Partial least squares

Menu:	QCExpert	Predictive methods	Partial least squares

Module PLS regression provides the user with one of the best computational tools for evaluating a pair of multidimensional variables, which is expected to have linear relationship inside one or the other multidimensional variable, and linear relationship between the two variables with each other. This computationally intensive methodology allows to explain and predict one of the variables using other group of variables. The PLS regression method found a large number of applications in the planning and management of quality in manufacturing technology, design and optimization of the characteristics of products in the development of new products, marketing studies, research in the evaluation of experiments, in clinical trials. An example might be modeling the relationship between technological parameters in the production and product quality parameters, or between the chemical composition and physical and biological characteristics. The typical questions of technological practice, which PLS can often answer include:

It has a purity of the raw material any effect on the strength of the product?
What happens if the temperature is increased in the process?
Can we increase the stability of the product by reducing the speed or rotation?
Which process parameters affect the most product strength?
How to set the value of procedural parameters to achieve the desired product characteristics?
What caused the decrease in the parameter?
In what and how subsequent production batches differ?
How to improve the stability / quality?
How to increase the strength / value / competitiveness?
Which input parameters are crucial for the quality?
Which process parameters are crucial for the quality?

Mathematical basics of the PLS regression method

Let us denote $\mathbf{X}(n\times p)$ the matrix (table) of measured values of p variables (columns) with n lines and denote $\mathbf{Y}(n\times q)$ the corresponding table with the same number of lines n but with q variables. Center all columns (subtract column average from each column).To extract maximum information from the p- q- dimensional matrices to a lower dimension space, we decompose **X** and **Y** to the product of the orthogonal matrices $\mathbf{T}(n\times k)$ and $\mathbf{U}(n\times k)$, with coefficient matrices **P** and **Q**

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}$$

while maximizing the correlation between **T** and **U**. Required dimension k, $1 < k \leq \min(p, q)$ is chosen by user, for example, on the basis of the squares sum decrease (scree plot), see below. Noise and irrelevant information contained „litter" in every measured data is swept into residual matrices **E** and **F**. Decomposition **U** = **TB** (where **B** is a square diagonal matrix) give us a tool for computing (estimating) **Y** from **X** but also **X** from **Y**, just by switching the **X** and **Y** data because the model PLS-R is symmetric.

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{Q}^T,$$

T is calculated from the new data **X**, $\mathbf{T} = \mathbf{X}\mathbf{P}^{-}$ ($\mathbf{P}^{-}$ indicate generalized Moore-Penrose pseudoinversion of a rectangular matrix **P**). Furthermore, there is an internal link between **X** and **Y**. By writing $\mathbf{W} = \mathbf{B}\mathbf{Q}^T$, we can rewrite the original pair of relations in the form

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{T}\mathbf{W} + \mathbf{F}$$

so that the data **X** and **Y** are linked through a common scores matrix T, which is actually orthogonalized original matrix **X** in generally smaller number of dimensions, with a maximum extracted information contained in the original **X**, with removed noise (which is moved to the matrix **E**), while the maximum covariance with the matrix **Y** is ensured. Using the relation

$$\mathbf{A} = \mathbf{P}^{-} \mathbf{B} \mathbf{Q}^{T}$$

we can reconstruct coefficients of a classical regression model with multivariate response $\mathbf{Y} = \mathbf{AX}$. Columns $\mathbf{a}_i$ of **A** contain linear coefficients (absolute terms are zero thanks to centering data) of the models $\mathbf{y}_i = \mathbf{X}\mathbf{a}_i$, where $\mathbf{y}_i$ is *i*-th column of matrix **Y**. The coefficients are not usually fully numerically identical to coefficients obtained by classical linear regression. They are generally biased, but shrinked, which means that they have lower variances, and are generally more stable.

As mentioned above, this method is looking for a relationship between the two phenomena described by multidimensional numerical vectors. A typical example is **X** matrix containing measured technological parameters in the production of individual units or batches and matrix **Y** containing the relevant physical parameters of finished products, their deviations from the specifications, etc. Another example is a matrix **X** containing climate and chemical descriptions of the various sites and **Y** matrix with biological parameters of micro-organisms, vegetation and fauna in these locations. There are many other applications in geology, biology, toxicology, chemistry, medicine, psychiatry, behavioral sciences, pharmacology, cosmetics, food, steel industry to name just a few. With PLS prediction we can then obtain estimates of unknown quantities **Y** on the basis of known values **X**.

Model validation

The prediction quality of a particular PLS model can be assessed on the basis of its ability to predict the value of **y** from the value of **x**. This is used in various validation procedures, sometimes called cross-validation. The principle of validation of the model is the same as in the case of neural networks. We „hide“ part of the data before computation of the PLS model. This hidden data are called test or validation data. For the rest of the data, called training data, we calculate parameters of the PLS model. Then the validation are „unhidden“ and used to and check whether the model correctly predicts validation data. Validation must have the same nature and range of values **x**, and therefore the same model as the data used for training. For the validation data we then construct diagnostic charts, which simply conclude whether the model is appropriate for all data. If the model describes well only the training data and not validation data, this usually means that we have little data (rows), or that we have chosen is too large proportion of validation data. A proportion between 10 and 40% of the validation data is usual.

Of course, even advanced PLS regression method is not miraculous and has certain restrictions, which is mainly assumption of linearity of all relationships and normality of error distribution. Along with the ability of prediction and graphical diagnostics, however, it provides a very powerful tool for analysis and prediction of multidimensional variables. In quality control, thanks to its prediction capability, PLS regression is an ideal tool for the quality planning, design of products, optimization of technologies and applied research.

Data and parameters

Module PLS regression needs two data tables, matrix **X** with *p* columns and **Y** with *q* columns selected as the dialog box items *Matrix X* and *Matrix Y*, see Fig. 1. The matrix columns must contain numeric data only, the number of rows must be the same for both **X** and **Y**. Each matrix must contain at least two columns. Columns of matrix **X** must not appear in the matrix **Y**. The limiting dimension *k* can be set by user. If the box *Dimension* is not checked, the maximum dimension is set to $k = \min(p, q)$. It is recommended to perform PLS in maximal dimension first, then optionally we can determine an appropriate value of *k* using the scree plot (see paragraph Graphical output below) and repeat the

computation again with new *k*. If the checkbox *Connect Biplot* is checked, consecutive points of the Biplot will be connected in the order of data in spreadsheet. This can help to follow a possible trajectory of the process. If the checkbox *X-Prediction* is checked, it is necessary to choose the same number of columns as in the field *Independent variable X*. These variables will be used to compute the predicted values of the dependent variable. The X for the prediction must have the same number of columns as the independent variable matrix **X**, but may have a different number of lines (at least 2 lines). A typical example of input matrices and data for prediction is given on Fig. 2.

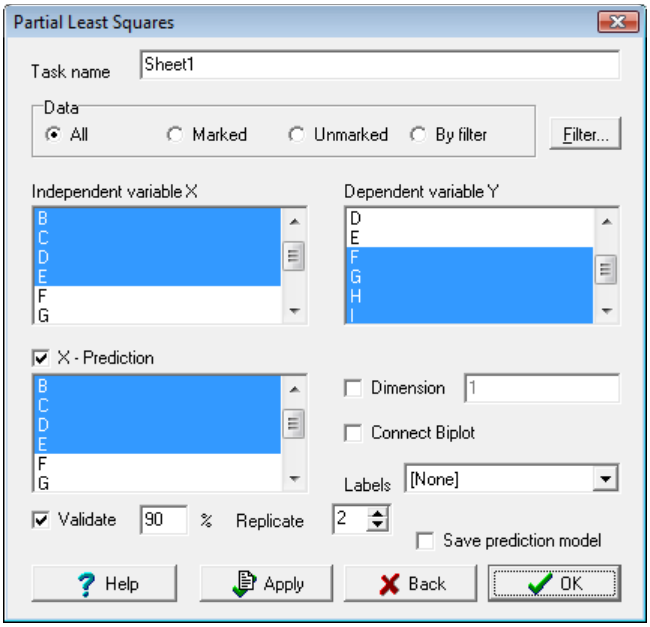


Fig. 1 Dialog box for PLS regression

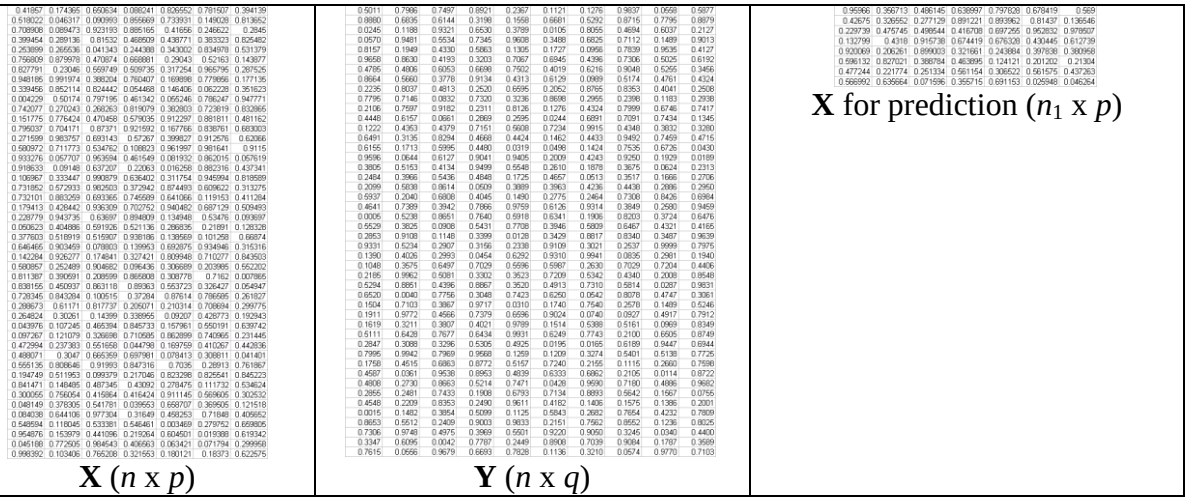


Fig. 2 Typical data for PLS regression

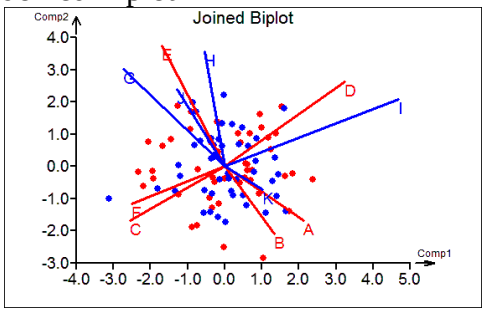
Protocol

Input data	
No of rows	Number of valid rows
No of columns	Number of columns of **X** a **Y** matrices.
Columns	Column names of both input matrices.
Chosen dimension	Dimension *k* for the PLS model chosen in the input dialog window. The dimension must be less or equal to min(*p*, *q*). Scree plot may be used as an

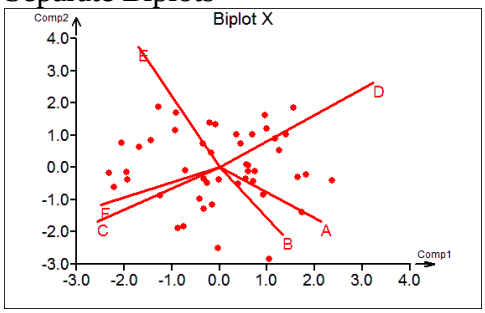
	aid to select suitable *k* if required.
PLS - coefficients, B	Diagonal elements of the matrix **B**.
Explained sum of squares	Table of the squares sum of residuals with growing dimension of the model, $i = 1, \ldots k$, these values are used for constructing scree plot.
No of components	Number of components (dimensions) used for the squares sum.
RSS	Residual square sum value, for 0 components the RSS is the total squares sum without a model.
Percent	% of the RSS
Explained %	(100 – %RSS).
Loadings X, P	Loadings matrix **P**.
Loadings Y, Q	Loadings matrix **Q**.
Regression coefficients, A	Matrix of regression coefficients a_{ij} formally similar to those in the separate classical multiple linear regression models $\mathbf{Y} = \mathbf{XA}$, or $\mathbf{y}_j = \Sigma a_{ij}\mathbf{x}_i$. The coefficient values are generally different from the classical coefficients, since they are based on the orthogonal component regression and therefore they are biased, shortened (with lower standard deviations) and more stable.
Prediction	Predicted values for the data selected in the field X-Prediction in the PLS dialog box. This part of output is not generated unless the checkbox is checked.

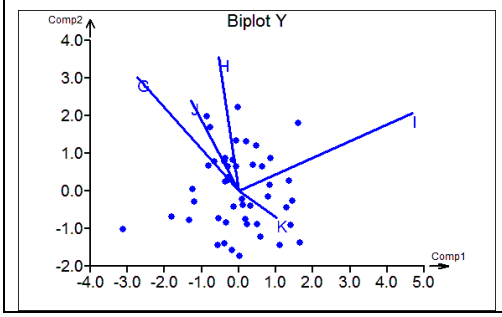
Graphs

Joined Biplot

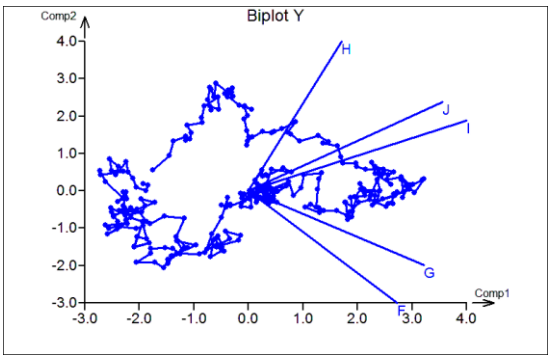


Separate Biplots

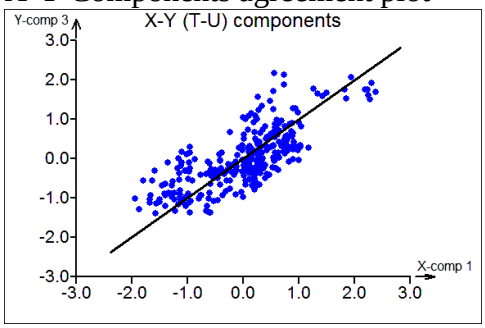
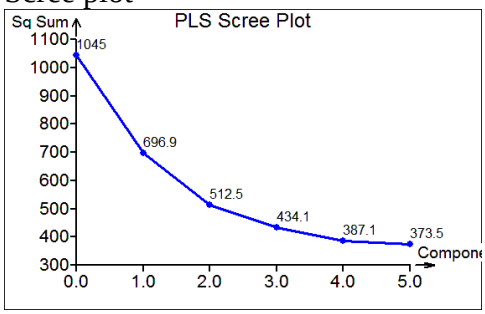
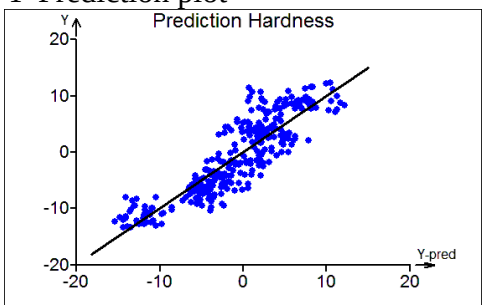
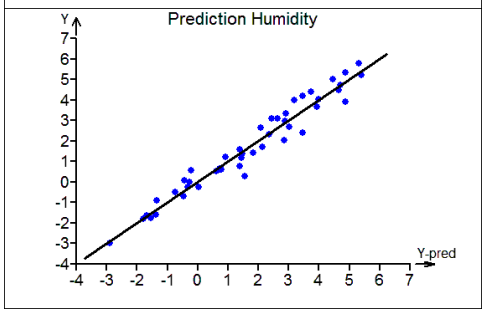
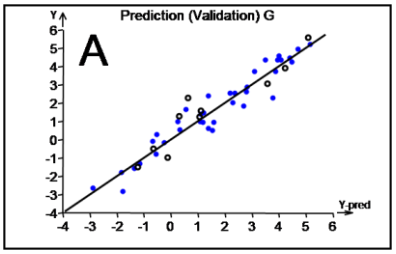
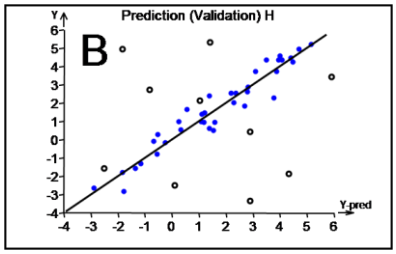


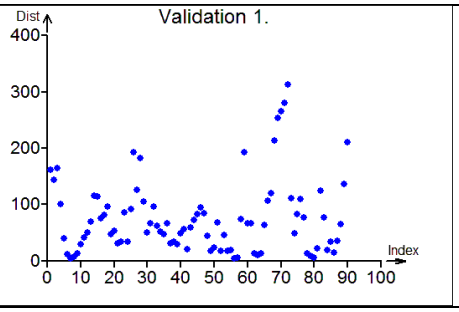


Bi-plot for the matrices X and Y in one plot, matrix X and Y separately. Biplot is a projection of multidimensional data in the plane (the best one in terms of least squares). Points represent rows, rays correspond to columns of the original data. To identify the data rows you can use the labels of points selected in the dialog. Close vector lines (rays) are likely to be mutually correlated. Points located in the direction of a ray will have bigger value of the respective variable. You should be aware that, due to the drastic reduction in the number of dimensions, particularly for larger *p*,*q* this information and guidance is rather a global assessment of the structure and possible links and relationships in the data. If the checkbox Connect Biplot was checked, the points in the plot are linked in chronological order, which sometimes makes it possible to identify and spot trends in the time-series or non-stationary process, as shown on the illustration below.



Connected biplot makes it possible to spot „wandering“ of the

	process in time.
X-Y Components agreement plot 	Plot of agreement between the columns of **T** and **U**. This graph shows the global success of a PLS model fitting. The closer the points to the line, the more successful a PLS model is.
Scree plot 	The effectiveness of the model expressed by reduction of unexplained (residual) sum of squares, depending on the number of factors included (columns of matrix **T** and **U**).
Y-Prediction plot  	This plot expresses compliance of dependent variables and model prediction. The closer are the points to the line, the better the fit. This plot is created for each dependent variable. Some variables can be predicted better, others worse. If the plot does not show a visible trend, the suitable model for this variable was probably not found, a model is not able to predict dependent variable. If Validation checkbox was selected, the validation (test) points in the plot are marked in red (in the example below marked empty circles). In the Fig A below both the training and the validation data are fitted well, showing the model is reliable and the dependence is real. However, if the validation data strongly disagree with other data, as in the Fig B below, the PLS model may be over fitted, describing only the training data, and probably is not usable for the prediction of new values. It is advisable to try to reduce the dimension of the model by entering numbers less than min(p, q) into the field dimension.   Good prediction Poor prediction
Validation residuals plot	Plot Validation is used to assess the quality of prediction of validation data. Unique very remote points may represent outlying measurements. On the Y-axis are Euclidean distances of the data from the model.


Validation 1.
Dist
400
300
200
100
0
Index
0 10 20 30 40 50 60 70 80 90 100