Canonical correlation

Menu:	QCExpert	Canonical correlation

This module looks for a general linear relationship between two multivariate variables X and Y with dimensions m_1, m_2. The variables are represented by m_1, m_2 columns in the data sheet. The relationship between X and Y is expressed as canonical correlation coefficients which are tested for statistical significance. If any (at least the first) canonical coefficient is significant, then we conclude that there is a proved relationship or influence between the set of variables X and a set of variables Y.

Canonical correlation is a more general method than is pairwise and multiple correlations in Correlation module based on projections into principal components and finding a linear combination of first and second variable, which has the maximum correlation coefficient. The method provides a test of statistical significance of canonical correlation, canonical correlation coefficients, canonical variables and other results. The aim is to identify the strongest statistical relationship between groups of variables, and help users to find the real causal relationships. The result of the calculation are new pairs of univariate variables A_i, B_i. Total number of these couples is m = min (m_1, m_2). The most important is usually the first couple

$$A_1 = a_{1,1}x_1 + a_{2,1}x_2 + \dots + a_{m1,1}x_{m1}$$
$$B_1 = b_{1,1}y_1 + b_{2,1}y_2 + \dots + b_{m2,1}y_{m2}$$

that is established, so that between these canonical variables A_1, B_1 was the maximum possible pair correlation coefficient. Every other canonical pair A_i, B_i is always orthogonal to all the previous canonical variables,

$$A_i^T.A_j = 0;\ B_i^T.B_j = 0 \ \text{ for } i \neq j.$$

If the test confirms statistical significance of correlations, it could be concluded that there is a statistically proven relationship between groups 1 and 2 on the specified level of significance α (usually α = 0.05). Signs of canonical correlation coefficients don’t matter.

Data and parameters

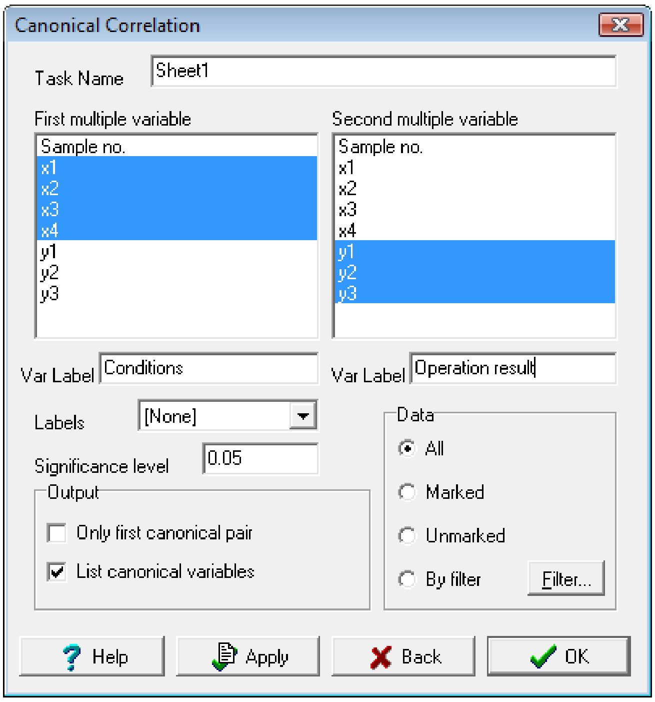
Two multidimensional selections are analyzed on the basis of data arranged in two groups of columns, as suggested in the following table. Two user-selected groups of columns usually characterized by two groups of variables, which we expect to correlate. Table 1 represents a first group of columns x1, x2, x3, x4, and a second group of columns y1, y2, y3. The number of variables m_1, m_2 in groups may be different and must be greater than 1. Here, $m_1 = 4$, $m_2 = 3$. Values in a row must always correspond to the same sample, a situation the patient, etc. All values must be present in each row. Rows with missing values will be ignored. Typical groups parameters may be, for example, 1st Group: chemical composition, 2nd Group: physical parameters, or effect; 1st Group: feedstock parameters, 2 variable: parameters of the product; 1st group: the results of psychological tests, 2 Group: marks at school, etc.

Table 1 Data structure for canonical correlation, variable 1 = (x1, x2, x3, x4); variable 2 = (y1, y2, y3)

Sample no.	x1	x2	x3	x4	y1	y2	y3
33	8.08	2.89	500	21	6.5	28	5.24
34	8.29	4.43	600	22	6.1	32	6.51
35	8.81	3.92	600	19	5.7	33	7.91
36	8.53	3.75	700	17	5.1	38	8.15
37	9.04	3.77	600	12	3.4	32	7.02

39	7.44	2.5	500	15.5	3.8	27	6.255
44	8.83	3.46	500	14.5	4.1	28	6.555
45	7.82	3.2	600	22	4.9	33	5.94
48	8.43	3.31	500	14.5	4.1	28	6.125
15	8.02	2.9	500	21.5	5	27	5.125
16	8.91	3.08	500	21	5.2	29	6.54
17	8.95	3.14	600	17.5	4.8	33	7.855
18	8.88	3.69	600	17	4.2	32	7.64
19	7.28	4.08	600	21	5.2	32	7.51

After opening the dialog *Canonical correlations* the columns of the first and second variable are selected. The columns may not overlap! A column selected in one group may not occur in the second group. We can enter a description of the first and second group and the level of significance (usual value level of significance is 0.05, i.e. 5%). In addition, you can specify the contents of the output report. If the box *Only first canonical pair* is checked only the first canonical couple variables A_1, B_1. If the box *List canonical variables* is not checked the values of canonical variables will not be listed regardless of the field *Only first canonical pair*, which is advantageous when we have many rows, which would produce too long output report.



Task Name: Sheet1

First multiple variable: Sample no., x1, x2, x3, x4, y1, y2, y3

Second multiple variable: Sample no., x1, x2, x3, x4, y1, y2, y3

Var Label: Conditions

Var Label: Operation result

Labels: [None]

Significance level: 0.05

Output

☐ Only first canonical pair

☑ List canonical variables

Data

● All

○ Marked

○ Unmarked

○ By filter Filter...

Fig. 1 Canonical correlations dialog

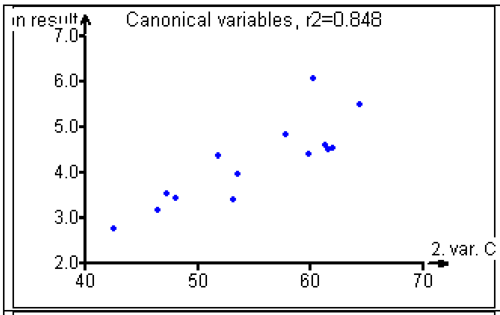
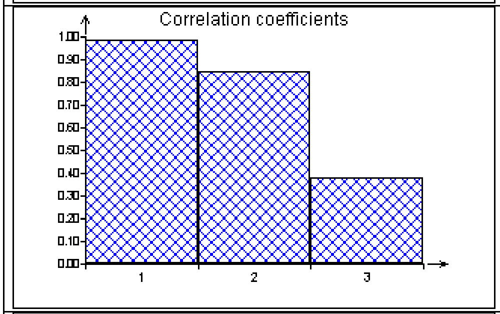
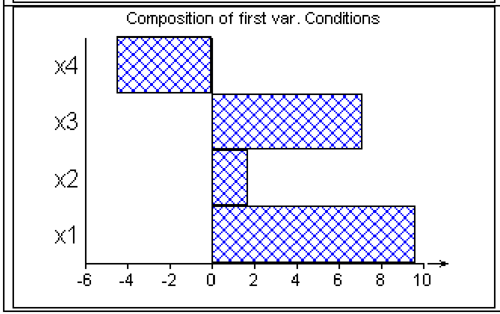
In the group *Data* we can choose a subset of data according to marked data in the data table or possibly define data by a filter. Press the *Apply* or *OK* button to run the calculation.

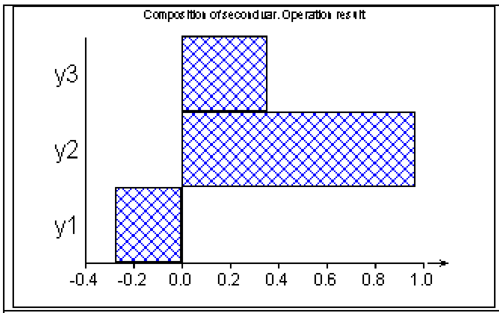
Protocol

Task name	Task name taken from dialog box
Data	Selected data
Basic characteristics	
First (Second) multiple variable	Characteristics of the original variables
Mean	Arithmetic averages of the columns

Std Deviation	Standard deviations of the columns
Canonical correlation coefficients	
Correlation *i*	Value of the i-th canonical correlation coefficient, $i = 1, \ldots, \min(m_1, m_2)$
Correlation significance	Statistical significance of the correlation coefficient at the confidence level α
Lambda Chi-squared p-value Conclusion	Test χ^2 statistic Chi-squared critical quantile p-value of the test statistic Verbal conclusion of the significance test (Significant or Insignificant)
Composition of canonical variables	Canonical coefficient a_{ij}, b_{ij}, composition of canonical variables, i-th canonical variable A_i is given by $A_i = \Sigma x_j.a_{ij}$.
First variable	Coefficients a_{ij}, for 1st canonical variable
Second variable	Coefficients b_{ij}, for 2nd canonical variable
Values of canonical variables	Values of the canonical variables A_i, B_i.

Graphs

	Graphic representation of all pairs of the canonical variables. The criterion of statistical significance is much tougher than for pair correlations, so even strong „looking“ correlation may not be statistically significant. The sign correlation does not matter.
	Bars chart of the absolute values of individual canonical correlation coefficients.
	Composition of the first canonical variable expressed by values $a_{1,1}$, $a_{2,1}$, …, $a_{m1,1}$

	Composition of the first canonical variable expressed by values $b_{1,1}$, $b_{2,1}$, …, $b_{m2,1}$