Back

Discriminant Function Analysis

General Purpose

Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups. For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide (1) to go to college, (2) to attend a trade or professional school, or (3) to seek no further training or education. For that purpose the researcher could collect data on numerous variables prior to students' graduation. After graduation, most students will naturally fall into one of the three categories. *Discriminant Analysis* could then be used to determine which variable(s) are the best predictors of students' subsequent educational choice.

A medical researcher may record different variables relating to patients' backgrounds in order to learn which variables best predict whether a patient is likely to recover completely (group 1), partially (group 2), or not at all (group 3). A biologist could record different characteristics of similar types (groups) of flowers, and then perform a discriminant function analysis to determine the set of characteristics that allows for the best discrimination between the types.

To index

Computational Approach

Computationally, discriminant function analysis is very similar to analysis of variance (*ANOVA*). Let us consider a simple example. Suppose we measure height in a random sample of 50 males and 50 females. Females are, on the average, not as tall as

males, and this difference will be reflected in the difference in means (for the variable *Height*). Therefore, variable height allows us to discriminate between males and females with a better than chance probability: if a person is tall, then he is likely to be a male, if a person is short, then she is likely to be a female.

We can generalize this reasoning to groups and variables that are less "trivial." For example, suppose we have two groups of high school graduates: Those who choose to attend college after graduation and those who do not. We could have measured students' stated intention to continue on to college one year prior to graduation. If the means for the two groups (those who actually went to college and those who did not) are different, then we can say that intention to attend college as stated one year prior to graduation allows us to discriminate between those who are and are not college bound (and this information may be used by career counselors to provide the appropriate guidance to the respective students).

To summarize the discussion so far, the basic idea underlying discriminant function analysis is to determine whether groups differ with regard to the mean of a variable, and then to use that variable to predict group membership (e.g., of new cases).

Analysis of Variance. Stated in this manner, the discriminant function problem can be rephrased as a one-way analysis of variance (ANOVA) problem. Specifically, one can ask whether or not two or more groups are *significantly different* from each other with respect to the mean of a particular variable. To learn more about how one can test for the statistical significance of differences between means in different groups you may want to read the Overview section to *ANOVA/MANOVA*. However, it should be clear that, if the means for a variable are significantly different in different groups, then we can say that this variable discriminates between the groups.

In the case of a single variable, the final significance test of whether or not a variable discriminates between groups is the *F* test. As described in *Elementary Concepts* and *ANOVA /MANOVA*, *F* is essentially computed as the ratio of the between-groups variance in the data over the pooled (average) within-group variance. If the between-group variance is significantly larger then there must be significant differences between means.

Multiple Variables. Usually, one includes several variables in a study in order to see which one(s) contribute to the discrimination between groups. In that case, we have a matrix of total variances and covariances; likewise, we have a matrix of pooled within-group variances and covariances. We can compare those two matrices via multivariate *F* tests in order to determined whether or not there are any significant differences (with regard to all variables) between groups. This procedure is identical to multivariate analysis of variance or

MANOVA. As in *MANOVA*, one could first perform the multivariate test, and, if statistically significant, proceed to see which of the variables have significantly different means across the groups. Thus, even though the computations with multiple variables are more complex, the principal reasoning still applies, namely, that we are looking for variables that discriminate between groups, as evident in observed mean differences.

To index

Stepwise Discriminant Analysis

Probably the most common application of discriminant function analysis is to include many measures in the study, in order to determine the ones that discriminate between groups. For example, an educational researcher interested in predicting high school graduates' choices for further education would probably include as many measures of personality, achievement motivation, academic performance, etc. as possible in order to learn which one(s) offer the best prediction.

Model. Put another way, we want to build a "model" of how we can best predict to which group a case belongs. In the following discussion we will use the term "in the model" in order to refer to variables that are included in the prediction of group membership, and we will refer to variables as being "not in the model" if they are not included.

Forward stepwise analysis. In stepwise discriminant function analysis, a model of discrimination is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups. That variable will then be included in the model, and the process starts again.

Backward stepwise analysis. One can also step backwards; in that case all variables are included in the model and then, at each step, the variable that contributes least to the prediction of group membership is eliminated. Thus, as the result of a successful discriminant function analysis, one would only keep the "important" variables in the model, that is, those variables that contribute the most to the discrimination between groups.

F **to enter,** *F* **to remove.** The stepwise procedure is "guided" by the respective *F* to enter and *F* to remove values. The *F* value for a variable indicates its statistical significance in the discrimination between groups, that is, it is a measure of the extent to which a variable makes a unique contribution to the prediction of group membership. If you are familiar with stepwise multiple regression procedures, then you may interpret the *F* to enter/remove values in the same way as in stepwise regression.

Capitalizing on chance. A common misinterpretation of the results of stepwise discriminant analysis is to take statistical significance levels at face value. By nature, the stepwise procedures will capitalize on chance because they "pick and choose" the variables to be included in the model so as to yield maximum discrimination. Thus, when using the stepwise approach the researcher should be aware that the significance levels do not reflect the true *alpha* error rate, that is, the probability of erroneously rejecting H_0 (the null hypothesis that there is no discrimination between groups).

To index

Interpreting a Two-Group Discriminant Function

In the two-group case, discriminant function analysis can also be thought of as (and is analogous to) multiple regression (see Multiple Regression; the two-group discriminant analysis is also called *Fisher linear discriminant analysis* after Fisher, 1936; computationally all of these approaches are analogous). If we code the two groups in the analysis as *1* and *2*, and use that variable as the dependent variable in a multiple regression analysis, then we would get results that are analogous to those we would obtain via *Discriminant Analysis*. In general, in the two-group case we fit a linear equation of the type:

$$Group = a + b_1*x_1 + b_2*x_2 + ... + b_m*x_m$$

where *a* is a constant and b_1 through b_m are regression coefficients. The interpretation of the results of a two-group problem is straightforward and closely follows the logic of multiple regression: Those variables with the largest (standardized) regression coefficients are the ones that contribute most to the prediction of group membership.

To index

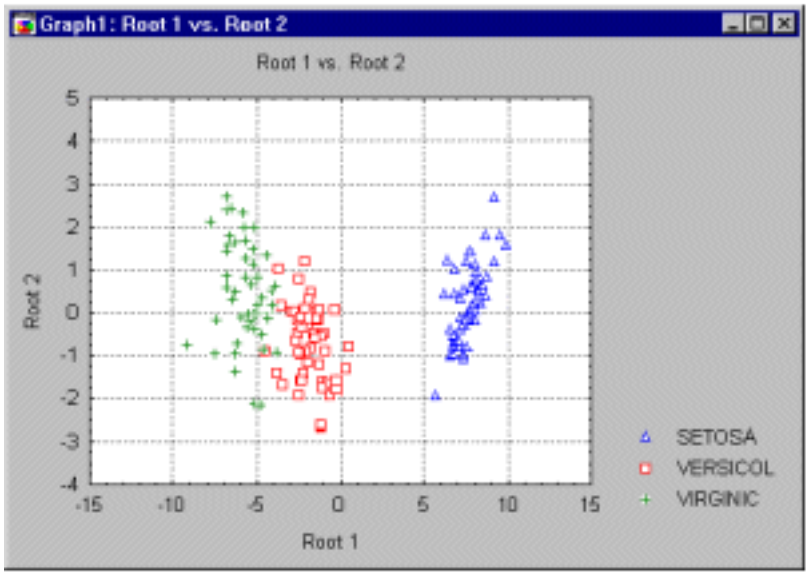
Discriminant Functions for Multiple Groups

When there are more than two groups, then we can estimate more than one discriminant function like the one presented above. For example, when there are three groups, we could estimate (1) a function for discriminating between group 1 and groups 2 and 3 combined, and (2) another function for discriminating between group 2 and group 3. For example, we could have one function that discriminates between those high school graduates that go to college

and those who do not (but rather get a job or go to a professional or trade school), and a second function to discriminate between those graduates that go to a professional or trade school versus those who get a job. The *b* coefficients in those discriminant functions could then be interpreted as before.

Canonical analysis. When actually performing a multiple group discriminant analysis, we do not have to specify how to combine groups so as to form different discriminant functions. Rather, you can automatically determine some optimal combination of variables so that the first function provides the most overall discrimination between groups, the second provides second most, and so on. Moreover, the functions will be independent or *orthogonal*, that is, their contributions to the discrimination between groups will not overlap. Computationally, you will perform a *canonical correlation* analysis (see also *Canonical Correlation*) that will determine the successive functions and canonical *roots* (the term root refers to the eigenvalues that are associated with the respective canonical function). The maximum number of functions will be equal to the number of groups minus one, or the number of variables in the analysis, whichever is smaller.

Interpreting the discriminant functions. As before, we will get *b* (and standardized *beta*) coefficients for each variable in each discriminant (now also called *canonical*) function, and they can be interpreted as usual: the larger the standardized coefficient, the greater is the contribution of the respective variable to the discrimination between groups. (Note that we could also interpret the *structure coefficients*; see below.) However, these coefficients do not tell us between which of the groups the respective functions discriminate. We can identify the nature of the discrimination for each discriminant (canonical) function by looking at the means for the functions across groups. We can also visualize how the two functions discriminate between groups by plotting the individual scores for the two discriminant functions (see the example graph below).



In this example, *Root* (function) *1* seems to discriminate mostly between groups *Setosa*, and *Virginic* and *Versicol* combined. In the vertical direction (*Root 2*), a slight trend of *Versicol* points to fall below the center line (*0*) is apparent.

Factor structure matrix. Another way to determine which variables "mark" or define a particular discriminant function is to look at the factor structure. The factor structure coefficients are the correlations between the variables in the model and the discriminant functions; if you are familiar with factor analysis (see *Factor Analysis*) you may think of these correlations as factor *loadings* of the variables on each discriminant function.

Some authors have argued that these structure coefficients should be used when interpreting the substantive "meaning" of discriminant functions. The reasons given by those authors are that (1) supposedly the structure coefficients are more stable, and (2) they allow for the interpretation of factors (discriminant functions) in the manner that is analogous to factor analysis. However, subsequent Monte Carlo research (Barcikowski & Stevens, 1975; Huberty, 1975) has shown that the discriminant function coefficients and the structure coefficients are about equally unstable, unless the n is fairly large (e.g., if there are 20 times more cases than there are variables). The most important thing to remember is that the discriminant function coefficients denote the unique (partial) contribution of each variable to the discriminant function(s), while the structure coefficients denote the simple correlations between the variables and the function(s). If one wants to assign substantive "meaningful" labels to the discriminant functions (akin to the interpretation of factors in factor analysis), then the structure coefficients should be used (interpreted); if one wants to learn what is each variable's unique contribution to the discriminant function, use the discriminant function coefficients (weights).

Significance of discriminant functions. One can test the number of roots that add *significantly* to the discrimination between group. Only those found to be statistically significant should be used for interpretation; non-significant functions (roots) should be ignored.

Summary. To summarize, when interpreting multiple discriminant functions, which arise from analyses with more than two groups and more than one variable, one would first test the different functions for statistical significance, and only consider the significant functions for further examination. Next, we would look at the standardized *b* coefficients for each variable for each significant function. The larger the standardized *b* coefficient, the larger is the respective variable's unique contribution to the discrimination specified by the respective discriminant function. In order to derive substantive "meaningful" labels for the discriminant functions, one can also examine the factor structure matrix with the correlations between the variables and the discriminant functions. Finally, we would look at the means for the significant discriminant functions in order to determine between which groups the respective functions seem to discriminate.

To index

Assumptions

As mentioned earlier, discriminant function analysis is computationally very similar to MANOVA, and all assumptions for *MANOVA* mentioned in ANOVA/MANOVA apply. In fact, you may use the wide range of diagnostics and statistical tests of assumption that are available to examine your data for the discriminant analysis.

Normal distribution. It is assumed that the data (for the variables) represent a sample from a multivariate normal distribution. You can examine whether or not variables are normally distributed with histograms of frequency distributions. However, note that violations of the normality assumption are usually not "fatal," meaning, that the resultant significance tests etc. are still "trustworthy." You may use specific tests for normality in addition to graphs.

Homogeneity of variances/covariances. It is assumed that the variance/covariance matrices of variables are homogeneous across groups. Again, minor deviations are not that important; however, before accepting final conclusions for an important study it is probably a good idea to review the within-groups variances and correlation matrices. In particular a scatterplot matrix can be produced and can be very useful for this purpose. When in doubt, try re-running the analyses excluding one or two groups that are of less interest. If the overall results (interpretations) hold up, you probably do not have a problem. You may also use the numerous tests available to examine whether or not this assumption is violated in your data. However, as mentioned in *ANOVA/MANOVA*, the multivariate Box

M test for homogeneity of variances/covariances is particularly sensitive to deviations from multivariate normality, and should not be taken too "seriously."

Correlations between means and variances. The major "real" threat to the validity of significance tests occurs when the means for variables across groups are correlated with the variances (or standard deviations). Intuitively, if there is large variability in a group with particularly high means on some variables, then those high means are not reliable. However, the overall significance tests are based on pooled variances, that is, the average variance across all groups. Thus, the significance tests of the relatively larger means (with the large variances) would be based on the relatively smaller pooled variances, resulting erroneously in statistical significance. In practice, this pattern may occur if one group in the study contains a few extreme outliers, who have a large impact on the means, and also increase the variability. To guard against this problem, inspect the descriptive statistics, that is, the means and standard deviations or variances for such a correlation.

The matrix ill-conditioning problem. Another assumption of discriminant function analysis is that the variables that are used to discriminate between groups are not completely redundant. As part of the computations involved in discriminant analysis, you will invert the variance/covariance matrix of the variables in the model. If any one of the variables is completely redundant with the other variables then the matrix is said to be *ill-conditioned* , and it cannot be inverted. For example, if a variable is the sum of three other variables that are also in the model, then the matrix is ill-conditioned.

Tolerance values. In order to guard against matrix ill-conditioning, constantly check the so-called tolerance value for each variable. This tolerance value is computed as *1 minus R-square* of the respective variable with all other variables included in the current model. Thus, it is the proportion of variance that is unique to the respective variable. You may also refer to Multiple Regression to learn more about multiple regression and the interpretation of the tolerance value. In general, when a variable is almost completely redundant (and, therefore, the matrix ill-conditioning problem is likely to occur), the tolerance value for that variable will approach 0.

To index

Classification

Another major purpose to which discriminant analysis is applied is the issue of predictive classification of cases. Once a model has been finalized and the discriminant functions have been derived, how well can we *predict* to which group a particular case belongs?

A priori **and** *post hoc* **predictions.** Before going into the details of different estimation procedures, we would like to make sure that this difference is clear. Obviously, if we estimate, based on some data set, the discriminant functions that best discriminate between groups, and then use the *same* data to evaluate how accurate our prediction is, then we are very much capitalizing on chance. In general, one will *always* get a worse classification when predicting cases that were not used for the estimation of the discriminant function. Put another way, *post hoc* predictions are always better than *a priori* predictions. (The trouble with predicting the future *a priori* is that one does not know what will happen; it is much easier to find ways to predict what we already know has happened.) Therefore, one should never base one's confidence regarding the correct classification of future observations on the same data set from which the discriminant functions were derived; rather, if one wants to classify cases predictively, it is necessary to collect new data to "try out" (cross-validate) the utility of the discriminant functions.

Classification functions. These are not to be confused with the discriminant functions. The classification functions can be used to determine to which group each case most likely belongs. There are as many classification functions as there are groups. Each function allows us to compute *classification scores* for each case for each group, by applying the formula:

$$S_i = c_i + w_{i1}*x_1 + w_{i2}*x_2 + ... + w_{im}*x_m$$

In this formula, the subscript i denotes the respective group; the subscripts $1, 2, ..., m$ denote the m variables; c_i is a constant for the i'th group, w_{ij} is the weight for the j'th variable in the computation of the classification score for the i'th group; x_j is the observed value for the respective case for the j'th variable. S_i is the resultant classification score.

We can use the classification functions to directly compute classification scores for some new observations.

Classification of cases. Once we have computed the classification scores for a case, it is easy to decide how to classify the case: in general we classify the case as belonging to the group for which it has the highest classification score (unless the *a priori* classification probabilities are widely disparate; see below). Thus, if we were to study high school students' post-graduation career/educational choices (e.g., attending college, attending a professional or trade school, or getting a job) based on several

variables assessed one year prior to graduation, we could use the classification functions to predict what each student is most likely to do after graduation. However, we would also like to know the *probability* that the student will make the predicted choice. Those probabilities are called *posterior* probabilities, and can also be computed. However, to understand how those probabilities are derived, let us first consider the so-called *Mahalanobis* distances.

Mahalanobis distances. You may have read about these distances in other parts of the manual. In general, the Mahalanobis distance is a measure of distance between two points in the space defined by two or more correlated variables. For example, if there are two variables that are uncorrelated, then we could plot points (cases) in a standard two-dimensional scatterplot; the Mahalanobis distances between the points would then be identical to the Euclidean distance; that is, the distance as, for example, measured by a ruler. If there are three uncorrelated variables, we could also simply use a ruler (in a 3-D plot) to determine the distances between points. If there are more than 3 variables, we cannot represent the distances in a plot any more. Also, when the variables are correlated, then the axes in the plots can be thought of as being *non-orthogonal*; that is, they would not be positioned in right angles to each other. In those cases, the simple Euclidean distance is not an appropriate measure, while the Mahalanobis distance will adequately account for the correlations.

Mahalanobis distances and classification. For each group in our sample, we can determine the location of the point that represents the means for all variables in the multivariate space defined by the variables in the model. These points are called group *centroids*. For each case we can then compute the Mahalanobis distances (of the respective case) from each of the group centroids. Again, we would classify the case as belonging to the group to which it is closest, that is, where the Mahalanobis distance is smallest.

Posterior classification probabilities. Using the Mahalanobis distances to do the classification, we can now derive probabilities. The probability that a case belongs to a particular group is basically proportional to the Mahalanobis distance from that group centroid (it is not exactly proportional because we assume a multivariate normal distribution around each centroid). Because we compute the location of each case from our prior knowledge of the values for that case on the variables in the model, these probabilities are called *posterior* probabilities. In summary, the posterior probability is the probability, based on our knowledge of the values of other variables, that the respective case belongs to a particular group. Some software packages will automatically compute those probabilities for all cases (or for selected cases only for cross-validation studies).

A priori classification probabilities. There is one additional factor that needs to be considered when classifying cases. Sometimes, we know ahead of time that there are more

observations in one group than in any other; thus, the *a priori* probability that a case belongs to that group is higher. For example, if we know ahead of time that 60% of the graduates from our high school usually go to college (20% go to a professional school, and another 20% get a job), then we should adjust our prediction accordingly: *a priori* , and all other things being equal, it is more likely that a student will attend college that choose either of the other two options. You can specify different *a priori* probabilities, which will then be used to adjust the classification of cases (and the computation of posterior probabilities) accordingly.

In practice, the researcher needs to ask him or herself whether the unequal number of cases in different groups in the sample is a reflection of the true distribution in the population, or whether it is only the (random) result of the sampling procedure. In the former case, we would set the *a priori* probabilities to be proportional to the sizes of the groups in our sample, in the latter case we would specify the *a priori* probabilities as being equal in each group. The specification of different *a priori* probabilities can greatly affect the accuracy of the prediction.

Summary of the prediction. A common result that one looks at in order to determine how well the current classification functions predict group membership of cases is the *classification matrix* . The classification matrix shows the number of cases that were correctly classified (on the diagonal of the matrix) and those that were misclassified.

Another word of caution. To reiterate, *post hoc* predicting of what has happened in the past is not that difficult. It is not uncommon to obtain very good classification if one uses the same cases from which the classification functions were computed. In order to get an idea of how well the current classification functions "perform," one must classify (*a priori*) *different* cases, that is, cases that were not used to estimate the classification functions. You can include or exclude cases from the computations; thus, the classification matrix can be computed for "old" cases as well as "new" cases. Only the classification of new cases allows us to assess the predictive validity of the classification functions (see also cross-validation); the classification of old cases only provides a useful diagnostic tool to identify outliers or areas where the classification function seems to be less adequate.

Summary. In general *Discriminant Analysis* is a very useful tool (1) for detecting the variables that allow the researcher to discriminate between different (naturally occurring) groups, and (2) for classifying cases into different groups with a better than chance accuracy.