Journal of Educational Statistics
Fall 1979, Volume 4, Number 3, pp. 247-263

THE TWO KINDS OF LINEAR DISCRIMINANT FUNCTIONS AND THEIR RELATIONSHIP

Bert F. Green

The Johns Hopkins University

Key words: Discriminate Functions, Classification

ABSTRACT

Fisher's two-group discriminant function has been generalized in two different ways for the case of three or more groups. Both generalizations have been called discriminant functions, leading to confusion in the literature. A reasonable nomenclature is canonical discriminant functions and classification functions. The precise functional relation between the two functions is derived, and the interpretation of the two functions is discussed. An illustrative example is provided.

INTRODUCTION

Two different sets of linear combinations of variables are useful in describing group differences and in classifying individuals into groups. The two sets of functions are both used when measures are available on several variables for each case in several groups. But the functions are quite different - and must be used differently to achieve the same ends. Unfortunately, both have been called discriminant functions in the literature. A student who fails to read the fine print will be confused by the literature and misled by the popular statistical computing packages. This paper explains the differences and derives

the precise relationship between the two functions.

Fisher (1938) developed the linear discriminant function as the linear combination of variables that best discriminates between two groups. This discriminant function is the linear combination with the largest value of t, and is the basis for Hotelling's (1931) multivariate T^2 statistic. The same function can also be used to classify cases, and hence provides a basis for discerning group membership of future cases. Cases with function values below some cutoff are classified in one group, values above the cutoff are put in the other group.

There are two ways to generalize this concept to more than two groups. One way is to generalize t to F, seeking the linear combination with the largest ratio of between-group to within-group variance (see, for example, Tatsuoka, 1971; Wilks, 1962). There is a set of orthogonal discriminant functions, each providing progressively less discrimination among the groups. Because of their close relation to canonical correlation, they can be called canonical discriminant functions. These functions are mainly useful in describing group differences. They show the major ways in which the group centroids vary. In fact, as will be shown, they can be viewed as principal components of the between-group covariance matrices, adjusted for within-group covariance structure.

A second and more pragmatic way to generalize the two-group discriminant function is to find a set of linear combinations, one for each group, that indicates the relative closeness of an individual case to each group centroid. These functions are important in classification; they provide a convenient linear basis for classifying new cases. The coefficients of these functions are seldom discussed descriptively, but in fact they are simply the group centroids, adjusted for the within-group covariance structure. They have been called discriminating functions (Nillson, 1965), discriminatory functions (Dixon, 1975), discriminant functions (Rao, 1965), and classification functions (Overall & Klett, 1972). Anderson (1958) presents them without any label. The term classification functions seems best; it is distinctive and describes their main use.

Much of the statistical literature is concerned with only the two-group problem, in which case the two types of functions are identical. When more than two groups are to be discriminated, they can always be decided in pairs. Doing so, in the context discussed in this paper, is equi-

valent to the use of the linear classification functions discussed above.

Some texts treat only the two-group problem, in which case both functions are identical; students could mistakenly conclude that both approaches are identical in general. Some tests treat only one of the two types of functions; others treat both, sometimes in different contexts and with no clear description of the relationship. The proceedings of one recent conference (Cacoullos, 1973) contain papers about both kinds of discriminant functions, with no acknowledgement of the differences. Good, general discussions can be found in Tatsuoka (1971), Overall & Klett (1972), and Nie, Hull, Jenkins, Steinbrenner, and Bent (1972).

A SUMMARY OF THE KNOWN RESULTS

The main facts about both canonical discriminant functions and classification functions are described here, and their interrelationship is explained. Details of the analysis are in a later section. Only the description and interpretative aspects of discriminant functions are discussed; the algebraic results are true for the samples, or for the population, if population values are at hand. Problems of estimation and inference are not treated here.

We start with observation on p variables, X_j, for a sample of N cases divided into m groups, with n_g cases in the g-th group. The analysis focusses on the group means, $\bar{x}_{gj}$. Since only mean differences are of interest, we really study the deviations of the group means from the grand mean, so it will be very convenient, and will lost nothing, to set the grand mean to zero in the sample.

Canonical discriminant analysis is based on the analysis of variance. Differences between groups are quantified as between-group mean squares; the analysis seeks the linear combination of variables $V = \Sigma_j a_j X_j$, for which the ratio of between-group to within-group mean square is largest. The between-group mean square (variance-covariance) matrix, B, has typical entry

$$b_{jk} = \frac{1}{m-1} \sum_g n_g (\bar{x}_{jg} - \bar{x}_{j.})(\bar{x}_{kg} - \bar{x}_{k.}) = \frac{1}{m-1} \sum_g n_g \bar{x}_{jg} \bar{x}_{kg},$$

where x_{ji} is the observation of the j-th variable for the i-th case, expressed as a deviation from the grand mean; $x_{jg} = n_g^{-1} \Sigma x_{ji}$ is the mean of variable j for the g-th group; and $x_{j.} = N^{-1} \Sigma_g n_g x_{jg} = 0$, by definition. Within each group there is a similar variance-covariance matrix, W_g, with typical entry

$$w_{jk(g)} = \frac{1}{(n_g - 1)} \sum_{i=1}^{n_g} (x_{ji} - \bar{x}_{jg})(x_{ki} - \bar{x}_{kg}).$$

The pooled within-group mean-square matrix is W, with typical entry

$$x_{jk} = \frac{1}{n-m} \sum_g (n_g - 1)\, w_{jk(g)}.$$

Many writers define discriminant functions in terms of sums of squares and cross-products (SSCP) matrices, that is, SSCP(B) = (m-1)B; SSCP(W) = (n-m)W. The use of mean squares is descriptively preferable.

The ratio of between-group to within-group variance can now be expressed as

$$\lambda = \frac{(a'Ba)}{(a'Wa)},$$

where a is a column vector of weights (a_j) for the variables x_j. Calculus shows that the particular a that maximizes λ, subject to the scale-specifying restriction that a'Wa = 1, is given by

$$Ba = \lambda Wa.$$

This equation is an eigenstructure equation and has as many solutions for λ and a as the rank of B, which we denote by q. There can be no more canonical functions than variables, so $q \leq p$; since the between-groups sums of squares have m - 1 degrees of freedom, there can be no more canonical functions than one less than the number of groups, $q \leq (m-1)$: so $q \leq \min(p, m - 1)$. The q different solution pairs, $(\lambda_r, a_r, r = 1, 2, \ldots q)$ are mutually orthogonal, in the sense that $a'_2 W a_1 = a'_1 W a_2 = 0$, etc. (In the rare case of equal roots, the corresponding a's are indeterminate but can always be chosen to be mutually orthogonal.)

Thus, although we sought only the one best function, a_1 and λ_1 we obtained a complete set of orthogonal functions. When the eigenvalues, which are the variance ratios, λ_r, are sorted from largest to smallest, the associated canonical discriminant functions account for successively less between-group variation, relative to within-group variation, and the entire set of g functions completely describe the between-group mean differences on the p variables. If λ_1 is large relative to the other λ_r's, then there is only one principal dimension of group difference, described by a_1. On the other hand, if several λ_r's are of notable size, then there are several orthogonal dimensions of important group differences.

Usually, interest centers on the λ's and a's, but the canonical discriminant function can also be used to classify cases. For that, we need scores, v_{ri}, for each case on the discriminant functions. Let x_i be a column vector of scores for Case i on the original variables. Let A be a p x q matrix which has as columns the weights for the successive discriminant functions a_r. Then the scores of Case i on the q canonical discriminant functions are given by $v_i = x_i'A$. The corresponding group means are $\bar{v}_g' = \bar{x}_g'A$, where $\bar{v}_g$ and $\bar{x}_g$ are column vectors of means for Group g. The squared distance, in discriminant space, of the i-th case from the g-th group centroid is

$$d^2_{ig} = (v_i - \bar{v}_g)'(v_i - \bar{v}_g).$$

A case can be classified with the group to which it is closest, or more elaborate Bayesian considerations can be entertained. For example, to take group size into account, we can assume that the within-group distributions of the v's are normal (and spherical), which leads to selecting the group for which $d^2_{ig} - \log_e p(g)$ is smallest, where p(g) is the proportion of total cases to be classified in the g-th group.

CLASSIFICATION FUNCTIONS

If the main purpose of the analysis is to obtain a basis for classifying individuals into groups, describing group differences may not be relevant. Classification may reasonably be based directly on the Mahalanobis distance. The squared Mahalanobis distance from Case i to the centroid of Group g is

$$m^2_{ig} = (x_i - \bar{x}_g)'W^{-1}(x_i - \bar{x}_g).$$

This expands to

$$m^2_{ig} = x_i'W^{-1}x_i - 2x_i'W^{-1}\bar{x}_g + \bar{x}_g'W^{-1}\bar{x}_g.$$

The first term is the squared distance of the case to the grand mean, which is the same for all groups, so it can be ignored when trying to classify Case i. The last term is the squared distance of the g-th group centroid to the grand mean which is a constant for all cases. The middle term is a linear function of the scores. Then the expression can be written as

$$m^2_{ig} = m^2_{io} + u_{ig},$$

where $m^2_{ig} = x_i'W^{-1}x_i$; $c_g = \bar{x}_g'W^{-1}\bar{x}_g$, and $u_{ig} = c_g - 2x_i'W^{-1}\bar{x}_g$. The vector of classification scores, u_i, with entries u_{ig}, g = 1, 2, ...m, is then defined as

$$u_i' = c_0' - 2x_i'C,$$

where c_0' is a row vector with entries c_g, and where C, the coefficients of the classification function, is defined in terms of $\bar{X}$, the p x m matrix of group means $\bar{x}_{jg}$, as

$$C = W^{-1}\bar{X}.$$

The same Bayesian considerations mentioned above can be used here.

If each group is assumed to be normally distributed,

with mean u_g and covariance matrix Σ, estimated by $\bar{x}_g$ and W, then the values of the Mahalanobis distance are proportional to the log likelihoods for the various groups, and the linear classification scores for a given case are linearly related to the log likelihoods. The linear relation includes a constant dependent on the case, so the scores reflect only relative likelihood across groups for a case. Also under these assumptions, as Tatsuoka (1971) points out, m^2_{ig} has a chi square distribution with p degrees of freedom.

The attraction of linear functions for classification is ease of computation. Consequently, they are now less attractive than they once were, since computing power is now so prevalent. Linear functions are still important in automatic pattern recognition in artificial intelligence (picture processing or speech recognition) where vast amounts of computation are needed, but for problems of modest size, it is almost as easy to compute d^2_{ig} or m^2_{ig}. Only a few seconds are involved, even with an many as 1,000 cases on 10 variables in 5 groups.

It is easily shown (e.g., Kshirsagar & Arseven , 1975) that

$$m^2_{ig} = d^2_{ig} + e_i^2,$$

where e_i^2 is the squared distance of Case i to the discriminant space, which is zero if q = p, but not if q < p. In any case e_i^2 is a constant across groups, so classifying by m^2, d^2, and u will yield identical results. But if we should want to compare values across cases, the classification function scores u are defective. A term, m^2_{io}, has been lost. Relative sizes of u, across groups for a given case, mean something, but relative sizes of u across cases is meaningless. Nor is it possible to tell from u whether a case is so far from every group that it should not be classified at all. Only m^2 can show that.

For either type of function to be useful, the variance-covariance structure must be similar within groups: the W_g must be alike. Specifically, the data are assumed to be samples from m populations with the same within-group covariance matrix. Then W is a good estimate of the common matrix. The analysis also requires that W be of full rank,

that is, $|W| \neq 0$. If $|W| = 0$, the offending variable or variables must be removed, or the variables must be transformed, so that W for the remaining variables has full rank. (For a complete discussion of the problem when $|W| = 0$, see McDonald, Tori, & Nishisato, [1979].)

When the within-group covariance matrices are different for different groups, neither the classification functions nor the discriminant functions give all the available information for classifying cases. Both kinds of functions retain their meaning, with respect to W, but better classification can be achieved by using Mahalanobis distances calculated for each group separately, using the W_g for that group. Classifying the available cases by that method will show how much can be gained by using the various W_g's instead of the pooled W.

THE RELATION BETWEEN THE TWO SETS OF FUNCTIONS

The two sets of linear function $v_i' = x_i'A$, and $u_i' = c_0' - 2x_i'C$ have weight matrices A and C respectively. Clearly the two sets of weights are quite different: A has only q columns, whereas C has m columns. There are always fewer canonical discriminant functions than classification functions. Yet just as plainly, there is a strong relation between A and C, since, used somewhat differently, they lead to exactly the same classifications.

First, note that if the original variables were uncorrelated within groups, and had unit variance (i.e., if W = I), then the relation would be easy to see. In that case, the eigenequation would reduce to $Ba = a\lambda$, which is the ordinary eigenequation that arises in principal component analysis. So the columns of A are the principal components of the between-groups covariance matrix, B. Further, the g-th column of C becomes simply $\bar{x}_g$. Thus C contains the group means, expressed as deviations from the centroid. Since the between-group covariance is computed from the weighted cross-products of these groups means, the columns of A are the principal components of the covariances of the columns of C.

When $W \neq I$, the relationship is a bit more complex. However, a transformation can readily be found to a set of uncorrelated variables (called Y_i below) which can be viewed as adjusting for the within-group correlation. One can say that the discriminant functions are the principal components

of the covariances of the group means, adjusted for within-group covariance. The transformation is a standard way to compute discriminant functions.

The relationship of C and A can be used to show (see below) that a simple linear function transforms A to C:

$$C = A\overline{V}$$

Further, this transform, $\overline{V}$, has an interesting interpretation: it contains the group means on the canonical discriminant functions. Thus, a good way to look at the classification coefficients C is in terms of their counterparts, $\overline{V}$, in discriminant coordinates. C is merely a p dimensional version of the q dimensional group differences.

A related fact is that the squared distances in discriminant space, d^2_{ig}, can be factored in just the same way as the Mahalanobis squared distance, m^2_{ig}. It can readily be shown that

$$d^2_{ig} = d^2_{i0} + u_{ig},$$

where $d^2_{io} = v_i'v_i$, the squared distance of the i-th case from the centroid, in discriminant space. Obviously, then, the only difference between u_{ig} and d^2_{ig} is d^2_{i0}, which is constant for case i across groups. It follows rather directly that

$$m^2_{ig} = d^2_{ig} + e^2_i = e^2_i + d^2_{i0} + 2u_{ig},$$

where e^2_i is the squared distance of the i-th case to the discriminant space. The squared Mahalanobis distance of Case i to Group g has three components: the squared distance to the discriminant space, the squared distance to the centroid within the discriminant space, and the classification function value for Group g.

Both C and A give information about the group structure. The group means, adjusted for W, are in C, and show on which variables a given group is especially high or low. The principal components of the between-group covariance matrix, adjusted for W, are in A; they show the principal ways in which the groups differ from each other. In an ordinary scientific investigation, both would help. If the nature of the intergroup difference is of no interest, then classi-

fication functions may be enough. Even then, if one wants to consider the possibility that a case may be so far away from any group that it should not be classified at all, then discriminant distance or Mahalanobis distance must be calculated.

An example is shown in Table I, involving four variables, three groups, and hence two canonical discriminant functions, and three classification functions. Table II contains intermediate results in the Y-transformed variables used in this detailed analysis.

TABLE I

Discriminant Function Example:
Four Variables, Three Groups

B

324	243	324	162
243	243	243	0
324	243	324	162
164	0	162	324

W

6	2	2	0
2	5	2	2
2	2	8	4
0	2	4	4

$\overline{X}$

0	6	-6
-3	6	-3
0	6	-6
6	0	6

C

1.50	0	-1.50
-2.25	1.50	0.75
-2.25	1.50	0.75
4.88	-2.25	-2.63

A

.250	.250
-.375	.125
-.375	.125
.813	.063

$\overline{V}$

6.	-3.	-3.
0	3.	-3.

C_0

-18.	-9.	-9.

λ

243.1	81.0

TABLE II

Auxiliary Matrices for Example

F

.408	-.160	-.088	.244
0	.480	-.117	-.261
0	0	.380	-.410
0	0	0	.839

G

2.500	0	0	0
.816	2.081	0	0
.816	.640	2.631	0
0	.961	1.286	1.192

B_y

54.016	26.484	27.077	4.931
26.484	27.008	9.862	-54.155
27.077	9.862	14.404	16.243
4.931	-54.155	16.243	228.669

A_y

0	.816
-.240	.400
-.058	.409
.969	.075

C_y

0	-1.441	0.351	5.815
2.450	1.922	1.053	-2.684
-2.450	-0.480	-1.404	-3.131

C_0

-18.
-9.
-9.

The Two-group Problem

When there are only two groups, the two columns of $\overline{X}$ are proportional:

$$x_{1j} = -\frac{n_2}{n_1}\overline{x}_{2j},$$

and it follows directly that the two-columns of C are proportional. Also, there is only one discriminant function; A has one column. $\overline{V}$ is 1 x 2. It follows that A is also proportional to each of the columns of C. Thus, for two groups, the classification function coefficients C are proportional to the discriminant function coefficients A.

The two-group classification functions have been treated extensively in the statistical literature. Of course, if there are m groups, there are m(m-1)/2 pairs of groups, and the members of each pair can be contrasted by the two-group classification formula for that pair. There are

$m(m-1)/2$ such functions. When $W_g = W$, this procedure is equivalent to using the linear classification functions described above. That is, with $m > 2$, the difference between the two linear classifications for groups i and j is the two-group linear classification for groups i and j separately (always assuming that one overall estimate of W is used throughout).

It is important to recognize that canonical discriminant functions may not provide the best classification, unless all of them are used. In the present example, the first discriminant function separates Group I from II and III, but does not distinguish at all between II and III. The second canonical function discriminates II from III almost perfectly (assuming multivariate normality). But the single function, $.707(v_1 - v_2)$, provides almost perfect discrimination among the groups. This merely demonstrates that maximum between-group variance is a different concept from maximum proportion of correct classification. (That is, the three numbers 6, -3, and 3 have larger variance than the three numbers 4.24, 0, and -4.24.) In this example, one function would serve to classify nearly everyone; two canonical discriminant functions are needed; and, of course, there are three classification functions, one per group.

DETAILS OF THE ANALYSIS

In the original (x) variables, the canonical discriminant weights A, a p x q matrix, satisfy

$$A'BA = D_\lambda$$

$$A'WA = I,$$

where D_λ is a q x q diagonal matrix with elements λ_r. The discrimiant scores are

$$v_i' = x_i'A,$$

where v_i and x_i are column vectors of scores for the i-th case. The p x m matrix $\bar{X}$ contains the group means x_{jg}; the classification function weights are

$$C = W^{-1}\bar{X};$$

and the classification scores for the i-th case are

$$u_i' = c_0' - 2x_i'C.$$

To transform to uncorrelated (Y) variables, we find a p x p matrix V such that

$$W = FF'.$$

Any grammian factoring, including principal components, will yield such an F. Then define

$$G = (F')^{-1}.$$

It follows that $G'F = GF' = F'G = FG' = 1$, and that

$$W^{-1} = GG'.$$

Then let

$$y_i' = x_i'G;\ x_i' = y_i'F'.$$

Matrices associated with the analysis of the Y variables are denoted by a y- subscript. It is readily verified that

$$B_y = G'BG;\ B = FB_yF',$$

$$W_y = G'WG = 1;\ W = FW_yF' = FF'.$$

The canonical discriminant weights for the Y variables satisfy

$$A_y'B_yA_y = D_\lambda$$

$$A'_yA_y = I.$$

Then the tranform between A and A_y is

$$A_y = F'A;\ A = GA_y.$$

Logically, the same discriminant scores should be obtained from an analysis of either the X version or the Y version of the same data. It is easily verified that

$$v_i' = x_i'A = y_i'A_y.$$

Similarly, the classification scores are

$$u_i' = c_0 - 2y_iC_y = c_0 - 2x_iC;$$

$$C = GC_y,\ C_y = F'C,\ C_y = \bar{Y}.$$

The entries in c_0 are the diagonals of $\bar{Y}'\bar{Y} = (\bar{X}'G)(G'\bar{X}) = \bar{X}'W^{-1}\bar{X}$, and are thus unchanged by the transform.

Equivalent results can be expressed in the discriminant variables (V):

$$B_v = D_\lambda$$

$$W_v = I$$

$$A_v = I$$

$$C_v = \bar{V}.$$

Since $v_i' = y_i'A_y$, or $v_i = A_y'y_i$, it follows that

$$\bar{V} = A_y'\bar{Y};\ \ \bar{V} = A_y'C_y.$$

To derive the relationship between C and A, note that by definition the between-groups covariance matrix B_y is obtained by

$$B_y = \bar{Y}D_m\bar{Y}',$$

where D_m is an m x m diagonal matrix, with entries $n_g/(m-1)$. But from above,

$$B_y = A_yD_\lambda A_y'.$$

It must therefore be the case that the singular value decomposition of $\bar{Y}D_m^{\frac{1}{2}}$ can be expressed as

$$\bar{Y}D_m^{\frac{1}{2}} = AD_\lambda^{\frac{1}{2}}Q';\ A'A = I,\ Q'Q = QQ' = I.$$

Since $\bar{Y} = C_y$, we have

$$C_y = A_y T$$

$$T = D_\lambda^{\frac{1}{2}} Q' D_m^{-\frac{1}{2}}.$$

Since $A_y' A_y = I$, premultiplying by A_y' yields

$$A_y' C_y = T.$$

Above we showed that $A_y' C_y = \overline{V}$, so we must have

$$\overline{V} = T.$$

Finally,

$$C_y = A_y \overline{V}; \; A_y' C_y = \overline{V}.$$

Since $C_y = F'C$ and $A_y = F'A$, we also have

$$C = A\overline{V}; \; A'WC = \overline{V}.$$

SUMMARY

Linear canonical discriminant functions, with coefficients A, and linear classification functions, with coefficients C, are not the same. They are related like principal components and deviation scores. A linear transformation $\overline{V}$ takes A into C. There is always one classification function for each group, and its coefficients are the deviations from the overall centroid of the means of the variables for that group, adjusted for W. There are no more than the smaller of p and m - 1 canonical discriminant functions, and their coefficients are the eigenvectors of the between-group covariance matrix, adjusted for W. In the special case of two groups, the classification functions are colinear with the canonical discriminant function.

ACKNOWLEDGEMENT

This article was originally presented at the August 1978 meeting of the Psychometric Society.

REFERENCES

Anderson, T.W. An introduction to multivariate statistical analysis. New York: Wiley, 1958.

Cacoullos, T. Discriminant analysis and applications. New York: Academic Press, 1973.

Dixon, W.J. (Ed.). BMD: Biomedical computer programs. Berkley, Calif.: University of California Press, 1975.

Fisher, R.A. The statistical utilization of multiple measurements. Annals of Eugenics, 1938, 8, 376-386.

Hotelling, H. The generalization of student's ratio. Annals of Mathematical Statistics, 1931, 2, 360-378.

Kshirsagar, A.M., & Arseven, E. A note on the equivalence of two discrimination procedures. The American Statistician, 1975, 29 (1), 38-39.

McDonald, R.P., Tori, Y., & Nishisato, S. Some results on proper eigenvalues and eigenvectors with applications to scaling. Psychometrika, 1979, 44, 211-228.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. SPSS: Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1972.

Nillson, H.J. Learning machines. New York, McGraw-Hill, 1965.

Overall, J.E., & Klett, C.J. Applied multivariate analysis. New York: McGraw-Hill, 1972.

Rao, C.R. Linear statistical inference and it applications. New York: Wiley, 1965.

Tatsuoka, M.M. Multivariate analysis: Techniques for educational and psychological research. New York: Wiley, 1971.

Wilks, S.S. Mathematical statistics. New York: Wiley, 1962.

AUTHOR

GREEN, BERT F. Address: Department of Psychology, Johns Hopkins University, Baltimore, MD 21218. Title: Professor of Psychology. Degrees: A.B., Yale University; M.A., Ph.D., Princeton University. Specialization: Quantitative Psychology