Analytica Chimica Acta, 185 (1986) 1—17
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

PARTIAL LEAST-SQUARES REGRESSION: A TUTORIAL

PAUL GELADI*[a] and BRUCE R. KOWALSKI

Laboratory for Chemometrics and Center for Process Analytical Chemistry, Department of Chemistry, University of Washington, Seattle, WA 98195 (U.S.A.)

(Received 15th July 1985)

SUMMARY

A tutorial on the partial least-squares (PLS) regression method is provided. Weak points in some other regression methods are outlined and PLS is developed as a remedy for those weaknesses. An algorithm for a predictive PLS and some practical hints for its use are given.

The partial least-squares regression method (PLS) is gaining importance in many fields of chemistry; analytical, physical, clinical chemistry and industrial process control can benefit from the use of the method. The pioneering work in PLS was done in the late sixties by H. Wold in the field of econometrics. The use of the PLS method for chemical applications was pioneered by the groups of S. Wold and H. Martens in the late seventies after an initial application by Kowalski et al. [1]. In spite of the large amount of literature that emerged from these groups, most articles describing PLS give algorithms and theory that are incomplete and often difficult to understand. Two recent articles [2, 3] show that PLS is a good alternative to the more classical multiple linear regression and principal component regression methods because it is more robust. Robust means that the model parameters do not change very much when new calibration samples are taken from the total population.

This article is meant as a tutorial. The reader is referred to texts on linear algebra [4, 5] if needed. The two most complete articles on PLS available at present are by S. Wold et al. [4, 6]. The nomenclature used in Kowalski [6] will be used here. Furthermore, all vectors will be column vectors. The corresponding row vectors will be designated as transposed vectors. The notation will be kept as rigorous as possible. Table 1 lists the notation used. The paragraphs on multiple linear regression, principal component analysis and principal component regression are included because they are necessary for a good understanding of PLS. They do not represent a complete treatment of these subjects.

---

[a]Present address: Chemometrics Group, Department of Organic Chemistry, Umeå University, S-901 87 Umeå, Sweden.

0003-2670/86/$03.50

TABLE 1

Symbols

$\| \ \|$	the Fröbenius or Euclidian norm
i	a dummy index for counting samples (objects)
j	a dummy index for counting independent (x) variables
k	a dummy index for counting dependent (y) variables
h	a dummy index for counting components or factors
n	the number of samples in the calibration (training) set
m	the number of independent (x) variables
p	the number of dependent (y) variables
a	the number of factors used ($\leq$ rank of $\mathbf{X}$)
r	the number of samples in a prediction (test) set
$\mathbf{x}$	a column vector of features for the independent variables (size $m \times 1$)
$\mathbf{y}$	a column vector of features for the dependent variables (size $p \times 1$)
$\mathbf{X}$	a matrix of features for the independent variables (size $n \times m$)
$\mathbf{Y}$	a matrix of features for the dependent variables (size $n \times p$)
$\mathbf{b}$	a column vector of sensitivities for the MLR method (size $m \times 1$)
$\mathbf{B}$	a matrix of sensitivities for the MLR method (size $m \times p$)
$\mathbf{t}_h$	a column vector of scores for the X block, factor h (size $n \times 1$)
$\mathbf{p}'_h$	a row vector of loadings for the X block, factor h (size $1 \times m$)
$\mathbf{w}'_h$	a row vector of weights for the X block, factor h (size $1 \times m$)
$\mathbf{T}$	the matrix of X scores (size $n \times a$)
$\mathbf{P}'$	the matrix of X loadings (size $a \times m$)
$\mathbf{u}_h$	a column vector of scores for the Y block, factor h (size $n \times 1$)
$\mathbf{q}'_h$	a row vector of loadings for the Y block, factor h (size $1 \times p$)
$\mathbf{U}$	the matrix of Y scores (size $n \times a$)
$\mathbf{Q}'$	the matrix of Y loadings (size $a \times p$)
$\mathbf{M}_h$	a rank 1 matrix, outer product of $\mathbf{t}_h$ and $\mathbf{p}'_h$ (size $n \times m$)
$\mathbf{E}_h$	the residual of X after subtraction of h components (size $n \times m$)
$\mathbf{F}_h$	the residual of Y after subtraction of h components (size $n \times p$)
b_h	the regression coefficient for one PLS component
$\mathbf{I}_n$	the identity matrix of size $n \times n$
$\mathbf{I}_m$	the identity matrix of size $m \times m$

Calibration (training) and prediction (test) steps

Chemical analysis usually consists of two steps. First, the characteristics of a method or instrument are investigated and an attempt is made to find a model for its behavior (a model is a relationship $Y = f(X)$ between two groups of variables, often called dependent Y and independent X). This is the calibration or training step. The data set used for this step is called a calibration or training set. The model parameters are called regression coefficients or sensitivities. The second step is the one in which the independent variables are obtained for one or more samples. These are used together with the sensitivities to predict values for the dependent variables. This is the prediction or test step. The data set used in this step is the prediction or test set.

The terms dependent block and independent block are introduced for the blocks of dependent and independent variables, respectively.

Mean-centering and scaling of variables

Before the model is developed, it is convenient to tailor the data in the calibration set in order to make the calculations easier. For ease of explanation, the values for each variable are used in the mean-centered form. The average value for each variable is calculated from the calibration set and then subtracted from each corresponding variable. In the rest of the text, all variables, both dependent and independent, are assumed to be mean-centered.

There are also different ways of scaling the variables. It should be pointed out that the dependent variables and the independent ones can be scaled differently because the sensitivities absorb the differences in scaling. There are essentially three ways of treating variables. In one, no scaling is needed when all the variables in a block are measured in the same units, as in spectrometry. In the second, variance scaling is used when the variables in a block are measured in different units (e.g., ppm, %, km); scaling is accomplished by dividing all the values for a certain variable by the standard deviation for that variable, so that the variance for every variable is unity. Thirdly, one can decide that certain variables are of less importance and hence should not influence the model very much; so they are given a smaller weight.

An illustration of scaling and mean centering is given in Fig. 1. In the further text, all variables are assumed to have some type of scaling, whichever is considered to be most appropriate.

MULTIPLE LINEAR REGRESSION (MLR)

The multiple linear regression (MLR) problem can be stated as follows. Features are measured for m variables x_j ($j = 1 - m$) and for a variable y with the goal to establish a linear (or first-order) relationship between them. This can be represented mathematically as

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_mx_m + e \qquad (1a)$$

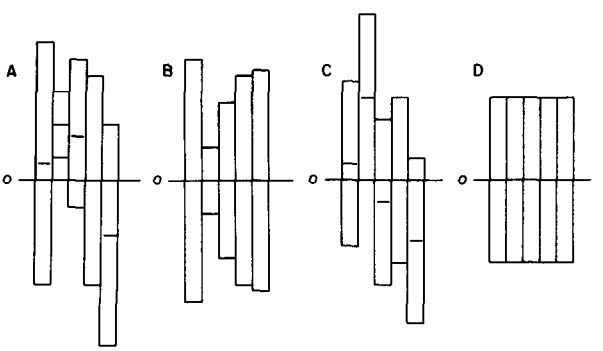


Fig. 1. Data preprocessing. The data for each variable are represented by a variance bar and its center. (A) Most raw data look like this. (B) The result after mean-centering only. (C) The result after variance-scaling only. (D) The result after mean-centering and variance-scaling.

$$y = \sum_{j=1}^{m} b_j x_j + e \tag{1b}$$

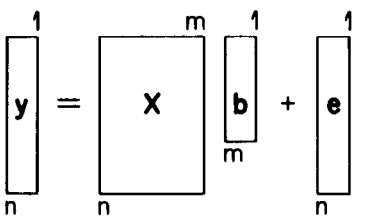
$$y = \mathbf{x}'\mathbf{b} + e \tag{1c}$$

In Eqn. 1(a), the x_j are called independent variables and y is the dependent variable, the b_j's are sensitivities and e is the error or residual. In Eqn. 1(c), y is a scalar, **b** is a column vector and $\mathbf{x}'$ is a row vector.

Equation 1 describes multilinear dependencies for only one sample. If one gets n samples, the y_i ($i = 1 - n$) can be written as a column vector **y**, **b** remains the same and the vectors, $\mathbf{x}_j'$, form the rows of a matrix **X**:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{2}$$

For a better understanding of these matrix equations, they are also given in graphical representation:



In this case, n is the number of samples and m the number of independent variables.

It is now possible to distinguish among three cases.

(1) $m > n$. There are more variables than samples. In this case, there is an infinite number of solutions for b, which all fit the equation. This is not what is wanted.

(2) $m = n$. The numbers of samples and of variables are equal. This situation may not be encountered often in practical situations. However, it gives a unique solution for **b** provided that **X** has full rank. This allows us to write

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} = 0 \tag{3}$$

e is called the residual vector. In this case, it is a vector of zeroes: 0.

(3) $m < n$. There are more samples than variables. This does not allow an exact solution for **b**. But one can get a solution by minimizing the length of the residual vector **e** in the following equation:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} \tag{4}$$

The most popular method for doing this is called the "least-squares method". The least-squares solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{5}$$

(Complete explanations are available elsewhere [5, 7, 8].) Equation 5 gives a hint towards the most frequent problem in MLR: the inverse of $\mathbf{X}'\mathbf{X}$ may

not exist. Collinearity, zero determinant and singularity are all names for the same problem. A good description of this situation is available [9].

At this point, it might appear that there always have to be at least as many samples as variables, but there are other ways to formulate this problem. One of them is to delete some variables in the case $m > n$. Many methods exist for choosing which variables to delete [7, 8].

Multiple linear regression with more than one dependent variable

A popular misconception is that MLR is only possible for one dependent variable. This is the case that is almost always found in textbooks. Also, most software packages run MLR in this way. It is easy to extend MLR for more dependent variables. The example given here is for two variables, but extension to more than two is straightforward.

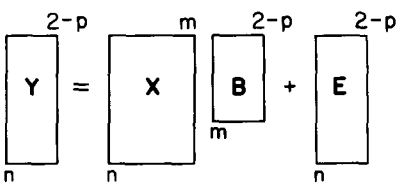
Suppose there are two dependent variables, $\mathbf{y}_1$ and $\mathbf{y}_2$. In this case, one can simply write two MLR's and find two vectors of sensitivities, $\mathbf{b}_1$ and $\mathbf{b}_2$:

$$\mathbf{y}_1 = \mathbf{X}\mathbf{b}_1 + \mathbf{e}_1; \quad \mathbf{y}_2 = \mathbf{X}\mathbf{b}_2 + \mathbf{e}_2 \tag{6}$$

But one can then put $\mathbf{y}_1$ and $\mathbf{y}_2$ side by side in a $n \times 2$ matrix and do the same for $\mathbf{b}_1$ and $\mathbf{b}_2$ and $\mathbf{e}_1$ and $\mathbf{e}_2$. So one gets

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \tag{7}$$

where $\mathbf{Y} = (\mathbf{y}_1\mathbf{y}_2)$, $\mathbf{B} = (\mathbf{b}_1\mathbf{b}_2)$ and $E = (\mathbf{e}_1\mathbf{e}_2)$. A more graphical representation for $2 - p$ dependent variables is



This is the general case that will be referred to in the further text.

Summary: MLR

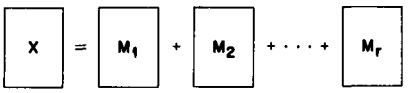
— For $m > n$, there is no unique solution unless one deletes independent variables.
— For $m = n$, there is one unique solution.
— For $m < n$, a least-squares solution is possible. For $m = n$ and $m < n$, the matrix inversion can cause problems.
— MLR is possible with more than one dependent variable.

PRINCIPAL COMPONENT ANALYSIS (PCA): NIPALS METHOD

Principal component analysis (PCA) is a method of writing a matrix $\mathbf{X}$ of rank r as a sum of r matrices of rank 1:

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3 + \ldots + \mathbf{M}_r \tag{8}$$

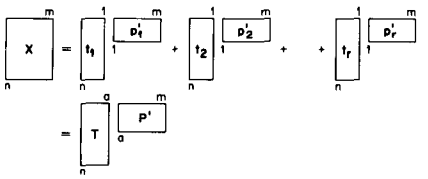
or in graphical representation:



(Rank is a number expressing the true underlying dimensionality of a matrix.) These rank 1 matrices, $\mathbf{M}_h$, can all be written as outer products of two vectors, a score $\mathbf{t}_h$ and a loading $\mathbf{p}_h'$:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \ldots\ldots\ldots + \mathbf{t}_a\mathbf{p}_a' \tag{9}$$

or the equivalent $\mathbf{X} = \mathbf{TP}'$ ($\mathbf{P}'$ is made up of the $\mathbf{p}'$ as rows and $\mathbf{T}$ of the $\mathbf{t}$ as columns) or graphically:



To illustrate what the $\mathbf{t}_h$ and $\mathbf{p}_h'$ mean, an example for two variables, in the two-dimensional plane, is shown in Fig. 2A. Extension to more dimensions is easy but difficult to show on paper. For the example in Fig. 2A, the principal component is the line of best fit for the data points that are shown in Fig. 2B. Best fit means that the sum of squares of x_1 and x_2 residuals is minimized. This is also the average of both regression lines. It goes from $-\infty$ to $+\infty$. The $\mathbf{p}_h'$ is a 1×2 row vector. Its elements, p_1 and p_2, are the direction cosines, or the projections of a unit vector along the principal component on the axes of

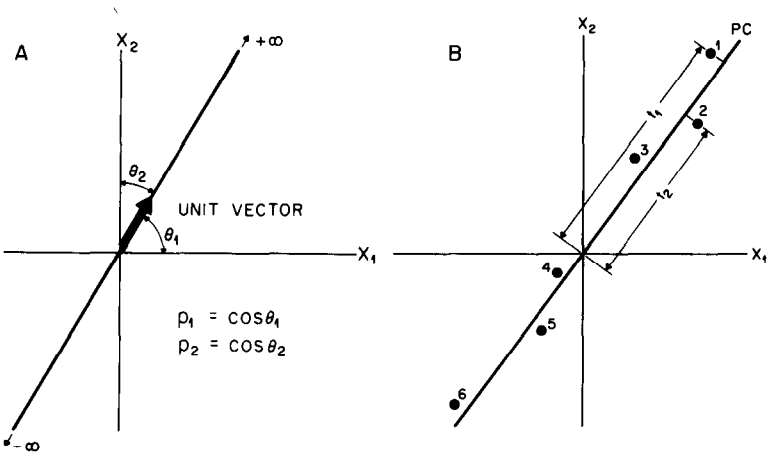


Fig. 2. A principal component in the case of two variables: (A) loadings are the angle cosines of the direction vector; (B) scores are the projections of the sample points (1—6) on the principal component direction. Note that the data are mean-centered.

the plot. The scores vector, t_h, is a $n \times 1$ column vector. Its elements are the coordinates of the respective points on the principal component line (Fig. 2B). For this example, it can easily be understood why one wants the length of the p'_h to be one $[\cos(\theta_1)^2 + \cos(\theta_2)^2 = \cos(\theta_1)^2 + \sin(\theta_1)^2 = 1]$; similar rules exist for more than two dimensions.

Generally, what one wants is an operator that projects the columns of X onto a single dimension and an operator that projects the rows of X onto a single dimension (see Fig. 3). In the first case, each column of X is represented by a scalar; in the second case, each row of X is represented by a scalar. In the rest of this section it will be shown that these operators are of a very simple nature.

Nonlinear iterative partial least squares (NIPALS) does not calculate all the principal components at once. It calculates t_1 and p'_1 from the X matrix. Then the outer product, $t_1 p'_1$, is subtracted from X and the residual E_1 is calculated. This residual can be used to calculate t_2 and p'_2:

$$E_1 = X - t_1 p'_1 \quad E_2 = E_1 - t_2 p'_2 \ldots$$

$$E_h = E_{h-1} - t_h p'_h \ldots \quad E_{\tanh(X)} = 0 = E_{\tanh(X)-1} - t_{\tanh(X)} p'_{\tanh(X)} \tag{10}$$

The NIPALS algorithm is as follows:

(1) take a vector x_j from X and call it t_h: $t_h = x_j$ (11)

(2) calculate p'_h: $p'_h = t'_h X / t'_h t_h$ (12)

(3) normalize p'_h to length 1: $p'_{h\,\text{new}} = p'_{h\,\text{old}} / \| p'_{h\,\text{old}} \|$ (13)

(4) calculate t_h: $t_h = X p_h / p'_h p_h$ (14)

(5) compare the t_h used in step 2 with that obtained in step 4. If they are the same, stop (the iteration has converged). If they still differ, go to step 2.

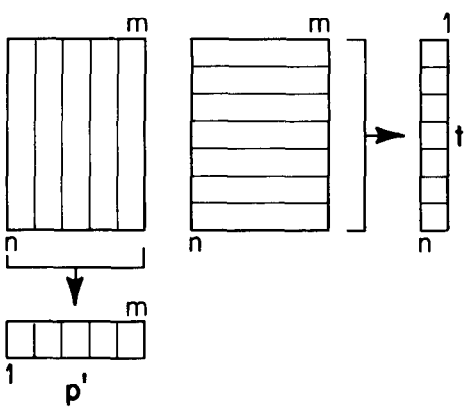


Fig. 3. Scores and loadings are obtained by projecting X into vectors. Loadings. each column of X is projected into an element of the vector p'. Scores: each row of X is projected into an element of the vector t.

(Note that after the first component is calculated, **X** in steps 2 and 4 has to be replaced by its residual.)

An explanation of how NIPALS works can be seen when one realizes that $\mathbf{t}_h'\mathbf{t}_h$ in Eqn. 12, $\|\mathbf{p}_h'\|$ in Eqn. 13 and $\mathbf{p}_h'\mathbf{p}_h$ in Eqn. 14 are scalars. These scalar constants are best combined in one general constant C. Then one can substitute Eqn. 12 into 14: $\mathbf{t}_h = \mathbf{X}\mathbf{p}_h$ and $\mathbf{p}_h' = \mathbf{t}_h'\mathbf{X}$ give $C\mathbf{p}_h' = (\mathbf{X}\mathbf{p}_h)'\mathbf{X}$, or $C\mathbf{p}_h' = \mathbf{p}_h'\mathbf{X}'\mathbf{X}$, or $(C\mathbf{I}_m - \mathbf{X}'\mathbf{X})\mathbf{p}_h = 0$; or one can substitute Eqn. 14 into 12 and get $(C'\mathbf{I}_n - \mathbf{X}\mathbf{X}')\mathbf{t}_h = 0$. These are the eigenvalue/eigenvector equations for $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ as used in the classical calculation. ($\mathbf{I}_n$ is the identity matrix of size $n \times n$; $\mathbf{I}_m$ is that of size $m \times m$.) The classical eigenvector and eigenvalue theory is well described by Strang [10].

It has been shown that on convergence, the NIPALS solution is the same as that calculated by the eigenvector formulae. The NIPALS method is convenient for microcomputers; it is also necessary for a good understanding of PLS. Otherwise, it does not matter what method one uses. In practical situations, NIPALS usually converges; in the case of non-convergence, two or more very similar eigenvalues exist. Then it does not matter which combination or rotation of eigenvectors is chosen. The reader is referred to Mardia et al. [5] for a more detailed discussion of PCA.

Summary: PCA

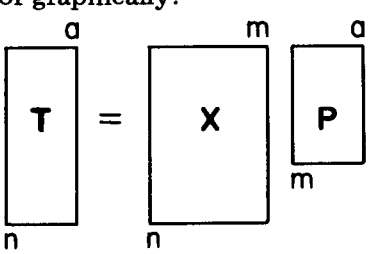
— A data matrix **X** of rank r can be decomposed to a sum of r rank 1 matrices.
— These rank 1 matrices are outer products of vectors called scores and loadings.
— The scores and loadings can be calculated pair-by-pair by an iterative procedure.

PRINCIPAL COMPONENT REGRESSION (PCR)

The results from the section on PCA can be used to explain the principal component transformation of a data matrix **X**. This is a representation of **X** as its scores matrix **T** (where dimensions having small eigenvalues are excluded). The transformation is

$$\mathbf{T} = \mathbf{X}\mathbf{P} \; (= \mathbf{T}\mathbf{P}'\mathbf{P} = \mathbf{T}\mathbf{I}_n) \tag{15}$$

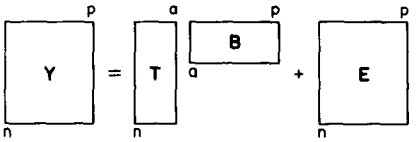
or graphically:



So now the MLR formula can be written as

$$\mathbf{Y} = \mathbf{T}\mathbf{B} + \mathbf{E} \text{ (solution: } \hat{\mathbf{B}} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}) \tag{16}$$

or graphically:



The variables of X are replaced by new ones that have better properties (orthogonality) and also span the multidimensional space of X. The inversion of T'T should give no problem because of the mutual orthogonality of the scores. Score vectors corresponding to small eigenvalues can be left out in order to avoid collinearity problems from influencing the solution [9].

PCR solves the collinearity problem (by guaranteeing an invertible matrix in the calculation of $\hat{\mathbf{B}}$) and the ability to eliminate the lesser principal components allows some noise (random error) reduction. However, PCR is a two-step method and thereby has the risk that useful (predictive) information will end up in discarded principal components and that some noise will remain in the components used for regression.

Detailed information on PCR is given by Mardia et al. [5] and Draper and Smith [7]. Gunst and Mason [8] give a slightly different definition of PCR.

Summary: PCR

— A data matrix can be represented by its score matrix.
— A regression of the score matrix against one or several dependent variables is possible, provided that scores corresponding to small eigenvalues are omitted.
— This regression gives no matrix inversion problems; it is well conditioned.

PARTIAL LEAST-SQUARES REGRESSION

Model building

The PLS model is built on the properties of the NIPALS algorithm. As mentioned in the PCR section, it is possible to let the score matrix represent the data matrix. A simplified model would consist of a regression between the scores for the X and Y block. The PLS model can be considered as consisting of outer relations (X and Y block individually) and an inner relation (linking both blocks).

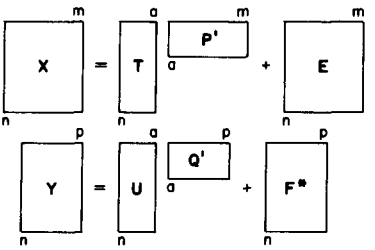
The outer relation for the X block (cf. PCA section) is

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \sum \mathbf{t}_h \mathbf{p}_h' + \mathbf{E} \tag{17}$$

One can build the outer relation for the Y block in the same way:

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F}^* = \sum \mathbf{u}_h \mathbf{q}_h' + \mathbf{F}^* \tag{18}$$

Graphically, Eqns. 17 and 18 can be shown as



The summations are from 1 to a. One can describe all the components and thus make $\mathbf{E} = \mathbf{F}^* = 0$ or not. How and why this is done is discussed below. It is the intention to describe Y as well as is possible and hence to make $\|\mathbf{F}^*\|$ as small as possible and, at the same time, get a useful relation between X and Y. The inner relation can be made by looking at a graph of the Y block score, u, against the X block score, t, for every component (Fig. 4). The simplest model for this relation is a linear one:

$$\hat{\mathbf{u}}_h = b_h \mathbf{t}_h \tag{19}$$

where $b_h = \mathbf{u}_h' \mathbf{t}_h / \mathbf{t}_h' \mathbf{t}_h$. The b_h play the role of the regression coefficients, b, in the MLR and PCR models.

This model, however, is not the best possible. The reason is that the principal components are calculated for both blocks separately so that they have a weak relation to each other. It would be better to give them information about each other so that slightly rotated components result which lie closer to the regression line of Fig. 4.

Over-simplified model: 2x PCA. An over-simplified model can be written in algorithmic form as in the NIPALS section.

For the X block: (1) take $\mathbf{t}_{start}$ = some $\mathbf{x}_j$; (2) $\mathbf{p}' = \mathbf{t}'\mathbf{X}/\mathbf{t}'\mathbf{t}$ $(= \mathbf{u}'\mathbf{X}/\mathbf{u}'\mathbf{u})$; (3) $\mathbf{p}'_{new} = \mathbf{p}'_{old}/\|\mathbf{p}'_{old}\|$; (4) $\mathbf{t} = \mathbf{X}\mathbf{p}/\mathbf{p}'\mathbf{p}$; (5) compare t in steps 2 and 4 and if they are equal stop, else go to 2.

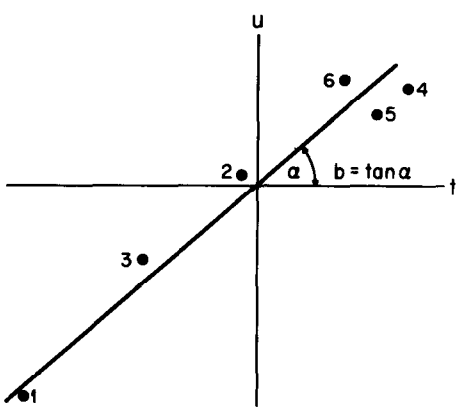


Fig. 4. The inner relation. A linear regression of u against t. Note that the data are mean-centered.

For the Y block, (1) take $\mathbf{u}_{start}$ = some $\mathbf{y}_j$; (2) $\mathbf{q}' = \mathbf{u}'\mathbf{Y}/\mathbf{u}'\mathbf{u}$ (= $\mathbf{t}'\mathbf{Y}/\mathbf{t}'\mathbf{t}$); (3) $\mathbf{q}'_{new} = \mathbf{q}'_{old}/\|\mathbf{q}'_{old}\|$; (4) $\mathbf{u} = \mathbf{Y}\mathbf{q}/\mathbf{q}'\mathbf{q}$; (5) compare $\mathbf{u}$ in steps 2 and 4 and if they are equal stop, else go to 2.

Improving the inner relation: exchange of scores. The above relations are written as completely separated algorithms. The way each can get information about the other is to let $\mathbf{t}$ and $\mathbf{u}$ change place in step 2. (Note the parts in parentheses in this step.) Thus, the two algorithms can be written in sequence: (1) take $\mathbf{u}_{start}$ = some $\mathbf{y}_j$; (2) $\mathbf{p}' = \mathbf{u}'\mathbf{X}/\mathbf{u}'\mathbf{u}$ ($\mathbf{w}' = \mathbf{u}'\mathbf{X}/\mathbf{u}'\mathbf{u}$); (3) $\mathbf{p}'_{new} = \mathbf{p}'_{old}/\|\mathbf{p}'_{old}\|$ ($\mathbf{w}'_{new} = \mathbf{w}'_{old}/\|\mathbf{w}'_{old}\|$); (4) $\mathbf{t} = \mathbf{X}\mathbf{p}/\mathbf{p}'\mathbf{p}$ ($\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}'\mathbf{w}$); (5) $\mathbf{q}' = \mathbf{t}'\mathbf{Y}/\mathbf{t}'\mathbf{t}$; (6) $\mathbf{q}'_{new} = \mathbf{q}'_{old}/\|\mathbf{q}'_{old}\|$; (7) $\mathbf{u} = \mathbf{Y}\mathbf{q}/\mathbf{q}'\mathbf{q}$; (8) Compare the $\mathbf{t}$ in step 4 with the one in the preceding iteration step. If they are equal (within a certain rounding error), stop; else go to 2. (In the case for which the Y block has only one variable, steps 5—8 can be omitted by putting $q = 1$.)

This algorithm usually converges very quickly to give rotated components for X and Y block.

Obtaining orthogonal X block scores. There is still a problem; the algorithm does not give orthogonal $\mathbf{t}$ values. The reason is that the order of calculations that was used for the PCA has been changed. Therefore, the $\mathbf{p}'$ are replaced by weights $\mathbf{w}'$ (see formulas in parentheses in previous subsection). An extra loop can be included after convergence to get orthogonal $\mathbf{t}$ values:

$$\mathbf{p}' = \mathbf{t}'\mathbf{X}/\mathbf{t}'\mathbf{t} \tag{20}$$

With $\mathbf{p}'_{new} = \mathbf{p}'_{old}/\|\mathbf{p}'_{old}\|$, it becomes possible to calculate the new $\mathbf{t}$: $\mathbf{t} = \mathbf{X}\mathbf{p}/\mathbf{p}'\mathbf{p}$, but this turns out to be just a scalar multiplication with the norm of the $\mathbf{p}'$ in Eqn. 20: $\mathbf{t}_{new} = \mathbf{t}_{old}\|\mathbf{p}'_{old}\|$. Orthogonal $\mathbf{t}$ values are not absolutely necessary, but they make the comparison with PCR easier. One must give the same rescaling to the weights, $\mathbf{w}'$, if the prediction is to be made without error: $\mathbf{w}'_{new} = \mathbf{w}'_{old}\|\mathbf{p}'_{old}\|$. Then $\mathbf{t}$ can be used for the inner relation as in Eqn. 19, and the residuals can be calculated from $\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}'_1$ and $\mathbf{F}_1^* = \mathbf{Y} - \mathbf{u}_1\mathbf{q}'_1$. In general,

$$\mathbf{E}_h = \mathbf{E}_{h-1} - \mathbf{t}_h\mathbf{p}'_h; \quad \mathbf{X} = \mathbf{E}_0 \tag{21}$$

$$\mathbf{F}_h^* = \mathbf{F}_{h-1}^* - \mathbf{u}_h\mathbf{q}'_h; \quad \mathbf{Y} = \mathbf{F}_0 \tag{22}$$

But in the outer relation for the Y block, $\mathbf{u}_h$ is replaced by its estimator, $\hat{\mathbf{u}}_h = \mathbf{b}_h\mathbf{t}_h$, and a mixed relation is obtained:

$$\mathbf{F}_h = \mathbf{F}_{h-1} - b_h\mathbf{t}_h\mathbf{q}'_h \tag{23}$$

(It is recalled that the aim is to make $\|\mathbf{F}_h\|$ small.) This mixed relation ensures the ability to use the model parameters for predicting from a test set. Furthermore, because the rank of Y is not decreased by 1 for each component, one can go on until the rank of the X block is exhausted. The complete algorithm is given in the Appendix together with an illustration of the matrices and vectors.

Summary: PLS

— There are outer relations of the form $\mathbf{X} = \mathbf{TP}' + \mathbf{E}$ and $\mathbf{Y} = \mathbf{UQ}' + \mathbf{F}^*$.
— There is an inner relation $\hat{\mathbf{u}}_h = b_h \mathbf{t}_h$.
— The mixed relation is $\mathbf{Y} = \mathbf{TBQ}' + \mathbf{F}$ where $\|\mathbf{F}\|$ is to be minimized.
— In the iterative algorithm, the blocks get each other's scores, this gives a better inner relation.
— In order to obtain orthogonal X scores, as in the PCA, it is necessary to introduce weights.

Properties of the PLS factors

For the user of PLS, it is obviously of interest to know what kind of properties to expect from it. The main properties can be summarized as follows.

The quantities $\mathbf{p}_h'$ and $\mathbf{q}_h'$ have unit length for each h: $\|\mathbf{p}_h'\| = \|\mathbf{q}_h'\| = 1$, or $\Sigma p_{hj}^2 = 1$ and $\Sigma q_{hj}^2 = 1$ for $j = 1$ to m.

Both $\mathbf{t}_h$ and $\mathbf{u}_h$ are centered around zero for each h: $\Sigma t_{hi} = 0$ and $\Sigma u_{hi} = 0$ for $i = 1$ to n.

The $\mathbf{w}_h'$ are orthogonal: $\mathbf{w}_i'\mathbf{w}_j = \delta_{ij}\|\mathbf{w}_i'\|^2$ where δ_{ij} is the Kronecker delta.

The $\mathbf{t}_h$ are orthogonal: $\mathbf{t}_i'\mathbf{t}_j = \delta_{ij}\|\mathbf{t}_i\|^2$.

It is useful to check that these properties hold for a number of data sets. These properties are also good indicators for computer rounding errors.

Prediction

The important part of any regression is its use in predicting the dependent block from the independent block. This is done by decomposing the X block and building up the Y block. For this purpose, $\mathbf{p}'$, $\mathbf{q}'$, $\mathbf{w}'$ and b from the calibration part are saved for every PLS factor. It should be noted that the new X block has r samples instead of n.

The independent blocks are decomposed and the dependent block is built up. For the X block, t is estimated by multiplying X by w as in the model building part

$$\hat{\mathbf{t}}_h = \mathbf{E}_{h-1}\mathbf{w}_h \tag{24}$$

$$\mathbf{E}_h = \mathbf{E}_{h-1} - \hat{\mathbf{t}}_h\mathbf{p}_h' \tag{25}$$

For the Y block:

$$\mathbf{Y} = \mathbf{F}_h = \sum b_h\hat{\mathbf{t}}_h\mathbf{q}_h' \tag{26}$$

where the summation is over h for all the factors (a) one wants to include and $\mathbf{X} = \mathbf{E}_0$, $\mathbf{Y} = \mathbf{F}_a$.

Number of components

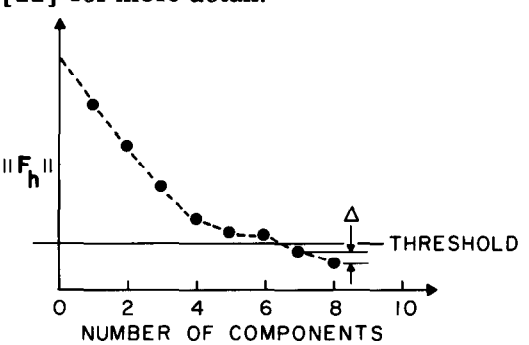
If the underlying model for the relation between X and Y is a linear model, the number of components needed to describe this model is equal to the model dimensionality. Nonlinear models require extra components to describe nonlinearities. The number of components to be used is a very important property of a PLS model.

Although it is possible to calculate as many PLS components as the rank of the X block matrix, not all of them are normally used. The main reasons for this are that the measured data are never noise-free and some of the smaller components will only describe noise, and that, as mentioned in earlier paragraphs, it is common to leave out small components because they carry the problems of collinearity.

This means that there must be one or several methods to decide when to stop. One possible criterion can be found in Eqn. 23, where the norm of $\mathbf{F}_h$ should be small. Figure 5 gives a plot of $\|\mathbf{F}_h\|$ vs. the number of components. It is possible to choose a threshold level and to stop when $\|\mathbf{F}_h\|$ goes below that threshold. Another possibility is to look at the difference between actual and previous $\|\mathbf{F}_h\|$ values and to stop when this becomes small compared to some previously established measurement error. A combination of threshold and difference methods would be preferable.

Sometimes the analysis of variance with *F*-test on the inner relation can be used to validate the model. In this case, one uses the *F*-test on the linear regression [7].

The above-mentioned methods are valuable for the model-building stage of PLS. If prediction is desired, another class of methods must be used to establish the number of components needed. These methods are called cross-validation. One can calculate a statistic for lack of prediction accuracy called PRESS (prediction residual sum of squares). Figure 6 gives a sample plot of the PRESS statistic against the number of components. It is obvious that one wants to use the number of components that gives a minimal PRESS. The location of this minimum is not always well defined. The evaluation of the number of components is analogous to the concept of detection limits, i.e., the smallest signal that can be detected in the presence of noise. See [11] for more detail.



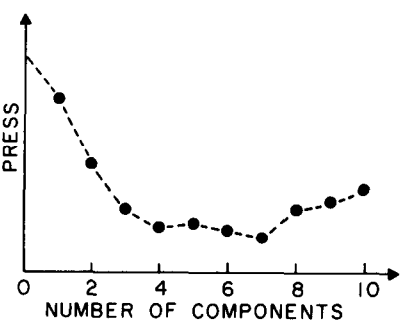


Fig. 5. $\|\mathbf{F}_h\|$ vs. the number of PLS components. A threshold and/or a difference criterion (see Δ in figure) can be used to stop the algorithm.

Fig. 6. Plot of PRESS against the number of components. This criterion evaluates the predictive power of the model. The number of components giving a minimum PRESS is the right number for the model that gives optimal prediction. In this example, models with 4—8 components would be acceptable.

Statistics

From the matrices of residuals $\mathbf{E}_h$ and $\mathbf{F}_h$, sums of squares can be calculated as follows: the total sum of squares over a matrix, the sums of squares over rows, and the sums of squares over columns. These sums of squares can be used to construct variance-like estimators. The statistical properties of these estimators have not undergone a rigorous mathematical treatment yet, but some properties can be understood intuitively.

The sum of squares of the $\mathbf{F}_h$ is the indicator of how good the model is (Eqn. 23). The sum of squares of $\mathbf{E}_h$ is an indicator of how much of the **X** block is not used in the model. In some cases, a substantial part of the **X** block does not participate in the model, which means that the independent variables have unexpected properties or large errors.

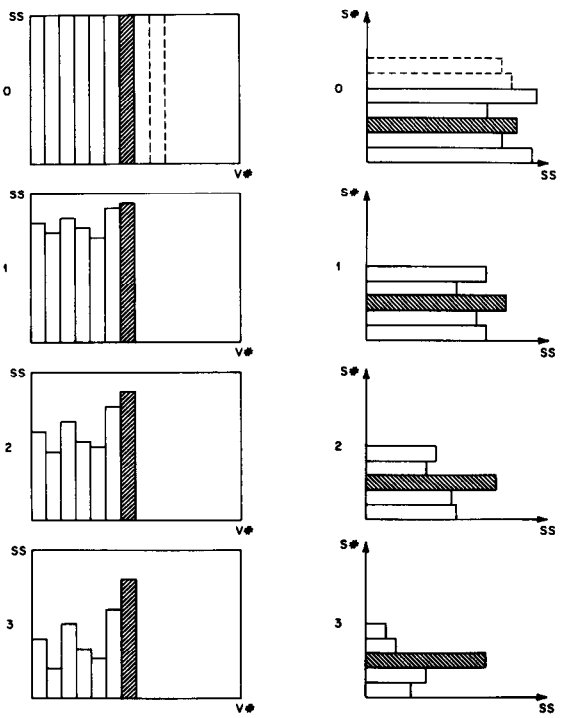


Fig. 7. Statistics for the variables. The data are shown as bars representing the sum of squares per variable (for the model building both X and Y variables; for the prediction only X variables). After 0 PLS components, the data is in mean-centered and variance-scaled form. As the number of PLS components increases, the information in each variable is exhausted. The hatched bar shows the behavior of a "special" variable, one that contributes little to the model.

Fig. 8. Statistics for the objects (samples). The data are shown as bars representing the sum of squares per object. As the number of PLS components increases, the sum of squares for each object decreases. The hatched bar shows the behavior of a "special" object, probably an outlier.

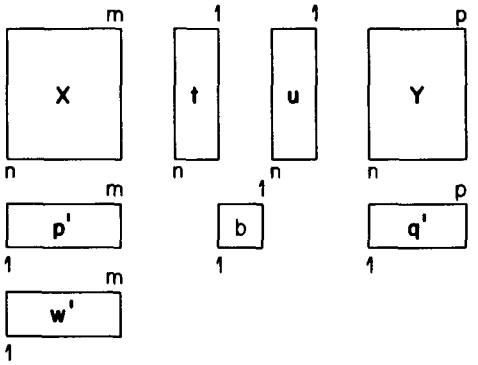


Fig. 9. A graphical representation of the matrices and vectors used in PLS.

Sums of squares over the columns indicate the importance of a variable for a certain component. Sums of squares over the rows indicate how well the objects fit the model. This can be used as an outlier detection criterion. Illustrations are given in Fig. 7 for variable statistics and in Fig. 8 for sample statistics. More on this can be found in the articles by S. Wold et al. as cited elsewhere [4, 6].

An advantage of PLS is that these statistics can be calculated for every component. This is an ideal means of following the model-building process. The evolution of these statistics can be followed (as shown in Figs. 7 and 8) as more and more components are calculated so that an idea of how the different objects and variables fit can be obtained. In combination with a criterion for model dimensionality, the statistics can be used to estimate which objects and variables contribute mainly to the model and which contribute mainly to the residual.

Conclusion

The topic of partial least squares is much larger than the material covered above. Some subjects not discussed at all or not in detail are: outlier detection, treatment of missing data, F and t statistics, classification/pattern recognition, leverage, selection of variables, data transformations, extensions to more blocks and hierarchical models, and lack of fit.

There are also other PLS algorithms. Each may have some advantage in a particular application. The algorithm given here is one of the most complete and elegant ones when prediction is important. An example of its application to simulated data is given in the next paper [11].

The authors thank Svante Wold and Harald Martens for their contributions to the PLS method. This paper results from a number of discussions at the Laboratory for Chemometrics. We thank all colleagues and visitors to the lab and especially Dave Veltkamp for their support and stimulating discussions. The Science Department of the Belgian Ministry of Foreign Affairs provided P. Geladi with a NATO travel grant 10/B/84/BE. Graphics were done by

Louise Rose. This work was supported by a grant from the Center for Process Analytical Chemistry, a National Science Foundation Cooperative Research Center at the University of Washington.

Appendix: The PLS algorithm

It is assumed that X and Y are mean-centered and scaled:

For each component: (1) take $\mathbf{u}_{start}$ = some $\mathbf{y}_j$.

In the X block: (2) $\mathbf{w}' = \mathbf{u}'\mathbf{X}/\mathbf{u}'\mathbf{u}$

(3) $\mathbf{w}'_{new} = \mathbf{w}'_{old}/\|\mathbf{w}'_{old}\|$ (normalization)

(4) $\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}'\mathbf{w}$

In the Y block: (5) $\mathbf{q}' = \mathbf{t}'\mathbf{Y}/\mathbf{t}'\mathbf{t}$

(6) $\mathbf{q}'_{new} = \mathbf{q}'_{old}/\|\mathbf{q}'_{old}\|$ (normalization)

(7) $\mathbf{u} = \mathbf{Y}\mathbf{q}/\mathbf{q}'\mathbf{q}$

Check convergence: (8) compare the t in step 4 with the one from the preceding iteration If they are equal (within a certain rounding error) go to step 9, else go to step 2. (If the Y block has only one variable, steps 5—8 can be omitted by putting $q = 1$, and no more iteration is necessary.)

Calculate the X loadings and rescale the scores and weights accordingly:

(9) $\mathbf{p}' = \mathbf{t}'\mathbf{X}/\mathbf{t}'\mathbf{t}$

(10) $\mathbf{p}'_{new} = \mathbf{p}'_{old}/\|\mathbf{p}'_{old}\|$ (normalization)

(11) $\mathbf{t}_{new} = \mathbf{t}_{old}\|\mathbf{p}'_{old}\|$

(12) $\mathbf{w}'_{new} = \mathbf{w}'_{old}\|\mathbf{p}'_{old}\|$

($\mathbf{p}'$, $\mathbf{q}'$ and $\mathbf{w}'$ should be saved for prediction; t and u can be saved for diagnostic and/or classification purposes).

Find the regression coefficient b for the inner relation:

(13) $b = \mathbf{u}'\mathbf{t}/\mathbf{t}'\mathbf{t}$

Calculation of the residuals. The general outer relation for the X block (for component h) is

$\mathbf{E}_h = \mathbf{E}_{h-1} - \mathbf{t}_h\mathbf{p}'_h$; $\mathbf{X} = \mathbf{E}_0$

The mixed relation for the Y block (for component h) is

$\mathbf{F}_h = \mathbf{F}_{h-1} - b_h\mathbf{t}_h\mathbf{q}'_h$; $\mathbf{Y} = \mathbf{F}_0$

From here, one goes to Step 1 to implement the procedure for the next component. (Note: After the first component, X in steps 2, 4 and 9 and Y in steps 5 and 7 are replaced by their corresponding residual matrices $\mathbf{E}_h$ and $\mathbf{F}_h$.)

Matrices and vectors are shown graphically in Fig. 9.

REFERENCES

1 B. Kowalski, R. Gerlach and H. Wold, Chemical Systems under Indirect Observation, in K. Jöreskog and H. Wold (Eds.), Systems under Indirect Observation, North-Holland, Amsterdam, 1982, pp. 191—209.
2 S. Wold, A. Ruhe, H. Wold and W. Dunn, SIAM J. Sci. Stat. Comput., 5 (1984) 735
3 M. Otto and W. Wegscheider, Anal. Chem., 57 (1985) 63.

4 S. Wold in, H. Martens and H. Russwurm (Eds.), Food Research and Data Analysis, Applied Science Publishers, London, 1983.
5 K. Mardia, J. Kent and J. Bibby, Multivariate Analysis, Academic Press, London, 1980.
6 S. Wold, in B. Kowalski (Ed.), Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, 1984.
7 N. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 1981.
8 R. Gunst and R. Mason, Regression Analysis and its Applications, M. Dekker, New York, 1980.
9 D. Belsley, E. Kuh and R. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York, 1980.
10 G. Strang, Linear Algebra and its Applications, Academic Press, New York, 1980.
11 P. Geladi and B. Kowalski, *Anal. Chim. Acta*, 185 (1986) 19.