The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation

Discriminant Function Analysis

Contributors: David W. Stockburger
Edited by: Bruce B. Frey
Book Title: The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation
Chapter Title: "Discriminant Function Analysis"
Pub. Date: 2018
Access Date: March 2, 2018
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks,
Print ISBN: 9781506326153
Online ISBN: 9781506326139
DOI: http://dx.doi.org/10.4135/9781506326139.n204
Print pages: 526-532

Discriminant function analysis is used to predict group membership based on a linear combination of interval predictor variables. The procedure begins with a set of observations, whereby both group membership and the values of the predictor variables are known, with the end result being a linear combination of the interval variables that allows prediction of group membership. The way in which the interval variables combine allows a greater understanding and simplification of a multivariate data set. Discriminant analysis, based on matrix theory, is an established technology that has the advantage of a clearly defined decision-making process. Machine learning techniques such as neural networks may be used alternatively for predicting group membership from similar data, often with more accurate predictions, as long as the statistician is willing to accept decision-making without much insight into the process.

For example, a researcher might have a large data set of information from a high school about its former students. Each student belongs to a single group: (a) did not graduate from high school, (b) graduated from high school or obtained a General Educational Development, and (c) attended. The researcher wishes to predict student outcome group using interval predictor variables such as grade point average, attendance, degree of participation in various extracurricular activities (e.g., band, athletics), weekly amount of screen time, and parental educational level. Given this complex multivariate data set and the discriminant function analysis procedure, the researcher can find a subset of variables that in a linear combination allows prediction of group membership. As a bonus, the relative importance of each variable in this subset is part of the output. Often researchers are satisfied with this understanding of the data set and stop at this point.

Discriminant function analysis is a sibling to multivariate analysis of variance as both share the same canonical analysis parent. Where multivariate analysis of variance received the classical hypothesis testing gene, discriminant function analysis often contains the Bayesian probability gene, but in many other respects, they are almost identical.

This entry explains the procedure by breaking it down into its component parts and then assembling them into a whole. The two main component parts in discriminant function analysis are implicit in the title: discriminating between groups and functional analysis. Because knowledge of how to discriminate between groups is necessary for an understanding of the later functional analysis, it is presented first.

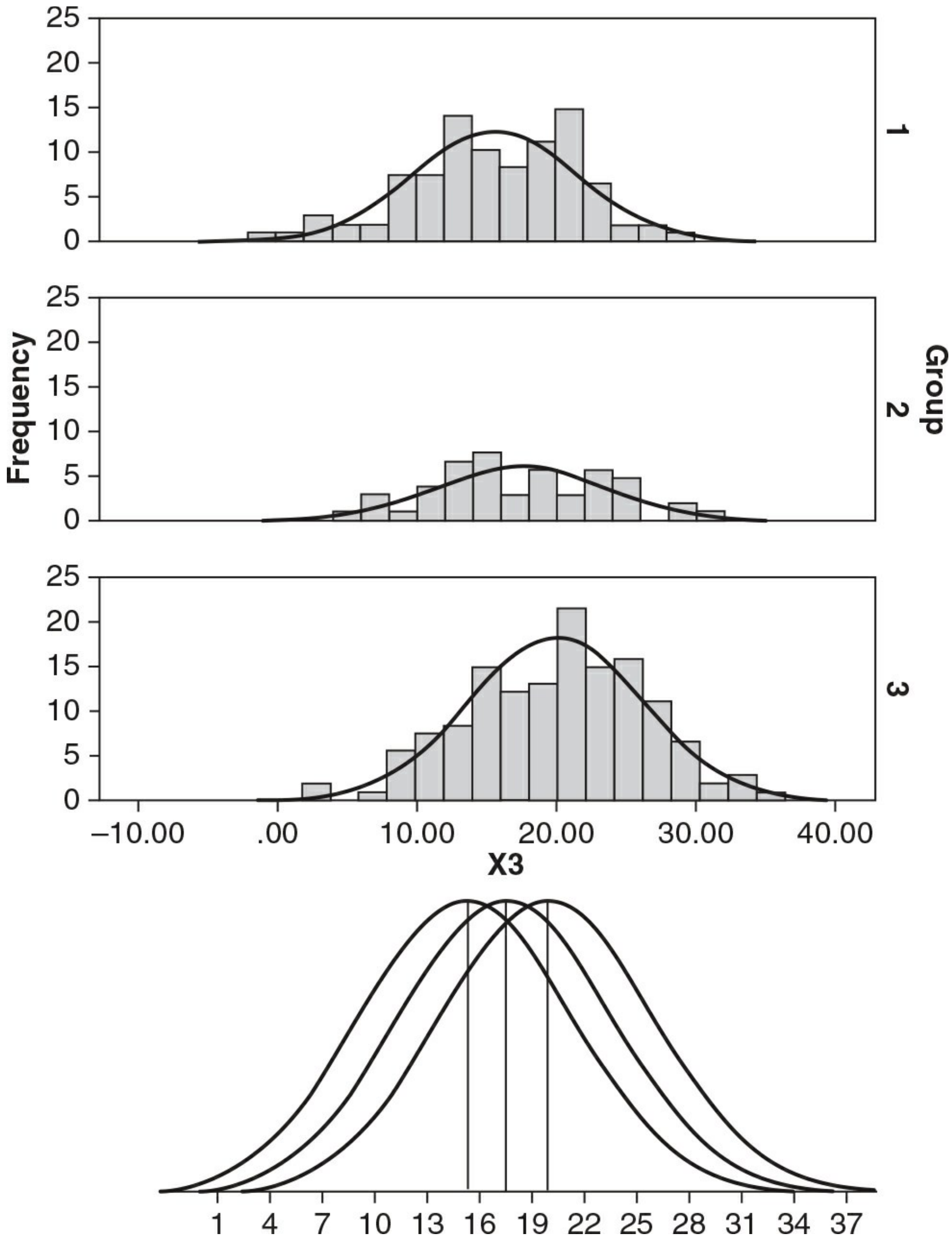
Discriminating Between Groups

Discriminating Between Groups With a Single Variable

The simplest case of discriminant function analysis is the prediction of group membership based on a single variable. An example might be the prediction of successful completion of high school based on the attendance record alone. For the rest of this section, the example uses three simulated groups with *N*s equal to 100, 50, and 150, respectively.

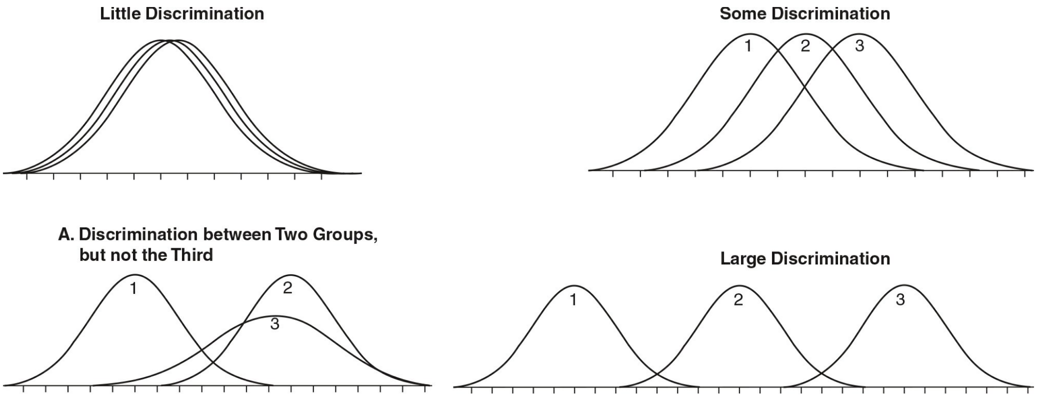
In the example (Figure 1), histograms are drawn separately for each of the three groups. Second, overlapping normal curve models are shown where the normal curve parameters μ and σ are estimated by the mean and standard deviation of the three groups. An analysis of variance shows that the three means are statistically different from each other, but only limited discrimination between groups is possible.

Figure 1 Modeling group membership



[Figure 2](#) shows various possibilities for overlapping group probability models, from little or no discrimination to almost perfect discrimination between groups. Note that the greater the difference between means relative to the within-group variability, the better the discrimination between groups.

Figure 2 Varieties of group discrimination

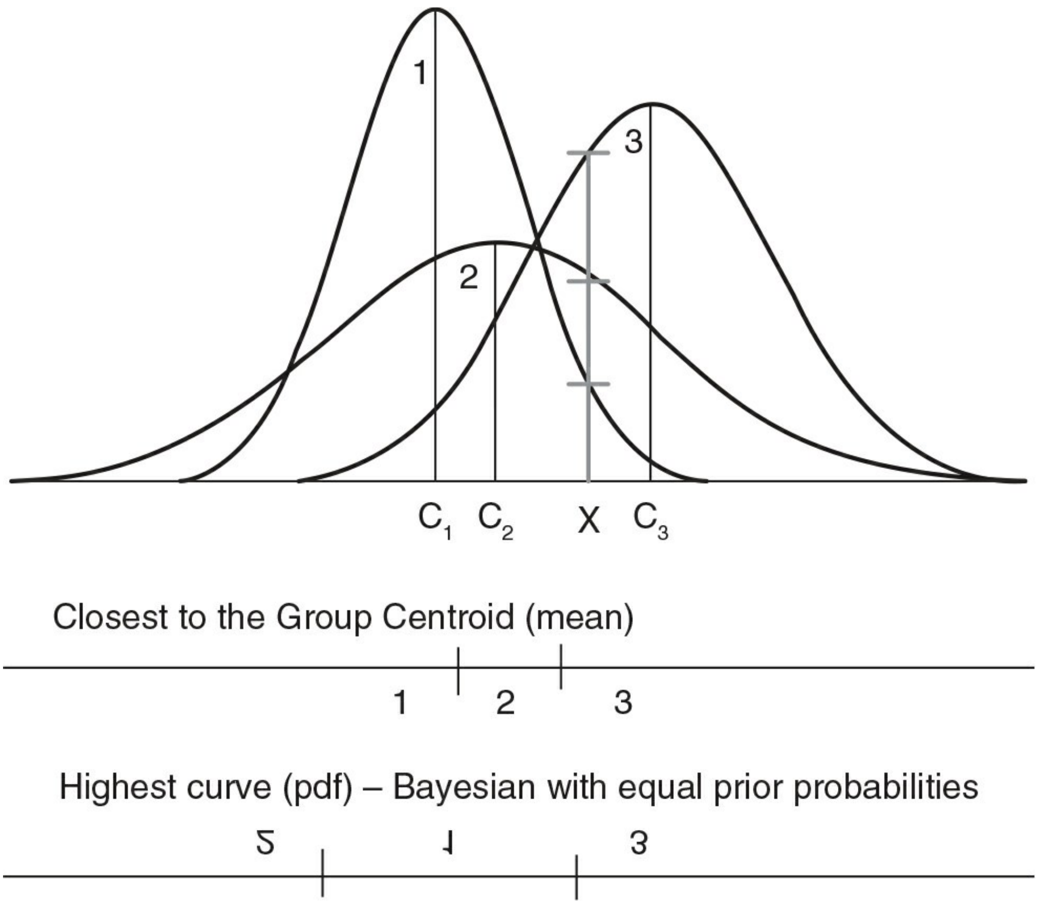


Given that means and standard deviations can be calculated for each group, different classification schemes can be devised to classify scores based on a single variable. One possibility is to simply measure the distance of a particular score from each of the group means and select the group that has the smallest distance. (In discriminant function analysis, group means are called *centroids*.) The advantage of this system is that no distributional assumptions are necessary.

Although not absolutely necessary to perform a discriminant function analysis, Bayes' theorem offers a distinct improvement over distance measures. Bayes' theorem modifies existing probabilities, called *prior probabilities*, into posterior probabilities using evidence based on the collected data. In the case of discriminant function analysis, prior probabilities are the likelihood of belonging to a particular group before the interval variables are known and are generally considered to be subjective probability estimates. Prior probabilities are symbolized as $P(G)$. For example, $P(G_1)$ is the Dyslexia prior probability of belonging to Group 1. In discriminant function analysis software programs (e.g., SPSS), the default option is to set all prior probabilities as equally likely. For example, if there were three groups, each of the three prior probabilities would be set to .33333.… Optionally, the prior probabilities can be set to the relative frequency of each group. In the example data with *N*s of 100, 50, and 150, the prior probabilities would be set to .333…, .16666…, and .75, respectively. Since prior probabilities are subjective, it would also be possible to set them based on cost of misclassification. For example, if misclassification as Group 1 membership is costly, the prior probability might be set to .10 rather than .333.

The probability models of the predictor variables for each group can be used to provide the conditional probability estimates of a score (D) given membership in a particular group, $P(D|G)$. Using the PDF of the probability model, the height of the curve at the data point can be used as an estimate of this probability. Figure 3 illustrates this concept at the data point x, where $P(D = x|G_1) < P(D = x|G_2) < P(D = x|G_3)$.

Figure 3 Classification based on probability models with different territorial maps along a single dimension



Bayes' theorem provides a means to transform prior probabilities into posterior probabilities given the conditional probabilities $P(D|G)$. Posterior probabilities are the probability of belonging to a group given the prior and conditional probabilities. In the case of discriminant function analysis, prior probabilities $P(G)$ are transformed into posterior probabilities of group membership given a particular score $P(G|D)$. The formula for computing $P(G|D)$ using Bayes' theorem is as follows:

$$P\left(G_j|D\right) = \frac{P\left(D|G_j\right)P\left(G_j\right)}{\sum_i^{\text{Groups}} P\left(D|G_i\right)P\left(G_i\right)}.$$

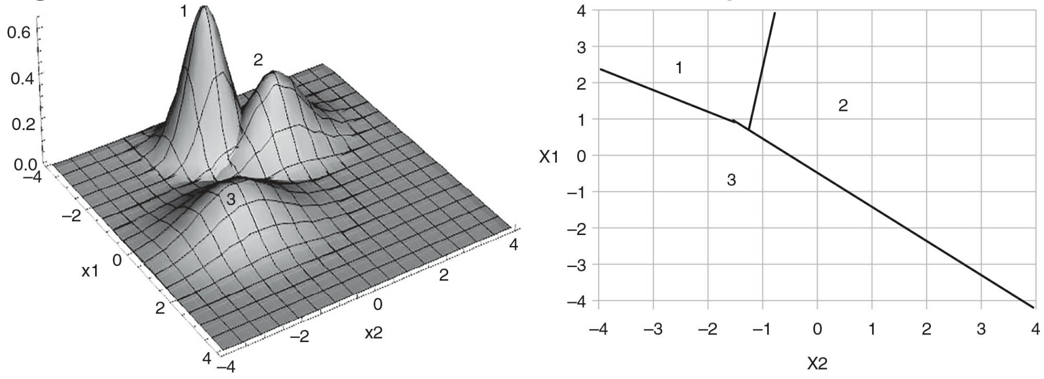
The Bayesian classification system works by computing the posterior probability at a given data point for each group and then selecting the group with the largest posterior probability.

If equal prior probabilities are used, then $P(G_i)$ is constant for all groups and can be canceled from the formula. Since the denominator is the same for all groups, the classification system will select the group with the largest $P(D|G)$. In the case of the normal curve examples of conditional distributions presented in Figure 3, at any given point on the *x* axis, the selected group would correspond to the group with the highest curve. This is reflected on the last territorial map on the figure. Note how different it is from the classification system based on distances from each mean. If unequal prior probabilities are used, then the posterior probabilities are weighted by the prior probabilities and the territorial maps will necessarily change.

Discriminating Between Groups With Multiple Variables

In some cases, especially with multiple groups and complex multivariate data, discrimination between groups along a single dimension is not feasible, and multiple dimensions must be used to ensure reasonably correct classification results. A visual representation of a fairly simple situation with two dimensions and three groups is presented in Figure 4. Note that better classification results can be obtained using two dimensions than any single dimension.

Figure 4 Bivariate normal distribution with territorial map



Conceptually, the classification methods are fairly straightforward extensions of the classification systems along a single dimension, although visual representations become much more problematic, especially in three or more dimensions.

Various methods of computing distances from the group centroids can be used, and the minimum distance can be used as a classification system. The advantage of using distance measures is that no distributional assumptions are necessary.

When using a Bayesian classification system, distributional assumptions are necessary. One common distributional assumption is a multivariate normal distribution. The requirements for a multivariate normal distribution are much more stringent and complex than for a univariate normal distribution and therefore harder to meet. For example, both X1 and X2 could be normally distributed, but the combination might not be a bivariate normal distribution. The multivariate normal assumption becomes even more problematic with many more variables. If the distributional assumptions are acceptable, then the Bayesian classification system proceeds in a manner like discriminating between groups with a single variable. The advantage to using a Bayesian classification system is that posterior probabilities of belonging to each group are available.

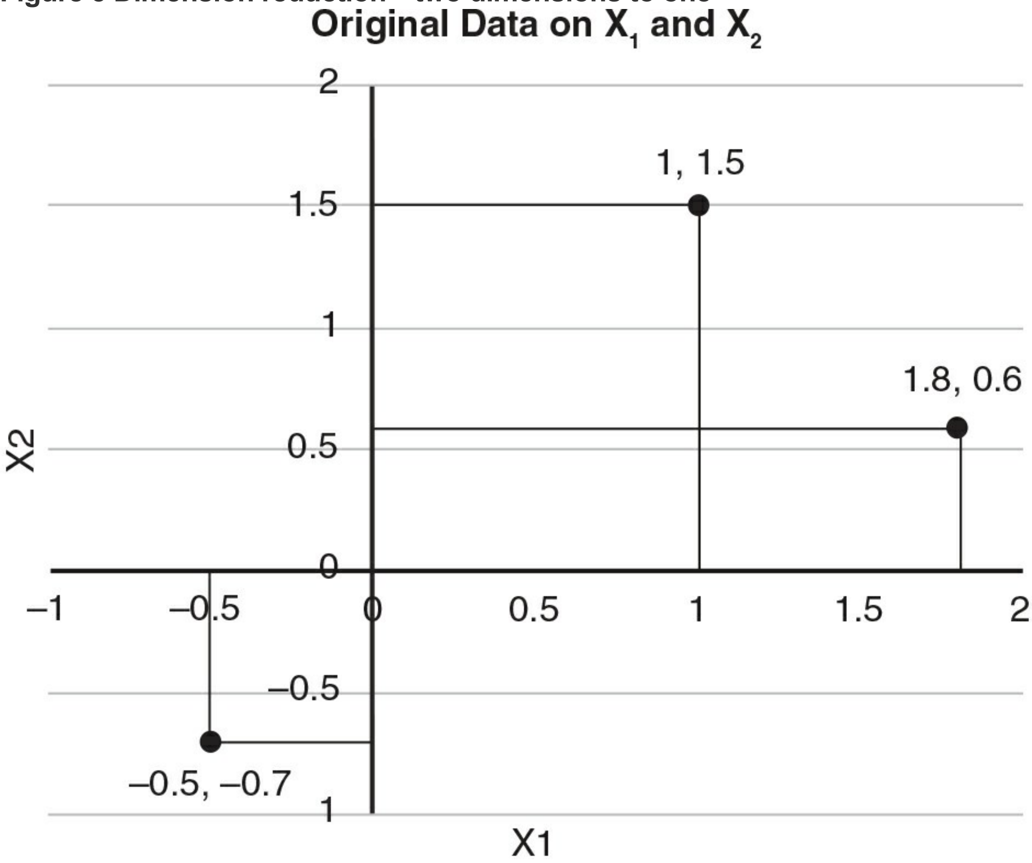
Linear Functions

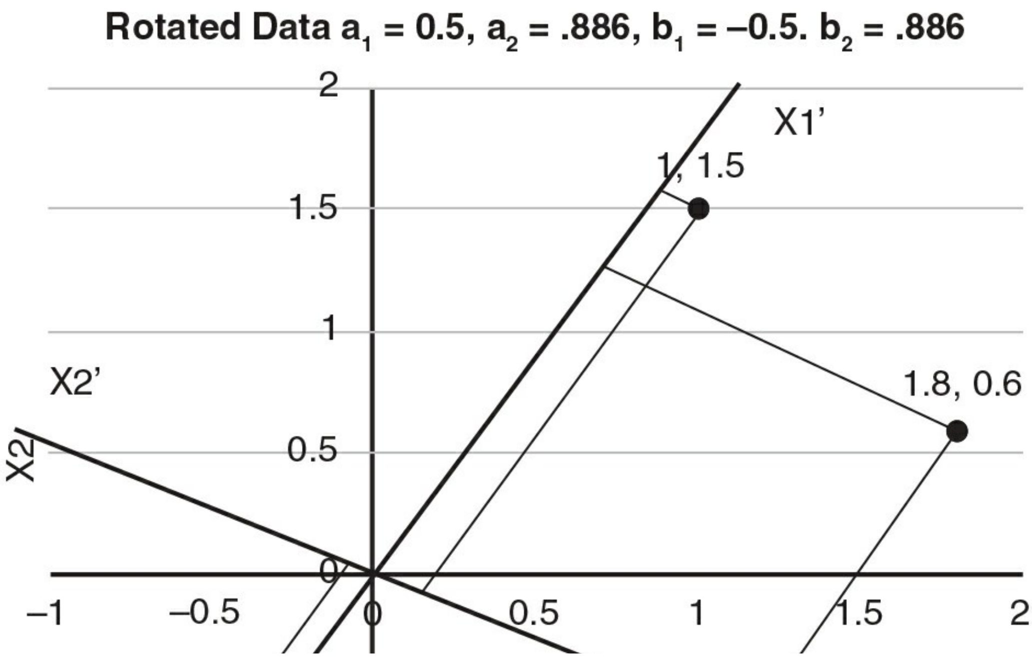
It is only when there are two or more predictor variables that the power of discriminant function analysis becomes apparent. Basically, the procedure discovers linear combinations of the predictor variables that best discriminate between the groups by using matrix operations that are available in canonical analysis. The matrix procedure discovers the linear combination of variables that minimizes the within-group variability and, in the process, maximizes the between-group variability. While a matrix presentation can be beautiful in its apparent simplicity, as some of the additional resources show, what is really occurring beneath the surface can be difficult to fathom if one is not familiar with matrix operations. Thus, this presentation visually focuses on the underlying concepts rather than a mathematically precise formulation.

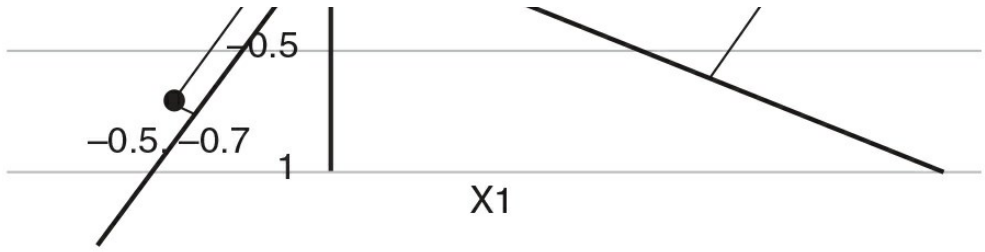
Changing Structure Using Linear Functions

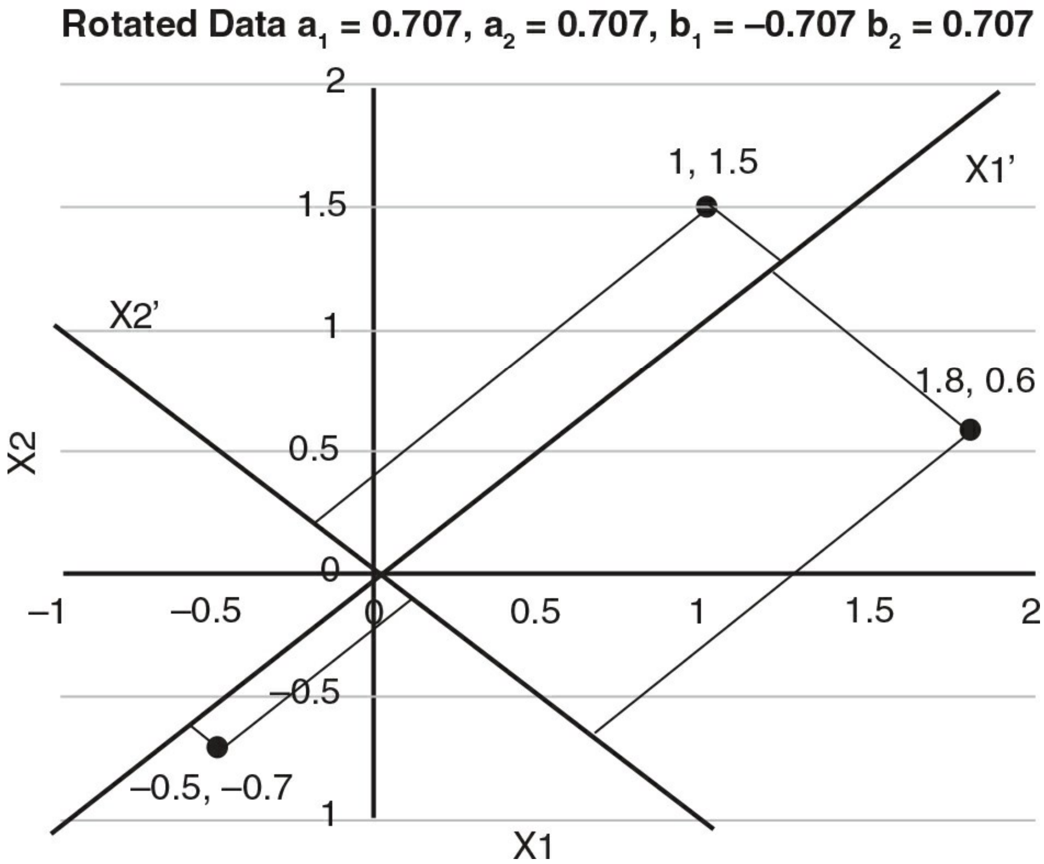
The effect of linear transformations can be observed in Figure 5. Three points, (X1, X2) = (1, 1.5), (1.8, 0.6), and (−0.5, −0.7), are first displayed on their original axis. Note that the population variability of X1 (1.36) and X2 (1.22) as projected onto their respective axes is approximately equal. The sum of the two variances is 2.58.

Figure 5 Dimension reduction—two dimensions to one









These points can be transformed by the following formulas:

$$a_1\mathrm{X1} + a_2\mathrm{X2}$$

$$b_1\mathrm{X1} + b_2\mathrm{X2}$$

$$a_1^2 + a_2^2 = 1,\ \ b_1^2 + b_2^2 = 1,$$

where

$$a_1 = 0.5,\ a_2 = .866,\ b_1 = 0.5,\ b_2 = -.866$$

and projected on the new axes X1′ and X2′ in Figure 5. The three points now become (1.80, 0.12), (−0.86, −0.08), and (1.42, 1.26) using the two transformations. The population variance for X1′ and X2′ is 2.06 and .52, respectively, along the new axes. Note that the sum of the two variances is 2.58, the same as for the original axes. Thus, by using linear transformations with constraints, the variance can be partitioned differently along different axes.

A second linear transformation using:

$$a_1 = 0.707,\ a_2 = .707,\ b_1 = 0.707,\ b_2 = -.707$$

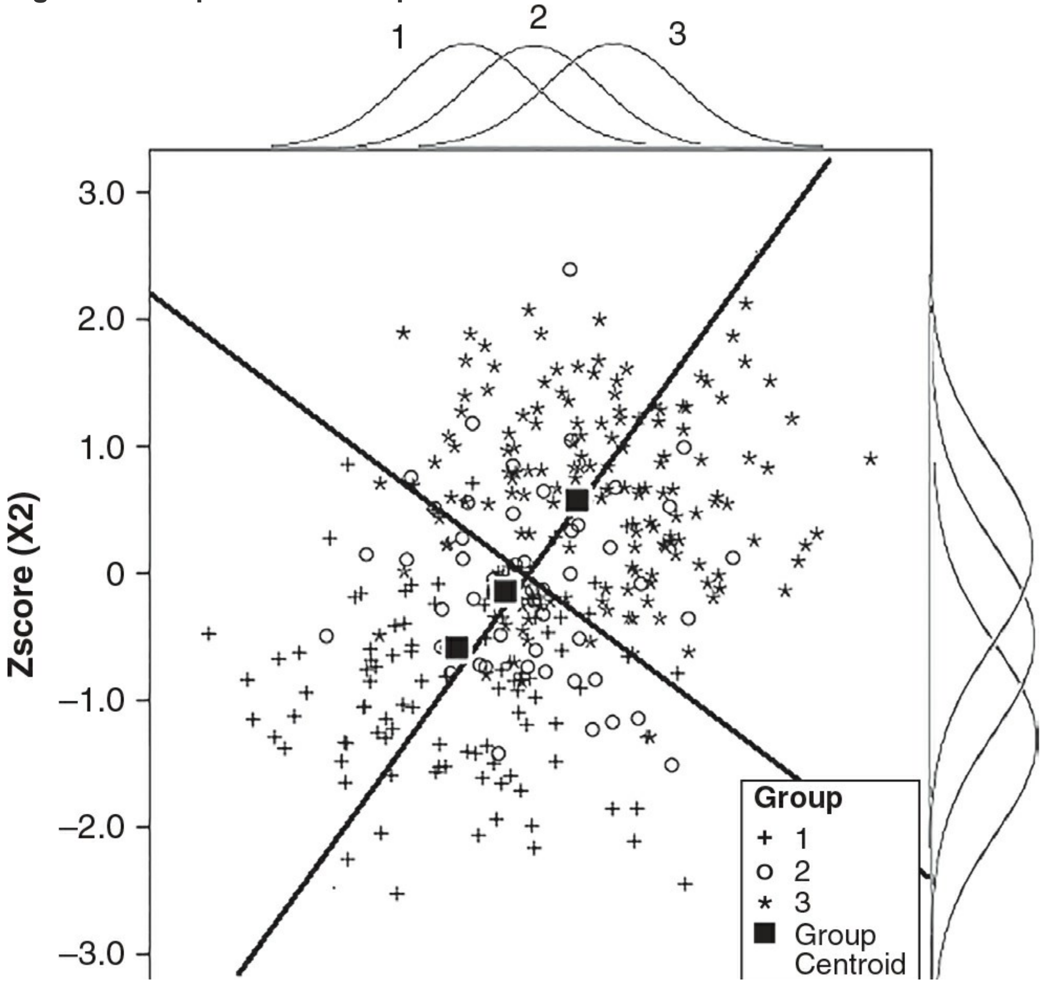
can also be observed in Figure 5. The three points now become (1.77, −0.36), (−0.85, 0.14), and (1.70, 0.85). The population variance for X1′ is 2.22 and for X2′ is .36 for a total of 2.58, the same as the previous axes.

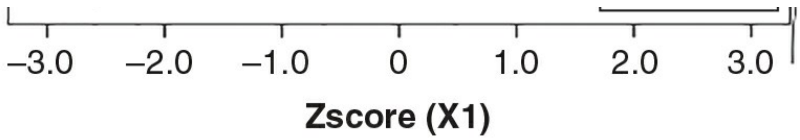
There are values for *a* and *b* such that the variability for X1′ is a maximum and X2′ is a minimum. That is what the matrix operations of discriminant function analysis provide. Basically, it finds an axis (a single dimension) in multidimensional space that maximizes the discrimination between groups. Given the first axis is set, it then finds a second axis (dimension) that maximizes the remaining discrimination between groups. The second axis is orthogonal to the first. The procedure continues until it runs out of groups or variables. At some point, the inclusion of dimensions provides very little additional discriminatory ability and allows the researcher to interpret a much smaller set of variables than the original multivariate data.

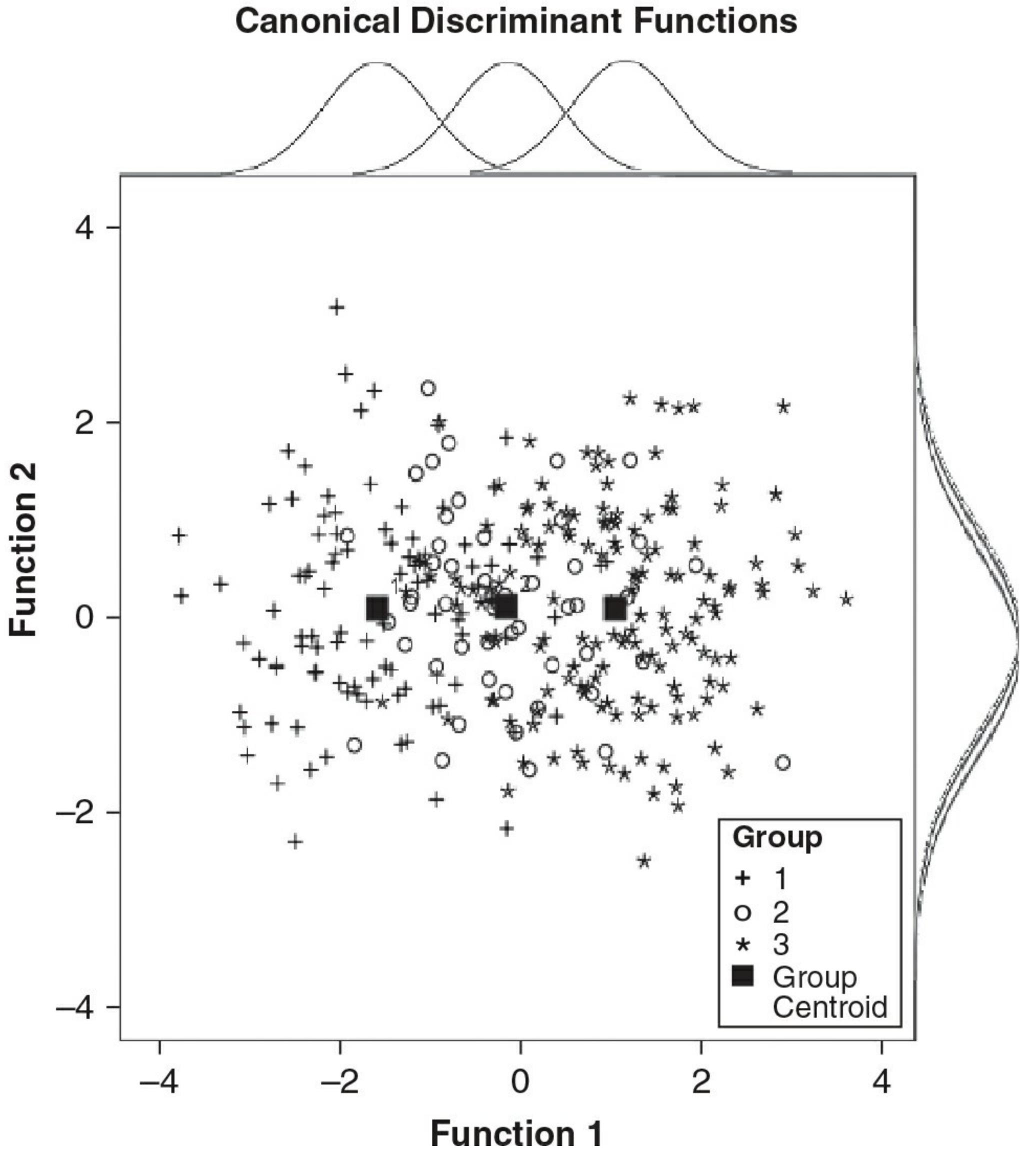
Rotating the Axes to Maximize Discrimination—An Example

An example of the application of discriminant function analysis may be the best manner to illustrate how the procedure works. In this example, there are three groups (1, 2, and 3) and two variables (X1 and X2). Because differential variability of the interval variables can affect the results greatly, the first step in the analysis is to standardize the variables. The scatterplot of the standardized variables for this example appears in Figure 6. The means for the three groups are plotted on the graph and are called group centroids.

Figure 6 Example data scatterplot before and after rotation







From the marginal distributions in Figure 6, it can be seen that individually, both X1 and X2 somewhat discriminate between the three groups, but the distributions have considerable overlap. Although it would be possible to sequentially apply Bayes' theorem using the two variables, discriminant function analysis first finds a linear combination of the two variables that best discriminates between all groups and then generates a second function that contains whatever is left over.

Applying discriminant function analysis to these data, the first decision is how many factors or dimensions are to be included in the analysis. Inferential and model building techniques are typically used to make the decision, but they are beyond the scope of this entry. For this example, a significant Wilks's λ test and squared canonical correlation greater than .10 (Pedhazur, 1973) suggest using a discriminant function analysis that would result in a single factor. The squared canonical correlation for each discriminant function can be interpreted as the proportion of variability that the discriminant function describes, similar to R^2in multiple regression. For the discriminant function analysis on the example data, this would result in a single factor.

Even though the analysis would suggest that only a single factor be analyzed, both factors are presented below for completeness sake. The two discriminant functions from the "Standardized Canonical Discriminant Function Coefficients" table are

$$\text{Factor 1} = 0.572 \times zX1 + 0.836 \times zX3$$
$$\text{Factor 2} = 0.821 \times zX1 - 0.549 \times zX3.$$

The bottom line is that Factor 1 would be computed for all records, and then a classification system would be employed to classify into appropriate groups. Most statistical packages optionally allow these additional discriminant variables to be created. The results of the applied classification system (equal prior and Bayesian decision process) can be seen in the contingency table. Note that the application of discriminant function analysis in the example resulted in a 71% correct classification.

A scatterplot of both discriminant functions is presented in Figure 6. Note the position of group centroids along the Factor 1 axis and the marginal discrimination of the two functions. The discriminant function coefficients are essentially the β weights of each variable for the discriminant function. They describe the relative importance of that variable in constructing the function, although they must be interpreted with caution, as they have similar issues as the interpretation of β weights in multiple regression. It can be seen in the example that X2 (0.836) contributes to the function to a greater extent than X1 (0.572).

To make predictions using the results of discriminant function analysis, the raw scores need to be standardized using the means and standard deviations of the original data set. Following that, the discriminant functions are computed for each record, and then the classification system is applied relative to the conditional distributions.

If there are more than two variables and two groups, the procedure results in additional discriminant functions equal to the lesser of the number of groups minus one or the number of interval variables. For example, when there are three groups and two interval variables, the procedure will produce two discriminant functions. In almost all cases, however, the procedure will reduce the dimensionality of the original data.

As with any multivariate system of analysis, the more the groups and variables, the greater the complexity of analysis. With three groups and three variables, the first discriminant function would be the line through the multidimensional space that minimized the within-group variance. The second line would be perpendicular to the first and would minimize the within-group residual variability from the first discriminant function. The third discriminant function would be a line perpendicular to the first two and again minimize the within-group residual variability from the first two discriminant functions.

Limitations

Discriminant function analysis has been around since its origin in 1936 with two defined groups by R. A. Fisher. It was later extended by others to include more than two groups. Because of the computational difficulty of the analysis, it was not extensively used until computers became widely available. It has the advantage of describing a complex decision process with a few parameters and producing results that can be interpreted.

The linear models of discriminant function analysis are also its main disadvantage, as many relationships in the real world are not linear. The use of programs that can be trained to use multiple “if-then” statements or neural networks that learn complex relationships with large data sets and estimation of thousands of parameters have eclipsed the use of linear models. The accuracy of these types of programs is generally greater than linear models but comes at a cost to the researcher of not understanding the “why” of the decisions.

A second major disadvantage of discriminant function analysis is the reliance on the assumption of multivariate normal distributions for classification. Although classification decisions can be made without reference to this assumption, when it is made, it is almost certain to be incorrect. How robust the system is with respect to this assumption can be checked with use of two data sets, one for training and one for testing.

Discriminant function analysis offers a powerful tool to discriminate between groups based on creating new variables, called discriminant functions, using linear models of existing interval variables. Measures of accuracy of prediction along with the manner in which the variables combine provide the statistician with a means of understanding multivariate data.

See alsoBayesian Statistics; Canonical Correlation; Logistic Regression; Multivariate Analysis of Variance

David W. Stockburger
http://dx.doi.org/10.4135/9781506326139.n204
10.4135/9781506326139.n204

Further Readings

Johnson, R. A., & Wichern, D. W. (1982). Applied multivariate statistical analysis (
3rd ed.
). Upper Saddle River, NJ: Prentice Hall.
Lantz, B. (2013). Machine learning with R. Birmingham, UK: Packt.
Pedhazur, E. J. (1973). Multiple regression in behavioral research explanation and prediction (
3rd ed.
). Fort Worth, TX: Holt, Rinehart and Winston.
Van de Geer, J. P. (1971). Introduction to multivariate analysis for the social sciences. W. H.