Multivariate analysis

Menu:	QCExpert	Multivariate Analysis

The Multivariate analysis module is useful for exploratory analysis of multivariate quantitative data. Further, it allows you to perform the principal components analysis. Multivariate data (formally a random sample with vector-valued observations) arise as a result of simultaneous measurement of several (*m*) variables on the same unit. For instance, several physical and/or chemical properties of one sample can be measured, several linear measurements can be taken on the same piece of product, or there might be several characteristics for any employee available. The number of the vector observations is denoted by *n*. To check whether data follow approximately multivariate normal normal distribution, a multivariate normality plot, symmetry plots can be used, together with interactive plots which can help to identify outliers. Such checks are useful for instance in connection with the Hotelling chart construction, where the multivariate normal distribution is assumed. The Andrews plot and Biplot can be used to explore data structure. The principal component analysis is based on coordinate transformation, chosen in such a way that the resulting coordinate system is orthogonal and as much of the original multivariate variability is retained by as few newly defined variables as possible. The amount of variability described by the principal components is displayed in the Screeplot. Composition of individual components in terms of the original variables is displayed in the Loadings plot. Another multivariate statistical concept is the Mahalanobis distance (MD), which is a multivariate analogue of distance between a given point and the mean, measured in standard deviation units. Such a type of distance measure has a direct probability interpretation. Large MD values occur with small probability. In the model checking context, they suggest an outlier. When outlier checking is the main goal, a robust version of MD (based on an M-estimate of the multivariate mean) might be useful.

Data and parameters

Data are organized in columns, each column corresponds to a variable. Number of values should be the same in all columns. Data rows with missing values in one or more variable will be omitted from computations. Minimum column number is 2. Minimum row number is 4. Column names should resemble actual variable names, e.g. Cr_content, Mn_content, Elasticity. Columns, corresponding to the variables you want to include in the analysis can be selected in the *Columns* field of the *Multivariate analysis* dialog panel, Fig. 1. All current data sheet columns are selected by default. Requested output items can be selected in the *Output* field. When the *Use correlation matrix* selection is checked, the principal component analysis is based on the correlation matrix, otherwise it is based on the covariance matrix. Covariance based analysis is recommended especially when the scale of the analyzed variables is vastly different.

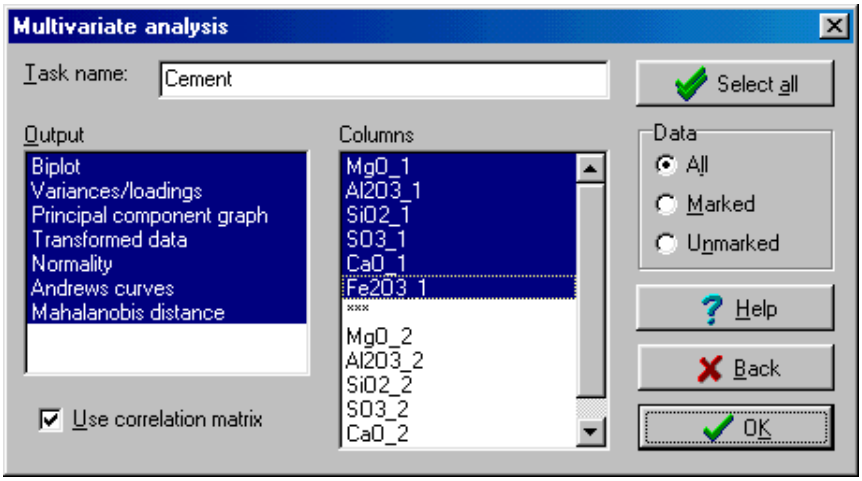


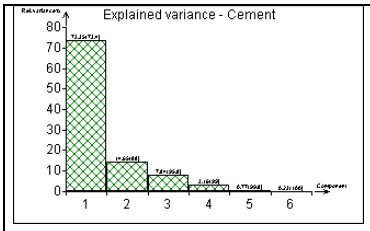
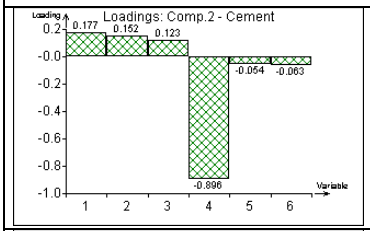
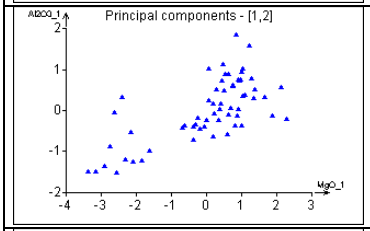
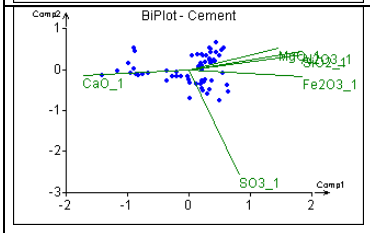
Fig. 1 Multivariate analysis dialog panel

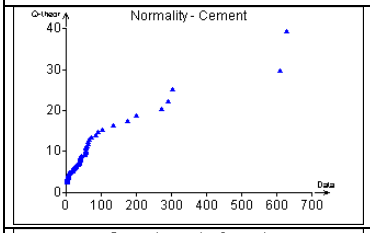
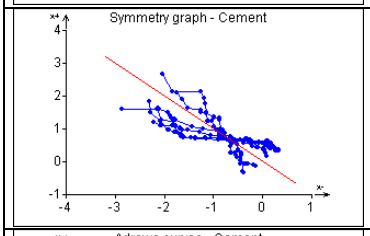
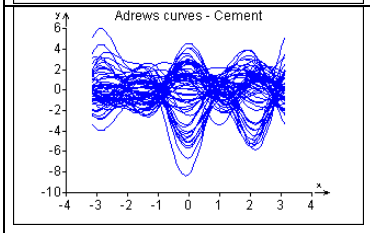
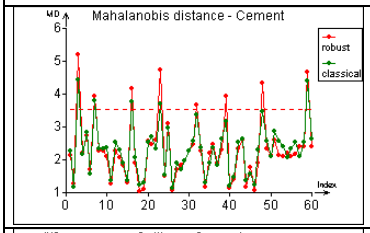
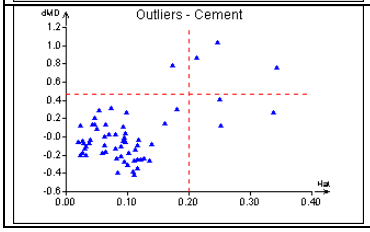
Protocol

Project name	Project name, as entered in the dialog panel.
Number of variables	Number of variables (columns).
Number of multivariate data points	Number of complete data rows.
Summary statistics	
Variable	Variable name.
Mean	Arithmetic mean.
Variance	Variance.
Standard deviation	Standard deviation.
Minimum	Minimum.
Maximum	Maximum.
Correlation matrix	Pairwise correlations for all variable pairs. The diagonal consists of ones, necessarily. When the *Use correlation matrix* option is checked, the correlation matrix is used for component analysis. See also the Correlation module.
Variable	Variable name.
Covariance matrix	Covariance for all possible variable pairs, arranged in matrix form. Variances appear on the main diagonal. The covariance matrix is used for component analysis when the *Use correlation matrix* is not checked.
Variable	Variable name.
Explained variability	Four characteristics describing how much variability is explained by the principal components. The components are always sorted in descending order of importance. The first component describes most of the multivariate variability, while the last component explains the smallest variance portion.
Principal component	Principal component number.
Eigenvalue	Correlation or covariance matrix eigenvalues (depending on whether the *Use correlation matrix* option in the Multivariate analysis dialog panel is checked).
Variance	Variance explained by a particular principal component. It corresponds to the variance of the projection of original data onto the principal component.
Standard deviation	Square root of the previous characteristic.
Rel. variance,%	Variance explained relatively, in percent of the total variance.
Cumulative variance,%	Cumulative variance, explained by the current principal component and all previous ones, expressed relatively, in percent of the total variance.
Eigenvectors	Correlation or covariance matrix eigenvectors (depending on whether the *Use correlation matrix* option in the Multivariate analysis dialog panel is checked).
Data column	Original variable name.
Loadings	Eigenvectors multiplied by square root of their corresponding eigenvalue. Component loadings can help you to interpret component structure. When a component is tight only to a subset of the original variables, the loadings should be small for all the variables not in the subset.
Data column	Original variable name.
Robust M-estimates	An M-type estimate of the vector of mean values (location vector), based on

	iterative procedure. It is a robust alternative to the vector of arithmetic means as the classical estimate. Its use should be considered when outliers presence is suspected.
Mahalanobis distance, MD	A distance measure, having a probability interpretation under multivariate normality. It is a multivariate analogue of distance between a point and the mean, measured in standard deviation units. Two versions are available in QC.Expert™.
Classical MD	A distance between the *i*-th point $\mathbf{x}_i$ and the the vector of expected vales, as estimated by the vector of means, $\mathbf{x}_p$. The distance measure respects the covariance matrix, which is estimated by **S**, it is given by $(\mathbf{x}_i - \mathbf{x}_p)^T(\mathbf{S})^{-1}(\mathbf{x}_i - \mathbf{x}_p)$.
Robust MD	A robust version of the Mahalanobis distance measure. The vector of location parameters is estimated by a robust, M-type estimator $\mathbf{x}_M$. The robust MD is then estimated by $(\mathbf{x}_i - \mathbf{x}_M)^T(\mathbf{S})^{-1}(\mathbf{x}_i - \mathbf{x}_M)$. Outliers have large MD value.
Transformed data	Original data expressed in the newly defined coordinates (obtained by transformation of the original coordinate system).

Graphs

	Screeplot. It displays variance, explained by individual principal components, expressed relatively, in percent of the overall variance. Components are always ordered in descending order of importance, so that the most important component appears first, the least important appears last. Principal component number appears on the *x* axis, while the percentage of explained variance appears on the *y* axis. Cumulative sums of explained variance appear as numbers above each of the columns. Individual variances form which the screeplot is constructed can be found in protocol.
	Loadings plot. One plot is produced for each of the components. Component loadings can help you to interpret component structure. When a component represents only a subset of the original variables, the loadings should be small for all the variables not in the subset. Thus, for example the first component may be related mainly to mechanical properties, the second to chemical composition, etc.
	Principal component plot. One plot is produced for each pair of components. These plots might be sometimes more useful than scatterplots for original variables pairs. (Simple scatterplots are produced e.g. in the Correlation module.)
	Biplot. It is a plot in which both the observations and the variables are represented in a two dimensional space (plane). The points correspond to data lines, while the lines correspond to data columns. Data points can be selected interactively, clicking on the corresponding points on the plot. When interpreting the plot, you should keep in mind that the plot is based on an underlying 2-dimensional approximation to the original data. For each datapoint, the approximation is proportional to the result of vector multiplication of individual lines and points (taken as vectors with the (0,0) origin). From there, it follows that lines close on the plot should correspond to correlated data columns. Row vectors (points) located in the direction of a data column vector (line), should exhibit higher absolute

	values of the variable corresponding to that column. On the other hand, you should realize that goodness of the approximation on which the plot is based, decreases as *m* (original data dimensionality) increases. Especially for large *m*, the dimensionality reduction down to two might be too drastic and the plot might not yield much of a useful insight.
	Multivariate normality plot. It is an analogue of the univariate normal Q-Q plot, used to check multivariate normality of the data. Multivariate normal data should plot close to the line. The plot is based on the F distribution valid for Mahalanobis distance under multivariate normality. To reduce problems with dependence between individual $\mathbf{x}_i$'s and their mean $\mathbf{x}$, jackknifing is used.
	Symmetry plot is analogous to the half-sum plot in univariate case. A point in the plot corresponds to a data point from the current data sheet, so that more than one cell is marked when one point is selected in the plot interactively. Under perfect symmetry, the points should plot on the red line $y=-x$.
	Andrews curves can be a useful tool for exploratory multivariate analysis. Each of the curves represents one data point (a point in *m*-dimensional space). A bunch of similar curves correspond to a cluster of data points similar to each other. Curves of different shape that the bulk of others suggest outliers. Individual data points (curves) can be selected interactively by clicking/dragging mouse. A selection can be cancelled by repeating the operation while holding the *Ctrl* key.
	Plot of both classical and robust Mahalanobis distance. The robust version (red) is more useful for detection of outliers. The classical version (green) is plotted for comparison. The points above the horizontal line (95% quantile of the appropriate null distribution) are suspect as outliers. Only the red points (obtained from the robust version) can be selected interactively.
	An alternative tool for data diagnostics. Elements of the projection matrix (see the Linear regression chapter) are plotted on the x-axis, while difference between robust and classical Mahalanobis distance are plotted on the y-axis. Points right from the vertical line or above the horizontal line are suspect.