SOME RESULTS ON THE MULTIPLE GROUP DISCRIMINANT PROBLEM

By

Peter A. Lachenbruch

Department of Biostatistics
University of North Carolina, Chapel Hill, N.C.

Institute of Statistics Mimeo Series No. 829

JULY 1972

SOME RESULTS ON THE MULTIPLE GROUP DISCRIMINANT PROBLEM

Peter A. Lachenbruch

Introduction. The problem of assigning an individual to one of two groups on the basis of a set of observations has been extensively studied since the introduction of the linear discriminant function (LDF) by Fisher in 1936 (4). The behavior of the LDF has been considered from the point of view of its distribution when samples are used to estimate the coefficients, (1,12), estimation of error rates (3,8,10)and robustness to various departures from the assumptions (5,6,9,11). Recently, Glick (7) has shown the asymptotic optimality of "plug in" estimators of correct classification rates. Relatively little has been done regarding the problem of assigning an observation to one of k>2 groups. In this study, two different approaches are reviewed, some of their properties studied, and compared.

The first approach assumes that the populations have density functions $f_i(\underset{\sim}{x})$ whose form is known, but whose parameters may not be known. Then, it can be shown that the rule which maximizes the mean probability of correct classification is to assign the unknown observation $\underset{\sim}{x}$ to population i (π_i) if $p_i f_i(\underset{\sim}{x}) = \max p_j f_j(x)$ where p_i is the a priori probability of an individual belonging to π_i. In this study, it is assumed that the p_i are equal so the rule becomes "assign $\underset{\sim}{x}$ to π_i if $f_i(\underset{\sim}{x}) = \max_j f_j(\underset{\sim}{x})$." If the parameters of the f_i are unknown, they may be suitably estimated, e. g. by maximum likelihood, and the rule applied to the estimated functions. Computer programs can be written to perform the calculations for any given set of density functions, and many are already available for the special case of normal populations with the same covariance matrices. In this case, the rule is equivalent to "assign $\underset{\sim}{x}$ to π_i if

$$(\underset{\sim}{x} - \tfrac{1}{2}\underset{\sim}{\mu}_i)' \Sigma^{-1} \underset{\sim}{\mu}_i = \max_j (\underset{\sim}{x} - \tfrac{1}{2}\underset{\sim}{\mu}_j)' \Sigma^{-1} \underset{\sim}{\mu}_j \quad ,$$

where $\underset{\sim}{\mu}_i$ is the mean of π_i and Σ is the common covariance matrix." In the sampling situation, $\underset{\sim}{\mu}_i$ is replaced by $\underset{\sim}{\bar{x}}_i$ and Σ is replaced by S, the usual maximum likelihood estimates. This method will be referred to as the Multiple Discriminant Function or MDF method.

The second approach to the multiple group porblem is based on an extension of Fisher's original approach to the two group problem. Let B be the between groups mean square and cross-product matrix and W be the within groups mean square and cross-product matrix; then, for two groups Fisher suggested finding the linear combination $\underset{\sim}{\lambda}'\underset{\sim}{x}$, which maximized $\phi = \underset{\sim}{\lambda}'B\underset{\sim}{\lambda} / \underset{\sim}{\lambda}'W\underset{\sim}{\lambda}$ which is the ratio of the between groups mean square to the within groups mean square of the linear combination $\underset{\sim}{\lambda}'\underset{\sim}{x}$. For two groups these two approaches lead to the same function. For more than two groups, the linear compounds which maximize ϕ are solutions of the equation $(B - \gamma W)\underset{\sim}{\lambda} = \underset{\sim}{0}$.

This familiar equation has solutions which are the eigenvectors of $W^{-1}B$, and there are at most min (k-1,p) of them where k is the number of populations, and p the number of variables. Since the $\{\underset{\sim}{\lambda}'\underset{\sim}{x}\}$ are linear transformations of normal variates, they too are normal (conditional on $\underset{\sim}{\lambda}$), and it is easy to show that the optimal assignment rule when using r eigenvectors is to assign to π_i if

$$\sum_{\ell=1}^{r} (\underset{\sim}{\lambda}_\ell'(\underset{\sim}{x} - \underset{\sim}{\mu}_i))^2 = \min_j \sum_{1}^{r} (\underset{\sim}{\lambda}_\ell' (\underset{\sim}{x}-\underset{\sim}{\mu}_j))^2$$

or

$$(\underset{\sim}{y}-\tfrac{1}{2}\underset{\sim}{\nu}_i)'\underset{\sim}{\nu}_i = \max_j (\underset{\sim}{y}-\tfrac{1}{2}\underset{\sim}{\nu}_j)'\underset{\sim}{\nu}_j$$

where

$$\underset{\sim}{y}' = (\underset{\sim}{\lambda}_1'\underset{\sim}{x}, \underset{\sim}{\lambda}_2'\underset{\sim}{x} \ldots \underset{\sim}{\lambda}_r'\underset{\sim}{x})$$

and

$$\underset{\sim}{\nu}_i' = (\underset{\sim}{\lambda}_1'\underset{\sim}{\mu}_i \ldots \underset{\sim}{\lambda}_r'\underset{\sim}{\mu}_i)$$

This method will be referred to as the eigenvector or EV method.

In either case the boundaries of the assignment regions are hyperplanes. For the MDF, the hyperplanes have dimension $p-1$ and for the EV they have dimension $r \leq \min(k-1,p) - 1$.

With the availability of computers, it is possible to use either method fairly easily. If p is large, it may be desirable to represent the data in fewer dimensions, which the EV approach does, although it is doubtful there is any great saving in computation since the compounds must be found. At any rate the EV method is used, particularly by psychometricians, so it does seem useful to compare its behavior to that of the MDF.

A number of questions may be asked about these functions.

1. Under what situations can the EV approach work well compared to the MDF method?
2. What is the effect of sample size on the behavior of the methods?
3. How does the number of groups affect their performance?
4. How does the number of variables affect their performance?
5. How would one answer questions 2-4 if one is concerned about the relative performance of the methods?

In the following pages we attempt to answer these questions using the estimated proportion of correct classification as our criterion. Further work will be needed to study the robustness of the methods when unequal covariances appear, or non-normal distributions are present.

Configuration of Population Means.

In the two group case, it is possible to draw a line between the two means of the populations. In the k - group case, this will not be possible in general. The means will lie in a space of dimension less than or equal to the minimum of k-1 and p. The placement of the mean values will have a substantial effect on the performance of the EV classification rule when not all of the eigenvectors are used. In this paper, we consider two extreme placements. We assume $\Sigma = I$. In the first, the means are collinear (C case) and equally spaced. If the parameters are known, the first eigenvector contains all the information about the population and is equivalent to the MDF. In this case the eigenvalues other than the first are all 0. The second case has the means arranged so the pairwise distance is the same for any two groups. This implies that the means are the vertices of a regular simplex (S case). This is the least favorable case for the EV method in the sense that all eigenvalues are equal and each vector accounts for 1/(k-1) of the variability.

It is clear that we can assume the means are arranged as

$$\underset{\sim}{\mu}_1' = (0,\ldots), \ \underset{\sim}{\mu}_2' = (\delta_{21},0,\ldots 0), \ \underset{\sim}{\mu}_3' = (\delta_{31}, \ \delta_{32},0,\ldots 0),$$

etc. If k>p+1, there may be k-p-1 means in which all components are used. For the case of collinear means, spaced δ units apart we may consider

$$\underset{\sim}{\mu}_1' = (0,..0)$$

$$\underset{\sim}{\mu}_2' = (\delta, \ 0,..0), \ \underset{\sim}{\mu}_3' = (2\delta, \ 0,\ldots 0) \text{ etc.}$$

The probability of correct classification if one uses the MDF procedure is

$$\Phi(\delta/2) - \Phi(-\delta/2) \text{ for } \pi_2 \ldots \pi_{k-1}$$

and $\Phi(\delta/2)$ for π_1 and π_k. This can be seen as follows. An observation from π_i will be correctly classified if $f_i(\underset{\sim}{x})$ is a maximum. The assignment rule is

assign to

$$\pi_1 \text{ if } -\infty < x_1 \leq \delta/2$$

$$\pi_2 \text{ if } \delta/2 < x_1 \leq (\delta/2)+\delta$$

$$\pi_3 \text{ if } \delta/2+\delta < x_1 \leq \delta/2+2\delta$$

$$\vdots$$

$$\pi_{k-1} \text{ if } \delta/2+(k-3)\delta < x_1 \leq \delta/2+(k-2)\delta$$

$$\pi_k \text{ if } \delta/2+(k-2)\delta < x_1 < \infty$$

where x_1 is the first component of $\underset{\sim}{x}$. The following table gives some values of the probabilities.

TABLE 1

Probabilities of Correct Classification for C Case

δ	$P_1=P_k$	$P_i (2 \leq i \leq k-1)$	$\overrightarrow{P}_3$	$\overline{P}_5$	$\overline{P}_{10}$
½	.60	.20	.47	.36	.28
1	.69	.38	.59	.50	.44
2	.84	.68	.79	.74	.71
3	.93	.86	.91	.89	.87

The last three columns are the mean probabilities of correct classification for k = 3,5, 10.

For this case, it is easy to obtain exact probabilities of correct classification. For more general placement of means, the problem becomes more difficult. This configuration of means is most favorable to the EV method since all of the between group variability occurs along one axis.

A less favorable configuration is one in which the means are placed at the vertices of a regular k-1 dimensional figure. This is an equilateral triangle in 2 dimensions, a pyramid in 3 dimensions, and in general is a regular simplex. One can show that the coordinates of the vertices of a regular simplex spaced δ units apart are $\underset{\sim}{\mu}_1' = (0,\ldots 0)$

$$\underset{\sim}{\mu}_\ell = \delta(a_{1\ell}, a_{2\ell}, \ldots a_{\ell\ell}, 0, \ldots 0)$$

where $$a_{\ell\ell} = ((1+\ell)/2\ell)^{\frac{1}{2}}$$

$$a_{i\ell} = 1/2\left(\frac{i+1}{2}\right)^{\frac{1}{2}} \qquad i<\ell$$

$$a_{i\ell} = 0 \qquad i>\ell$$

Table 2 gives coordinates for the vertices of a simplex spaced one unit apart for up to 10 dimensions. This is the least favorable case for the EV method since the eigenvalues are all equal and each vector accounts for 1/k-1 of the between group variability. The MDF procedure assigns an observation to π_i if

$$z_i = (\underset{\sim}{x}_i - \underset{\sim}{\mu}_i/2)' \ \Sigma^{-1}\underset{\sim}{\mu}_i$$

is a maximum (in our case we will have $\Sigma = I$). To find the probability of correct classification for observation from π_j we must find

$$P(z_i > z_j \,|\, \underset{\sim}{x}\varepsilon\pi_i)$$

for all j. We shall study this probability for observations for π_1, but our methods are applicable for all π_i.

TABLE 2

Coordinates of Simplicies

		i									
	Vector	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	0	0	0	0	0	0	0
2	3	½	$\sqrt{3/4}$	0	0	0	0	0	0	0	0
3	4	½	$\frac{1}{2\sqrt{3}}$	$\sqrt{4/6}$	0	0	0	0	0	0	0
4	5	½	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\sqrt{5/8}$	0	0	0	0	0	0
5	6	½	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\frac{1}{2\sqrt{10}}$	$\sqrt{6/10}$	0	0	0	0	0
6	7	½	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\frac{1}{2\sqrt{10}}$	$\frac{1}{2\sqrt{15}}$	$\sqrt{7/12}$	0	0	0	0
7	8	½	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\frac{1}{2\sqrt{10}}$	$\frac{1}{2\sqrt{15}}$	$\frac{1}{2\sqrt{21}}$	$\sqrt{8/14}$	0	0	0
8	9	½	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\frac{1}{2\sqrt{10}}$	$\frac{1}{2\sqrt{15}}$	$\frac{1}{2\sqrt{21}}$	$\frac{1}{2\sqrt{28}}$	$\sqrt{9/16}$	0	0
9	10	½	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\frac{1}{2\sqrt{10}}$	$\frac{1}{2\sqrt{15}}$	$\frac{1}{2\sqrt{21}}$	$\frac{1}{2\sqrt{28}}$	$\frac{1}{2\sqrt{36}}$	$\sqrt{10/18}$	0
	Mean	½	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\frac{1}{2\sqrt{10}}$	$\frac{1}{2\sqrt{15}}$	$\frac{1}{2\sqrt{21}}$	$\frac{1}{2\sqrt{28}}$	$\frac{1}{2\sqrt{36}}$	$\frac{1}{2\sqrt{45}}$	0

Suppose $\underset{\sim}{x}\varepsilon\pi_1$. We can show that the joint distribution of $\underset{\sim}{z}' = (z_1 \ldots z_k)$ is normal with mean $\underset{\sim}{\mu}_z' = (0, -\delta^2/2, -\delta^2/2, \ldots -\delta^2/2)$ covariance matrix

$$A=(a_{ij})=\begin{pmatrix} 0 & 0 & .. & 0 \\ 0 & \delta^2 & .. & \delta^2/2 \\ : & : & :: & : \\ 0 & \delta^2/2 & .. & \delta^2 \end{pmatrix}$$

$$a_{11} = 0$$

$$a_{ii} = \delta^2 \quad i \geq 2$$

$$a_{1j} = a_{i1} = 0$$

$$a_{ij} = \delta^2/2 \quad i \neq j \quad i>1 \quad j>1$$

Thus, the probability of correct classification of $\underset{\sim}{x}$ is

$$P(z_1>z_2,\ z_1>z_3, \ldots, z_1>z_k)$$

$$= P(z_2<0, z_3<0 \ldots z_k<0)$$

Let $y_1 \ldots y_k$ be independent $N(0,\delta^2/2)$, and let $x_i=y_i+y_k$. Then, the event $(z_2<0, \ldots z_k<0)$ is equivalent to the event $(x_1<\delta^2/2, \ldots x_{k-1}<\delta^2/2)$. Since the y_i are independent, we have

$$P(x_1<\delta^2/2, \ldots x_{k-1}<\delta^2/2) = P(y_1<\delta^2/2-y_k, \ldots y_{k-1}<\delta^2/2-y_k)$$

$$=\int_{-\infty}^{\infty} P(y_1<\delta^2/2-y_k) \ldots P(y_{k-1}<\delta^2/2-y_k) d_{yk}$$

$$= \int_{-\infty}^{\infty} \Phi\left(\frac{\delta^2/2-y_k}{\sqrt{\delta^2/2}}\right)^{k-1} f(y_k)dy_k$$

This integral was mentioned by Bechhofer and Sobel (2). Table 3 gives these probabilities for various values of δ and k. Note that the first column k=2 is the value for two group discriminant analysis.

TABLE 3

Correct Classification Probabilities

	K									
δ	2	3	4	5	6	7	8	9	10	11
0.0	0.500	0.333	0.250	0.200	0.166	0.142	0.125	0.111	0.100	0.091
0.10	0.520	0.354	0.269	0.217	0.182	0.156	0.137	0.123	0.111	0.102
0.20	0.540	0.374	0.288	0.235	0.198	0.171	0.151	0.135	0.123	0.113
0.30	0.560	0.395	0.308	0.253	0.215	0.187	0.166	0.149	0.136	0.125
0.40	0.579	0.416	0.328	0.272	0.233	0.204	0.182	0.164	0.149	0.138
0.50	0.599	0.438	0.349	0.292	0.252	0.222	0.199	0.180	0.164	0.152
0.60	0.618	0.459	0.370	0.312	0.271	0.241	0.216	0.197	0.180	0.167
0.70	0.637	0.481	0.392	0.333	0.291	0.260	0.235	0.214	0.197	0.183
0.80	0.655	0.503	0.414	0.355	0.312	0.280	0.254	0.233	0.215	0.200
0.90	0.674	0.525	0.436	0.377	0.333	0.300	0.274	0.252	0.234	0.219
1.00	0.691	0.546	0.458	0.399	0.355	0.322	0.295	0.273	0.254	0.238
1.10	0.709	0.568	0.481	0.421	0.377	0.343	0.316	0.293	0.274	0.258
1.20	0.726	0.589	0.504	0.444	0.400	0.365	0.338	0.315	0.295	0.278
1.30	0.742	0.610	0.526	0.467	0.423	0.388	0.360	0.337	0.317	0.300
1.40	0.758	0.631	0.549	0.490	0.446	0.411	0.383	0.359	0.339	0.322
1.50	0.773	0.651	0.571	0.513	0.469	0.434	0.405	0.382	0.361	0.344
1.60	0.788	0.671	0.593	0.536	0.493	0.458	0.429	0.405	0.384	0.366
1.70	0.802	0.690	0.615	0.559	0.516	0.481	0.452	0.428	0.407	0.389
1.80	0.816	0.709	0.636	0.582	0.539	0.505	0.476	0.452	0.431	0.412
1.90	0.829	0.728	0.657	0.605	0.563	0.529	0.500	0.475	0.454	0.436
2.00	0.841	0.745	0.678	0.627	0.586	0.552	0.524	0.499	0.478	0.460
2.10	0.853	0.762	0.698	0.648	0.609	0.576	0.548	0.524	0.502	0.484
2.20	0.864	0.779	0.717	0.670	0.631	0.599	0.572	0.548	0.527	0.508
2.30	0.875	0.794	0.736	0.690	0.653	0.622	0.595	0.572	0.551	0.532
2.40	0.885	0.809	0.754	0.710	0.674	0.644	0.618	0.595	0.575	0.557
2.50	0.894	0.824	0.771	0.729	0.695	0.666	0.641	0.619	0.599	0.581
2.60	0.903	0.837	0.788	0.748	0.715	0.687	0.663	0.642	0.623	0.605
2.70	0.911	0.850	0.804	0.766	0.735	0.708	0.685	0.664	0.646	0.629
2.80	0.919	0.862	0.819	0.783	0.753	0.728	0.706	0.686	0.668	0.652

Table 3 continued...

δ	2	3	4	5	6	7	8	9	10	11
2.90	0.926	0.874	0.833	0.800	0.771	0.747	0.726	0.707	0.690	0.674
3.00	0.933	0.885	0.847	0.815	0.788	0.765	0.745	0.727	0.711	0.696
3.10	0.939	0.895	0.859	0.830	0.805	0.783	0.764	0.746	0.731	0.717
3.20	0.945	0.904	0.871	0.844	0.820	0.800	0.781	0.765	0.750	0.737
3.30	0.951	0.913	0.883	0.857	0.835	0.816	0.798	0.783	0.769	0.756
3.40	0.955	0.921	0.893	0.870	0.849	0.831	0.814	0.800	0.786	0.774
3.50	0.960	0.929	0.903	0.881	0.862	0.845	0.830	0.816	0.803	0.792
3.60	0.964	0.936	0.912	0.892	0.874	0.859	0.844	0.831	0.819	0.808
3.70	0.968	0.942	0.921	0.902	0.886	0.871	0.858	0.846	0.834	0.824
3.80	0.971	0.948	0.929	0.912	0.897	0.883	0.871	0.859	0.849	0.839
3.90	0.974	0.954	0.936	0.920	0.907	0.894	0.883	0.872	0.862	0.853
4.00	0.977	0.959	0.943	0.928	0.916	0.904	0.894	0.884	0.875	0.867
4.10	0.980	0.963	0.949	0.936	0.924	0.914	0.904	0.895	0.887	0.879
4.20	0.982	0.967	0.954	0.943	0.932	0.923	0.914	0.906	0.898	0.891
4.30	0.984	0.971	0.959	0.949	0.939	0.931	0.923	0.915	0.908	0.901
4.40	0.986	0.974	0.964	0.954	0.946	0.938	0.931	0.924	0.918	0.912
4.50	0.988	0.977	0.968	0.960	0.952	0.945	0.938	0.932	0.926	0.921
4.60	0.989	0.980	0.972	0.964	0.957	0.951	0.945	0.939	0.934	0.929
4.70	0.991	0.982	0.975	0.968	0.962	0.957	0.951	0.946	0.941	0.937
4.80	0.992	0.985	0.978	0.972	0.967	0.962	0.957	0.952	0.948	0.944
4.90	0.993	0.987	0.981	0.976	0.971	0.966	0.962	0.958	0.954	0.950
5.00	0.994	0.988	0.983	0.979	0.974	0.970	0.966	0.963	0.959	0.956

Comparing the results of Table 3 with the sample results in Table 4, we see that the behavior is as expected. For the samples of size 25, the probability of correct classification is overestimated by more than for samples of size 100. This bias is not unexpected. For larger p, the bias is somewhat more severe.

The integration was done using the ten point Hermitian quadrature routine from the IBM Scientific Subroutine Package.

One may think of the mean pairwise distance as a measure of divergence among the k populations. This is not altogether satisfactory, but it provides a simple measure to compare the two cases. For the regular simplex, all means are δ units apart, so the mean pairwise distance is δ. For the collinear case, suppose the spacing is γ. Then the mean distance is

$$\frac{1}{\binom{k}{2}} \sum_{i>j} \gamma(i-j) = \frac{\binom{k}{2}\binom{k+1}{3}\gamma}{\binom{k}{2}} = \binom{k+1}{3}\gamma$$

So if $\gamma=\delta$, the mean pairwise distance is greater for the collinear case than for the regular simplex configuration. This, of course, is why better discrimination is possible with collinear means spaced evenly δ units apart, than with means at the vertices of a regular simplex.

It may be of some interest to compare the probability of correct classification for the same mean between group distance for the simplex and collinear cases. Table 3a gives some comparisons.

TABLE 3a

Probabilities of Correct Classification for the Same Mean Between Population Distance

k	C Spacing	S Spacing	Collinear P(Correct Classification)	Simplex P(Correct Classification)
3	1	1.33	.59	.617
	2	2.67	.79	.846
	3	4.00	.91	.959
5	1	2.00	.50	.627
	2	4.00	.74	.928
	3	6.00	.89	.995
10	1	3.67	.44	.829
	2	7.33	.71	.999
	3	11.00	.87	1.000

Thus, for large k, the regular simplex arrangement leads to higher probabilities of correct classification.

Behavior of the Methods

To evaluate the performance of these methods we used sampling experiments since the distribution theory for these procedures even for the two group case is exceedingly complicated. The estimated mean probability of correct classification was used as the criterion. The estimate was obtained by resubstituting the original observations in the calculated MDF or EV. This is known to be biased, and comparing the observed quantities with the theoretically calculable ones (namely, the collinear case) we can estimate the approximate extent of this bias. Our results are given for up to 3 eigenvectors as this is the maximum number that seem to be used. All programs were written by the author using the IBM Scientific Subroutine Package when appropriate.

The sampling experiments consisted of taking 10 samples for each value of k, p, the number of variables, n, the number of observations in each of the k groups, δ, the between mean distance, and the two configurations of means. The values of these were:

$k = 5$ or 10
$p = 4$ or 10
$n = 25$ or 100
$\delta = 1,2,$ or 3
Type = Collinear or simplex

Since the simplex requires at least k-1 dimensions, it was not possible to arrange the means in a 9 dimensional simplex when $p = 4$. Thus, the simplex cases $p=4$, $k=10$, are not given.

For each combination, multivariate normal samples were taken, and estimates of the probability of correct classification by the MDF and EV methods were found. For the EV method, the use of 1, 2, or 3 eigenvectors was studied. The means and standard deviations of the probability of correct classification for each method are given in Table 4. The subscript on EV refers to the number of eigenvectors used. Because the EV method had almost the same standard deviation for all three cases, their standard deviations were pooled.

TABLE 4

Estimated Correct Classification Rates

δ=1

				Mean				Standard Deviation	
Type	p	k	n	MDF	EV_1	EV_2	EV_3	MDF	EV
C	4	5	25	.52	.50	.52	.53	.07	.06
		5	100	.50	.49	.50	.50	.03	.03
		10	25	.48	.45	.47	.47	.03	.03
		10	100	.45	.44	.44	.44	.02	.02
	10	5	25	.60	.53	.58	.61	.05	.03
		5	100	.52	.50	.51	.51	.02	.02
		10	25	.56	.47	.50	.52	.02	.03
		10	100	.49	.46	.47	.48	.01	.01
S	4	5	25	.42	.32	.39	.42	.02	.03
		5	100	.39	.30	.35	.38	.02	.02
		10	25			Not Applicable			
		10	100						
	10	5	25	.48	.32	.41	.46	.03	.04
		5	100	.41	.29	.35	.38	.01	.02
		10	25	.31	.16	.21	.25	.03	.02
		10	100	.26	.15	.18	.21	.01	.01

δ=2

Type	p	k	n	MDF	EV_1	EV_2	EV_3	MDF	EV
C	4	5	25	.75	.75	.76	.75	.05	.04
		5	100	.74	.74	.75	.75	.02	.02
		10	25	.73	.72	.73	.73	.03	.03
		10	100	.71	.70	.71	.71	.02	.01
	10	5	25	.80	.75	.78	.79	.03	.04
		5	100	.76	.75	.76	.75	.02	.02
		10	25	.77	.73	.75	.76	.02	.02
		10	100	.75	.73	.74	.74	.01	.01
S	4	5	25	.64	.39	.51	.59	.03	.03
		5	100	.63	.36	.48	.57	.01	.02
		10	25			Not Applicable			
		10	100						
	10	5	25	.68	.40	.55	.63	.05	.03
		5	100	.64	.36	.48	.57	.02	.02
		10	25	.53	.20	.29	.35	.03	.02
		10	100	.48	.18	.24	.29	.01	.01

Table 4 continued

				δ=3				Standard Deviation	
Type	p	k	n	MDF	EV_1	EV_2	EV_3	MDF	EV
C	4	5	25	.92	.91	.91	.92	.02	.02
		5	100	.90	.90	.91	.91	.02	.02
		10	25	.91	.90	.90	.91	.02	.02
		10	100	.90	.90	.90	.90	.01	.01
	10	5	25	.92	.92	.92	.93	.02	.02
		5	100	.88	.87	.87	.88	.01	.01
		10	25	.87	.86	.86	.86	.02	.02
		10	100	.90	.90	.90	.90	.01	.01
S	4	5	25	.84	.45	.62	.74	.04	.03
		5	100	.83	.42	.59	.72	.02	.03
		10	25						
		10	100		Not Applicable				
	10	5	25	.86	.46	.66	.77	.02	.04
		5	100	.84	.43	.61	.73	.02	.03
		10	25	.74	.23	.36	.46	.03	.03
		10	100	.72	.20	.31	.40	.01	.01

The following conclusions may be drawn from these data.

1. As expected, the EV approach did much better for the C cases than for the S cases. Averaged over all other factors, using one eigenvector correctly classified 0.70 of the C cases, but only .28 of the S cases. The differences between the MDF and EV methods are fairly small for the C groups, but considerably greater for the S cases. In the S cases no combination of eigenvectors was good for p=10, for p=4 occasionally 3 eigenvectors were satisfactory.

2. Increasing the sample size led to an observed decrease in the apparent probability of correct classification. While an increase in sample size should improve the functions, and thus imply an increase in the probability of correct classification, the method of estimating error, namely resubstituting the observations in the functions is known to have a positive bias, and increasing the sample size reduces this bias. The effect is about the same for all groups. This problem could have been avoided by use of an index sample technique in which the functions are tested on an independent set of observations.

3. An increase in k, the number of groups, leads to a decrease in the mean probability of correct classification because there are more chances for erroneous assignments. Tables 1 and 3 show the situation. Comparing them to Table 4, we observe close agreement between the theoretical and sample cases, with the bias of the resubstitution method being noticeable.

4. An increase in the number of variables, p, gives an increase in the probability of correct classification. This is interesting since no further information is supplied in these extra variables. The explanation lies in the fact that a set of data can be fit more exactly when more variables are used. The monotone increase in R^2 in multiple-regression problems is a familiar example of the same phenomenon.

Discussion:

If the means are collinear or nearly so, the EV method works about as well as the MDF method. If the number of variables is large, this may represent some economy in computation, particularly if the function is to be used in assigning future observations. However, if the above conditions do not hold, the MDF is much better than the EV and should be used.

Unfortunately, no firm statements on the effects of sample size can be made, because of the problems mentioned earlier. It appears that both the MDF and EV methods are affected to about the same degree.

The number of groups is inversely related to the performance of both methods and is about the same for both methods. The following table gives the performance figures. These are averaged over all other factors (i. e. δ,n, and p).

TABLE 5

Probability of correct classification by

number of groups, type of population, and method

		MDF	EV_1	EV_2	EV_3
C	k=5	.74	.72	.73	.73
	k=10	.71	.69	.70	.70
S	k=5	.64	.38	.50	.58
	k=10	.41	.18	.25	.30

The apparent improvement in performance as number of variables increases is greater for the MDF than any of the EV procedures. Table 6 gives the details.

TABLE 6

Probability of correct classification by number of variables, type of population and method

		MDF	EV_1	EV_2	EV_3
C	p=4	.71	.70	.71	.71
	p=10	.74	.70	.72	.73
S	p=4	.47	.28	.37	.43
	p=10	.58	.28	.39	.46

The eigenvectors are often used to describe the populations graphically. Many of the computer programs offer plots of the first two eigenvectors (sometimes referred to as canonical vectors) as an option to the user. These can be valuable adjuncts to the analysis when the means are roughly in a two dimensional subspace. The extent to which this holds is usually measured by the proportion of the total variance "explained" by the first two eigenvectors.

This study has considered the behavior of the MDF and EV methods under "ideal" conditions. That is, the variables were normally distributed and the covariance matrices were the same. Nothing has been done regarding the behavior of these methods when non-normality is present the populations have different covariance matrices, or different sample sizes in the populations are used. These remain problems for future study.

Acknowledgements

This research was supported by Research Career Development Award HD-46344. Computing time was made possible by the Triangle Universities Computing Center and by support from the University of North Carolina. Discussions with Judith O'Fallon, M.L. Eaton, L.L. Kupper and P.K. Sen have been most helpful.

References

1. Anderson, T.W. (1951) "Classification by multivariate analysis." Psychometrika, 16, pp. 31-50.

2. Bechhofer, R.E. and Sobel, M. (1954) "A single sample multiple decision procedure for ranking variances of normal populations. Ann. Math. Stat., 25, pp. 273-289.

3. Dunn, O.J., (1971) "Some expected values for probabilities of correct classification in discriminant analysis," Technometrics, 13, pp. 345-353.

4. Fisher, R.A. (1936) "The use of multiple measurements in taxonomic problems." Ann. Eugen., 7, pp. 179-188.

5. Gilbert, E. (1968) "On discrimination using qualitative variables." JASA, 63, pp. 1399-1412.

6. Gilbert, E. (1969) "The effect of unequal variance-covariance matrices on Fisher's linear discriminant function." Biometrics, 25, pp. 505-515.

7. Glick, N. (1972) "Sample based classification procedures derived from density estimators." JASA, 67, pp. 116-122.

8. Hills, M. (1966) "Allocation rules and their error rates." JRSS(B), 28, pp. 1-31.

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

10. Lachenbruch, P.A. and Mickey, M.R. (1968) "Estimation of error rates in discriminant analysis." Technometrics, 10, pp. 1-11.

11. Lachenbruch, P.A., Sneeringer, C. and Revo, L.T. (1973) "Robustness of the linear and quadratic discriminant functions to certain types of non-normality." Communications in Statistics, 1, to appear.

12. Okamoto, M. (1963) "An asymptotic expansion for the distribution of the linear discriminant function." Ann. Math. Stat., 34, pp. 1286-1301.