# Partial Least Squares

## 1 Background

Regression by means of projections to latent structures (PLS, also known as partial least squares) is a useful alternative to the linear multiple regression model fitted by "least squares" if:

- the number of $x$-variables is relatively high compared with the number of observations;
- the $x$-variables are correlated;
- there is more than one $y$-variable (*response* variable) and these variables are correlated.

Hence the PLS method is popular in industries that collect correlated data on many $x$-variables, known as *predictors*. For example, multivariate calibration in analytical chemistry; spectroscopy in chemometrics; and quantitative structure activity relationships (QSAR) in drug design.
The PLS method extracts orthogonal linear combinations of predictors, known as *factors*, from the predictor data that explain variance in both the predictor variables *and* the response variable(s).
In general, a PLS analysis consists of the stages:

1. Calculate a PLS model using a high number of factors (more than is likely to be required);
2. Determine the number of factors to include in a fitted model by either:
   - analysing information calculated during the process of extracting factors;
   - calculating a prediction accuracy estimate based on, e.g., cross-validation;
3. Fit the model with the determined number of factors by calculating parameter estimates of the linear regression;
4. Given a set of predictors and responses used to fit a PLS model, and a suitable number of factors to use to calculate parameter estimates, estimate response values to new predictor data.

## 2 Description

Let $X_1$ be the mean-centred $n$ by $m$ data matrix $X$ of $n$ observations on $m$ predictor variables. Let $Y_1$ be the mean-centred $n$ by $r$ data matrix $Y$ of $n$ observations on $r$ response variables.
The first of $k$ factors PLS methods extract from the data predicts both $X_1$ and $Y_1$ by regressing on a column vector of $n$ scores $t_1$:

$$\begin{aligned} \hat{X}_1 &= t_1 p_1^T \\ \hat{Y}_1 &= t_1 c_1^T, \qquad \text{with } t_1^T t_1 = 1, \end{aligned}$$

where the column vectors of $m$ $x$-loadings $p_1$ and $r$ $y$-loadings $c_1$ are calculated in the least squares sense:

$$\begin{aligned} p_1^T &= t_1^T X_1 \\ c_1^T &= t_1^T Y_1. \end{aligned}$$

The $x$-score vector $t_1 = X_1 w_1$ is the linear combination of predictor data $X_1$ that has maximum covariance with the $y$-scores $u_1 = Y_1 c_1$, where the $x$-weights vector $w_1$ is the normalised first left singular vector of $X_1^T Y_1$.
The method extracts subsequent PLS factors by repeating the above process with the residual matrices:

$$\begin{aligned} X_i &= X_{i-1} - \hat{X}_{i-1} \\ Y_i &= Y_{i-1} - \hat{Y}_{i-1}, \quad i = 2, 3, \ldots, k, \end{aligned}$$

and with orthogonal scores:

$$t_i^T t_j = 0, \quad j = 1, 2, \ldots, i-1.$$

Optionally, in addition to being mean-centred, the data matrices $X_1$ and $Y_1$ may be scaled by standard deviations of the variables.
The parameter estimates $B$ for a $l$-factor orthogonal scores PLS model with $m$ predictor variables and $r$ response variables are given by,

$$B = W \left( P^T W \right)^{-1} C^T, \quad B \in \mathbb{R}^{m \times r},$$

where $W$ is the $m$ by $k$ ($\geq l$) matrix of $x$-weights; $P$ is the $m$ by $k$ matrix of $x$-loadings; and $C$ is the $r$ by $k$ matrix of $y$-loadings for a fitted PLS model.
The parameter estimates $B$ are for centred, and possibly scaled, predictor data $X_1$ and response data $Y_1$. Parameter estimates may also be given for the predictor data $X$ and response data $Y$. Variable influence on projection (VIP) statistics, see Wold (1994), can be calculated for the parameter estimates.

## 3 Distribution

The software is available in two forms:

- interested parties may request from NAG a library of functions containing the PLS functionality; full documentation on these functions will also be provided. Please note that the PLS functions are supplied as pre-release versions and NAG reserves the right to change without notice the parameters lists, any names, and the documentation.
- a *Simfit* module. *Simfit* is a Windows GUI-driven data analysis package, see its homepage for more detail. The module is available on request from NAG. In order to use the *Simfit* application you will need a copy of Mark 21 of the NAG Fortran Library, product code FLDLL214ZL, a trial version of which is available from this web-page.

## 4 Reference

Wold S (1994) PLS for Multivariate Linear Modeling *QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry* van de Waterbeemd H (Editor) Verlag-Chemie.