassification when $\mathbf{x}_o$ really belongs to Π_i is

$$\begin{aligned} E_i &= 1 - \text{chance of correct classification} \\ &= 1 - \Pr\{C_i > C_j, \text{ all } j \neq i\}. \end{aligned} \tag{2.6}$$

Observe that

$$\mathrm{E}(C_i \mid \mathbf{x}_o \in \Pi_i) = \tfrac{1}{2} \boldsymbol{\mu}_i' \Sigma^{-1} \boldsymbol{\mu}_i = \tfrac{1}{2} \phi_{ii}, \quad \text{say,} \tag{2.7}$$

$$\mathrm{E}(C_i \mid \mathbf{x}_o \in \Pi_j) = \boldsymbol{\mu}_i' \Sigma^{-1} \boldsymbol{\mu}_j - \tfrac{1}{2} \boldsymbol{\mu}_i' \Sigma^{-1} \boldsymbol{\mu}_i = \phi_{ij} - \tfrac{1}{2} \phi_{ii} \quad (i \neq j) \tag{2.8}$$

and

$$\begin{aligned} \operatorname{var}(C_i) &= \phi_{ii}, \qquad \operatorname{cov}(C_i, C_j) = \phi_{ij}, \\ \operatorname{corr}(C_i, C_j) &= \frac{\phi_{ij}}{(\phi_{ii}\phi_{jj})^{1/2}} = \rho_{ij}, \quad \text{say} \qquad (i, j = 1, \ldots, g). \end{aligned} \tag{2.9}$$

The C_is have a multivariate normal distribution with the above parameters. The chance of misclassification is thus

$$E_i = 1 - \int \Psi(\mathbf{C})\, d\mathbf{C}, \tag{2.10}$$

where $\mathbf{C}$, the vector of the C_is, has a multivariate normal distribution with parameters given earlier, and the single integral sign represents a $(g-1)$-fold integral over the region $C_i > C_j$ $(j = 1, \ldots, g;\ j \neq i)$.

3. A Crude Approximation to E_i

To obtain a crude approximation first we ignore the correlations ρ_{ij} between the C_is . Then, using the distribution of $C_i - C_j$,

$$\Pr\{C_i < C_j\} = \Pr\{C_i - C_j < 0\} = \Phi(-\tfrac{1}{2}\Delta_{ij}), \tag{3.1}$$

where Φ denotes the cumulative distribution function (cdf) of a $N(0,1)$ variable and

$$\Delta_{ij}^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{3.2}$$

Therefore,

$$E_i \approx 1 - \prod_{j=1, j\neq i}^{g} \left(1 - \Phi(-\tfrac{1}{2}\Delta_{ij})\right). \tag{3.3}$$

This approximation is based on Kimball's (1951) inequality as stated in Hochberg & Tamhane (1987 p.63); according to them this is at least as good as the Bonferroni bound used for similar results in multiple comparisons in analysis of variance. The total chance of misclassification is then approximately

$$E = \frac{1}{g}\sum_{i=1}^{g} E_i. \tag{3.4}$$

If we now replace every Δ_{ij} in (3.3) by

$$\Delta_{\min} = \min(\Delta_{ij}), \tag{3.5}$$

so that E_i is replaced by the maximum of the E_is (to avoid underestimation), (3.4) reduces to

$$1 - \left(1 - \Phi(-\tfrac{1}{2}\Delta_{\min})\right)^{g-1}, \tag{3.6}$$

which could further be approximated by

$$E \approx (g-1)\Phi(-\tfrac{1}{2}\Delta_{\min}). \tag{3.7}$$

In practice, we have to use

$$\hat{E} \approx (g-1)\Phi\Big(-\tfrac{1}{2}\min\Big(\sqrt{\hat{\Delta}^2_{ij}}\Big)\Big), \tag{3.8}$$

where $\hat{\Delta}^2_{ij}$ is the unbiased estimate of Δ^2_{ij} given by

$$\hat{\Delta}^2_{ij} = \frac{N-g-d-1}{N-g} D^2_{ij} - \frac{2d}{\bar{n}}, \tag{3.9}$$

where

$$D^2_{ij} = f(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' S^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) \qquad (i,j = 1,\ldots,g), \tag{3.10}$$

and $\bar{n}$ is the average sample size.

To see how 'bad' this approximation is, we carried out a simulation study for $g = 4$ populations, with $d = 4$, $\Sigma = I$ and for the following cases:

1. $\boldsymbol{\mu}_1' = [1,0,0,0]$, $\boldsymbol{\mu}_2' = [0,1,0,0]$, $\boldsymbol{\mu}_3' = [0,0,1,0]$, $\boldsymbol{\mu}_4' = [0,0,0,0]$ and $n_i = 50$ $(i = 1,\ldots,4)$;
2. $\boldsymbol{\mu}_1' = [1,1,0,0]$, $\boldsymbol{\mu}_2' = [0,1,1,0]$, $\boldsymbol{\mu}_3' = [0,0,1,1]$, $\boldsymbol{\mu}_4' = [0,0,0,0]$ and $n_i = 50$ $(i = 1,\ldots,4)$;
3. $\boldsymbol{\mu}_1' = [1,2,0,0]$, $\boldsymbol{\mu}_2' = [0,1,2,0]$, $\boldsymbol{\mu}_3' = [0,0,1,2]$, $\boldsymbol{\mu}_4' = [0,0,0,0]$ and $n_i = 50$ $(i = 1,\ldots,4)$;
4. $\boldsymbol{\mu}_1' = [1,2,3,0]$, $\boldsymbol{\mu}_2' = [0,1,2,3]$, $\boldsymbol{\mu}_3' = [3,0,1,2]$, $\boldsymbol{\mu}_4' = [1,1,1,1]$ and $n_i = 50$ $(i = 1,\ldots,4)$;
5. $\boldsymbol{\mu}_i'$ $(i = 1,\ldots,4)$ as in Case 1, $n_1 = 30$, $n_2 = 40$, $n_3 = 50$, $n_4 = 60$;.
6. $\boldsymbol{\mu}_i'$ $(i = 1,\ldots,4)$ as in Case 2, with sample sizes as in Case 5;
7. $\boldsymbol{\mu}_i'$ $(i = 1,\ldots,4)$ as in Case 3, with sample sizes as in Case 5;
8. $\boldsymbol{\mu}_i'$ $(i = 1,\ldots,4)$ as in Case 4, with sample sizes as in Case 5.

The choice of these $\boldsymbol{\mu}$s and n_is is similar to that of Schervish (1981) in his simulation studies.

For each case, the sample observations were classified by Rao's scoring method and the true chance of misclassification was estimated by counting the number of misclassified observations. Then it was compared with E and $\hat{E}$ given by (3.7) and (3.8) respectively. This whole process was then repeated 200 times, so we have 200 values of p, E and $\hat{E}$ computed from samples of sizes n_i given above, for each of the eight cases. This study showed that E or $\hat{E}$ overestimates p by at least a factor 2 and thus a crude approximation is $\frac{1}{2}E$ or $\frac{1}{2}\hat{E}$. The mean square errors (MSE) of $\frac{1}{2}E$ or $\frac{1}{2}\hat{E}$, when compared with p, from these 200 iterations ranged from 0.000727 to 0.002345. In the next section, the correlations between the C_is are not ignored. Instead the correlation matrix is approximated by another equicorrelation matrix.

4. The 'GOS' Method of Estimation

This method is based on Gibbons, Olkin & Sobel's (1977) method of selecting and ordering normal populations, so we call it the 'GOS' method. They considered g variables u_i $(i = 1,\ldots,g)$ (their notation is different) with a common variance σ^2 and a common correlation coefficient ρ. Denote the ordered means of u_i by

$$\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[g]}. \tag{4.1}$$

Their goal was to select the variable with the largest mean with a specified assurance p^* of the probability of correct selection, when

$$\delta = \mu_{[g]} - \mu_{[g-1]} \geq \text{ a preassigned quantity } \delta^*, \quad (\delta^* > 0,\ g^{-1} < p^* < 1). \tag{4.2}$$

Their procedure is to choose a random sample of an appropriate size m from the distribution of the vector variable

$$\mathbf{u}' = (u_1,\ldots,u_g). \tag{4.3}$$

The sample means $\bar{u}_1,\ldots,\bar{u}_g$ are then ordered. The ordered means are denoted by $\bar{u}_{[i]}$ $(i = 1,\ldots,g)$. The variable u_j corresponding to the largest sample mean $\bar{u}_{[g]}$ is then selected. The sample size m for this procedure is calculated from the expression

$$m = \left(\frac{\tau\sigma}{\delta^*}\right)^2 (1-\rho), \tag{4.4}$$

where σ and ρ are either known or are estimated from some prior data and τ is obtained by their tables A.1 and A.2 (pp.400, 401 of Gibbons *et al.* 1977, partially reproduced here, with permission, as Table 1; the ranges of g and τ are sufficient in most practical situations). Their tables give τ for values of g from 2 to 25 and p^* from 0.75 to 0.999, and also give the p^* for $g = 2$ to 50; τ goes from 0 to 5 at intervals of 0.2.

It is interesting to see the similarities and differences between their problem and ours. We have the scores C_i $(i = 1,\ldots,g)$ given by (2.4) for a new observation $\mathbf{x}_o$, and our sample size is only $m = 1$, and we assign the observation $\mathbf{x}_o$ to the population with maximum score. The GOS method determines the sample size for a specified chance of correct selection. We have a predetermined sample size and wish to find the chance $p = 1 - p^*$ of misclassification (or correct selection). They assume σ^2 and ρ to be the same. In our case, the variances of the C_i are ϕ_{ii} and the correlations are ρ_{ij} as given by (2.10); they are not the same. Thus, as an approximation, we propose to replace the ϕ_{ii}s and replace ρ_{ij}s $(i \neq j)$ by common values. This approximation gives good results only when the ϕ_{ii}s and the ρ_{ij}s are not 'too different'. This type of approximation was first tried by Penrose (1947) in the construction of Fisher's linear discriminant function and was found to be very satisfactory. Putting $m = 1$ in (4.4), we obtain

$$\tau^2 = \frac{\delta^{*2}}{\sigma^2(1-\rho)}. \tag{4.5}$$

TABLE 1
Values of p for fixed τ

							τ						
g	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4
2	0.500	0.444	0.389	0.336	0.286	0.240	0.198	0.161	0.129	0.102	0.079	0.060	0.045
3	0.667	0.609	0.548	0.487	0.426	0.366	0.310	0.258	0.211	0.170	0.134	0.104	0.079
4	0.750	0.696	0.637	0.575	0.512	0.448	0.386	0.326	0.271	0.221	0.177	0.139	0.107
5	0.800	0.750	0.695	0.635	0.571	0.506	0.441	0.378	0.318	0.262	0.212	0.168	0.131
6	0.833	0.788	0.736	0.678	0.616	0.551	0.484	0.419	0.355	0.296	0.242	0.193	0.152
7	0.857	0.815	0.766	0.711	0.650	0.586	0.519	0.452	0.387	0.324	0.267	0.215	0.170
8	0.875	0.836	0.790	0.737	0.678	0.615	0.548	0.480	0.413	0.349	0.289	0.234	0.186
9	0.889	0.852	0.809	0.758	0.701	0.639	0.573	0.505	0.437	0.371	0.309	0.252	0.201
10	0.900	0.866	0.824	0.776	0.720	0.659	0.594	0.526	0.457	0.390	0.326	0.268	0.215
15	0.933	0.907	0.874	0.833	0.785	0.729	0.668	0.602	0.533	0.463	0.394	0.329	0.269
20	0.950	0.928	0.900	0.865	0.822	0.772	0.714	0.651	0.583	0.512	0.442	0.374	0.309
25	0.960	0.942	0.917	0.886	0.847	0.800	0.746	0.685	0.619	0.549	0.478	0.408	0.341
50	0.980	0.958	0.936	0.914	0.892	0.870	0.828	0.777	0.718	0.653	0.584	0.512	0.440

							τ						
g	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.2	4.4	4.6	4.8	5.0
2	0.033	0.024	0.017	0.012	0.008	0.005	0.004	0.002	0.001	0.001	0.001	0.001	0.001
3	0.059	0.043	0.031	0.022	0.015	0.010	0.007	0.004	0.003	0.002	0.001	0.001	0.001
4	0.081	0.060	0.044	0.031	0.022	0.015	0.010	0.007	0.004	0.003	0.002	0.001	0.001
5	0.100	0.075	0.055	0.039	0.028	0.019	0.013	0.008	0.005	0.003	0.002	0.001	0.001
6	0.117	0.088	0.065	0.047	0.033	0.023	0.015	0.010	0.007	0.004	0.003	0.002	0.001
7	0.131	0.100	0.074	0.054	0.038	0.026	0.018	0.012	0.008	0.005	0.003	0.002	0.001
8	0.145	0.110	0.082	0.060	0.043	0.030	0.020	0.014	0.009	0.006	0.004	0.002	0.001
9	0.157	0.120	0.090	0.066	0.047	0.033	0.023	0.015	0.010	0.006	0.004	0.002	0.001
10	0.169	0.130	0.098	0.072	0.052	0.036	0.025	0.017	0.011	0.007	0.004	0.003	0.002
15	0.215	0.168	0.129	0.096	0.070	0.050	0.035	0.024	0.016	0.010	0.007	0.004	0.002
20	0.250	0.198	0.153	0.116	0.085	0.062	0.043	0.030	0.020	0.013	0.008	0.005	0.003
25	0.279	0.223	0.174	0.133	0.099	0.072	0.051	0.035	0.024	0.016	0.010	0.006	0.004
50	0.370	0.304	0.244	0.191	0.147	0.109	0.080	0.057	0.039	0.026	0.017	0.011	0.007

For our variables C_i, the differences in the means are

$$\mathrm{E}(C_i) - \mathrm{E}(C_j) = \tfrac{1}{2}\Delta_{ij}^2 \qquad (i,j = 1,\ldots,g), \tag{4.6}$$

the variances are ϕ_{ii} and the correlations are ρ_{ij}. Using averages of these quantities, we take our τ^2 to be

$$\tau^2 = \frac{(\frac{1}{2}\bar{\Delta}_{ij}^2)^2}{\bar{\phi}_{ii}(1-\bar{\rho}_{ij})}. \tag{4.7}$$

We also considered 'suitable' values such as $\min(\Delta_{ij}^2)$, $\max(\phi_{ii})$ and $\min(\rho_{ij})$ to get a conservative τ^2, but our simulation studies show that the choice of (4.7)

gives better results. Using (4.7) for τ^2, we refer to the GOS table and get the chance of correct selection as p^* and therefore the chance of misclassification is $p = 1 - p^*$. Using the same cases and sample sizes for our simulation as described in Section 3, we compare p with the true chance. The MSE of p from 200 iterations ranges from 0.000894 to 0.00919 for the eight cases, showing that p gives a reasonable estimate of the chance of misclassification. In practice, Δ_{ij}^2 must be estimated and, therefore, one has to use

$$\hat{\tau}^2 = \frac{\left[\text{Average}\left(\frac{N-g-d-1}{N-g} D_{ij}^2 - \frac{2d}{\bar{n}}\right)\right]^2}{4[\text{Average}(\hat{\phi}_{ii})][1 - \text{Average}(\hat{\rho}_{ij})]}, \tag{4.8}$$

where

$$N = \sum_{i=1}^{g} n_i, \quad \bar{n} = \frac{N}{g}, \quad \hat{\phi}_{ii} = (f-d-1)\bar{\mathbf{x}}_i' S^{-1} \bar{\mathbf{x}}_i - \frac{d}{\bar{n}},$$

$$\hat{\phi}_{ij} = (f-d-1)\bar{\mathbf{x}}_i' S^{-1} \bar{\mathbf{x}}_j - \frac{d}{\bar{n}} \quad (i \neq j), \qquad \hat{\rho}_{ij} = \frac{\hat{\phi}_{ij}}{\sqrt{\hat{\phi}_{ii}\hat{\phi}_{jj}}}. \tag{4.9}$$

5. Conclusion

Estimation of the chance of misclassification in the case of more than two groups is a difficult problem. One obvious method is the extension of the Lachenbruch (1965) hold-out method in the case of two populations, but it requires substantial computer use to drop every observation turn by turn and recalculate the scores to assign the observation 'held out'. We have given a reasonable approximation to the integral that gives the chance of misclassification by replacing the multivariate normal distribution of the C_is by an equicorrelated equivariance distribution and using the GOS table. Often, in practice, the scoring method of Rao is modified by replacing the variable $x_1, \ldots, x_d$ by a smaller number of canonical correlations (Kshirsagar & Arsenven, 1975). This refinement is worthwhile in the case of a large number of variables as in social sciences or in economics. Further details are in Park (1994).

References

GIBBONS, J.D., OLKIN, I. & SOBEL, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology.* New York: John Wiley and Sons, Inc.

HOCHBERG, Y. & TAMHANE, A.C. (1987). *Multiple Comparison Procedures.* New York: John Wiley and Sons, Inc.

KIMBALL, A.W. (1951). On dependent tests of significance in the analysis of variance. *Ann. Math. Statist.* **22**, 600–602.

KSHIRSAGAR, A.M. & ARSENVEN, E. (1975). A note on the equivalency of two discriminant procedures. *Amer. Statist.* **29**, 38–39.

LACHENBRUCH, P.A. (1965). Estimation of Error Rates in Discriminant Analysis. PhD dissertation, University of California, Los Angeles.

McLACHLAN, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

PARK, P.S. (1994). Errors of Misclassification in Discriminant Analysis. PhD thesis, University of Michigan, Dept of Biostatistics, Ann Arbor, Michigan.

PENROSE, L.S. (1947). Some notes on discrimination. *Ann. Eugen.* **13**, 228.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd edition. New York: John Wiley and Sons, Inc.

SCHERVISH, M.J. (1981). Asymptotic expansions for the means and variances of error rates. *Biometrika* **68**, 295–299.