DISCRIMINANT FUNCTION ANALYSIS (DA)

John Poulsen and Aaron French

Key words: assumptions, further reading, computations, standardized coefficents, structure matrix, tests of signficance

Introduction

Discriminant function analysis is used to determine which continuous variables discriminate between two or more naturally occurring groups. For example, a researcher may want to investigate which variables discriminate between fruits eaten by (1) primates, (2) birds, or (3) squirrels. For that purpose, the researcher could collect data on numerous fruit characteristics of those species eaten by each of the animal groups. Most fruits will naturally fall into one of the three categories. Discriminant analysis could then be used to determine which variables are the best predictors of whether a fruit will be eaten by birds, primates, or squirrels.

Logistic regression answers the same questions as discriminant analysis. It is often preferred to discriminate analysis as it is more flexible in its assumptions and types of data that can be analyzed. Logistic regression can handle both categorical and continuous variables, and the predictors do not have to be normally distributed, linearly related, or of equal variance within each group (Tabachnick and Fidell 1996).

Discriminant function analysis is multivariate analysis of variance (MANOVA) reversed. In MANOVA, the independent variables are the groups and the dependent variables are the predictors. In DA, the independent variables are the predictors and the dependent variables are the groups. As previously mentioned, DA is usually used to predict membership in naturally occurring groups. It answers the question: can a combination of variables be used to predict group membership? Usually, several variables are included in a study to see which ones contribute to the discrimination between groups.

Discriminant function analysis is broken into a 2-step process: (1) testing significance of a set of discriminant functions, and; (2) classification. The first step is computationally identical to MANOVA. There is a matrix of total variances and covariances; likewise, there is a matrix of pooled within-group variances and covariances. The two matrices are compared via multivariate F tests in order to determine whether or not there are any significant differences (with regard to all variables) between groups. One first performs the multivariate test, and, if statistically significant, proceeds to see which of the variables have significantly different means across the groups.

Once group means are found to be statistically significant, classification of variables is undertaken. DA automatically determines 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. The first function picks up the most variation; the second function picks up the greatest part of the unexplained variation, etc... Computationally, a canonical correlation analysis is performed that will determine the successive functions and canonical roots. Classification is then possible from the canonical functions. Subjects are classified in the groups in which they had the highest classification scores. The maximum number of discriminant functions will be equal to the degrees of freedom, or the number of variables in the analysis, whichever is smaller.

Standardized coefficients and the structure matrix

Discriminant functions are interpreted by means of standardized coefficients and the structure matrix. Standardized beta coefficients are given for each variable in each discriminant (canonical) function, and the larger the standardized coefficient, the greater is the contribution of the respective variable to the discrimination between groups. 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 function by looking at the means for the functions across groups. Group means are centroids. Differences in location of centroids show dimensions along which groups differ. We can, thus, visualize how the two functions discriminate between groups by plotting the individual scores for the two discriminant functions.

Another way to determine which variables 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.

The discriminant function coefficients denote the unique contribution of each variable to the discriminant function, while the structure coefficients denote the simple correlations between the variables and the functions.

Summary

To summarize, when interpreting multiple discriminant functions, which arise from analyses with more than two groups and more than one continuous variable, the different functions are first tested for statistical significance. If the functions are statistically significant, then the groups can be distinguished based on predictor variables. Standardized b coefficients for each variable are determined 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 identify which independent variables help cause the discrimination between dependent variables, one can also examine the factor structure matrix with the

correlations between the variables and the discriminant functions. Finally, the means for the significant discriminant functions are examined in order to determine between which groups the respective functions seem to discriminate. (For more detail, see Computations below.)

Assumptions:

Discriminant function analysis is computationally very similar to MANOVA, and all assumptions for MANOVA apply.

Sample size: Unequal sample sizes are acceptable. The sample size of the smallest group needs to exceed the number of predictor variables. As a "rule of thumb", the smallest sample size should be at least 20 for a few (4 or 5) predictors. The maximum number of independent variables is n - 2, where n is the sample size. While this low sample size may work, it is not encouraged, and generally it is best to have 4 or 5 times as many observations and independent variables..

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 not "fatal" and the resultant significance test are still reliable as long as non-normality is caused by skewness and not outliers (Tabachnick and Fidell 1996).

Homogeneity of variances/covariances: DA is very sensitive to heterogeneity of variance-covariance matrices. Before accepting final conclusions for an important study, it is a good idea to review the within-groups variances and correlation matrices. Homoscedasticity is evaluated through scatterplots and corrected by transformation of variables..

Outliers: DA is highly sensitive to the inclusion of outliers. Run a test for univariate and multivariate outliers for each group, and transform or eliminate them. If one group in the study contains extreme outliers that impact the mean, they will also increase variability. 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.

Non-multicollinearity: If one of the independent variables is very highly correlated with another, or one is a function (e.g., the sum) of other independents, then the tolerance value for that variable will approach 0 and the matrix will not have a unique discriminant solution. There must also be low multicollinearity of the independents. To the extent that independents are correlated, the standardized

discriminant function coefficients will not reliably assess the relative importance of the predictor variables.

Logistic regression may offer an alternative to DA as it usually involves fewer violations of assumptions.

Tests of significance:

There are several tests of signficance, but we only present Wilks' lambda here. Wilks' lambda is used in an ANOVA (F) test of mean differences in DA, such that the smaller the lambda for an independent variable, the more that variable contributes to the discriminant function. Lambda varies from 0 to 1, with 0 meaning group means differ (thus the more the variable differentiates the groups), and 1 meaning all group means are the same. The F test of Wilks' lambda shows which variables' contributions are significant.

Interpreting the discriminant functions The structure matrix table in SPSS shows the correlations of each variable with each discriminant function. These simple Pearsonian correlations are called structure coefficients or correlations or discriminant loadings. When the dependent has more than two categories there will be more than one discriminant function. The correlations then serve like factor loadings in factor analysis -- that is, by identifying the largest absolute correlations associated with each discriminant function the researcher gains insight into how to name each function.

For Further Reading:

Cooley, W.W. and P. R. Lohnes. 1971. Multivariate Data Analysis. John Wiley & Sons, Inc.

George H. Dunteman (1984). Introduction to multivariate analysis. Thousand Oaks, CA: Sage Publications. Chapter 5 covers classification procedures and discriminant analysis.

Klecka, William R. (1980). Discriminant Analysis. Quantitative Applications in the Social Sciences Series, No. 19. Thousand Oaks, CA: Sage Publications.

Lachenbruch, P. A. (1975). Discriminant Analysis. NY: Hafner. For detailed notes on computations.

Morrison, D.F. 1967. Multivariate Statistical Methods. McGraw-Hill: New York. A general textbook explanation.

Overall, J.E. and C.J. Klett. 1972. Applied Multivariate Analysis. McGraw-Hill: New York.

Press, S. J. and S. Wilson (1978). Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, Vol. 73: 699-705.

Tabachnick, B.G. and L.S. Fidell. 1996. Using Multivariate Statistics. Harper Collins College Publishers: New York. Tabachnick and Fidell compare and contrast statistical packages, and can be used with a modicum of pain to understand SPSS result print-outs.

Webpages:

www.statsoft.com/textbook/stathome.html Statsoft provides descriptions and explanations of many different statistical techniques.

www2.chass.ncsu.edu/garson/pa765/discrim.htm This website offers a good, readable treatment of DA. It also offers very understandable explanations of how to read result print-outs from SPSS and SAS. Other analyses like logistic regression and log-linear models can be found here.

Computations:

Fundamental equations for DA are the same as for MANOVA:

First, create cross-products matrices for between-group differences and within-groups differences, $SS_{total} = SS_{bg} + SS_{wg}$. The determinants are calculated for these matrices and used to calculate a test statistic – either Wilks' Lambda or Pillai's Trace.

Wilks' Lambda follows the equation:

$$\Lambda = \left| \frac{S_{wg}}{S_{bg} + S_{wg}} \right|$$

Next an *F* ratio is calculated as in MANOVA:

$$F_{approximate}(df_1, df_2) = \left(\frac{1 - y}{y} \right) \left(\frac{df_2}{df_1} \right)$$

For cases where n is equal in all groups:

$$y = \Lambda^{1/s} \quad p = \#\ \text{of predictor variables}$$

$$s = \sqrt{\frac{p^2(df_{effect})^2 - 4}{p^2 + (df_{effect})^2 - 5}}$$

df_{error} = number of groups times (n-1): k(n - 1)

$$df_1 = p(df_{effect})$$

$$df_2 = s\left[(df_{error}) - \frac{p - df_{effect} + 1}{2}\right] - \left[\frac{p(df_{effect}) - 2}{2}\right]$$

df_{effect} = number of groups minus one (k - 1)

For unequal n between groups, this is modified only by changing the df_{error} to equal the number of data points in all groups minus the number of groups (N – k). If the experimental *F* exceeds a critical *F*, then the experimental groups can be distinguished based on the predictor variables. The number of discriminant functions used in the analysis is equal to the number of predictor variables or the degrees of freedom, whichever is smaller.

The discriminant function score for the i th function is:

$$D_i = d_{i1}Z_1 + d_{i2}Z_2 + \ldots + d_{ip}Z_p$$

Where z = the score on each predictor, and d_i= discriminant function coefficient. The discriminant function score for a case can be produced with raw scores and unstandardized discriminant function scores. The discriminant function coefficients are, by definition, chosen to maximize differences between groups. The mean over all the discriminant function coefficients is zero, with a SD equal to one.

The mean discriminant function coefficient can be calculated for each group – these group means are called Centroids, which are created in the reduced space created by the discriminant function reduced from the initial predictor variables. Differences in the location of these centroids show the dimensions along which the groups differ.

Once the discriminant functions are determined groups are differentiated, the utility of these functions can be examined via their ability to correctly classify each data point to their *a priori* groups. Classification functions are derived from the linear discriminant functions to achieve this purpose. Different classification functions are used and equations exist that are best suited for equal or unequal samples in each group.

For cases with an equal sample size for each group the classification function coefficient (C_j) is equal to the sum of:

$$C_j = c_{j0} + c_{j1}x_1 + c_{j2}x_2 + \ldots + c_{jp}x_p$$

for the jth group, j = 1...k, x = raw scores of each predictor, c_{jo} = a constant. If W = within-group variance-covariance matirix, and M = column matrix of means for group j, then the constant $c_{jo} = (-1/2)C_jM_j$.

For unequal sample size in each group:

$$C_j = c_{jo} + \sum_{i=1}^{p} c_{ij}x_i + \ln\left(n_j / N\right)$$

n_j = size in group j, N = total sample size.

This page was last updated on