JOURNAL OF CHEMOMETRICS, VOL. 2, 211–228 (1988)

PLS REGRESSION METHODS

AGNAR HÖSKULDSSON
The Engineering Academy of Denmark, DIAM, Building 358, 2800 Lyngby, Denmark

SUMMARY

In this paper we develop the mathematical and statistical structure of PLS regression. We show the PLS regression algorithm and how it can be interpreted in model building. The basic mathematical principles that lie behind two block PLS are depicted. We also show the statistical aspects of the PLS method when it is used for model building. Finally we show the structure of the PLS decompositions of the data matrices involved.

KEY WORDS Partial least squares Component regression Model building Covariance decomposition

INTRODUCTION

Partial least squares or PLS regression is used in many applied sciences. Wold[1] gives a survey of PLS methods with emphasis on social and economic sciences. In chemometrics these methods have been found valuable in numerous situations.

An important question is 'What situations are typical of those where PLS methods can be expected to be good for modelling purposes?'. They are the ones where there are many variables but not necessarily many samples or observations. This is a common situation in many laboratories. Typically it may take some time to get a new sample, but each sample may give a large amount of information (variables).

The next important question is 'Why can one expect PLS regression methods to perform better than multiple linear regression, ridge regression and other well known regression techniques?'. The answer is the stability of predictors derived from PLS methods. It turns out that the essential crieteria for the predictability of models is the number of variables included in the models. The uncertainty of the estimated parameters quickly becomes the dominating factor in the variability of predictors. Thus it is important to keep the number of variables as low as possible. In PLS components are selected that give 'maximal' reduction in the covariance $\mathbf{X}^T\mathbf{Y}$ of the data. In that sense PLS will give the minimum number of variables that is necessary. Criteria that give penalties on the number of variables, like the Akaike criteria, or those where the model performance is evaluated, like the Mallows Cp criteria, all give rise to more variables than the PLS method.

An integral part of the PLS method is the way in which the associated data analysis is done. With the aid of careful data analysis, outliers and groups of data or variables can be detected. An account of the data analysis aspect of PLS methods, with examples from chemistry, is given in the tutorial papers of Geladi and Kowalski[2] and Geladi.[3]

0886–9383/88/030211–18$09.00

Received 25 August 1987
Revised 20 November 1987

One of the reviewers has drawn the author's attention to the works of Helland[4] and Manne,[5] which also treat structural questions of PLS regression. Some of the results of this paper may be found in their works, although the emphasis of this work is very different from theirs. Some of the results of this work are also mentioned in Wold *et al.*[6]

PLS REGRESSION TECHNIQUES

The PLS regression algorithm

The basic algorithm of PLS regression as developed by Wold *et al.*[7] is as follows.

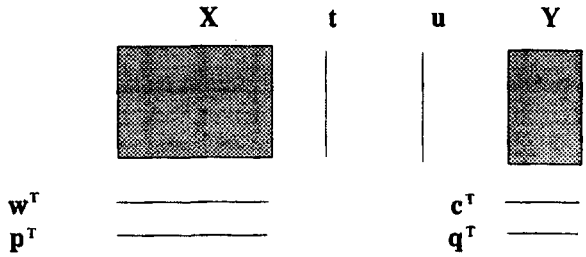
The starting point is two data matrices **X** and **Y**. **X** is an $N \times M$ matrix and **Y** an $N \times K$ matrix. No assumption is needed concerning the dimensions of **X** and **Y**. Before the algorithm starts, the matrices may be scaled or centred. Scaling can correspond to working with correlation matrices, and centring to subtracting mean values from each of the column values.

The algorithm is as follows.

1. Start: set **u** to the first column of **Y**
2. $\mathbf{w} = \mathbf{X}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$
3. Scale **w** to be of length one
4. $\mathbf{t} = \mathbf{X}\mathbf{w}$
5. $\mathbf{c} = \mathbf{Y}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$
6. Scale **c** to be of length one
7. $\mathbf{u} = \mathbf{Y}^T\mathbf{c}/(\mathbf{c}^T\mathbf{c})$
8. If convergence then 9 else 2
9. *X*-loadings: $\mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$
10. *Y*-loadings: $\mathbf{q} = \mathbf{Y}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$
11. Regression (**u** upon **t**): $b = \mathbf{u}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$
12. Residual matrices: $\mathbf{X} \rightarrow \mathbf{X} - \mathbf{t}\mathbf{p}^T$ and $\mathbf{Y} \rightarrow \mathbf{Y} - b\mathbf{t}\mathbf{c}^T$

The next set of iterations starts with the new **X** and **Y** matrices as the residual matrices from the previous iteration. The iterations can continue until a stopping criteria is used or **X** becomes the zero matrix.

The dimensions of the vectors in the algorithm can be depicted as follows:



It is difficult to know about the numerical properties of this algorithm when we look at the definitions above. The situation becomes clearer when we relate, for a given dimension, the vectors at step n with the corresponding vectors at step $n-1$. Let us first look at $\mathbf{u}_n$:

$$\begin{aligned}\mathbf{u}_n &= \mathbf{Y}\mathbf{c}_n/(\mathbf{c}_n^T\mathbf{c}_n)\\ &= \mathbf{Y}\mathbf{Y}^T\mathbf{t}_n/(\mathbf{c}_n^T\mathbf{c}_n)\ (\mathbf{t}_n^T\mathbf{t}_n)\\ &= \mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}_n/[(\mathbf{c}_n^T\mathbf{c}_n)\ (\mathbf{t}_n^T\mathbf{t}_n)\ (\mathbf{w}_n^T\mathbf{w}_n)]\\ &= \mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{X}^T\mathbf{u}_{n-1}/[(\mathbf{c}_n^T\mathbf{c}_n)\ (\mathbf{t}_n^T\mathbf{t}_n)\ (\mathbf{w}_n^T\mathbf{w}_n)\ (\mathbf{u}_{n-1}^T\mathbf{u}_{n-1})] \qquad (1)\end{aligned}$$

Similarly we derive for the vectors $\mathbf{c}_n$,$\mathbf{t}_n$ and $\mathbf{w}_n$

$$\begin{aligned}\mathbf{c}_n &= \mathbf{Y}^T\mathbf{X}\mathbf{X}\mathbf{Y}\mathbf{c}_{n-1}/[(\mathbf{t}_n^T\mathbf{t}_n)\ (\mathbf{w}_n^T\mathbf{w}_n)\ (\mathbf{u}_{n-1}^T\mathbf{u}_{n-1})\ (\mathbf{c}_{n-1}^T\mathbf{c}_{n-1})]\\ \mathbf{t}_n &= \mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t}_{n-1}/[(\mathbf{w}_n^T\mathbf{w}_n)\ (\mathbf{u}_{n-1}^T\mathbf{u}_{n-1})\ (\mathbf{c}_{n-1}^T\mathbf{c}_{n-1})\ (\mathbf{t}_{n-1}^T\mathbf{t}_{n-1})]\\ \mathbf{w}_n &= \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}_{n-1}/[(\mathbf{u}_{n-1}^T\mathbf{u}_{n-1})\ (\mathbf{c}_{n-1}^T\mathbf{c}_{n-1})\ (\mathbf{t}_{n-1}^T\mathbf{t}_{n-1})\ (\mathbf{w}_{n-1}^T\mathbf{w}_{n-1})] \qquad (2)\end{aligned}$$

These equations show that the algorithm performs in a similar way to the power method of determining the largest eigenvalue for a matrix. The algorithm will, like the power method, converge rapidly in almost all practical cases. The only complication which may arise is when there are equal eigenvalues that are the largest. But this situation contains no convergence problem, only that the selected component needs not be unique. Experience indicates that in most practical cases convergence is attained in less than ten iterations.

At convergence we can write

$$\begin{aligned}\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{X}^T\mathbf{u} &= a\mathbf{u}\\ \mathbf{Y}^T\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{c} &= a\mathbf{c}\\ \mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} &= a\mathbf{t}\\ \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} &= a\mathbf{w}\end{aligned}$$

The power method tells us that a is the maximum eigenvalue of the eigenvalue problem. The vectors $\mathbf{u}$, $\mathbf{c}$, $\mathbf{t}$ and $\mathbf{w}$ are thus the eigenvectors of the appropriate matrices corresponding to the maximum eigenvalue.

The PLS regression algorithm can thus be viewed as follows. For each set of residual matrices $\mathbf{X}$ and $\mathbf{Y}$ we compute the maximum eigenvalue and associated eigenvectors of the matrices

$$\mathbf{X}^T\mathbf{Y}\ \mathbf{Y}^T\mathbf{X} \qquad \mathbf{Y}^T\mathbf{X}\ \mathbf{X}^T\mathbf{Y} \qquad \mathbf{X}\ \mathbf{X}^T\mathbf{Y}\ \mathbf{Y}^T \qquad \mathbf{Y}\ \mathbf{Y}^T\mathbf{X}\ \mathbf{X}^T$$

and the eigenvectors are used to compute new residual matrices. In the computations only the eigenvector of $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ is needed. The others may be computed as in the algorithm.

It is an important aspect of the PLS algorithm that we only select one pair of eigenvectors at each step (dimension). A detailed study of the algorithm shows that this is appropriate.

The geometry of the PLS regression algorithm

We shall derive here the geometrical properties of the vectors and matrices involved in the PLS algorithm. We shall use the notation

$$\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k) \qquad (3)$$

to denote the $\mathbf{w}$ vectors arrived at in the algorithm. Thus $\mathbf{W}$ is an $N \times k$ matrix with columns consisting of the $\mathbf{w}$ vectors; similarly for the other set of vectors.

The basic properties are derived from the way the residual matrix $\mathbf{X}_i$ is computed from the

previous residual matrices. We have

$$\begin{aligned}\mathbf{X}_i &= \mathbf{X}_{i-1} - \mathbf{t}_{i-1}\mathbf{p}_{i-1}^{\mathrm{T}} \\ &= \mathbf{X}_{i-1} - \mathbf{t}_{i-1}\,\mathbf{t}_{i-1}^{\mathrm{T}}\,\mathbf{X}_{i-1}/(\mathbf{t}_{i-1}^{\mathrm{T}}\mathbf{t}_{i-1}) \\ &= [\mathbf{I} - \mathbf{t}_{i-1}\,\mathbf{t}_{i-1}^{\mathrm{T}}/(\mathbf{t}_{i-1}^{\mathrm{T}}\mathbf{t}_{i-1})\,]\,\mathbf{X}_{i-1} \\ &= [\mathbf{I} - \mathbf{t}_{i-1}\,\mathbf{t}_{i-1}^{\mathrm{T}}/(\mathbf{t}_{i-1}^{\mathrm{T}}\mathbf{t}_{i-1})]\,[\mathbf{X}_{i-2} - \mathbf{t}_{i-2}\,\mathbf{t}_{i-2}^{\mathrm{T}}\mathbf{X}_{i-2}/(\mathbf{t}_{i-2}^{\mathrm{T}}\mathbf{t}_{i-2})] \qquad (4)\end{aligned}$$

The basic properties follow from this equation, which relates one residual matrix to another.

Property 1

The vectors **w** are mutually orthogonal:

$$(\mathbf{w}_i,\mathbf{w}_j) = \mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = 0 \quad \text{for } i \neq j \qquad (5)$$

Proof. Suppose that $i < j$. Then we can write, as above,

$$\mathbf{X}_j = \mathbf{Z}\,[\mathbf{X}_i - \mathbf{t}_i\,\mathbf{t}_i^{\mathrm{T}}/(\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i)\,\mathbf{X}_i] \qquad (6)$$

where **Z** is some matrix. We shall show that

$$\mathbf{X}_j\,\mathbf{w}_i = 0 \quad \text{for } j > i \qquad (7)$$

This follows from equation (6) and

$$[\mathbf{X}_i - \mathbf{t}_i\,\mathbf{t}_i^{\mathrm{T}}/(\mathbf{t}_i^{\mathrm{T}}\,\mathbf{t}_i)\,\mathbf{X}_i]\,\mathbf{w}_i = \mathbf{t}_i - \mathbf{t}_i\,\mathbf{t}_i^{\mathrm{T}}/(\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i)\,\mathbf{t}_i = 0$$

This gives

$$\mathbf{w}_j^{\mathrm{T}}\mathbf{w}_i = \mathbf{w}_j^{\mathrm{T}}\mathbf{X}_j^{\mathrm{T}}\mathbf{Y}_j\mathbf{Y}_j^{\mathrm{T}}\mathbf{X}_j\mathbf{w}_i/a_j = 0$$

which is equation (5).

Property 2

The vectors $\mathbf{t}_i$ are mutually orthogonal:

$$(\mathbf{t}_i,\,\mathbf{t}_j) = \mathbf{t}_i^{\mathrm{T}}\mathbf{t}_j = 0 \quad \text{for } i \neq j \qquad (8)$$

Proof. We suppose that $i < j$. Similarly to (4) we can write

$$\begin{aligned}\mathbf{X}_j &= \mathbf{X}_{j-1} - \mathbf{X}_{j-1}\mathbf{w}_{j-1}\mathbf{t}_{j-1}^{\mathrm{T}}\mathbf{X}_{j-1}/(\mathbf{t}_{j-1}^{\mathrm{T}}\mathbf{t}_{j-1}) \\ &= \mathbf{X}_{j-1}\,[\mathbf{I} - \mathbf{w}_{j-1}\mathbf{t}_{j-1}^{\mathrm{T}}\,\mathbf{X}_{j-1}/(\mathbf{t}_{j-1}^{\mathrm{T}}\,\mathbf{t}_{j-1})\,] \\ &= \mathbf{X}_{i+1}\,\mathbf{Z} \\ &= [\mathbf{X}_i - \mathbf{t}_i\,\mathbf{t}_i^{\mathrm{T}}\,\mathbf{X}_i/(\mathbf{t}_i^{\mathrm{T}}\,\mathbf{t}_i)\,]\,\mathbf{Z}\end{aligned}$$

where **Z** is some matrix product which is not used here. This gives

$$\mathbf{t}_i^{\mathrm{T}}\,\mathbf{X}_j = 0 \quad \text{for } i < j \qquad (9)$$

and therefore

$$\mathbf{t}_i^T \mathbf{t}_j = \mathbf{t}_i^T \mathbf{X}_j \mathbf{w}_j = 0 \quad \text{for } i < j$$

The vectors $\mathbf{t}_i$ are thus an orthogonal basis in the space generated by the columns of $\mathbf{X}$. If we suppose that the rank of $\mathbf{X}$ is k, we can write the $\mathbf{X}$ matrix as

$$\mathbf{X} = \sum_1^k \mathbf{t}_i \mathbf{p}_i^T + \mathbf{X}_0 \tag{10}$$

where $\mathbf{X}_0$ is a matrix which is orthogonal to $\mathbf{Y}$. In the following we will suppose that $\mathbf{X}_0$ is zero, because $\mathbf{X}_0$ does not contribute anything to $\mathbf{Y}$

Property 3

The vectors $\mathbf{w}_i$ *are orthogonal to the vectors* $\mathbf{p}_j$ for $i < j$:

$$(\mathbf{w}_i, \mathbf{p}_j) = \mathbf{w}_i^T \mathbf{p}_j = 0 \quad \text{for } i < j \tag{11}$$

Proof. From equation (9) we obtain

$$\mathbf{w}_i^T \mathbf{p}_j = \mathbf{w}_i^T \mathbf{X}_j^T \mathbf{t}_j / (\mathbf{t}_j^T \mathbf{t}_j) = 0 \quad \text{for } i < j$$

Property 4

The vectors $\mathbf{p}_i$ are orthogonal in the kernel space of $\mathbf{X}$:

$$(\mathbf{p}_i, \mathbf{p}_j)_x = \mathbf{p}_i^T (\mathbf{X}^T \mathbf{X})^- \mathbf{p}_j = 0 \quad \text{for } i \neq j \tag{12}$$

Proof. From equation (10) we get

$$\mathbf{X}^T \mathbf{X} = \Sigma\, (\mathbf{t}_i, \mathbf{t}_i)\, \mathbf{p}_i\, \mathbf{p}_i^T = \mathbf{P}\, \mathbf{L}\, \mathbf{P}^T$$

where $\mathbf{L}$ is a diagonal matrix containing the coefficients $(\mathbf{t}_i, \mathbf{t}_i)$. From the theory of kernel spaces (see e.g. Aronszajn[8]) equation (12) follows. Note that $\mathbf{A}^-$ is the generalized inverse of a matrix $\mathbf{A}$.

This is as much as can be said about orthogonality of the $\mathbf{w}$,$\mathbf{t}$, $\mathbf{u}$ and $\mathbf{p}$ vectors. In general we have

$$\mathbf{p}_i^T \mathbf{w}_j \neq 0 \quad \text{for } i < j$$

These four properties follow from the way the residual matrices $\mathbf{X}_i$ are constructed. Basically they do not depend on the way a new $\mathbf{t}$ vector is constructed. Having determined a new $\mathbf{t}$ vector, a new subspace is constructed which is orthogonal to the $\mathbf{t}$ vector.

No special orthogonality properties are available among the vectors $(\mathbf{u}_i)$, $(\mathbf{c}_i)$ and $(\mathbf{q}_i)$ themselves. But of course these vectors, being the images of appropriate vectors, satisfy some orthogonality conditions relative to some matrices.

The residual matrices can be written in the following way, where we suppose that $\mathbf{c}_i$ has been

scaled to be of length one:

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i \mathbf{p}_i^T \qquad \mathbf{Y}_{i+1} = \mathbf{Y}_i - b_i \mathbf{t}_i \mathbf{c}_i^T$$
$$\mathbf{Y}_{i+1}^T \mathbf{X}_{i+1} = \mathbf{Y}_i^T \mathbf{X}_i - b_i (\mathbf{t}_i^T \mathbf{t}_i) \mathbf{c}_i \mathbf{p}_i^T$$
$$\mathbf{X}_{i+1}^T {}^T\mathbf{X}_{i+1} = \mathbf{X}_i^T \mathbf{X}_i - (\mathbf{t}_i^T \mathbf{t}_i) \mathbf{p}_i \mathbf{p}_i^T$$
$$\mathbf{Y}_{i+1}^T \mathbf{Y}_{i+1} = \mathbf{Y}_i^T \mathbf{Y}_i - b_i^2 (\mathbf{t}_i^T \mathbf{t}_i) \mathbf{c}_i \mathbf{c}_i^T$$

Here $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{Y}_1 = \mathbf{Y}$ are the original data matrices.

In the PLS algorithm the $\mathbf{c}$ vectors are scaled to be of length one. Suppose that we do not scale the $\mathbf{c}$ vectors. Then

$$\mathbf{u}^T \mathbf{t} = \mathbf{c}^T \mathbf{Y}^T \mathbf{t}/(\mathbf{c}^T\mathbf{c}) = \mathbf{c}^T (\mathbf{Y}^T \mathbf{t})/(\mathbf{c}^T\mathbf{c}) = \mathbf{c}^T\mathbf{c}\, (\mathbf{t}^T\mathbf{t})/(\mathbf{c}^T\mathbf{c}) = \mathbf{t}^T\mathbf{t}$$

Therefore, if we do not scale the $\mathbf{c}$ vectors, we have

$$b_i = 1 \quad \text{for all } i \tag{13}$$

In the following we will suppose that the $\mathbf{c}$ vectors are not scaled. Since $b = |\mathbf{c}|$, the formulae are easily modified if the coefficients b are wanted in the formulae.

Properties of the PLS regression algorithm

Extraction of components in X and Y space

Suppose that the $\mathbf{X}$ matrix is rotated. This corresponds to multiply $\mathbf{X}$ from the right by an orthogonal matrix. Denote by $\mathbf{S}$ such a matrix,

$$\mathbf{S} = \mathbf{X}\mathbf{O}_x$$

Then we have

$$\Sigma\, \mathbf{s}_{ij}^2 = \text{tr}\, (\mathbf{S}^T\mathbf{S}) = \text{tr}\, (\mathbf{X}^T\mathbf{X}) = \Sigma\, \mathbf{x}_{ij}^2$$

This shows that the total variation is unchanged under orthogonal transformation (rotation). Similarly we can rotate the $\mathbf{Y}$ matrix,

$$\mathbf{Z} = \mathbf{Y}\mathbf{O}_y$$

Denote by $\mathbf{s}_i$ and $\mathbf{z}_i$ the columns of $\mathbf{S}$ and $\mathbf{Z}$, and suppose that there are more x-variables than y-variables, $n > m$. An important question is how close can the $\mathbf{s}$ vectors come to the $\mathbf{z}$ vectors? One way to look at the question is to consider the squared distances between the vectors,

$$\mathbf{I} = \sum_1^m |\mathbf{s}_i - \mathbf{z}_i|^2 + \sum_{m+1}^n |\mathbf{s}_i|^2 \tag{14}$$

The latter sum reflects the assumption that there are fewer components in Y space than in X space. The argumentation here is symmetric, so we may suppose the reverse.

Suppose we wish to determine the orthogonal matrices in such a way, that (1) is minimized. Intuitively it means that we rotate X and Y space such that the corresponding components in X and Y space are as close as possible to each other.

Interpretation 1

Suppose that two orthogonal matrices are determined such that (14) is minimized. Then the vectors **t** and **u** in the PLS algorithm satisfy

$$\mathbf{t} = \mathbf{s}_1 \qquad \mathbf{u} = \mathbf{z}_1$$

Proof. Equation (14) can be written as

$$I = \text{tr}(\mathbf{X}^T\mathbf{X}) + \text{tr}(\mathbf{Y}^T\mathbf{Y}) - 2\text{tr}(\mathbf{X}^T\mathbf{Y}\mathbf{O}_y\mathbf{O}_x^T) \tag{15}$$

In the next section it is shown that I is minimized when $\mathbf{O}_y$ and $\mathbf{O}_x$ are the orthogonal matrices in the singular value representation of $\mathbf{X}^T\mathbf{Y}$:

$$\mathbf{X}^T\mathbf{Y} = \Sigma \mathbf{e}_i\, \mathbf{g}_i\mathbf{f}_i^T = \mathbf{G}\,\mathbf{E}\,\mathbf{F}^T \tag{16}$$

It is also shown that **t** is the first column of **G** and **u** the first column of **F**.

The minimum value of I in equation (14) is

$$\begin{aligned} \min I &= \text{tr}(\mathbf{X}^T\mathbf{X}) + \text{tr}(\mathbf{Y}^T\mathbf{Y}) - 2\text{tr}((\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})^{1/2}) \\ &= \text{tr}(\mathbf{X}^T\mathbf{X}) + \text{tr}(\mathbf{Y}^T\mathbf{Y}) - 2\Sigma e_i \end{aligned}$$

In the PLS algorithm only one pair of components is selected. The value of I is thus reduced by $2e_1$. Only one pair of components is chosen because $2e_2$ is smaller than $2e_1$ at the next step of the algorithm. The biggest reduction in I is obtained by taking one component at a time, computing the residual matrices and then again taking one pair of components.

The PLS algorithm can be viewed as a stepwise procedure, where a pair of components in X and Y space are selected which are closest to each other. The procedure is continued as long as there are significant components. In this way the space of projection of **Y** and **X** is filled out with components that are mutually orthogonal and as close as possible to some components in Y space.

This interpretation of the PLS algorithm is very appealing. In doing the regression of Y onto X, one tries to fill out the X and Y space at the same time as selecting the components. One stops the iteration when there are no components in X or Y space which are close to each other, either because one has filled out X or there is nothing more in X space that can be used to describe Y.

Components with maximal covariance

Consider two components **f** and **g** in X and Y space:

$$\mathbf{f} = \mathbf{X}\mathbf{d} \qquad |\mathbf{d}| = 1$$

$$\mathbf{g} = \mathbf{Y}\mathbf{e} \qquad |\mathbf{e}| = 1$$

The sample covariance between the two components is given by

$$\text{Cov}(\mathbf{f},\mathbf{g}) = \mathbf{f}^T\,\mathbf{g}/N.$$

If one is searching for two components in X and Y space, it is always a good choice to choose two that have the maximal covariance among all components in X and Y space. The PLS algorithm does this.

Interpretation 2

The vectors **w** and **c** in the PLS algorithm satisfy the maximization.

$$[\mathrm{Cov}(\mathbf{t},\mathbf{u})]^2 = [\mathrm{Cov}(\mathbf{Xw},\mathbf{Yc})]^2 = \max [\mathrm{Cov}(\mathbf{f},\mathbf{g})]^2 \quad \text{for } |\mathbf{d}| = |\mathbf{e}| = 1 \tag{17}$$

Proof. Consider the singular value decomposition of $\mathbf{X}^T\mathbf{Y}$:

$$\mathbf{X}^T\mathbf{Y} = \Sigma\, a_i\, \mathbf{f}_i\, \mathbf{g}_i^T$$

The singular value a_1 has the interpretation

$$(a_1)^2 = \max (\mathbf{d}^T\mathbf{X}^T\mathbf{Y}\mathbf{e})^2 \quad \text{for } |\mathbf{d}| = |\mathbf{e}| = 1$$

with maximum attained at $\mathbf{d} = \mathbf{f}_1$ and $\mathbf{e} = \mathbf{g}_1$. In the PLS algorithm we have $\mathbf{w} = \mathbf{f}_1$ and $\mathbf{c} = \mathbf{g}_1$, which proves equation (17).

The components **t** and **u** in the PLS algorithm have the interpretation that they are the components in X and Y space that have maximal covariance among all components in X and Y space. The PLS algorithm only selects one pair of components at a time, because the covariance of the second pair is smaller than the maximal covariance at the next iteration.

PLS as regression on orthogonal components

In linear regression the variables or components are selected on the basis of the covariance matrix $\mathbf{X}^T\mathbf{X}$. If variables are selected, the regression can be carried out by any method equivalent to the Cholesky factorization of $\mathbf{X}^T\mathbf{X}$. If components are selected, we usually first determine the transformation matrix **O** and compute the components

$$\mathbf{F} = \mathbf{XO} \tag{18}$$

and then do regression of **Y** onto **F**. In principal component regression **O** is the matrix of eigenvectors of $\mathbf{X}^T\mathbf{X}$.

In many applied situations it is proper to weight the covariance matrix, i.e. to replace $\mathbf{X}^T\mathbf{X}$ with $\mathbf{X}^T\mathbf{V}\mathbf{X}$, where **V** is some positive definite matrix. A common example is the one where the residual covariance matrix is not a constant times the identity matrix. Examples of this situation are when the residuals are correlated or when the variances of one part of the residuals are different from the variances of some other parts of the residuals. In this case **V** is the inverse of the estimate of the residual covariance matrix.

Another view of weighting is to look at the 'size' of the **Y** matrix. The matrix

$$\mathbf{V} = \mathbf{Y}\,\mathbf{Y}^T$$

can be viewed as the size of the data in the **Y** matrix. It is reasonable to use this **V** as a weighting matrix and consider

$$\mathbf{X}^T\,\mathbf{V}\,\mathbf{X} = \mathbf{X}^T\,\mathbf{Y}\,\mathbf{Y}^T\,\mathbf{X}$$

With this method of weighting we weight according to the size of the data in the **Y** matrix: values in **Y** close to zero give small weight and large values give large weight. In the extreme, if rows in **Y** are zero, they and the associated rows in **X** are not used in the analysis.

This method of weighting is appealing. It is reasonable, for example, to give small weight to data when they are at the noise level. We do not want to aim at predicting the noise. On the other hand we want to give a high weight to the part of the data where the signal is.

In doing the regression, one can follow the method of principal component regression analysis, but now with the weighted covariance matrix. First we compute the eigenvalues and vectors of the weighted covariance matrix

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X} = \mathbf{O}\,\mathbf{D}\,\mathbf{O}^{\mathrm{T}} \tag{19}$$

and then compute the matrix of the associated components as in equation (18)

$$\mathbf{F} = \mathbf{X}\mathbf{O} \tag{20}$$

and do regression of $\mathbf{Y}$ on $\mathbf{F}$. This is in fact what the PLS algorithm does, apart from the fact that it chooses only one component at a time.

Interpretation 3

Consider the PLS algorithm at the ith step. The ith step amounts to selecting as $\mathbf{w}_i$ the eigenvector in equation (19) with $\mathbf{X}$ replaced by $\mathbf{X}_i$, and similarly for $\mathbf{Y}$, that is associated with the largest eigenvalue, computing $\mathbf{t}_i$ as in equation (20), $\mathbf{t}_i = \mathbf{X}_i\mathbf{w}_i$, and doing the regression of $\mathbf{Y}_i$ onto $\mathbf{t}_i$. The projection of $\mathbf{Y}_i$ is

$$\mathbf{t}_i\,\mathbf{c}_i^{\mathrm{T}}$$

the residual $\mathbf{Y}$ matrix is

$$\mathbf{Y}_{i+1} = \mathbf{Y}_i - \mathbf{t}_i\,\mathbf{c}_i^{\mathrm{T}}$$

and the residual covariance matrix is

$$\mathbf{Y}_{i+1}^{\mathrm{T}}\,\mathbf{Y}_{i+1} = \mathbf{Y}_i^{\mathrm{T}}\,\mathbf{Y}_i - \mathbf{c}_i\,\mathbf{c}_i^{\mathrm{T}}\,(\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i)$$

The projection is unchanged if $\mathbf{Y}_i$ is replaced by $\mathbf{Y}$.

Proof. Let $\mathbf{w}_i$ be the eigenvector associated with the largest eigenvalue:

$$\mathbf{X}_i^{\mathrm{T}}\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}\mathbf{X}_i\,\mathbf{w}_i = a_i\,\mathbf{w}_i \qquad \mathbf{t}_i = \mathbf{x}_i\mathbf{w}_i$$

The projection of $\mathbf{Y}_i$ onto $\mathbf{t}_i$ is

$$\mathbf{t}_i^{\mathrm{T}}\mathbf{Y}_i/(\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i)\,\mathbf{t}_i = \mathbf{t}_i\mathbf{c}_i^{\mathrm{T}}$$

and the residual covariance matrix is

$$\begin{aligned}\mathbf{Y}_{i+1}^{\mathrm{T}}\,\mathbf{Y}_{i+1} &= (\mathbf{Y}_i - \mathbf{t}_i\,\mathbf{c}_i^{\mathrm{T}})^{\mathrm{T}}\,(\mathbf{Y}_i - \mathbf{t}_i\,\mathbf{c}_i^{\mathrm{T}}) \\ &= \mathbf{Y}_i^{\mathrm{T}}\,\mathbf{Y}_i - \mathbf{c}_i\,\mathbf{t}_i^{\mathrm{T}}\mathbf{Y}_i - \mathbf{Y}_i^{\mathrm{T}}\mathbf{t}_i\mathbf{c}_i^{\mathrm{T}} + \mathbf{c}_i\,\mathbf{c}_i^{\mathrm{T}}\,(\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i) \\ &= \mathbf{Y}_i^{\mathrm{T}}\,\mathbf{Y}_i - \mathbf{c}_i\,\mathbf{c}_i^{\mathrm{T}}\,(\mathbf{t}_i^{\mathrm{T}}\,\mathbf{t}_i).\end{aligned}$$

Since $\mathbf{Y}_i^{\mathrm{T}}\mathbf{t}_i = \mathbf{Y}^{\mathrm{T}}\mathbf{t}_i$, the projection of $\mathbf{Y}_i$ onto $\mathbf{t}_i$ is the same as the projection of $\mathbf{Y}$ onto $\mathbf{t}_i$

The PLS algorithm can thus be viewed as a stepwise regression on orthogonal components, where at each step we determined the component as the one that gives the maximum

'variance' of the weighted covariance matrix, and the weighting is done so that Y-values close to the noise level get small weight and $\mathbf{Y}$-values representing signals get high weight.

The regression of $\mathbf{Y}$ on $\mathbf{T}_i = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_i)$ can be represented as

$$\mathbf{Y} = \Sigma^i \mathbf{t}_j \mathbf{c}_j^{\mathrm{T}} + \mathbf{Y}_{i+1}$$

and the residual covariance matrix can be written as

$$\mathbf{Y}_{i+1}^{\mathrm{T}} \mathbf{Y}_{i+1} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \Sigma^i \mathbf{c}_j^{\mathrm{T}} \mathbf{c}_j (\mathbf{t}_j^{\mathrm{T}} \mathbf{t}_j)$$

MATHEMATICS OF PLS

The PLS method is based on the singular value decomposition of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$:

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \Sigma\, a_i \mathbf{f}_i \mathbf{g}_i^{\mathrm{T}}$$

where $(\mathbf{f}_i)$ and $(\mathbf{g}_i)$ are orthonormal vectors of appropriate dimension and (a_i) are the singular values arranged in decreasing order. The largest singular value a_1 has the interpretation

$$a_1^2 = \max (\mathbf{f}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{g})^2 \quad \text{with } |\mathbf{f}| = |\mathbf{g}| = 1 \tag{21}$$

Here $|\mathbf{f}|$ is the length of the $\mathbf{f}$ vector. This is interpretation 2 of the PLS method. Equation (21) is easily proved by the Cauchy–Schwartz inequality.

In proving interpretation 1, suppose that there are fewer columns in $\mathbf{Y}$ than in $\mathbf{X}$. $\mathbf{Y}$ can be expanded by zeros to be of the same size as $\mathbf{X}$, and similarly $\mathbf{O}_y$ can be expanded to be an orthogonal matrix of proper size. Then

$$I = |\mathbf{X}\mathbf{O}_x - \mathbf{Y}\mathbf{O}_y|^2 = \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) + \mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) - 2\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{O}_y\mathbf{O}_x^{\mathrm{T}}) \tag{22}$$

Minimization of I is equivalent to the maximization of

$$\max \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{V}) \tag{23}$$

where the maximization is over all orthogonal matrices $\mathbf{V}$. A straightforward argument gives as a value of $\mathbf{V}$ as maximization

$$\mathbf{V} = \mathbf{G}\,\mathbf{F}^{\mathrm{T}}$$

where $\mathbf{G}$ and $\mathbf{F}$ are the matrices in the singular value decomposition of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. This gives

$$\mathbf{O}_x = \mathbf{F} \qquad \mathbf{O}_y = \mathbf{G}$$

as one possible solution to the minimization of equation (22). This proves interpretation 1. The minimum value of I is

$$\begin{aligned} I_{\min} &= \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) + \mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) - 2\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X})^{1/2} \\ &= \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) + \mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) - 2\Sigma\, a_i \end{aligned}$$

The basic motivation for the way PLS works comes from the following inequality:

$$s_i^2(\mathbf{A} - \mathbf{B}) \geq s_{i+k}^2(\mathbf{A}) = a_{i+k}^2$$

where $s_i(\mathbf{A})$ is the ith singular value of a matrix $\mathbf{A}$ and $\mathbf{B}$ is a matrix of rank $\mathbf{k}$. This inequality is proved, for example, by Rao.[9] When we apply this inequality, we get

$$s_1^2(\mathbf{X}_{i+1}{}^{\mathrm{T}}\mathbf{Y}) = s^2(\mathbf{X}_i{}^{\mathrm{T}}\mathbf{Y} - \mathbf{p}_i \mathbf{t}_i^{\mathrm{T}} \mathbf{Y}) \geq s_2^2(\mathbf{X}_i^{\mathrm{T}}\mathbf{Y})$$

This shows that the largest singular value at step $i + 1$ is larger than the second largest singular value at step i. For this reason only one component is selected at a time in the PLS method.

The singular values also have the interpretation

$$(s_1(\mathbf{X}_i^T\mathbf{Y}))^2 = \max \mathbf{f}^T\mathbf{X}_i^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_i\mathbf{f} \quad \text{for } |\mathbf{f}| = 1$$

Consider the following inequality:

$$\begin{aligned}(s_1(\mathbf{X}_{i+1}{}^T\mathbf{Y}))^2 &= \max \mathbf{f}^T\mathbf{X}_{i+1}{}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_{i+1}\mathbf{f} \\ &= \max \mathbf{f}^T(\mathbf{I} - \mathbf{p}_i\mathbf{w}_i^T)\mathbf{X}_i^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_i(\mathbf{I} - \mathbf{w}_i\mathbf{p}_i^T)\mathbf{f} \\ &\leqslant \max \mathbf{f}^T\mathbf{X}_i^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_i\mathbf{f} \\ &= (s_1(\mathbf{X}_i^T\mathbf{Y}))^2 \quad \text{for } |\mathbf{f}| = 1\end{aligned}$$

From this inequality it follows that what is selected from $\mathbf{X}_i^T\mathbf{Y}_i = \mathbf{X}_i^T\mathbf{Y}$ always decreases. Therefore one can select components until $s_1(\mathbf{X}_i\mathbf{Y})$ is small according to some criterion.

PLS COMPONENTS IN REGRESSION

PLS and linear regression

One of the most important tasks in regression analysis is to reduce the number of independent variables. The variances of regression coefficients increase when new variables are introduced (more precisely they never decrease). Therefore it is a typical situation in practice that it is advantageous to discard variables. PLS can be viewed as a good method to do regression analysis, because the components are selected so that they describe the dependent variables in a certain sense. The main issue is that the PLS method is able to keep the number of variables low. We shall consider here more closely the regression formulation of the PLS method.

A linear regression model can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{B} + \mathbf{e} \tag{24}$$

Here $\mathbf{B}$ are the regression coefficients, $\mathbf{X}$ is the design matrix and $\mathbf{e}$ is the residual. In the following we will use the notation

$$\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k) \qquad \mathbf{P}_i = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_i)$$

i.e. $\mathbf{P}_i$ is the first i columns of $\mathbf{P}$; similarly for other matrices. We suppose here that $\mathbf{X}$ is an $N \times k$ matrix of rank k. Then we can write

$$\mathbf{X} = \Sigma^k \mathbf{t}_i \mathbf{p}_i^T = \mathbf{T}\mathbf{P}^T$$

Let $\mathbf{R}$ be the inverse of $\mathbf{P}^T$: $\mathbf{R} = (\mathbf{P}^T)^{-1}$. We shall consider equation (24) more closely. Suppose now that we have selected i PLS components. The PLS transformation is

$$\mathbf{T} = \mathbf{X}\mathbf{R} \tag{25}$$

and the regression equation (24) can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{R}\mathbf{P}^T\mathbf{B} + \mathbf{e} = \mathbf{T}\mathbf{C} + \mathbf{e} \quad \text{with } \mathbf{C} = \mathbf{P}^T\mathbf{B}$$

Thus if values of the independent variables are given as $\mathbf{x}_0$, the predictor is computed as

$$\mathbf{y} = \mathbf{x}_0^T\mathbf{R}\mathbf{C} \tag{26}$$

The columns of **R** are easily computed during the iterations. We have

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 \qquad \text{or} \qquad \mathbf{r}_1 = \mathbf{w}_1$$

$$\mathbf{t}_2 = \mathbf{X}(\mathbf{I} - \mathbf{w}_1\,\mathbf{p}_1^{\mathrm{T}})\mathbf{w}_2 \qquad \text{or} \qquad \mathbf{r}_2 = \mathbf{w}_2 - \mathbf{w}_1\,(\mathbf{p}_1^{\mathrm{T}}\mathbf{w}_2) = \mathbf{w}_2 - \mathbf{r}_1\,(\mathbf{p}_1^{\mathrm{T}}\mathbf{w}_2)$$

In general we have

$$\mathbf{r}_i = \mathbf{w}_i - \mathbf{r}_{i-1}\,(\mathbf{p}_{i-1}^{\mathrm{T}}\mathbf{w}_i)$$

from which the columns of **R** can be computed successively.

One dependent variable

It is instructive to look at the case where there is only one dependent variable. In this case the vectors involved can be computed directly from the data.

The **Y** matrix is now of rank one and so is the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$ that has **w** as an eigenvector. It is given by

$$\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}/a \quad \text{with } a = (\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y})^{1/2}$$

and the eigenvalue is

$$l = \mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = a^2$$

The vectors **t**, **c** and **u** are

$$\mathbf{t} = \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}/a \qquad \mathbf{c} = \mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}/(a\,\mathbf{t}^{\mathrm{T}}\mathbf{t}) = a/(\mathbf{t}^{\mathrm{T}}\mathbf{t}) \qquad \mathbf{u} = \mathbf{c}\mathbf{Y}$$

The regression coefficient b is equal to c and the approximate variance of b is

$$\mathrm{Var}(b) = s^2/(\mathbf{t}^{\mathrm{T}}\mathbf{t})$$

Thus an approximate t-test that the ith coefficient b_i is zero is

$$b_i/(\mathrm{Var}(b_i))^{1/2} = b_i\,(\mathbf{t}^{\mathrm{T}}\mathbf{t})^{1/2}/(\mathbf{Y}_{i+1}^{\mathrm{T}}\mathbf{Y}_{i+1}/(N-i-1))$$

with $\mathbf{Y}_{i+1}^{\mathrm{T}}\mathbf{Y}_{i+1}$ the residual variance. The number of degrees of freedom for the t-test is $N-i-1$. It should be emphasized that the t-test is only an approximate t-test. The regression coefficient can in fact be written as

$$b = (\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}})^{3/2}/(\mathbf{Y}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^2\mathbf{Y})$$

which shows that it is not an easy task to obtain the exact distribution of b.

The total variation can be written as a sum of the contributions of the individual components:

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \Sigma^i\, b_j^2\,(\mathbf{t}_j^{\mathrm{T}}\,\mathbf{t}_j) + \mathbf{Y}_{i+1}^{\mathrm{T}}\,\mathbf{Y}_{i+1}$$
$$= \Sigma^i\, l_j/(\mathbf{t}_j^{\mathrm{T}}\mathbf{t}_j) + \mathbf{Y}_{i+1}^{\mathrm{T}}\mathbf{Y}_{i+1}$$

The components $\mathbf{t}_j$ are orthogonal to one another and orthogonal to the residual vector. Thus

$$b_j^2\,(\mathbf{t}_j^{\mathrm{T}}\mathbf{t}_j)/(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})$$

represents the proportions of variance of **Y** that is accounted for by the jth component.

The projection matrix

The regression analysis amounts to projecting the **Y** matrix onto a subspace of **X** generated by the **t** vectors. The projection can be written as

$$\mathbf{Y}_{A+1} = \mathbf{P}_T\mathbf{Y} = (\Sigma^A \mathbf{t}_i \mathbf{t}_i^T/(\mathbf{t}_i^T\mathbf{t}_i))\ \mathbf{Y}$$

and the projection matrix is

$$\begin{aligned}\mathbf{P}_T &= \Sigma\, \mathbf{t}_i\, \mathbf{t}_i^T/(\mathbf{t}_i^T\mathbf{t}_i)\\ &= \Sigma\, X_i\, \mathbf{w}_i\, \mathbf{t}_i^T/(\mathbf{t}_i^T\mathbf{t}_i)\\ &= \Sigma\, X\, r_i\, \mathbf{t}_i^T/(\mathbf{t}_i^T\mathbf{t}_i)\end{aligned}$$

The matrix $\mathbf{P}_T$ is idempotent and symmetric. Thus we have for $\mathbf{P}_T = (p_{ij})$

(a) $\Sigma_i\, p_{ij}^2 = p_{jj}$

(b) $0 \leq p_{jj} \leq 1$

(c) $\Sigma\, p_{jj} = A$ = number of **t** vectors.

These properties show that $\mathbf{P}_T$ can be analyzed in the same way as the hat-matrix in linear regression analysis. For example, a value of p_{jj} close to one indicates an outlier. For a detailed account of how to study the projection matrix see Belsley *et al.*[10]

Mean square error

The variance for a stochastic variable Z is given by

$$\mathrm{Var}(Z) = E(Z^2) - (E(Z))^2$$

If $Z = Y - \hat{Y}$, we can write the formula as

$$E(|Y - \hat{Y}|^2) = (E(Y) - E(\hat{Y}))^2 + \mathrm{Var}(Y - \hat{Y}) \tag{27}$$

Verbally this formula can be written as

mean square error = squared bias + residual variance

This formula is of fundamental importance in linear modelling of data. The reason is that there is a lower bound on the mean square error. After having included a certain number of variables or components, the mean square error does not get any smaller when new variables or components are introduced into the model. Introducing more variables generally reduces the bias, but after a certain number of variables the residual variance will increase or the mean square error may even increase.

One can recommend estimating the bias. This can be achieved in the following manner. Carry out the PLS estimation until the regression coefficients become zero or close to zero. The cross-validation tells one when to stop extracting components. The difference in estimated Y from where cross-validation tells one to stop and all the PLS components will be an estimate of the bias. Normally the cross-validation tells one to stop before all components have been extracted. The last components may have regression coefficients different from zero, but do not contribute to the predictive power of the model. Wold[11] has given a good account of cross-validation with the purpose of detecting when to stop extracting components from data.

Variance of predictors

It is instructive to consider the variance of the predictor in standard multiple regression. Let the predictor be

$$y = b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

If the theoretical covariance matrix is proportional to the identity matrix, the variance of the predictor is

$$\mathrm{Var}(y) = s^2 \mathbf{x}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}$$

It is easy to show that $\mathbf{x}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}$ always increases by increasing k. That is, if s^2 can be considered constant, the variance will always increase when new variables are introduced into the equation. s^2, being the average of squared residuals, typically stabilizes after introducing a few variables or components. This is an example of the fact that it is important to keep the number of variables or components small.

We shall now consider the case of prediction in PLS models. Let $\mathbf{x}$ be an observation for which we want to do a prediction. We first consider the PLS transformation

$$\mathbf{z}^{\mathrm{T}} = \mathbf{x}^{\mathrm{T}}\mathbf{R}$$

The predicted y value is now

$$y = z_1 c_1 + z_2 c_2 + \ldots + z_A c_A = \mathbf{z}^{\mathrm{T}}\mathbf{c}$$

If we assume that $\mathrm{Var}(\mathbf{c}) = s^2\,\mathrm{diag}(1/(\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i))$, we get as variance

$$\mathrm{Var}(y) = s^2 \Sigma_1^A z_i^2/(\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i)$$

where $s^2 = \mathbf{Y}_{A+1}^{\mathrm{T}}\mathbf{Y}_{A+1}/(N-A-1)$.

TEST FOR DIMENSIONALITY

An important question is to establish a test for the number of PLS components. We shall consider here a likelihood ratio test for testing the number of components. We will assume that (x,y) follow a multivariate normal distribution with a covariance matrix $\mathbf{S}$ of dimension $(n + m) \times (n + m)$. Let $\mathbf{S}$ be partitioned as follows:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{21} & \mathbf{S}_{22} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$$

Here $\mathbf{S}_{11}$ is the covariance matrix of x of dimension $n \times n$. The test that we consider relates to the dimension of $\mathbf{S}_{21}$, the covariance matrix between x and y. We shall develop a likelihood ratio test that there are k singular value components. This implies that the covariance matrix $\mathbf{S}_{12}$ is of rank k and thus can be written as

$$\mathbf{S}_{21} = \Sigma_1^k \mathbf{f}_i \mathbf{g}_i^{\mathrm{T}} \tag{28}$$

for some independent vectors $(\mathbf{f}_i)$ and $(\mathbf{g}_i)$. When there are k significant singular value components, there will be at least k PLS components.

The procedure can be used as follows. We test the hypothesis sequentially for $k = 0,1,\ldots,\min(n,m)$. The first value of k for which we accept the hypothesis by the likelihood ratio test can be used as an estimate of the dimensionality of $\mathbf{S}_{21}$. When k is zero, there are no significant components in $\mathbf{S}_{21}$, and therefore no (more) significant PLS components.

We will suppose here, as usual, that the data matrices $\mathbf{X}$ and $\mathbf{Y}$ have been adjusted for the appropriate mean values. The sample covariance matrix can be written as

$$\mathbf{V} = \begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Y} \\ \mathbf{Y}^T\mathbf{X} & \mathbf{Y}^T\mathbf{Y} \end{bmatrix}$$

In the proof of the following theorem we will use the Cholesky partition of a positive definite matrix into a product of a lower triangular matrix and its transpose. For the $\mathbf{V}$ matrix we can write $\mathbf{V} = \mathbf{F}\mathbf{F}^T$ or, as partitioned,

$$\mathbf{V} = \begin{bmatrix} \mathbf{F}_{11} & \mathbf{0} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{11}^T & \mathbf{F}_{21}^T \\ \mathbf{0} & \mathbf{F}_{22}^T \end{bmatrix}$$

Here $\mathbf{F}_{11}$ and $\mathbf{F}_{22}$ are lower triangular matrices. The matrix $\mathbf{F}_{21}$ may be computed as

$$\mathbf{F}_{21} = \mathbf{Y}^T\mathbf{X}(\mathbf{F}_{11}^T)^{-1} \tag{29}$$

Theorem

Let $\mathbf{Y}^T\mathbf{X}$ have the singular value decomposition

$$\mathbf{Y}^T\mathbf{X} = \Sigma\, \mathbf{h}_i\, \mathbf{k}_i^T$$

and let

$$\mathbf{s}_i^T = \mathbf{k}_i^T(\mathbf{F}_{11}^T)^{-1} \qquad \mathbf{v}_i = \mathbf{F}_{22}^{-1}\, \mathbf{h}_i$$

Then under the hypothesis the estimate of $\mathbf{S}_{21}$ is the first k singular value components of $\mathbf{Y}^T\mathbf{X}$ and the likelihood ratio criterion is

$$L_{H0}/L_{\max} = |\mathbf{I} + (\Sigma_{k+1}\, \mathbf{v}_i\, \mathbf{s}_i^T)\, (\Sigma_{k+1}\, \mathbf{v}_i\, \mathbf{s}_i^T)^T|^{-N/2}$$

Proof

The likelihood ratio criterion is based on considering the multivariate normal density for N observations

$$L(\mathbf{S}) = |\mathbf{S}|^{-N/2} \exp[-\mathrm{tr}(\mathbf{S}^{-1}\mathbf{V})/2]$$

An unrestricted maximum is

$$\max L(\mathbf{S}) = |\mathbf{V}|^{-N/2}\, N^{qN/2} \exp(-qN/2) \qquad q = n + m \tag{30}$$

To determine the maximum of $L(\mathbf{S})$ under the hypothesis, we partition $\mathbf{S}$ in the same way as $\mathbf{V}$. Let the Cholesky decomposition of $\mathbf{S}$ be given by $\mathbf{G}$ matrices with the same size and indices as the $\mathbf{F}$ matrices. Then we can write $\mathrm{tr}(\mathbf{S}^{-1}\mathbf{V})$ as

$$\mathrm{tr}(\mathbf{S}^{-1}\mathbf{V}) = \mathrm{tr}(\mathbf{X}^T\mathbf{X}\mathbf{S}_{11}^{-1}) + \mathrm{tr}\{\, (\mathbf{G}_{22}\mathbf{G}_{22}^T)^{-1} \times [\mathbf{F}_{22}\mathbf{F}_{22}^T + (\mathbf{F}_{21} - \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{F}_{11})(\mathbf{F}_{21} - \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{F}_{11})^T]\, \}$$

Since

$$|\mathbf{S}| = |\mathbf{S}_{11}|\,|\mathbf{G}_{22}\mathbf{G}_{22}^{\mathrm{T}}|$$

the likelihood function $L(\mathbf{S})$ factors into a product of two normal densities. The maximization of each factor of $L(\mathbf{S})$ gives

$$\begin{aligned} N\mathbf{S}_{11} &= \mathbf{X}^{\mathrm{T}}\mathbf{X} \\ N\mathbf{G}_{22}\mathbf{G}_{22}^{\mathrm{T}} &= \mathbf{F}_{22}\mathbf{F}_{22}^{\mathrm{T}} + (\mathbf{F}_{21} - \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{F}_{11})(\mathbf{F}_{21} - \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{F}_{11})^{\mathrm{T}} \\ &= \mathbf{F}_{22}\mathbf{F}_{22}^{\mathrm{T}} + (\mathbf{F}_{21} - \mathbf{G}_{21})(\mathbf{F}_{21} - \mathbf{G}_{21})^{\mathrm{T}} \\ &= \mathbf{F}_{22}\mathbf{F}_{22}^{\mathrm{T}} + (\mathbf{Y}^{\mathrm{T}}\mathbf{X} - \mathbf{S}_{21})(\mathbf{F}_{11}\mathbf{F}_{11}^{\mathrm{T}})^{-1}\,(\mathbf{Y}^{\mathrm{T}}\mathbf{X} - \mathbf{S}_{21})^{\mathrm{T}} \end{aligned}$$

In the second equality we have used the result from the first maximization, which gives $\mathbf{G}_{11} = \mathbf{F}_{11}$. In the last equality we have used equation (29).

The value of the likelihood function with these equations is

$$L(\mathbf{S}) = N^{qN/2}\,|\mathbf{V}|^{-N/2}\exp(-qN/2)\,|\mathbf{I} + \mathbf{Z}|^{-N/2} \tag{31}$$

where

$$\mathbf{Z} = \mathbf{F}_{22}^{-1}\,(\mathbf{Y}^{\mathrm{T}}\mathbf{X} - \mathbf{S}_{21})(\mathbf{F}_{11}\mathbf{F}_{11}^{\mathrm{T}})^{-1}\,(\mathbf{Y}^{\mathrm{T}}\mathbf{X} - \mathbf{S}_{21})^{\mathrm{T}}(\mathbf{F}_{22}^{\mathrm{T}})^{-1}$$

If we write $\mathbf{Z} = \mathbf{W}\mathbf{W}^{\mathrm{T}}$, we have

$$|\mathbf{I} + \mathbf{Z}| = \Pi\,(1 + \mathrm{ch}_j(\mathbf{Z})) = \Pi\,(1 + \mathrm{ch}_j^2(\mathbf{W}))$$

From equation (23) it follows that

$$\mathrm{ch}_j^2(\mathbf{W}) \geqslant \mathrm{ch}_{j+k}^2(\mathbf{F}_{22}^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{F}_{11}^{\mathrm{T}})^{-1})$$

with equality attained when $\mathbf{S}_{21}$ is the first k singular value components of $\mathbf{Y}^{\mathrm{T}}\mathbf{X}$. With this choice of $\mathbf{S}_{21}$ we have the maximum value of $L(\mathbf{S})$ under the hypothesis. Taking the ratio between equations (30) and (31), we get the results.

The above likelihood ratio test is related to the likelihood ratio test for the number of canonical correlation components. The treatment of the test statistic is similar in both cases. Thus if

$$\begin{aligned} \Lambda &= (\max_H L(\mathbf{S})/\max L(\mathbf{S}))^{2/N} \\ &= |\mathbf{I} + (\Sigma_{k+1}\mathbf{v}_i\,\mathbf{s}_i^{\mathrm{T}})(\Sigma_{k+1}\mathbf{v}_i\,\mathbf{s}_i^{\mathrm{T}})^{\mathrm{T}}|^{-1} \end{aligned}$$

we have that

$$T_k = -[N - (n + m + 1)/2]\log\Lambda$$

is asymptotically χ^2-distributed with $(n - k)\,(m - k)$ degrees of freedom. The accuracy of this approximation has been studied by, for example, Glynn and Muirhead.[12]

In the PLS context this provides us with a reasonably good test that there are significant PLS components, because the PLS component is the first singular value component. It can be used both at the start of the iteration in order to get information on the number of singular value components of $\mathbf{Y}^{\mathrm{T}}\mathbf{X}$, and at each iteration to see if there are significant components left. The modifications needed to take into account that in PLS we work with residual matrices are a complicated matter which is not considered here. Thus the above formulae should be used for guidance only.

THE STRUCTURE OF PLS REGRESSION

PLS iterations as a modelling technique

The PLS regression consists of three parts that can be formulated as follows. For each dimensions the steps are:

1. Determine **w**
2. Compute **t**, **c** and **u** as in the PLS algorithm
3. Adjust **X** and **Y** as in the PLS algorithm to reduce the dimension of **X**.

The coefficient b is a scaling factor. If **c** is not scaled to be of length one, $b = |\mathbf{c}|$.

In the PLS algorithm the vector **w** is determined as the eigenvector associated with the largest eigenvalue:

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\,\mathbf{w} = l\,\mathbf{w} \tag{32}$$

The interpretation of the PLS algorithm follows from this method of determining **w**. The orthogonality properties of $(\mathbf{t}_i)$ and $(\mathbf{w}_i)$ (and also of $(\mathbf{p}_i)$) follow from 2 and 3 above. Thus they are also valid for other ways of determining **w**.

In principal component regression **w** is determined as the eigenvector associated with the largest eigenvalue of the covariance matrix for **X**:

$$\mathbf{X}^T\mathbf{X}\,\mathbf{w} = l\,\mathbf{w} \tag{33}$$

A third way to determine **w** is to consider the maximization

$$\max \mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} \qquad \text{subject to} \qquad \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} = 1$$

This is attained when **w** is the eigenvector corresponding to the largest eigenvalue of the eigenvalue problem:

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\,\mathbf{w} = l\,\mathbf{X}^T\mathbf{X}\,\mathbf{w} \tag{34}$$

This choice of **w** would give the maximal reduction in the residual covariance matrix of **Y**:

$$\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T\mathbf{Y}/(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w})$$

i.e. it would give a component **t** such that the residual variance matrix is minimized.

In all three cases the set of vectors $(\mathbf{t}_i)$, $(\mathbf{w}_i)$ and $(\mathbf{p}_i)$ would share the same orthogonality properties, and all can be analyzed in a similar way. Statistical tests and analyses can be based on analogous approximate procedures.

There are other choices of **w** which are interesting. If **w** is chosen so that the ith co-ordinate of **w** is one and the others are zero, then this coice corresponds to selecting the jth variable to enter the model. This kind of choice is reasonable when some variables are more important than the others, and one wants the important variables to enter the model with their full value, although that may not be optimal according to the criteria of PLS.

Thus PLS consists basically of two parts. The first part is the statistical one of how to do the PLS regression. The second part is basically a technique to analyse data by successively orthogonalizing the data matrix **X**. At each step or component one looks at the regression coefficient b and tests it for significance, or evaluates the predictive power of the model by cross-validation, and then orthogonalizes **X** so that what is left is orthogonal to what has previously been selected.

PLS decompositions

We summarize here the decompositions that we arrive at by PLS. We suppose that A components have been extracted. The matrices $\mathbf{X}$ and $\mathbf{Y}$ can be written as

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \ldots + \mathbf{t}_A \mathbf{p}_A^T + \mathbf{X}_{A+1}$$
$$\mathbf{Y} = \mathbf{t}_1 \mathbf{c}_1^T + \mathbf{t}_2 \mathbf{c}_2^T + \ldots + \mathbf{t}_A \mathbf{c}_A^T + \mathbf{Y}_{A+1}$$

The variance and covariance matrices can be written as

$$\mathbf{X}^T\mathbf{X} = \mathbf{p}_1 \mathbf{p}_1^T (\mathbf{t}_1^T\mathbf{t}_1) + \ldots + \mathbf{p}_A \mathbf{p}_A^T (\mathbf{t}_A^T\mathbf{t}_A) + \mathbf{X}_{A+1}^T\mathbf{X}_{A+1}$$
$$\mathbf{X}^T\mathbf{Y} = \mathbf{p}_1 \mathbf{c}_1^T (\mathbf{t}_1^T\mathbf{t}_1) + \ldots + \mathbf{p}_A \mathbf{c}_A^T (\mathbf{t}_A^T\mathbf{t}_A) + \mathbf{X}_{A+1}^T\mathbf{Y}_{A+1}$$
$$\mathbf{Y}^T\mathbf{Y} = \mathbf{c}_1 \mathbf{c}_1^T (\mathbf{t}_1^T\mathbf{t}_1) + \ldots + \mathbf{c}_A \mathbf{c}_A^T (\mathbf{t}_A^T\mathbf{t}_A) + \mathbf{Y}_{A+1}^T\mathbf{Y}_{A+1}$$

and the residual covariance matrix for $\mathbf{Y}$ is

$$(\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}_{A+1}^T\mathbf{Y}_{A+1} = \mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$$
$$= \mathbf{Y}^T\mathbf{Y} - \{\mathbf{c}_1 \mathbf{c}_1^T (\mathbf{t}_1^T\mathbf{t}_1) + \ldots + \mathbf{c}_A \mathbf{c}_A^T(\mathbf{t}_A^T\mathbf{t}_A)\}$$

All these equations follow from the fact that the $\mathbf{t}_1$ are orthogonal to $\mathbf{X}_{A+1}$ and $\mathbf{Y}_{A+1}$ by construction.

The size or the relative size can be used to give an index of how well PLS has performend. For example, it is useful to get in terms of a percentage how much the covariance in the data has been reduced by computing

$$100\ \mathrm{tr}(\mathbf{X}_{A+1}^T\mathbf{Y}_{A+1}\mathbf{Y}_{A+1}^T\mathbf{X}_{A+1})/\mathrm{tr}(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})$$

Similarly by computing

$$100\mathrm{tr}(\mathbf{X}_{A+1}^T\mathbf{X}_{A+1})/\mathrm{tr}(\mathbf{X}^T\mathbf{X}) \qquad \text{and} \qquad 100\{1 - \mathrm{tr}(\mathbf{Y}_{A+1}^T\mathbf{Y}_{A+1})/\mathrm{tr}(\mathbf{Y}^T\mathbf{Y})\}$$

we get the percentage variance of $\mathbf{X}$ that was used and the percentage variance of $\mathbf{Y}$ that the model can account for.

REFERENCES

1. H. Wold, *Encyclopedia of Statistical Sciences*, Wiley, New York (1984).
2. P. Geladi and B. Kowalski, *Anal. Chim. Acta* **185**, (1986).
3. P. Geladi, *Anal. Chim. Acta* **185**, 19 (1986).
4. I. S. Helland, *Report 21* Department of Mathematics and Statistics, Agricultural University of Norway (1986).
5. R. Manne, *Chemometrics and Intell. Lab. Syst.*, 187 (1987).
6. S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom and H. Wold, *Bull. 46th ISI* (1987).
7. S. Wold, C. Albano, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjöström, *Multivariate Data Analysis in Chemistry, in Chemometrics, Mathematics and Statistics in Chemistry*, p. 17, ed. by B. R. Kowalski, Reidel Publishing Company, Dordrecht (1984).
8. N. Aronszajn, *Trans. Amer. Math. Soc.* **68**, 337 (1950).
9. C. R. Rao, *J. Multivariate Anal.* **9**, 362 (1979).
10. D. A. Belsley, E. Kuh and R. E. Welsch, *Regression Diagnostics*, Wiley, New York (1980).
11. S. Wold, *Technometrics* **20**, 397 (1978).
12. W. J. Glynn and R. J. Muirhead, *J. Multivariate Anal.* **8**, 468 (1978).