Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Discriminant analysis is based on comparing multivariate observations made with different groups of subjects in order to define the distances between the groups, and also to assign new observations to appropriate groups.

From the main SIMFIT menus choose [Statistics], [Multivariate, [Discriminant analysis] then read in the default test file `g03daf.tf1` which has the following data.

1	1.1314	2.4596
1	1.0986	0.2624
1	0.6419	-2.3026
1	1.3350	-3.2189
1	1.4110	0.0953
1	0.6419	-0.9163
2	2.1163	0.0000
2	1.3350	-1.6094
2	1.3610	-0.5108
2	2.0541	0.1823
2	2.2083	-0.5108
2	2.7344	1.2809
2	2.0412	0.4700
2	1.8718	-0.9163
2	1.7405	-0.9163
2	2.6101	0.4700
3	2.3224	1.8563
3	2.2192	2.0669
3	2.2618	1.1314
3	3.9853	0.9163
3	2.7600	2.0281

This data set has three groups, indicated by the nondecreasing integers in columns 1, for three types of Cushing's syndrome, the variables in columns 2 and 3 are logarithms of urinary excretion rates (mg/hr) for two steroid metabolites.

The following options are then available.

- Calculate the group sample means and the pooled sample means.
  The numerical values for the vectors of means can be displayed.
- Test for equality of the vectors of population means.
  If required, this can be done using the MANOVA options provided by SIMFIT.
- Test if all population variance-covariance matrices are equal.
  The results from discriminant analysis will differ depending on whether it is assumed that the variables all have the same population covariance matrix (as estimated from the pooled samples) or different covariance matrices (as estimated from the group samples).
- Calculate distances between the groups.
  The Mahalanobis distance matrix D^2_{ij} can be calculated assuming equal or unequal variance-covariance matrices.
- Plot the groups.
  The centroids can also be plotted to indicate the center of gravity of the groups while, for cases with more than two variables, the principal components can be plotted instead.

The results from such a systematic investigation are now presented.

First of all here are the group and pooled means followed by a MANOVA test for equality of means.

Table 1. Mean vectors

Group 1	1.0433	-0.6034
Group 2	2.0073	-0.2060
Group 3	2.7097	1.5998
Pooled	1.8991	0.1104

Table 2. MANOVA test for H_0: population mean vectors are equal

Number of groups	3				
Number of variables	2				
Number of observations	21				
Statistic	Value	Transform	$NDOF$	p	
Wilks lambda	0.3144	6.660	4, 34	0.0005	*Reject H_0 at 1% significance level*
Roys largest root	1.801				
Lawley-Hotelling T	1.937	8.231	4, 17	0.0006	*Reject H_0 at 1% significance level*
Pillais trace	0.7625				

A mean vector for a group is simply the vector consisting of sample means for each variable within that group, and the results suggest that the population mean vectors for these three groups are not the same, so that regarding the subjects as forming three distinct groups seems to be justified in this case.

The results in this next table for testing if the population variance-covariance matrices for the groups are identical suggests that we should consider rejecting the null hypothesis.

Table 3. Testing H_0: population variance-covariance matrices are equal

Number of groups	3	
Number of observations	21	
Number of variables	2	
Test statistic C	19.24	
Degrees of freedom	6	
$P(\chi^2 \geq C)$	0.0038	*Reject H_0 at 1% significance level*

Note that, in the following values for the Mahalanobis distances D_i^2 between groups, assuming a common population variance-covariance matrix leads to a symmetric distance matrix, but for unequal variance-covariance matrices the distance matrix is not symmetric.

Table 4. Mahalanobis distances

D_{ij}^2 assuming equal CV

0.00000	3.58476	11.7998
3.58476	0.00000	3.25922
11.7998	3.25922	0.00000

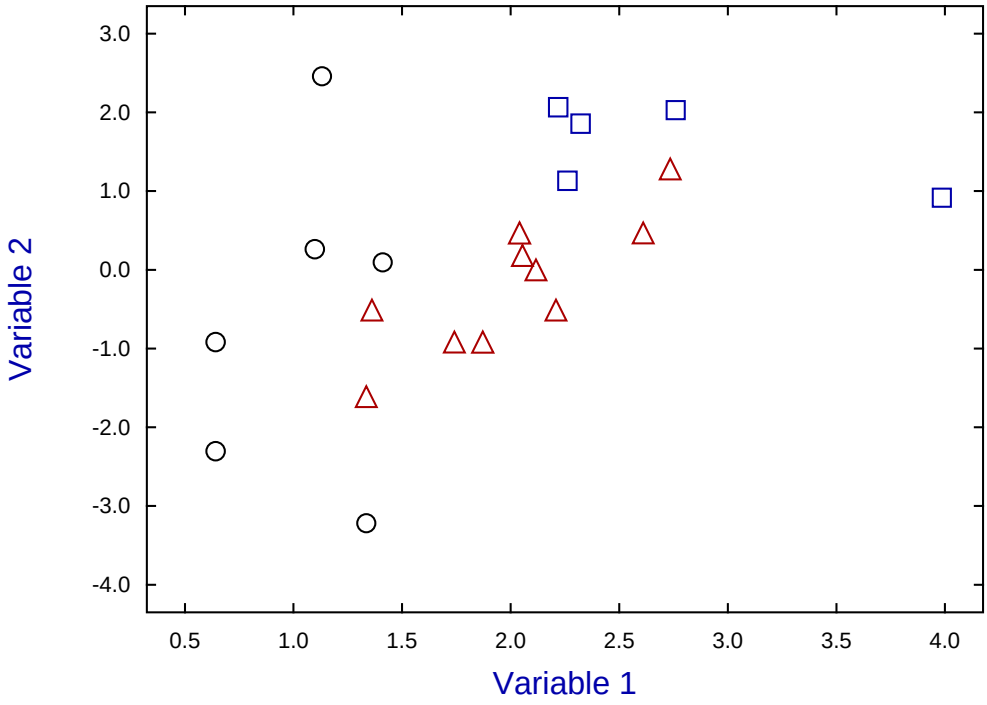
D_{ij}^2 assuming unequal CV

0.00000	9.55703	51.9737
8.51398	0.00000	25.2973
25.1215	4.71142	0.00000

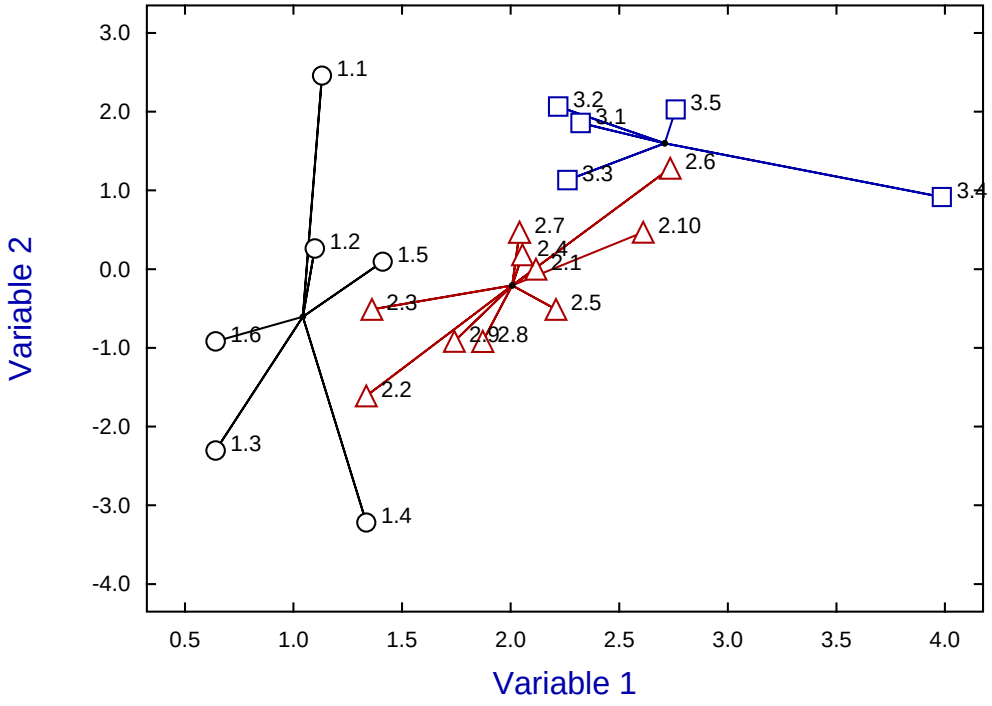
Finally, the next figure displays the observations for the groups followed by the same data but with centroids added together with spokes to emphasize the groups.

In the present case there are only two variables so these can be used as axes but, for more than two variables, the option to plot principal components should be used

Data for Three Groups



Data for Three Groups with Labels and Centroids



Theory

Defining the mean vector

Consider a group of X of size n with m variables as in

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

for then the vector of column means

$$\bar{X} = \begin{bmatrix} \bar{x}_1, & \bar{x}_2, & \cdots & \bar{x}_m \end{bmatrix}^T$$

where $\bar{x}_j$ is the mean of column j is generally referred to as the mean vector as in Table 1 and Table 2. Alternatively, regarding the points as having unit mass, this is would be the center of mass or centroid of the data regarded as a multivariate swarm. For two mean vectors to be equal requires all corresponding component means to be equal.

Testing for equality of covariance matrices

The results from analyzing `g03daf.tf1` in Table 3 refer to using Box's test to analyze for equality of population covariance matrices. This depends on n the overall sample size, m the number of variables, g the number of groups, n_i the sample size in group i, S the pooled variance-covariance matrix with determinant $|S|$, S_i the within-group variance-covariance matrices with determinants $|S_i|$, and the likelihood ratio test statistic C defined by

$$C = M \left\{ (n-g) \log |S| - \sum_{i=1}^{g} (n_i - 1) \log |S_i| \right\}.$$

Here the multiplying factor M is

$$M = 1 - \frac{2m^2 + 3m - 1}{6(m+1)(g-1)} \left(\sum_{i=1}^{g} \frac{1}{n_i - 1} - \frac{1}{n-g} \right)$$

and, for large n, C is approximately distributed as χ^2 with $m(m+1)(g-1)/2$ degrees of freedom. Just as tests for equality of variances are not very robust, this test should be used with caution, and then only with large samples, i.e. $n_i >> m$.

The squared Mahalanobis distance between two groups

The squared Mahalanobis distance D_{ij}^2 between two group means $\bar{x}_i$ and $\bar{x}_j$ referred to in Table 4 can be defined as either

$$D_{ij}^2 = (\bar{x}_i - \bar{x}_j)^T S^{-1} (\bar{x}_i - \bar{x}_j)$$
$$\text{or } D_{ij}^2 = (\bar{x}_i - \bar{x}_j)^T S_j^{-1} (\bar{x}_i - \bar{x}_j)$$

depending on whether the covariance matrices are assumed to be equal, when the pooled estimate S is used and $D_{ij}^2 = D_{ji}^2$, or unequal when the group estimate S_j is used and $D_{ij}^2 \neq D_{ji}^2$. This distance is a useful quantitative measure of similarity between groups, but often there will be extra measurements which can then be appended to the data file, as with `g03daf.tf1`, so that the distance between measurement k and group j can be calculated as either

$$D_{kj}^2 = (x_k - \bar{x}_j)^T S^{-1} (x_k - \bar{x}_j)$$
$$\text{or } D_{kj}^2 = (x_k - \bar{x}_j)^T S_j^{-1} (x_k - \bar{x}_j).$$