Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.

https://simfit.org.uk
https://simfit.silverfrost.com

Canonical variates is a technique used to transform multivariate data into new coordinates in order to highlight differences and similarities between groups of observations.

If MANOVA investigation suggests that at least one group mean vector differs from the the rest, it is usual to proceed to canonical variates analysis, although this technique can be also be used for data exploration when the assumption of multivariate normality with equal covariance matrices is not justified.

Example 1

From the main SIMFIT menu choose [Statistics], [Multivariate], then [Canonical variates], and read in the test file `manova1.tf4` from the `C:\Program Files\simfit\dem` folder, which contains the following data for three groups of three subjects, each with observations on three variables.

```
1   13.3   10.6   21.2
1   13.4    9.4   21.0
1   12.9   10.0   20.5
2   13.6   10.2   21.0
2   13.2    9.6   20.1
2   12.2    9.9   20.7
3   14.2   10.7   21.1
3   13.9   10.4   19.8
3   13.9   11.0   19.1
begin{values}
    14.0   11.0   22.0
    12.0    9.0   19.0
    13.0    9.0   20.0
end{values}
```

Column 1 has the group numbers as nondecreasing integers, while columns 2, 3, and 4 are observations on three variables. Because canonical variates are also used to see how closely additional unassigned observations compare to the defined groups of the training set, these can be added as additional values to the end of the data file as shown.

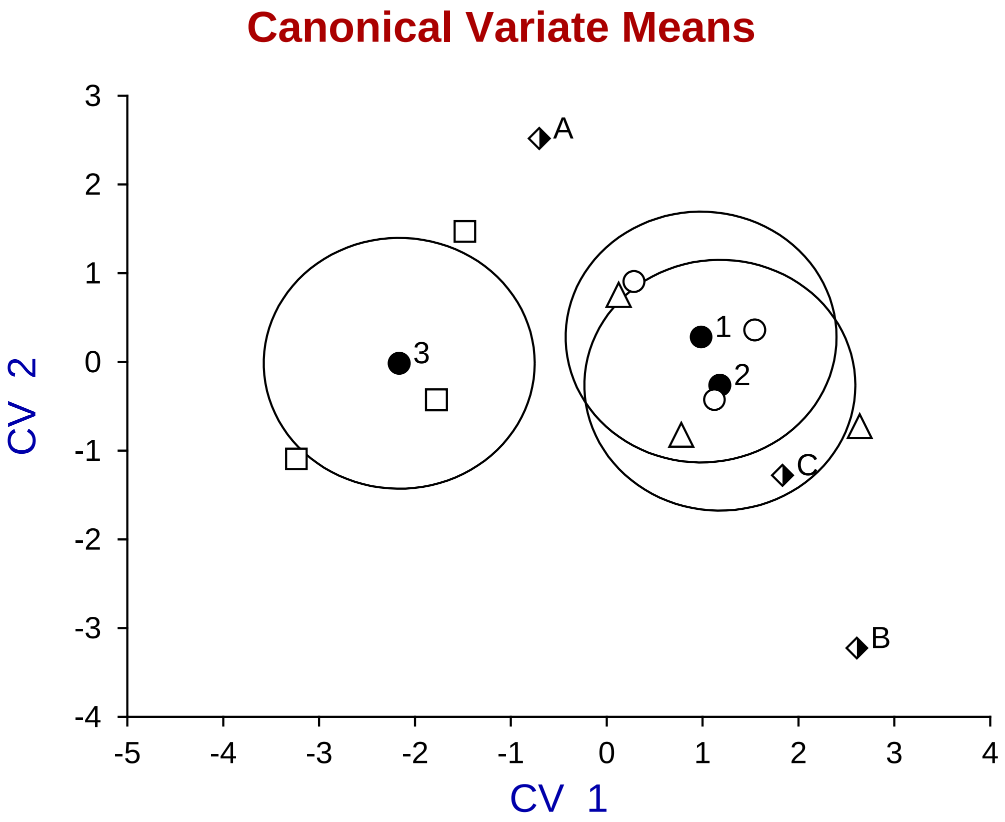
The table below shows the results from analyzing these data, and this is followed by a summary that will be explained in more detail later.

```
Results from analysis of manova.tf4: rank = 3
Correlations   Eigenvalues   Proportions    χ2       NDOF      p
  0.8826         3.5238        0.9795     7.9032      6      0.2453
  0.2623         0.0739        0.0205     0.3564      2      0.8368
Canonical variate means
  0.9841         0.2797
   1.181        -0.2632
  -2.165        -0.01642
Canonical coefficients
  -1.707         0.7277
  -1.348         0.3138
  0.9327          1.220
```

- The number of correlations is the larger of the rank and the number of groups less one.

- The eigenvalues are for the within group sum of squares matrix, and these are used to estimate the proportion of variation explained by the canonical variates.
- The chi-square statistic is used to decide the number of canonical variates required to represent the data.
- The degrees of freedom and a p value for the significance of this chi-square statistic are presented.

Perhaps the most useful application of this technique is to plot the group means together with the data and 95% confidence regions in canonical variate space in order to visualize how close or how far apart the groups are. This is done for the first two canonical variates in the next figure.



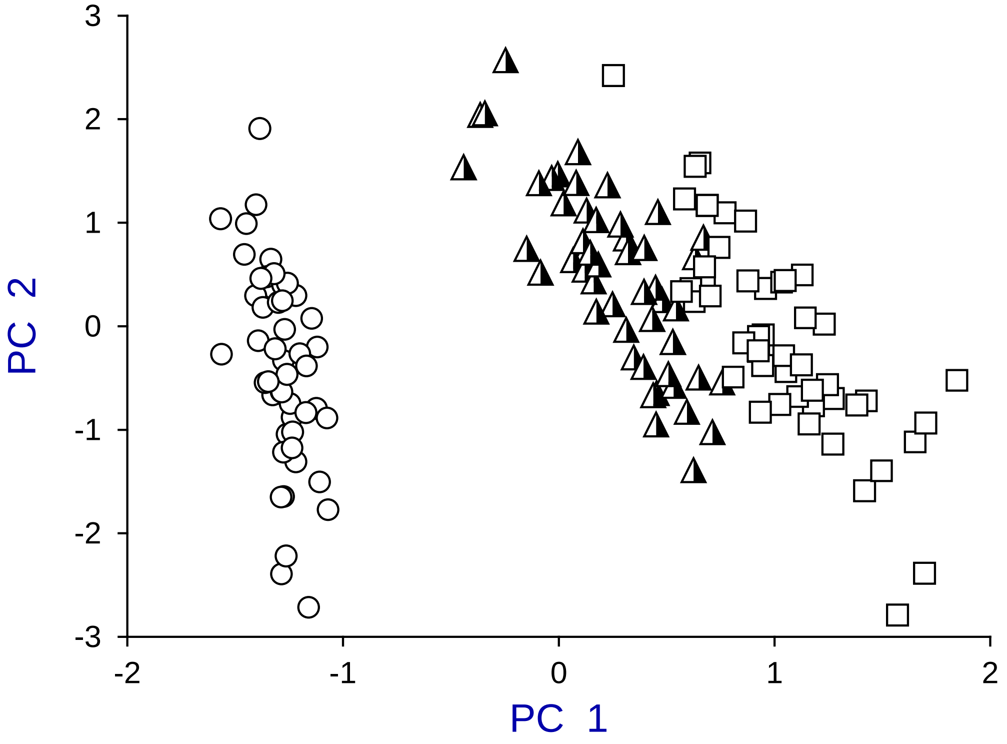
Using the option to edit plot parameters the above figure was constructed to show the labeled canonical variate means (filled circles) together with 95% confidence confidence regions, along with the original data and the three additional observations. Finally the ranges of data plotted were adjusted in order to display the confidence ranges as circles instead of ellipses.

From this graph it is evident that groups 1 and 2 (circles and triangles) are similar but both groups are distinct from group 3 (squares). Additional observation C (half filled diamonds) can be assigned the groups 1 and 2 but additional observations A and B do not belong to any of the training sets.

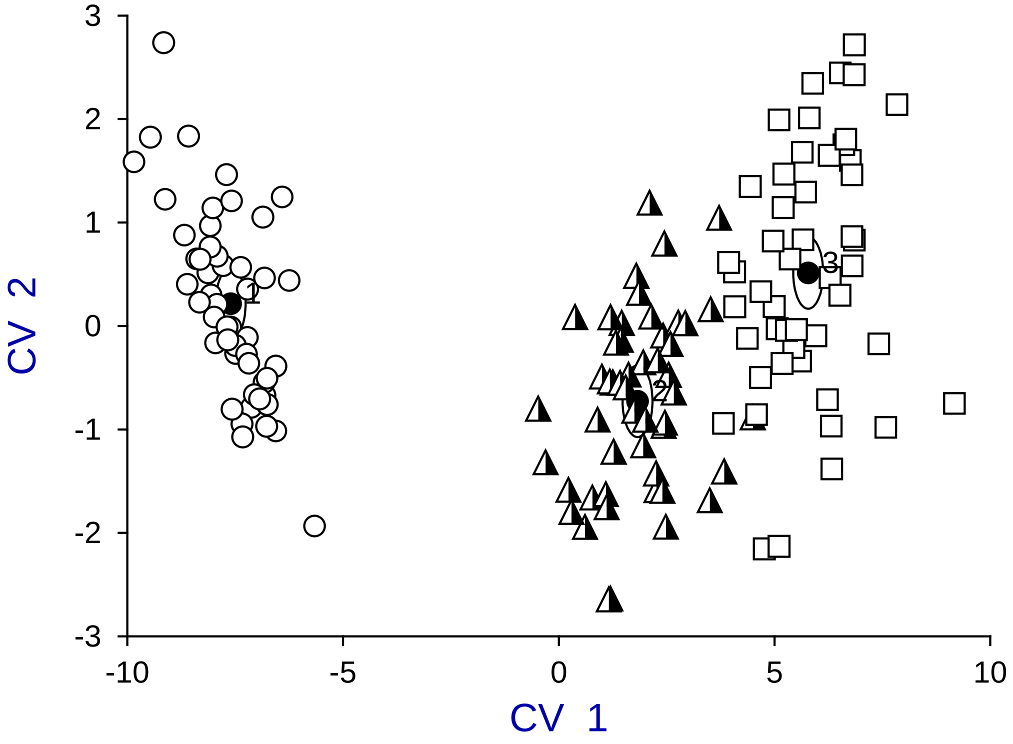
Example 2

To appreciate the use of canonical variates to distinguish groups with a larger data set, consider the next figures, which illustrate the famous Fisher Iris data set contained in `manova1.tf5` using the first two principal components, and also the first two canonical variates for comparison.

Principal Components for Iris Data



Canonical Variates for Iris Data



Theory

First of all, note that canonical variates, unlike principal components, are not simply obtained by a distance preserving rotation: the transformation is non-orthogonal and best represents the Mahalanobis distance between groups.

The confidence range

In the first figure of Example 1 we see the group means identified by the filled symbols labeled as 1, 2 and 3, each surrounded by a 95% confidence region, which in this case is circular as equally scaled physical distances are plotted along the axes. The canonical variates are uncorrelated and have unit variance so, assuming normality, the $100(1-\alpha)\%$ confidence region for the population mean is a circle radius

$$r = \sqrt{\chi^2_{\alpha,2}/n_i},$$

where group i has n_i observations and $\chi^2_{\alpha,2}$ is the value exceeded by $100\alpha\%$ of a chi-square distribution with 2 degrees of freedom.

Note that, alternatively, a circle radius $\sqrt{\chi^2_{\alpha,2}}$ can be plotted as this defines a tolerance region, i.e. the region within which $100(1-\alpha)\%$ of the whole population is expected to lie.

The additional observations

Also, the test file `manova1.tf4` has three other observations appended which are to be compared with the main groups in order to assign group membership, that is, to see to which of the main groups 1, 2 and 3 the extra observations should be assigned. The half-filled diamonds representing these are identified by the labels A, B and C which, like the identifying numbers 1, 2, and 3, are plotted automatically by SIMFIT to identify group means and extra data. In this case, as the data sets are small, the transformed observations from groups 1, 2 and 3 are also shown as circles, triangles and squares respectively, which is easily done by saving the coordinates from the plotted transforms of the observations in ASCII text files which are then added interactively as extra data files to the means plot.

The calculation of canonical variates

The aim of canonical variate analysis is to find the transformations a_i that maximize F_i, the ratios of B (the between group sum of squares and products matrices) to W (the within-group sum of squares and products matrix), i.e.

$$F_i = \frac{a_i^T B a_i/(g-1)}{a_i^T W a_i/(n-g)}$$

where there are g groups and n observations with m covariates each, so that $i = 1, 2, \ldots, l$ where l is the lesser of the number of groups minus one and the rank of the data matrix. The canonical variates are obtained by solving the symmetric eigenvalue problem

$$(B - \lambda^2 W)x = 0,$$

where the eigenvalues λ_i^2 define the ratios F_i, and the eigenvectors a_i corresponding to the λ_i^2 define the transformations. So, just as with principal components, a scree diagram of the eigenvalues in decreasing order indicates the proportion of the ratio of between-group to within-group variance captured by the canonical variates.

Note that the previous results table lists the rank k of the data matrix, the number of canonical variates $l = \min(k, g-1)$, the eigenvalues λ_i^2, the canonical correlations $\lambda_i^2/(1+\lambda_i^2)$, the proportions $\lambda_i^2/\sum_{j=1}^{l}\lambda_j^2$, the group means, the loadings, and the results of a chi-square test.

The number of canonical variates

It is important to realize that the first two canonical variates may be insufficient to represent the data adequately. A scree diagram can be plotted to estimate the minimum number required, or the eigenvalues, proportions, or chi-square statistics calculated from the data can be used.

For instance. If the data are assumed to be from a common multivariate distribution, then to test for a significant dimensionality greater than some level i, the statistic

$$\chi^2 = (n - 1 - g - (k - g)/2) \sum_{j=i+1}^{l} \log(1 + \lambda_j^2)$$

has an asymptotic chi-square distribution with $(k-i)(g-1-i)$ degrees of freedom. If the test is not significant for some level h, then the remaining tests for $i > h$ should be ignored. It should be noted that the group means and loadings are calculated for data after column centering and the canonical variates have within group variance equal to unity. Also, if the covariance matrices $\beta = B/(g-1)$ and $\omega = W/(n-g)$ are used, then $\omega^{-1}\beta = (n-g)W^{-1}B/(g-1)$, so eigenvectors of $W^{-1}B$ are the same as those of $\omega^{-1}\beta$, but eigenvalues of $W^{-1}B$ are $(g-1)/(n-g)$ times the corresponding eigenvalues of $\omega^{-1}\beta$.

In the iris plot of Example 2 there are only two canonical variates, so the canonical variates diagram is fully representative of the data set, and both techniques illustrate the distinct separation of group 1 (circles = setosa) from groups 2 (triangles = versicolor) and 3 (squares = virginica), and the lesser separation between groups 2 and 3.

Users of these techniques should always remember that, as eigenvectors are only defined up to an arbitrary scalar multiple and different matrices may be used in the principal component calculation, principal components and canonical variates may have to be reversed in sign and re-scaled to be consistent with calculations reported using software other than SimFIT. To see how to compare extra data to groups involved in the calculations, the test file `manova1.tf4` should be examined.