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

Canonical correlation is used to explore the correlations between selected columns of a matrix by calculating transformations into lower-dimensional subspaces where the transformed variables have maximum correlation, and can thus be quantified and visualized

Consider a n by m matrix A with elements a_{ij} as follows

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

where a subset of n_x columns (i.e. x-variables) will be defined as X, another disjoint subset of n_y columns (i.e. y-variables) will be defined as Y, while n_s columns may be suppressed (i.e. not used in the analysis). Clearly

$$m = n_x + n_y + n_s \text{ where } n_x \geq 1, n_y \geq 1 \text{ and } n_s \geq 0.$$

Example 1

From the main SimFIT menu choose [Statistics], [Multivariate], then [Canonical correlation] and observe the format for the test file `g03adf.tf1` shown below.

```
80.0    58.4    14.0    21.0
75.0    59.2    15.0    27.0
78.0    60.3    15.0    27.0
75.0    57.4    13.0    22.0
79.0    59.5    14.0    26.0
78.0    58.1    14.5    26.0
75.0    58.0    12.5    23.0
64.0    55.5    11.0    22.0
80.0    59.2    12.5    22.0
begin{indicators}
 -1      1       1      -1
end{indicators}
```

The final section after the data matrix specifies the meaning of the above data as follows.

- Column 1: variable 1 ($y(1)$ in this case as indicator(1) = -1)
- Column 2: variable 2 ($x(1)$ in this case as indicator(2) = 1)
- Column 3: variable 3 ($x(2)$ in this case as indicator(3) = 1)
- Column 4: variable 4 ($y(2)$ in this case as indicator(4) = -1)

In other words, the red data values are Y variables while the blue values are X variables. Note that, in this example, there are no variables to be suppressed by setting the corresponding indicator to zero, but in any case the assignment of columns to types X or Y or suppressed can also be done interactively. Analysis leads to the next table of results.

Results from analysis of data in test file `g03adf.tf1`
Variables: $yxxy$
Number of X variables = 2, Number of Y variables = 2, Number unused = 0
Minimum of rank of X and rank of Y = 2

Correlations	Eigenvalues	Proportions	χ^2	$NDOF$	p
0.9570	0.91591	0.8746	14.391	4	0.0061
0.3624	0.13133	0.1254	0.77438	1	0.3789

CVX: Canonical coefficients for centralized X

-0.4261	1.034
-0.3444	-1.114

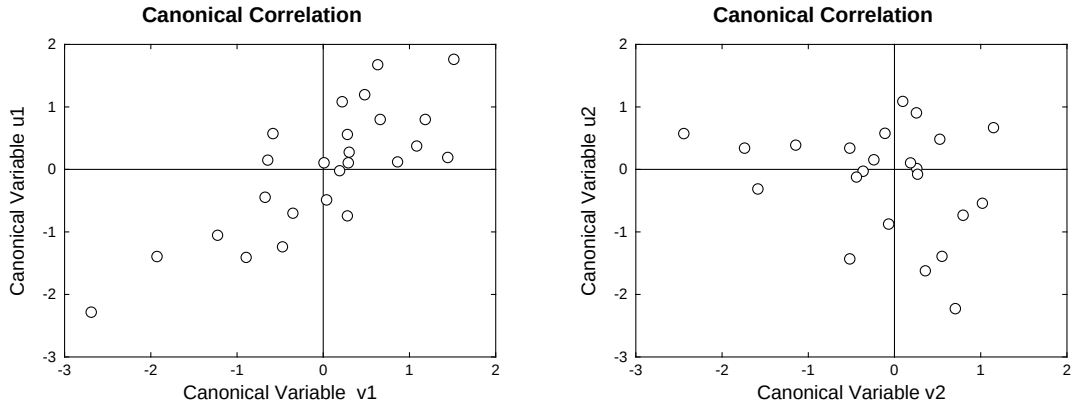
CVY: Canonical coefficients for centralized Y

-0.1415	0.1504
-0.2384	-0.3424

In this table the eigenvalues are proportional to the correlation explained by the corresponding canonical variable, while the χ^2 values and corresponding p values indicate the significance of the successive canonical variables. The results indicate that, with these data, the first canonical variate is sufficient to summarize the correlations between the X and Y variables. Scree diagrams can also be plotted for this purpose.

Example 2

The figure below illustrates two possible graphical displays for the canonical



variates defined by the SIMFIT test file `matrix.tf5`, where columns 1 and 2 are designated the Y sub-matrix, while columns 3 and 4 hold the X matrix. Note that, as eigenvectors do not have unique signs, it is often necessary to reverse the signs of canonical variates for plotting in order to agree with graphs calculated by alternative software. This feature, and also the ability to label the components in such diagrams according to labels added to the data file, is also supported.

Theory

This technique is employed when a n by m data matrix includes at least two groups of variables, say n_x variables of type X, and n_y variables of type Y, measured on the same n subjects, so that $m \geq n_x + n_y$. The idea is to find two transformations, one for the X variables to generate new variables V, and one for the Y variables to generate new variables U, with l components each for $l \leq \min(n_x, n_y)$, such that the canonical variates u_1, v_1 calculated from the data using these transformations have maximum correlation, then u_2, v_2, and so on. Now the variance-covariance matrix of the X and Y data can be partitioned as

$$\begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}$$

and it is required to find transformations that maximize the correlations between the X and Y data sets. Actually, the equations

$$(S_{xy}S_{yy}^{-1}S_{yx} - R^2S_{xx})a = 0$$
$$(S_{yx}S_{xx}^{-1}S_{xy} - R^2S_{yy})b = 0$$

have the same nonzero eigenvalues as the matrices $S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx}$ and $S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}$, and the square roots of these eigenvalues are the canonical correlations, while the eigenvectors of the two above equations define the canonical coefficients, i.e. loadings.

Note that the eigenvalues are proportional to the correlation explained by the corresponding canonical variates, so a scree diagram can be plotted to determine the minimum number of canonical variates needed to adequately represent the data. This diagram plots the eigenvalues together with the average eigenvalue, and the canonical variates with eigenvalues above the average should be retained. Alternatively, assuming multivariate normality, the likelihood ratio test statistics

$$-2\log\lambda = -(n - (k_x + k_y + 3)/2)\sum_{j=i+1}^{l}\log(1 - R_j^2)$$

can be calculated for $i = 0, 1, \ldots, l-1$, where $k_x \leq n_x$ and $k_y \leq n_y$ are the ranks of the X and Y data sets and $l = \min(k_x, k_y)$. These are asymptotically chi-square distributed with $(k_x - i)(k_y - i)$ degrees of freedom, so that the case $i = 0$ tests that none of the l correlations are significant, the case $i = 1$ tests that none of the remaining $l - 1$ correlations are significant, and so on. If any of these tests in sequence are not significant, then the remaining tests should, of course, be ignored.

The previous figure illustrates two possible graphical displays for the canonical variates defined by `matrix.tf5`, where columns 1 and 2 are designated the Y sub-matrix, while columns 3 and 4 hold the X matrix. The canonical variates for X are constructed from the n_x by n_{cv} loading or coefficient matrix CVX, where $CVX(i, j)$ contains the loading coefficient for the ith x variable on the jth canonical variate u_j. Similarly CVY) is the n_y by n_{cv} loading coefficient matrix for the ith y variable on the jth canonical variate v_j. More precisely, if cvx_j is column j of CVX, and cvy_j is column j of CVY, while $x(k)$ is the vector of centralized X observations for case k, and $y(k)$ is the vector of centralized Y observations for case k, then the components $u(k)_j$ and $v(k)_j$ of the n vector canonical variates u_j and v_j are

$$v(k)_j = cvx_j^T x(k), \quad k = 1, 2, \ldots, n$$
$$u(k)_j = cvy_j^T y(k), \quad k = 1, 2, \ldots, n.$$

It is important to realize that the canonical variates for U and V do not represent any sort of regression of Y on X, or X on Y, they are just new coordinates chosen to present the existing correlations between the original X and Y in a new space where the correlations are then ordered for convenience as

$$R^2(u_1, v_1) \geq R^2(u_2, v_2) \geq \ldots \geq R^2(u_l, v_l).$$

Clearly, the left hand plot shows the highest correlation, that is, between u_1 and v_1, whereas the right hand plot illustrates weaker correlation between u_2 and v_2. Note that further linear regression and correlation analysis can also be performed on the canonical variates if required, and also the loading matrices can be saved to construct canonical variates using the SimFIT matrix multiplication routines, and vectors of canonical variates can be saved directly from plots like those displayed.