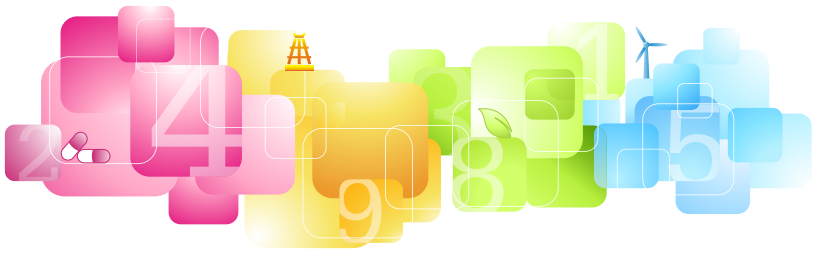
CAMO

Glossary of Multivariate Statistical Methods

An invaluable overview of the main terms and methods used in Multivariate Data Analysis

Over 150 key terms and concepts explained

Useful for both beginners and experienced practitioners



camo.com

Bring data to life

Thanks for downloading our Glossary of Multivariate Statistical Methods.

Multivariate data analysis (MVA) can be complicated, especially for people new to the field, so we've put together this handy glossary, written by world leading experts with many years experience in MVA.

The amount of data collected and stored in the world is growing at an incredible pace. Today, an increasing number of industries have recognized they need more powerful tools than classical statistics to make sense of their complex data. This is where multivariate analysis is proving invaluable in helping identify patterns that provide deeper insights into the masses of data available.

Since our formation in 1984, CAMO Software has been a pioneer in the field of MVA, working with many of the largest companies and most prestigious research institutes in the world. We believe that while the data you analyze might be complicated, the software tool you use should not be. That's why our leading multivariate and design of experiments software, The Unscrambler® X is known for its ease of use, outstanding graphics and state of the art analytical methods.

We hope you find this glossary useful, and would be happy to discuss your multivariate analysis requirements in more detail. You can find other useful resources at www.camo.com

The Unscrambler® X

Free TestDrive

Accuracy

The accuracy of a measurement method is its faithfulness, i.e. how close the measured value is to the actual value. Accuracy differs from precision, which has to do with the spread of successive measurements performed on the same object.

Alternating Least Squares (MCR-ALS)

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is an iterative approach (algorithm) for finding the matrices of concentration profiles and pure component spectra from a data table X containing the spectra (or instrumental measurements) of several unknown mixtures of a few pure components.

The number of compounds in X can be determined using PCA or can be known beforehand. In Multivariate Curve Resolution, it is standard practice to apply MCR-ALS to the same data with varying numbers of components (2 or more).

Analysis of variance (ANOVA)

Classical method to assess the significance of effects by decomposition of a response's variance into explained parts, related to variations in the predictors, and a residual part which summarizes the experimental error.

The main ANOVA results are: Sum of Squares (SS), number of Degrees of Freedom (DF), Mean Square (MS=SS/DF), F-value, p-value.

The effect of a design variable on a response is regarded as significant if the variations in the response value due to variations in the design variable are large compared with the experimental error. The significance of the effect is given as a p-value: usually, the effect is considered significant if the p-value is smaller than 0.05 (5%).

B-Coefficient

See Regression Coefficient.

Bias

Systematic difference between predicted and measured values. The bias is computed as the average value of the residuals.

Bilinear modeling

Bilinear modeling (BLM) is one of several possible approaches for data compression.

The bilinear modeling methods are designed for situations where collinearity exists among the original variables. Common information in the original variables is used to build new variables, that reflect the underlying (“latent”) structure. These variables are therefore called latent variables. The latent variables are estimated as linear functions of both the original variables and the observations, thereby the name bilinear.

PCA, PCR and PLS are bilinear methods.

Box-Behnken design

A class of experimental designs for response surface modeling and optimization, based on only 3 levels of each design variable. The mid-levels of some variables are combined with extreme levels of others. The combinations of only extreme levels (i.e. cube samples of a factorial design) are not included in the design.

Box-Behnken designs are always rotatable. On the other hand, they cannot be built as an extension of an existing factorial design, so they are more often recommended when changing the ranges of variation for some of the design variables after a screening stage, or when it is necessary to avoid too extreme situations.

Calibration

Stage of data analysis where a model is fitted to the available data, so that it describes the data as well as possible.

After calibration, the variation in the data can be expressed as the sum of a modeled part (structure) and a residual part (noise).

Calibration samples

Samples on which the calibration is based. The variation observed in the variables measured on the calibration samples provides the information that is used to build the model.

If the purpose of the calibration is to build a model that will later be applied on new samples for prediction, it is important to collect calibration samples that span the variations expected in the future prediction samples.

Category variable

A category variable is a class variable, i.e. each of its levels is a category (or class, or type), without any possible quantitative equivalent.

Examples: type of catalyst, choice among several instruments, wheat variety, material identification, etc.

Center sample

Sample for which the value of every design variable is set at its mid-level (halfway between low and high).

Center samples have a double purpose: introducing one center sample in a screening design enables curvature checking, and replicating the center sample provides a direct estimation of the experimental error.

Real center samples can be included when all design variables are continuous.

For design containing category variables real center point do not exist, however it is possible to generate faced center point taking the middle range values for the continuous variables and selecting a level for the category variables.

Central composite design

A class of experimental designs for response surface modeling and optimization, based on a two-level factorial design on continuous design variables. Star samples and center samples are added to the full factorial design to provide the intermediate levels necessary for fitting a quadratic model.

Central composite designs have the advantage that they can be built as an extension of a previous factorial design, if there is no reason to change the ranges of variation of the design variables.

If the default star point distance to center is selected, these designs are rotatable.

Classification

Data analysis method used for predicting class membership. Classification can be seen as a predictive method where the response is a category variable. The purpose of the analysis is to be able to predict which category a new sample belongs to. Classification methods implemented in The Unscrambler® include SIMCA, SVM classification, LDA, and PLS-discriminant analysis.

Classification can for instance be used to determine the geographical origin of a raw material from the levels of various impurities, or to accept or reject a product depending on its quality.

To run a SIMCA classification, one needs:

- One or several PCA models (one for each class) based on the same variables;
- Values of those variables collected on known or unknown samples.

Each new sample is projected onto each PCA model. According to the outcome of this projection, the sample is either recognized as a member of the corresponding class, or rejected.

Clustering

Clustering is a classification method that does not require any prior knowledge about the available samples. The basic principle consists in grouping together in a “cluster” several samples which are sufficiently close to each other.

The clustering methods available in The Unscrambler® include the K-means algorithm; the behavior of the algorithm may be tuned by choosing among various ways of computing the distance between samples. Hierarchical clustering can also be run, as can clustering using Ward's method.

Collinearity

Linear relationship between variables. Two variables are collinear if the value of one variable can be computed from the other, using a linear relation. Three or more variables are collinear if one of them can be expressed as a linear function of the others.

Variables which are not collinear are said to be linearly independent. Collinearity - or near-collinearity, i.e. very strong correlation - is the major cause of trouble for MLR models, whereas projection methods like PCA, PCR and PLS handle collinearity well.

Confusion matrix

The confusion matrix is a matrix used for visualization for classification results from supervised methods such as support vector machine classification or linear discriminant analysis classification. It carries information about the predicted and actual classifications of samples, with each row showing the instances in a predicted class, and each column representing the instances in an actual class.

Constrained design

Experimental design involving multilinear constraints between some of the designed variables. There are two types of constrained designed: classical mixture designs and D-optimal designs.

Correlation

A unit less measure of the amount of linear relationship between two variables.

The correlation is computed as the covariance between the two variables divided by the square root of the product of their variances. It varies from –1 to +1.

Positive correlation indicates a positive link between the two variables, i.e. when one increases, the other has a tendency to increase too. The closer to +1, the stronger this link.

Negative correlation indicates a negative link between the two variables, i.e. when one increases, the other has a tendency to decrease. The closer to –1, the stronger this link.

Correlation loadings

Loadings plot marking the 50% and 100% explained variance limits. Correlation loadings are helpful in revealing variable correlations.

Correlation Optimized Warping (COW)

COW is a method for aligning data where the signals exhibit shifts in their position along the x axis. This transform is a technique often use for time-shifting chromatographic spectra.

Covariance

A measure of the linear relationship between two variables.

The covariance is given on a scale which is a function of the scales of the two variables, and may not be easy to interpret. Therefore, it is usually simpler to study the correlation instead.

Cross validation

Validation method where some samples are kept out of the calibration and used for prediction. This is repeated until all samples have been kept out once. Validation residual variance can then be computed from the prediction residuals.

In segmented cross validation, the samples are divided into subgroups or “segments”. One segment at a time is kept out of the calibration. There are as many calibration rounds as segments, so that predictions can be made on all samples. A final calibration is then performed with all samples.

In full cross validation, only one sample at a time is kept out of the calibration per iteration.

Cubic effect

When analyzing the results from designed experiments, cubic effects can be included in the model to handle complex cases of nonlinear effects or multiple interactions between the X-variables.

Also called third order effects, they comprise:

- Interactions between 3 design parameters (A*B*C),
- Cubic terms of the design variables (A^3) and
- Combined effects (A^2*B).

Curvature

Curvature means that the true relationship between response variations and predictor variations is nonlinear.

In screening designs, curvature can be detected by introducing a center sample.

Data compression

Concentration of the information carried by several variables onto a few underlying variables.

The basic idea behind data compression is that observed variables often contain common information, and that this information can be expressed by a smaller number of variables than originally observed. PCA and PLS are forms of data compression.

Data mining

This is the practice of studying large amounts of data to find patterns or trends. MVA is a form of data mining.

Degrees of freedom

The number of degrees of freedom of a phenomenon is the number of independent ways this phenomenon can be varied.

Degrees of freedom are used to compute variances and theoretical variable distributions. For instance, an estimated variance is said to be “corrected for degrees of freedom” if it is computed as the sum of square of deviations from the mean, divided by the number of degrees of freedom of this sum.

Dendrogram

A dendrogram (from Greek dendron “tree”, -gramma “drawing”) is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.

Design analysis

Calculation of the effects of design variables on the responses. It consists mainly of Analysis of Variance (ANOVA), various significance tests, and multiple comparisons, response surface generation whenever they apply.

Design variable

Experimental factor for which the variations are controlled in an experimental design.

Distribution

Shape of the frequency diagram of a measured variable or calculated parameter. Observed distributions can be represented by a histogram.

Some statistical parameters have a well-known theoretical distribution which can be used for significance testing.

D-optimal design

Experimental design generated by a D-optimal algorithm. A D-optimal design takes into account the multilinear relationships existing between design variables, and thus works with constrained experimental regions. There are two types of D-optimal designs depending on their initial points: D-optimal mixture designs which are based on subsimplexes and general D-optimal designs which are based on subfactorial designs.

Downweight

A weighting option which allows one to remove the influence of a variable on a model by giving it a very low weight in a PCA, PCR or PLS model. The variable is still displayed, showing how it correlates to other variables. In previous versions of The Unscrambler® this weighting option was referred to as passify.

Experimental design

This is also referred to as Design of Experiments.

Plan for experiments where input variables are varied systematically within predefined ranges, so that their effects on the output variables (responses) can be estimated and checked for significance.

Experimental designs are built with a specific objective in mind, namely screening, screening with interaction, or optimization.

The number of experiments and the way they are built depends on the objective and on the operational constraints.

Experimental error

Random variation in the response that occurs naturally when performing experiments.

An estimation of the experimental error is used for significance testing, as a comparison to structured variation that can be accounted for by the studied effects.

Experimental error can be measured by replicating some experiments and computing the standard deviation of the response over the replicates. It can also be estimated as the residual variation when all “structured” effects have been accounted for.

Explained variance

Share of the total variance which is accounted for by the model.

Explained variance is computed as the complement to residual variance, divided by total variance. It is expressed as a percentage.

For instance, an explained variance of 90% means that 90% of the variation in the data is described by the model, while the remaining 10% are noise (or error).

F-distribution

Fisher distribution is the distribution of the ratio between two variances.

The F-distribution assumes that the individual observations follow an approximate normal distribution.

Fractional factorial design

A reduced experimental plan often used for screening of many variables. It gives as much information as possible about the main effects of the design variables with a minimum of experiments. Some fractional designs also allow two-variable interactions to be studied. This depends on the resolution of the design.

In fractional factorial designs, a subset of a full factorial design is selected so that it is still possible to estimate the desired effects from a limited number of experiments.

The degree of fractionality of a factorial design expresses how fractional it is, compared with the corresponding full factorial.

F-ratio

The F-ratio is the ratio between explained variance (associated to a given predictor) and residual variance. It shows how large the effect of the predictor is, as compared with random noise.

By comparing the F-ratio with its theoretical distribution (F-distribution), one obtains the significance level (given by a p-value) of the effect.

Full factorial design

Experimental design where all levels of all design variables are combined.

Such designs are often used for extensive study of the effects of few variables, especially if some variables have more than two levels. They are also appropriate in screening with interaction designs, to study both main effects and interactions, especially if no Resolution V design is available.

Histogram

A plot showing the observed distribution of data points. The data range is divided into a number of bins (i.e. intervals) and the number of data points that fall into each bin is summed up.

The height of the bar in the histograms shows how many data points fall within the data range of the bin.

Hotelling T^2 ellipse

This 95% confidence ellipse can be included in scores plots and reveals potential outliers, lying outside the ellipse.

Hotelling T^2 statistics

A linear function of the leverage that can be compared to a critical limit according to an F-test. This statistic is useful for the detection of outliers at the modeling or prediction stage.

Influence

A measure of how much impact a single data point (or a single variable) has on the model. The influence depends on the leverage and the residuals.

Inlier

A prediction sample far away from the calibration samples in the regression model. Local “holes” or areas with low density in terms of calibration samples can result in a situation where some prediction samples are detected as inliers.

Inner relation

In PLS regression models, scores in X are used to predict the scores in Y and from these predictions, the estimated

Interaction effects

There is an interaction between two design variables when the effect of the first variable depends on the level of the other. This means that the combined effect of the two variables is not equal to the sum of their main effects.

An interaction that increases the main effects is a synergy. If it goes in the opposite direction, it can be called an antagonism.

Intercept

(Also called Offset). The point where a regression line crosses the ordinate (Y-axis).

K-means

An algorithm for data clustering. The samples will be grouped into K (user-determined number) clusters based on a specific distance measurement, so that the sum of distances between each sample and its cluster centroid is minimized.

Lack of fit

In Response Surface Analysis, the ANOVA table includes a special chapter which checks whether the regression model describes the true shape of the response surface. Lack of fit means that the true shape is likely to be different from the shape indicated by the model.

If there is a significant lack of fit, one can investigate the residuals and try a transformation.

Latent variable

A variable that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed and directly measured.

LDA

See Linear Discriminant Analysis.

Least square criterion

Basis of classical regression methods, that consists in minimizing the sum of squares of the residuals. It is equivalent to minimizing the average squared distance between the original response values and the fitted values.

Leverage

A measure of how extreme a data point or a variable is compared to the majority.

In PCA, PCR and PLS, leverage can be interpreted as the distance between a projected point (or projected variable) and the model center. In MLR, it is the object distance to the model center.

Average data points have a low leverage. Points or variables with a high leverage are likely to have a high influence on the model.

Leverage correction

A quick method to simulate model validation without performing any actual predictions.

It is based on the assumption that samples with a higher leverage will be more difficult to predict accurately than more central samples. Thus a validation residual variance is computed from the calibration sample residuals, using a correction factor which increases with the sample leverage.

Note! For MLR, leverage correction is strictly equivalent to full cross-validation. For other methods, leverage correction should only be used as a quick-and-dirty method for a first calibration, and a proper validation method should be employed later on to estimate the optimal number of components correctly.

Linear Discriminant Analysis (LDA)

LDA is the simplest of all possible classification methods that are based on Bayes' formula. The objective of LDA is to determine the best fit parameters for classification of samples by a developed model.

Linear model

Regression model including as X-variables the linear effects of each predictor. The linear effects are also called main effects. Linear models are used in the analysis of Plackett-Burman and Resolution III fractional factorial designs. Higher resolution designs allow the estimation of interactions in addition to the linear effects.

Loading weights

Loading weights are estimated in PLS regression. Each X-variable has a loading weight along each model component.

The loading weights show how much each predictor (or X-variable) contributes to explaining the response variation along each model component. They can be used, together with the Y-loadings, to represent the relationship between X- and Y-variables as projected onto one, two or three components (line plot, 2-D scatter plot and 3-D scatter plot respectively).

Loadings

Loadings are estimated in bilinear modeling methods where information carried by several variables is concentrated onto a few components. Each variable has a loading along each model component.

The loadings show how well a variable is taken into account by the model components. Loadings can be used to understand how much each variable contributes to the meaningful variation in the data, and to interpret variable relationships. They are also useful to interpret the meaning of each model component.

Lower quartile

The lower quartile of an observed distribution is the variable value that splits the observations into 25% lower values, and 75% higher values. It can also be called 25% percentile.

L-shaped PLS Regression (L-PLS)

As opposed to bilinear modeling such as PLS where the data are arranged in such a way that the information obtained on a dependent variable Y is related to some independent measures X, L-PLS can be used in cases where the Y data may have descriptors of its columns, organized in a third table Z (containing the same number of columns as in Y).

The three matrices X, Y and Z can together be visualized in the form of an L-shaped arrangement. Such data analysis has potential widespread use in areas such as consumer preference studies, medical diagnosis and spectroscopic applications.

Main effect

Average variation observed in a response when a design variable goes from its low to its high level.

The main effect of a design variable can be interpreted as linear variation generated in the response, when this design variable varies and the other design variables have their average values.

Mean

Average value of a variable over a specific sample set. The mean is computed as the sum of the variable values, divided by the number of samples.

The mean gives a value around which all values in the sample set are distributed. In Statistics results, the mean can be displayed together with the standard deviation.

Mean centering

Subtracting the mean (average value) from a variable, for each data point.

Median

The median of an observed distribution is the variable value that splits the distribution in its middle: half the observations have a lower value than the median, and the other half have a higher value. It can also be called 50% percentile.

Missing values

Whenever the value of a given variable for a given sample is unknown or not available, this results in a hole in the data. Such holes are called missing values, and in The Unscrambler® corresponding cell of the data table are left empty.

In some cases, it is only natural to have missing values — for instance when the concentration of a compound (Y) in a new sample is supposed to be predicted from its spectrum (X).

Sometimes it would be nice to reconstruct the missing values, for instance when applying a data analysis that does not handle missing values well, like MLR, kernel-PLS or wide-kernel. One may choose to fill missing values by using the command Tasks - Transform - Missing Values….

Mixture components

Ingredients of a mixture.

There must be at least three components to define a mixture. A unique component cannot be called mixture.

Two components mixed together do not require a Mixture design to be studied: study the variation in quantity of one of them as a classical process variable.

Mixture constraint

Multilinear constraint between Mixture variables. The general equation for the Mixture constraint is where the Xi represent the ingredients of the mixture, and S is the total amount of mixture. In most cases, S is equal to 100%.

Mixture design

Special type of experimental design, applying to the case of a mixture constraint. There are three types of classical Mixture designs: Simplex-Lattice design, Simplex-Centroid design, and Axial design. Mixture designs that do not have a simplex experimental region are generated D-optimally; they are called D-optimal mixture designs.

Model

Mathematical equation summarizing variations in a data set.

Models are built so that the structure of a data table can be understood better than by just looking at all raw values.

Statistical models consist of a structure part and an error part. The structure part (information) is intended to be used for interpretation or prediction, and the error part (noise) should be as small as possible for the model to be reliable.

Model center

The model center is the origin around which variations in the data are modeled. It is the (0,0) point on a scores plot.

If the variables have been centered, samples close to the average will lie close to the model center.

Model check

In Response Surface Analysis, a section of the ANOVA table checks how useful the interactions and squares are, compared with a purely linear model. This section is called model check.

If one part of the model is not significant, it can be removed so that the remaining effects are estimated with a better precision.

Multiple comparison tests

Tests showing which levels of a category design variable can be regarded as causing real differences in response values, compared to other levels of the same design variable.

For continuous or binary design variables, analysis of variance is sufficient to detect a significant effect and interpret it. For category variables, a problem arises from the fact that, even when analysis of variance shows a significant effect, it is impossible to know which levels are significantly different from others. This is why multiple comparisons have been implemented. They are to be used once analysis of variance has shown a significant effect for a category variable.

Multiple Linear Regression (MLR)

A method for relating the variations in a response variable (Y-variable) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

An important assumption for the method is that the X-variables are linearly independent, i.e. that no linear relationship exists between the X-variables. When the X-variables carry common information, problems can arise due to exact or approximate collinearity.

Multivariate Curve Resolution (MCR)

A method that resolves unknown mixtures into n pure components. The number of components and their concentrations and instrumental profiles are estimated in a way that explains the structure of the observed data under the chosen model constraints.

Multivariate analysis

Multivariate analysis (MVA) is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest.

Source: Wikipedia

Nipals

In statistics, Non-linear Iterative Partial Least Squares (NIPALS) is an algorithm for computing the first few components in a principal component or partial least squares analysis. For very high-dimensional data sets, such as those generated in the ‘omics sciences (e.g., genomics, metabolomics) it is usually only necessary to compute the first few principal components.

Source: Wikipedia

Noise

Random variation that does not contain any information. The purpose of multivariate modeling is to separate information from noise.

Non-linearity

Deviation from linearity in the relationship between a response and its predictors.

Non-negativity

In MCR, the Non-negativity constraint forces the values in a profile to be equal to or greater than zero.

Normal distribution

Frequency diagram showing how independent observations, measured on a continuous scale, would be distributed if there were an infinite number of observations and no factors caused systematic effects.

A normal distribution can be described by two parameters:

- A theoretical mean, which is the center of the distribution;
- A theoretical standard deviation, which is the spread of the individual observations around the mean.

Normal probability plot

The normal probability plot (or N-plot) is a 2-D plot which displays a series of observed or computed values in such a way that their distribution can be visually compared to a normal distribution.

The observed values are used as abscissa, and the ordinate displays the corresponding percentiles on a special scale. Thus if the values are approximately normally distributed around zero, the points will appear close to a straight line going through (0,50%).

A normal probability plot can be used to check the normality of the residuals (they should be normal; outliers will stick out), and to visually detect significant effects in screening designs with few residual degrees of freedom.

Optimization

Finding the settings of design variables that generate optimal response values.

Orthogonal

Two variables are said to be orthogonal if they are completely uncorrelated, i.e. their correlation is 0. In PCA and PCR, the principal components are orthogonal to each other.

Factorial designs, Plackett-Burman designs, Central Composite designs and Box-Behnken designs are built in such a way that the studied effects are orthogonal to each other.

Orthogonal design

Designs built in such a way that the studied effects are orthogonal to each other, are called orthogonal designs.

Examples: Factorial designs, Plackett-Burman designs, Central Composite designs and Box-Behnken designs.

D-optimal designs and classical mixture designs are not orthogonal.

Outlier

An observation (outlying sample) or variable (outlying variable) which is abnormal compared to the major part of the data.

Extreme points are not necessarily outliers; outliers are points that apparently do not belong to the same population as the others, or that are badly described by a model.

Outliers should be investigated before they are removed from a model, as an apparent outlier may be due to an error in the data.

Overfitting

For a model, overfitting is a tendency to describe too much of the variation in the data, so that not only consistent structure is taken into account, but also some noise or noninformative variation.

Overfitting should be avoided, since it usually results in a lower quality of prediction. Validation is an efficient way to avoid model overfitting.

Partial Least Squares regression

See PLS regression.

Passified

See Downweight

In previous versions of The Unscrambler®, the term passify was used when a variable was weighted by multiplying by a very small number. The variable was said to be Passified, meaning that it loses all influence on the model, but it is not removed from the analysis.

The term for this type of weighting has been changed to Downweight.

PCA

See Principal Component Analysis.

PCR

See Principal Component Regression.

PCS

See Principal Component.

Percentile

The X% percentile of an observed distribution is the variable value that splits the observations into X% lower values, and 100-X% higher values.

Quartiles and median are percentiles. The percentiles are displayed using a box-plot.

Plackett-Burman design

A very reduced experimental plan used for a first screening of many variables. It gives information about the main effects of the design variables with the smallest possible number of experiments.

No interactions can be studied with a Plackett-Burman design, and moreover, each main effect is confounded with a combination of several interactions, so that these designs should be used only as a first stage, to check whether there is any meaningful variation at all in the investigated phenomena.

PLS Discriminant Analysis (PLS-DA)

Classification method based on modeling the differences between several classes with PLS.

If there are only two classes to separate, the PLS model uses one response variable, which codes for class membership as follows: -1 for members of one class, +1 for members of the other one.

If there are three classes or more, the PLS model uses one response variable (-1/+1 or 0/1, which is equivalent) coding for each class.

PLS regression

A method for relating the variations in one or several response variables (Y-variables) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

This method performs particularly well when the various X-variables express common information, i.e. when there is a large amount of correlation, or even collinearity.

Partial Least Squares Regression is a bilinear modeling method where information in the original X-data is projected onto a small number of underlying (“latent”) variables called PLS components. The Y-data are actively used in estimating the “latent” variables to ensure that the first components are those that are most relevant for predicting the Y-variables. Interpretation of the relationship between X-data and Y-data is then simplified as this relationship in concentrated on the smallest possible number of components.

By plotting the first PLS components one can view main associations between X-variables and Y-variables, and also interrelationships within X-data and within Y-data.

PLS1

Version of the PLS method with only one Y-variable.

PLS2

Version of the PLS method in which several Y-variables are modeled simultaneously, thus taking advantage of possible correlations or collinearity between Y-variables.

Precision

The precision of an instrument or a measurement method is its ability to give consistent results over repeated measurements performed on the same object. A precise method will give several values that are very close to each other.

Precision can be measured by standard deviation over repeated measurements.

If precision is poor, it can be improved by systematically repeating the measurements over each sample, and replacing the original values by their average for that sample.

Precision differs from accuracy, which has to do with how close the average measured value is to the target value.

Prediction

Computing response values from predictor values, using a regression model.

The following are needed to make predictions:

- A regression model (PCR, PLS or MLR), calibrated on X- and Y-data;
- New X-data collected on samples which should be similar to the ones used for calibration.

The new X-values are fed into the model equation (which uses the regression coefficients), and predicted Y-values are computed.

Predictor

Variable used as input in a regression model. Predictors are usually denoted X-variables.

Principal component (PC)

Principal Components (PCs) are composite variables, i.e. linear functions of the original variables, estimated to contain, in decreasing order, the main structured information in the data. A PC is the same as a score vector, and is also called a latent variable or a factor.

Principal components are estimated in PCA and PCR. PLS components are also denoted Pcs.

Principal Component Analysis (PCA)

PCA is a bilinear modeling method which gives an interpretable overview of the main information in a multidimensional data table.

The information carried by the original variables is projected onto a smaller number of underlying (“latent”) variables called principal components. The first principal component covers as much of the variation in the data as possible. The second principal component is orthogonal to the first and covers as much of the remaining variation as possible, and so on.

By plotting the principal components, one can view interrelationships between different variables, and detect and interpret sample patterns, groupings, similarities or differences.

Principal Component Regression (PCR)

PCR is a method for relating the variations in a response variable (Y-variable) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

This method performs particularly well when the various X-variables express common information, i.e. when there is a large amount of correlation, or even collinearity.

Principal Component Regression is a two-step method. First, a Principal Component Analysis is carried out on the X-variables. The principal components are then used as predictors in a Multiple Linear Regression.

Process variable

Experimental factor for which the variations are controlled in an experimental design, and to which the mixture variable definition does not apply.

Projection

Principle underlying bilinear modeling methods such as PCA, PCR and PLS.

In those methods, each sample can be considered as a point in a multidimensional space. The model will be built as a series of components onto which the samples - and the variables - can be projected. Sample projections are called scores, variable projections are called loadings.

The model approximation of the data is equivalent to the orthogonal projection of the samples onto the model. The residual variance of each sample is the squared distance to its projection.

Pure components

In MCR, an unknown mixture is resolved into n pure components. The number of components and their concentrations and instrumental profiles are estimated in a way that explains the structure of the observed data under the chosen model constraints.

P-value

The p-value measures the probability that a parameter estimated from experimental data should be as large as it is, if the real (theoretical, non-observable) value of that parameter were actually zero. Thus, p-value is used to assess the significance of observed effects or variations: a small p-value means a small risk of mistakenly concluding that the observed effect is real.

The usual limit used in the interpretation of a p-value is 0.05 (or 5%). If p-value < 0.05, the observed effect can be presumed to be significant and is not due to random variations.

p-value is also called “significance level”.

Q-residual limits

The Q-residual limits for components 0-A are computed as a function of the remaining eigenvalues A+1:Amax, where Amax is the maximum number of components that can be calculated, limited by the number of samples or variables.

When PCA is computed by the SVD algorithm all eigenvalues are returned, and Q-residuals can be estimated. When the NIP algorithm is chosen, only a few components are normally estimated, thus Q-residual limits are not available.

Similarly for PLS regression, the Q-residual limits are correct only if the maximum number of factors is computed, i.e. all the variance in X is modeled.

As the Q-residual limit is a function of the eigenvalue to the power of 3, one may get a reasonable estimate if more than 95% of the X-variance is explained in the model although the number of factors is less than the maximum.

Quadratic model

Regression model including as X-variables the linear effects of each predictor, all two-variable interactions, and the square effects with a quadratic model, the curvature of the response surface can be approximated in a satisfactory way.

Quantile plot

The Quantile plot represents the distribution of a variable in terms of percentiles for a given population. It shows the minimum, the 25% percentile (lower quartile), the median, the 75% percentile (upper quartile) and the maximum.

Random effect

Effect of a variable for which the levels studied in an experimental design can be considered to be a small selection of a larger (or infinite) number of possibilities.

Examples:

- Effect of using different batches of raw material;
- Effect of having different persons perform the experiments.

The alternative to a random effect is a fixed effect.

Reference sample

Sample included in a designed data table to compare a new product under development to an existing product of a similar type.

The design file will contain only response values for the reference samples, whereas the input part (the design part) is missing (m).

Regression coefficient

In a regression model equation, regression coefficients are the numerical coefficients that express the link between variation in the predictors and variation in the response.

Regression

Generic name for all methods relating the variations in one or several response variables (Y-variables) to the variations of several predictors (X-variables), with explanatory or predictive purposes.

Regression can be used to describe and interpret the relationship between the X-variables and the Y-variables, and to predict the Y-values of new samples from the values of the X-variables.

Repeated measurement

Measurement performed several times on one single experiment or sample.

The purpose of repeated measurements is to estimate the measurement error, and to improve the precision of an instrument or measurement method by averaging over several measurements.

Replicate

Replicates are experiments that are carried out several times. The purpose of including replicates in a data table is to estimate the experimental error.

Replicates should not be confused with repeated measurements, which give information about measurement error. In cross validation, replicates should be excluded as a group.

Residual

A measure of the variation that is not taken into account by the model.

The residual for a given sample and a given variable is computed as the difference between observed value and fitted (or projected, or predicted) value of the variable on the sample.

Residual variance

The mean square of all residuals, sample- or variable-wise.

This is a measure of the error made when observed values are approximated by fitted values, i.e. when a sample or a variable is replaced by its projection onto the model.

The complement to residual variance is explained variance.

Response surface analysis

Regression analysis, often performed with a quadratic model, in order to describe the shape of the response surface precisely.

This analysis includes a comprehensive ANOVA table, various diagnostic tools such as residual plots, and two different visualizations of the response surface: contour plot and landscape plot.

Note: Response surface analysis can be run on designed or non-designed data. However it is not available for Mixture Designs; use PLS instead.

Response variable

Observed or measured parameter which a regression model tries to predict.

Responses are usually denoted Y-variables.

RMSEC

Root Mean Square Error of Calibration. A measurement of the average difference between predicted and measured response values, at the calibration stage.

RMSEC can be interpreted as the average modeling error, expressed in the same units as the original response values.

RMSED

Root Mean Square Error of Deviations. A measurement of the average difference between the abscissa and ordinate values of data points in any 2-D scatter plot.

RMSEP

Root Mean Square Error of Prediction. A measurement of the average difference between predicted and measured response values, at the prediction or validation stage.

RMSEP can be interpreted as the average prediction error, expressed in the same units as the original response values.

R-square

The R-square of a regression model is a measure of the quality of the model. Also known as coefficient of determination, it is computed as 1 - (Residual Y-variance), or (Explained Y-variance)/100. For Calibration results, this is also the square of the correlation coefficient between predicted and measured values, and the R-square value is always between 0 and 1. The closer to 1, the better.

The R-square is displayed among the plot statistics of a Predicted vs. Reference plot. When based on the calibration samples, it tells about the quality of the fit. When computed from the validation samples (similar to the “adjusted R-square” found in the literature) it tells about the predictive ability of the model.

Sample

Object or individual on which data values are collected, and which builds up a row in a data table. In experimental design, each separate experiment is a sample.

Scatter effects

In spectroscopy, scatter effects are effects that are caused by physical phenomena, like particle size, rather than chemical properties. They interfere with the relationship between chemical properties and shape of the spectrum. There can be additive and multiplicative scatter effects.

Additive and multiplicative effects can be removed from the data by different methods. Multiplicative Scatter Correction removes the effects by adjusting the spectra from ranges of wavelengths supposed to carry no specific chemical information.

Scores

Scores are estimated in bilinear modeling methods where information carried by several variables is concentrated onto a few underlying variables. Each sample has a score along each model component.

The scores show the locations of the samples along each model component, and can be used to detect sample patterns, groupings, similarities or differences.

Screening

First stage of an investigation, where information is sought about the effects of many variables. Since many variables have to be investigated, only main effects, and optionally interactions, can be studied at this stage.

There are specific experimental designs for screening, such as factorial or Plackett-Burman designs.

Sensitivity to pure components

In MCR computations, sensitivity to pure components is one of the parameters influencing the convergence properties of the algorithm. It can be roughly interpreted as how dominating the last estimated primary principal component is (the one that generates the weakest structure in the data), compared to the first one.

The higher the sensitivity, the more pure components will be extracted.

Significance level

See P-value.

Significant

An observed effect (or variation) is declared significant if there is a small probability that it is due to chance.

SIMCA classification

Classification method based on disjoint PCA modeling.

SIMCA focuses on modeling the similarities between members of the same class. A new sample will be recognized as a member of a class if it is similar enough to the other members; else it will be rejected.

Simplex

Specific shape of the experimental region for a classical mixture design. A Simplex has N corners but N-1 independent variables in a N-dimensional space. This results from the fact that whatever the proportions of the ingredients in the mixture, the total amount of mixture has to remain the same: the Nth variable depends on the N-1 other ones. When mixing three components, the resulting simplex is a triangle.

Singular Value Decomposition (SVD)

In linear algebra, the singular value decomposition (SVD) is an important factorization of a rectangular real or complex matrix, with many applications in signal processing and statistics. Applications which employ the SVD include computing the pseudoinverse, least squares fitting of data, matrix approximation, and determining the rank, range and null space of a matrix.

Source: Wikipedia

SNV

See Standard Normal Variate.

Square effect

Average variation observed in a response when a design variable goes from its center level to an extreme level (low or high).

The square effect of a design variable can be interpreted as the curvature observed in the response surface, with respect to this particular design variable.

Standard deviation

SDev is a measure of a variable's spread around its mean value, expressed in the same unit as the original values.

Standard deviation is computed as the square root of the mean square of deviations from the mean.

Standard error of performance (SEP)

Variation in the precision of predictions over several samples.

SEP is computed as the standard deviation of the residuals.

Standard Normal Variate (SNV)

SNV is a transformation usually applied to spectroscopic data, which centers and scales each individual spectrum (i.e. a sample-oriented standardization). It is sometimes used in combination with detrending (DT) to reduce multicollinearity, baseline shift and curvature in spectroscopic data.

Standardization

Widely used preprocessing that consists in first centering the variables, then scaling them to unit variance.

The purpose of this transformation is to give all variables included in an analysis an equal chance to influence the model, regardless of their original variances.

In The Unscrambler® standardization can be performed automatically when computing a model, by choosing 1/SDev as variable weights.

Star samples

In optimization designs of the Central Composite family, star samples are samples with mid-values for all design variables except one, for which the value is extreme. They provide the necessary intermediate levels that will allow a quadratic model to be fitted to the data.

Star samples can be centers of cube faces, or they can lie outside the cube, at a given distance (larger than 1) from the center of the cube — see Star Points Distance To Center.

Student's t-distribution

Frequency diagram showing how independent observations, measured on a continuous scale, are distributed around their mean when the mean and standard deviation have been estimated from the data and when no factor causes systematic effects.

When the number of observations increases towards an infinite number, the Student's t-distribution becomes identical to the normal distribution.

A Student's t-distribution can be described by two parameters: the mean value, which is the center of the distribution, and the standard deviation, which is the spread of the individual observations around the mean. Given those two parameters, the shape of the distribution further depends on the number of degrees of freedom, usually n-1, if n is the number of observations.

Test samples

Additional samples which are not used during the calibration stage, but only to validate an already calibrated model.

The data for those samples consist of X-values (for PCA) or of both X- and Y-values (for regression). The model is used to predict new values for those samples, and the predicted values are then compared to the observed ones.

Test set validation

Validation method based on the use of different data sets for calibration and validation. During the calibration stage, calibration samples are used. Then the calibrated model is used on the test samples, and the validation residual variance is computed from their prediction residuals.

Training samples

See Calibration samples.

T-value

The t-value is computed as the ratio between deviation from the mean accounted for by a studied effect, and standard error of the mean.

By comparing the t-value with its theoretical distribution (Student's t-distribution), one obtains the significance level of the studied effect.

Uncertainty limits

Limits produced by Uncertainty Testing, helping one assess the significance of the X-variables in a regression model. Variables with uncertainty limits that do not cross the "0" axis are significant.

Uncertainty test

Martens' Uncertainty Test is a significance testing method implemented in The Unscrambler® which assesses the stability of PCA or Regression results. Many plots and results are associated to the test, allowing the estimation of the model stability, the identification of perturbing samples or variables, and the selection of significant X-variables. The test is performed with Cross Validation, and is based on the Jack-knifing principle.

Underfit

A model that leaves aside some of the structured variation in the data is said to underfit.

U-scores

The scores found by PLS in the Y-matrix.

See Scores for more details.

Validation samples

See Test samples.

Validation

Validation means checking how well a model will perform for future samples taken from the same population as the calibration samples. In regression, validation also allows for estimation of the prediction error in future predictions.

The outcome of the validation stage is generally expressed by a validation variance. The closer the validation variance is to the calibration variance, the more reliable the model conclusions.

When explained validation variance stops increasing with additional model components, it means that the noise level has been reached. Thus the validation variance is a good diagnostic tool for determining the proper number of components in a model.

Validation variance can also be used as a way to determine how well a single variable is taken into account in an analysis. A variable with a high explained validation variance is reliably modeled and is probably quite precise; a variable with a low explained validation variance is badly taken into account and is probably quite noisy.

Three validation methods are available in The Unscrambler® X

- Test set validation
- Cross validation
- Leverage correction

Variable

Any measured or controlled parameter that has varying values over a given set of samples.

A variable determines a column in a data table.

Variance

A measure of a variable's spread around its mean value, expressed in square units as compared to the original values.

Variance is computed as the mean square of deviations from the mean. It is equal to the square of the standard deviation.

Weighting

A technique to modify the relative influences of the variables on a model. This is achieved by giving each variable a new weight, i.e. multiplying the original values by a constant which differs between variables. This is also called scaling.

The most common weighting technique is standardization, where the weight is the standard deviation of the variable. Other weighting options in The Unscrambler® are constant, and down weighted.

camo.com

Bring data to life

NORWAY

Nedre Vollgate 8,
N-0158
Oslo
Tel: (+47) 223 963 00
Fax: (+47) 223 963 22

USA

One Woodbridge Center
Suite 319, Woodbridge
NJ 07095
Tel: (+1) 732 726 9200
Fax: (+1) 973 556 1229

INDIA

14 & 15, Krishna Reddy
Colony, Domlur Layout
Bangalore - 560 071
Tel: (+91) 80 4125 4242
Fax: (+91) 80 4125 4181

AUSTRALIA

PO Box 97
St Peters
NSW, 2044
Tel: (+61) 4 0888 2007

JAPAN

Shibuya 3-chome Square Bldg 2F
3-5-16 Shibuya Shibuya-ku
Tokyo, 150-0002
Tel: (+81) 3 6868 7669
Fax: (+81) 3 6730 9539