Partial least squares is also known as regression by projection to latent structures, or simply PLS, and it is sometimes useful when a n by r matrix of responses Y, with $r \geq 1$, is observed with a n by m matrix of predictor variables X, with $m > 1$, and one or more of the following conditions may apply:

1. There is no deterministic model to express the r columns of Y as functions of the m columns of the matrix X.
2. The number of columns of X is too large for convenient analysis, or the number of observations n is not significantly greater than the number of predictor variables m, e.g. the rank of X is less than m.
3. The X variables may be correlated and/or the Y variables may be correlated.

The idea behind PLS is to express the X and Y matrices in terms of sets of k factors, with $k \leq m$, derived from the matrices by projection and regression techniques. The X scores would have maximum covariance with the Y scores, and the principal problem is to decide on a sufficiently small dimension l, with $l \leq k$, that would be needed to represent the relationship between Y and X adequately. Having obtained satisfactory expressions for approximating X and Y using these factors, they can then be used to treat X as a training matrix, then predict what new Y would result from a new n by m matrix Z that is expressed in the same variables as the training matrix X. Hence the use of this technique in multivariate calibration, or quantitative structure activity relationships (QSAR).

Data format

From the main SimFIT menu choose [A/Z], open program **linfit**, then select [PLS] and inspect the two test files. The file `g02laf.tf1` contains the following 15 by 15 matrix of X data

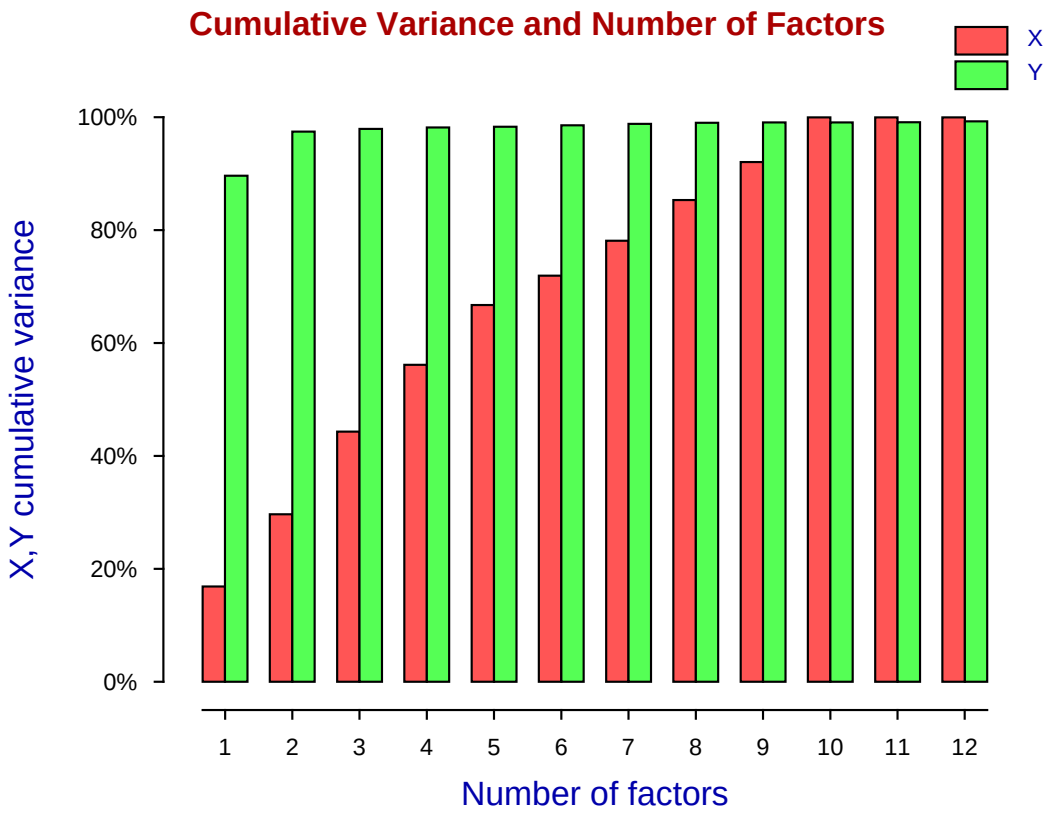
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	1.9607	-1.6324	0.5746	1.9607	-1.6324	0.5740	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	1.9607	-1.6324	0.5746	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	1.9607	-1.6324	0.5746	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	2.8369	1.4092	-3.1398	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.7548	3.6521	0.8524	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	-1.2201	0.8829	2.2253
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	2.4064	1.7438	1.1057	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
2.2261	-5.3648	0.3049	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-4.1921	-1.0285	-0.9801	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-4.9217	1.2977	0.4473	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	2.2261	-5.3648	0.3049	2.2261	-5.3648	0.3049	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.9217	1.2977	0.4473	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.1921	-1.0285	-0.9801	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398

while the test file `g02laf.tf2` contains the following 15 by 1 matrix with Y data.

Y_1
0.00
0.28
0.20
0.51
0.11
2.73
0.18
1.53
-0.10
-0.52
0.40
0.30
-1.00
1.57
0.59

Choosing the number of factors

Select a maximum of 12 factors then plot the cumulative variance plot as in the next figure.



This graph shows that most of the variance in the Y data (green bars) can be explained by only two factors while at least six to eight factors are required to account for a significant proportion of the variance in the X data (red bars).

The main point of PLS analysis is to choose the number of factors that subsequently will be used in the predictive procedures, as the number of factors will be the dimension of the subspace used in the projection of the X data into a space of smaller dimension.

It must be stressed that these factors are like principal components in that every factor is a linear combination of all the variables, and SIMFIT provides several ways to determine the influence of the original variables in the factors.

Contribution of variables to the projection

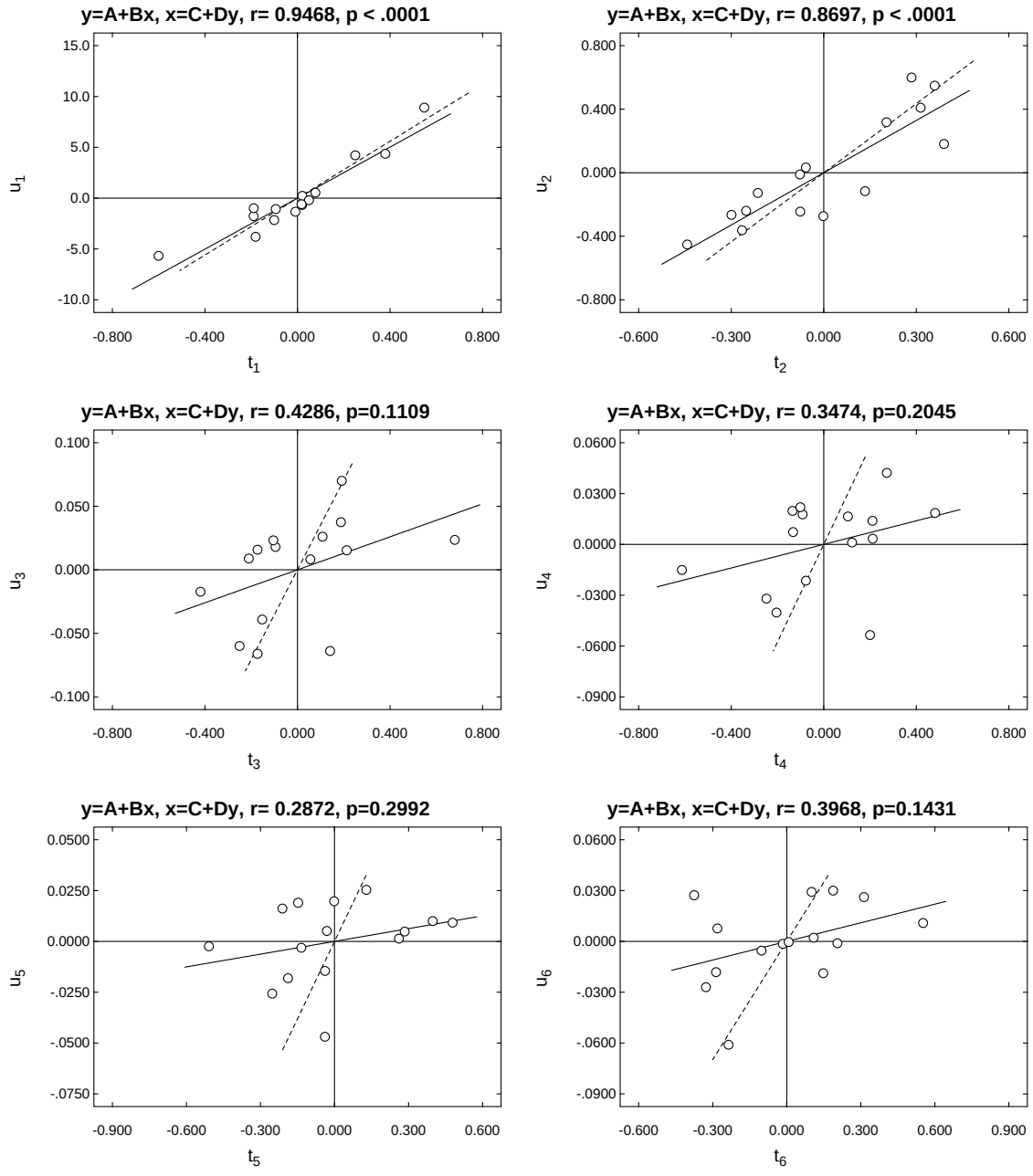
For example, the next table shows the variable influence on projection (VIP) results which indicate that variables 7, 8, 9, 10, and 11 seem to make the most significant contribution to the factors. That is because the sum of squared VIP values equals the number of X variables and a large $VIP(i)$ value indicates that variable i has an important influence on projection.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
VIP	0.611	0.318	0.751	0.504	0.272	0.359	1.577	2.435	1.132	1.223	1.180	0.884	0.213	0.213	0.213
							*	*	*	*	*				

Correlation between scores

Another way to assess the number of factors required to adequately represent the model is to examine the correlation between the scores as, unlike with principal components which are select to maximize variance, PLS factors are selected to maximize covariance between factors.

In the next figure are plotted the successive correlations between the X and Y scores. Each plot shows the best fit linear regression for the u_i i.e. Y scores on the t_i i.e. X scores, and also the best fit linear regression of the X scores on the Y scores, together with the correlation coefficients r and and significance levels p.



Clearly the scores corresponding to the first two factors are highly correlated, but thereafter the correlation is very weak.

Predicting Y given new X

Once a model has been selected with an appropriate number of factors, a set of parameters can be calculated to express Y as a function of the X matrix. In other words, the original X and Y matrices can be regarded as a training set, then a new X data matrix can be input to predict a new Y matrix of responses.

For instance, select a model with 7 factors and then read in the default test file `g02laf.tf3` as a Z matrix which has the following data.

Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	1.9607	-1.6324	0.5746	1.9607	-1.6324	0.574	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	1.9607	-1.6324	0.5746	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	1.9607	-1.6324	0.5746	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	2.8369	1.4092	-3.1398	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.7548	3.6521	0.8524	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	-1.2201	0.8829	2.2253
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	2.4064	1.7438	1.1057	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
2.2261	-5.3648	0.3049	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-4.1921	-1.0285	-0.9801	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-4.9217	1.2977	0.4473	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	2.2261	-5.3648	0.3049	2.2261	-5.3648	0.3049	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.9217	1.2977	0.4473	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398
-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.1921	-1.0285	-0.9801	0.0744	-1.7333	0.0902	2.8369	1.4092	-3.1398

This can then be used to predict a new set of responses, say $\tilde{Y}$, with the following values.

$\tilde{Y}$
0.1408592
0.2991056
0.1383467
0.2967188
0.1304868
2.6278021
0.1862098
1.4820439
-0.1140479
-0.4532927
0.4052814
0.2922055
-1.0294603
1.7814955
0.5962458

When using this technique it is important to realize that the training matrices and the prediction matrices must be centered and scaled using exactly the same mean vectors and scaling factors otherwise biased predictions will result. That is why it is best to submit the data without centering and scaling then SIMFIT will automatically use the centering and scaling from the training set with the prediction data and will map the predicted results back into the original coordinates. This will be explained with more detail in the theoretical section that follows.

Theory

The idea behind PLS is to express the n by m matrix X and n by r matrix Y in terms of sets of k factors, with $k \leq m$, derived from the matrices by projection and regression techniques. The X scores would have maximum covariance with the Y scores, and the principal problem is to decide on a sufficiently small dimension l, with $l \leq k$, that would be needed to represent the relationship between Y and X adequately.

If X_1 is the centered matrix obtained from X by subtracting the X column means, and Y_1 is obtained from Y by subtracting the Y column means, then the first factor is obtained by regressing on a column vector of n normalized scores t_1, as in

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

where the column vectors of m x-loadings p_1 and r y-loadings c_1 are calculated by least squares, i.e.

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

The x-score vector $t_1 = X_1 w_1$ is the linear combination of 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 normalized first left singular vector of $X_1^T Y_1$. The further $k - 1$ orthogonal factors are then calculated successively using

$$\begin{aligned} X_i &= X_{i-1} - \hat{X}_{i-1} \\ Y_i &= Y_{i-1} - \hat{Y}_{i-1}, \ i = 2, 3, \ldots, k \\ t_i^T t_j &= 0, \ j = 1, 2, \ldots, i-1. \end{aligned}$$

Once a set of k factors has been calculated, these can be used to generate the parameter estimates necessary to predict a new Y matrix from a Z matrix, given the original training matrix X. Usually k would be an upper limit on the number of factors to consider, and the m by r parameter estimates matrix B required for l factors, where $l \leq k$, would be given by

$$B = W(P^T W)^{-1} C^T.$$

Here W is the m by k 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. Note that B calculated in this way is for the centered matrices X_1 and Y_1, but parameter estimates appropriate for the original data are also calculated.

Before proceeding further it is important to emphasize a complication which can arise when predicting a new Y matrix using the parameter estimates. In most multivariate techniques it is immaterial whether the data are scaled and centered before submitting a sample for analysis, or whether the data are scaled and centered internally by the software. In the case of PLS, the Y predicted will be incorrect if the data are centered and scaled independently before analysis, but then the Z matrix for prediction is centered and scaled using its own column means and variances.

So there are just two ways to make sure PLS predicts correctly.

1. You can submit X and Y matrices that are already centered and scaled, but then you must submit a Z matrix that has not been centered and scaled using its own column means and standard deviations, but one that has been processed by subtracting the original X column means and scaled using the original X column standard deviations.

2. Do not center or scale any data. Just submit the original data for analysis, request automatic centering and scaling if necessary, but allow the software to then center and scale internally.

As the first method is error prone and will predict scaled and centered predictions, which could be confusing, the advice to PLS users would be:

> *Do not center or scale any training sets, or Z-data for predicting new Y, before PLS analysis.*
> *Always submit raw data and allow the software to perform centering and scaling.*
> *That way predictions will be in coordinates corresponding to the original Y-coordinates.*

Several techniques are available to decide how many factors l out of the maximum calculated k should be selected when using a training set for prediction.

For instance, the previous figure displaying cumulative variance was obtained by using test file `g02laf.tf1` with 15 rows and 15 columns as the source of X prediction data, and test file `g02laf.tf2` with 15 rows and just 1 column as the source of Y response data, then fitting a PLS model with up to a maximum of $k = 12$ factors. It illustrates how the cumulative percentage of variance in X and a column of Y is accounted for the factor model as the number of factors is steadily increased. It is clear that two factors are sufficient to account for the variance of the single column of Y in this case but more, probably about 6 to 8, are required to account for the variance in the X matrix, i.e. we should choose $6 \leq l \leq 8$.

Alternatively, the previous figures showing the successive correlations between the X and Y scores should be inspected. Each plot shows the best fit linear regression for the u_i i.e. Y scores on the t_i i.e. X scores, and also the best fit linear regression of the X scores on the Y scores, together with the correlation coefficients r and and significance levels p. Clearly the scores corresponding to the first two factors are highly correlated, but thereafter the correlation is very weak.

Note that the PLS model can be summarized as follows

$$\begin{aligned} X &= \bar{X} + TP^T + E \\ Y &= \bar{Y} + UC^T + F \\ U &= T + H \end{aligned}$$

where E, F, and H are matrices of residuals.

So the SIMFIT PLS routines also allow users to study such residuals, to see how closely the fitted model predicts the original Y data for increasing numbers of factors before the number of factors to be used routinely is decided. Various tests for goodness of fit can be derived from these residuals and, in addition, variable influence on projection (VIP) statistics can also be calculated.