MODELING, IDENTIFICATION AND CONTROL, 1998, VOL. 19, NO. 3, 117–140
doi:10.4173/mic.1998.3.1

The partial least squares algorithm: a truncated Cayley-Hamilton series approximation used to solve the regression problem

DAVID DI RUSCIO†

In this paper it is shown that the PLS algorithm for univariate data is equivalent to using a truncated Cayley-Hamilton polynomial expression of degree $1 \leq a \leq r$ for the matrix inverse $(X^TX)^{-1} \in \mathbb{R}^{r\times r}$ used to compute the LS solution. Furthermore, the a coefficients in this polynomial are computed as the LS optimal solution (minimizing parameters) to the prediction error. The resulting solution is non-iterative. The solution can be expressed in terms of one matrix inverse and is given by $B_{PLS} = K_a(K_a^TX^TXK_a)^{-1}K_a^TX^TY$ where $K_a \in \mathbb{R}^{r\times r}$ is the controllability (Krylov) matrix for the pair (X^TX, X^TY).

The iterative PLS algorithm for computing the orthogonal weighting matrix W_a which is presented in the literature is in this paper shown to be equivalent to computing an orthonormal basis (using, e.g., the QR algorithm) for the column space of K_a. The PLS solution can then equivalently be computed as $B_{PLS} = W_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY$, where W_a is the Q-orthogonal matrix from the QR decomposition $K_a = W_aR$.

Furthermore, we have presented an optimal and non-iterative truncated Cayley-Hamilton polynomial LS solution for multivariate data. This solution is found as the minimizing solution of a prediction error criterion.

1. Introduction

The Partial Least Squares (PLS) algorithm and its solution has got great attention and is widely used in chemometrics, which is defined as *The use of mathematics and statistics on chemical data* in Martens and Næs (1989).

PLS was introduced by Wold (1975), (1985) as an algorithm for, e.g., computing a solution B_{PLS} for the regression coefficients B in a linear model $Y = XB + E$ from known data matrices X and Y. The PLS algorithm is analyzed in, among others, Martens and Næs (1985), Manne (1987), Lorber *et al.* (1987), Helland (1988), Höskuldsson (1988), Frank and Friedman (1993), Burnham *et al.* (1996) and ter Braak and de Jong (1998).

While PLS have been used in many applications in chemometrics, there have been few applications to system parameter identification. PLS have traditionally been used on data from steady state systems, and for the problem of constructing a predictor for the output of a system. However, PLS was used in subspace (dynamic) system identification in Di Ruscio (1997) in order to compute a basis for the observability matrix which is the basic in subspace system identification of dynamic systems.

PLS is in the literature presented as an iterative algorithm, i.e., partial or piece-wise linear regression. One of the main contributions in the following paper is to give a new

Revised May 12, 1998.

†Department of Process Automation, Telemark Institute of Technology, N-3914 Porsgrunn, Norway. Tel: + 47 35 57 51 61. Fax: + 47 35 57 52 50. Email: David.Di.Ruscio@hit.no

interpretation and description of the basic PLS solution. We will show that the basic PLS algorithm is non-iterative and can be computed as the minimizing solution to a prediction error criterion. This is believed to be of interest for the community working with system identification in general, as well as to chemometricians.

We will try to give a discussion of the PLS algorithm which is as simple as possible. We believe that this only can be done by introducing as few definitions and variables as possible. In the literature the PLS algorithm and its solution is usually presented in terms of so called score vectors, loading vectors, weighting vectors and various iterative orthogonalization (deflation) processes, in addition to the solution for the matrix of regression coefficients. This work has shown that it exists as a very simple non-iterative algorithm for computing the PLS solution. It can be shown that the PLS solution can be expressed in terms of some weighting vectors only. We will therefore concentrate our discussion on these weights. A discussion and definitions of the score vectors, loading vectors, etc, which usually are defined in connection with the PLS algorithm, can be found elsewhere.

The rest of this paper is organized as follows. Some basic system definitions are presented in Section 2. A basic preliminary result concerning the latent variable LS solution is presented in Section 3. The PLS algorithm is reviewed and some new results are presented in Section 4. The main contributions concerning the interpretation of the PLS solution are presented in Sections 5 and 6. A real world example from the pulp and paper industry is presented in Section 9 and some conclusions follow in Section 11.

2. Notation, basic- and system-definitions

Define $y_k \in \mathbb{R}^m$ as a vector of output variables at observation number k. The output variables are sometimes referred to as *response variables*. Similarly a vector $x_k \in \mathbb{R}^r$ of input variables is defined. It is assumed that the vector of output variables y_k are linearly related to the vector of input variables x_k as follows

$$y_k = B^T x_k + e_k, \tag{1}$$

where e_k is a vector of white noise with covariance matrix $\mathrm{E}(e_k e_k^T)$ and k is the number of observations.

With N observations $k = 1, \ldots, N$ we define an output (or response) data matrix $Y \in \mathbb{R}^{N \times m}$ and an input data matrix $X \in \mathbb{R}^{N \times r}$ as follows

$$Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix}, \quad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}. \tag{2}$$

The data matrices Y and X are assumed to be known.

The linear relationship (1) can be written as the following linear matrix equation

$$Y = XB + E, \tag{3}$$

where B is a matrix with regression coefficients. $E \in \mathbb{R}^{N \times m}$ is in general an unknown matrix of noise vectors, defined as follows

$$E = \begin{bmatrix} e_1^T \\ \vdots \\ e_N^T \end{bmatrix}. \tag{4}$$

The linear relationship between the output (response) and the input data (or regressors) is an important assumption and restriction for the PLS as well as any LS algorithm to work.

We will in this work analyze systems with multiple output variables in the data matrix Y. This is often referred to as multivariable (or multivariate) systems.

If we only are interested in the matrix of regression coefficients B, then one should note that it (for steady state systems) usually is sufficient to consider one output at a time and only investigate single output systems. This means that the multivariable LS problem can be solved from m single ouput LS problems, i.e., each column in B is estimated from a separated LS problem.

Note also that (instead of modeling one output variable at a time) equation (3) can be transformed to an equivalent model with one ouput in different ways. Two possible models with one ouput, which are equivalent to the multivariable model (3) are presented as follows

$$\mathrm{cs}(Y) = (I_m \otimes X)\mathrm{cs}(B) + \mathrm{cs}(E), \tag{5}$$

$$\mathrm{cs}(Y^T) = (X \otimes I_m)\mathrm{cs}(B^T) + \mathrm{cs}(E^T), \tag{6}$$

where $\mathrm{cs}(\cdot)$ is the column string operator and $\otimes$ is the Kronecker product. $\mathrm{cs}(Y) \in \mathbb{R}^{Nm}$ is a column vector constructed from Y by stacking each column in Y on another. We also have $(I_m \otimes X) \in \mathbb{R}^{Nm \times rm}$ and $\mathrm{cs}(B) \in \mathbb{R}^{rm}$.

Note that (6) can be constructed directly from (1) by first writing (1) as

$$y_k = (x_k^T \otimes I_m)\mathrm{cs}(B^T) + e_k \tag{7}$$

and then combine all N equations ($k = 1, \ldots, N$) into a matrix equation of the form (3).

However, for the sake of completeness we will in general consider multivariable/ multivariate (multiple output) systems of the form (3).

One important application of the PLS algorithm is to compute projections. An example is the problem of computing the projection of the row space of Y^T onto the row space of X^T.

In the literature PLS is usually presented as two algorithms, PLS1 and PLS2. PLS1 is concerned with univariate $Y \in \mathbb{R}^N$, and PLS2 is concerned with multivariate $Y \in \mathbb{R}^{N \times m}$. We will follow these definitions.

3. Preliminary results

We will in this paper consider Least Squares solutions which may be regularized approximations to the Ordinary Least Squares solution, as defined below.

3.1. *Definition*

Consider a Least Squares solution of the form

$$B_{\mathrm{M}} = W_a p^* \tag{8}$$

where $W_a \in \mathbb{R}^{r \times a}$ is a weighting matrix, a is the number of significant components (latent variables) which is restricted to $1 \le a \le r$ and $p^* \in \mathbb{R}^a$ is the LS optimal solution to

$$p^* = \arg\min_p \|Y + X\underbrace{W_a p}_{B_{\mathrm{M}}}\|_F^2. \tag{9}$$

Furthermore, the optimal p and the LS solution B_M corresponding to the particular weighting matrix W_a, are given by*

$$B_M = W_a(W_a^T X^T X W_a)^{-1} W_a^T X^T Y \tag{10}$$

and

$$p^* = (W_a^T X^T X W_a)^{-1} W_a^T X^T Y \tag{11}$$

where we assume that $(W_a^T X^T X W_a)^{-1}$ is non-singular for some $1 \leq a \leq r$.

Note that any square non-singular matrix W_r gives the OLS solution $B_{OLS} = (X^T X)^{-1} X^T Y$. Hence, M = OLS in Equation (10).

Furthermore, choosing $W_a = V_1$ where $V_1 \in \mathbb{R}^{r \times a}$ is the a first columns in the right singular vector matrix V from the SVD,

$$X = USV^T = [U_1 \;\; U_2] \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} [V_1 \;\; V_2]^T$$

where $U_1 \mathbb{R}^{N \times a}$ and $S_1 \mathbb{R}^{a \times a}$, gives the PCR solution (truncated SVD solution), $B_{PCR} = V_1 S_1^{-1} U_1^T Y$.

We will in this paper show that the PLS solution can be defined similarly.

Note also that W_a span the column space (range) of the solution B_M. W_a could therefore have been defined as the range $\mathcal{R}(B_M)$ of the solution, instead of a weighting matrix.

4. The PLS solution

The PLS algorithm for computing a solution to the regression problem is presented by Wold (1975) and (1985). This algorithm is an extension of the NIPALS (power iteration) algorithm for computing principal components published in Wold (1966). We will also refer to Frank and Friedman (1993) for a review and pseudo code presentation of Wold's PLS algorithm. We will give below a different *ad-hoc* description of the PLS algorithm which has some similarities to the description by Helland (1988).

The normal equation is of central importance in LS problems and its solutions. Therefore it makes sense to study the PLS algorithm with the normal equation as a starting point. The normal equation $X^T Y = X^T X B(W)$ substituted for a LS solution $B(W) = W(W^T X^T X W)^{-1} W^T X^T Y$ yields

$$X^T Y = X^T X W (W^T X^T X W)^{-1} W^T X^T Y. \tag{12}$$

where we in the following discussion assume univariate $Y \in \mathbb{R}^N$. The extension to the multivariate case will be clarified later. The first weight vector w_1 in the PLS weighting matrix W can be taken directly as the correlation $w_1 = X^T Y$ when Y is a vector. When Y is a matrix then w_1 can be taken as the left singular vector of $X^T Y$ which corresponds to the largest singular value. This is equivalent to putting w_1 equal to the eigenvector corresponding to the largest eigenvalue of the matrix $X^T Y Y^T X$. Power iteration is a convenient tool for this computation.

The PLS algorithm was probably derived in a rather ad hoc manner, Helland (1988). Having this in mind, it is not too strange to choose a weight vector $w_1 = X^T Y$. For the sake of convenience w_1 is often scaled, e.g., the choice $w_1 = X^T Y / \|X^T Y\|_F$ gives an orthonormal weight vector, i.e., $w_1^T w_1 = 1$. However, as also pointed out in Helland (1988), this scaling is not necessary. In order not to complicate the discussion we choose not to use scaled weight vectors.

Substituting this and $W_1 = w_1$ into the normal equation gives us a residual

$$w_2 = w_1 - X^T X B_1 \text{ where } B_1 = W_1(W_1^T X^T X W_1)^{-1} W_1^T w_1 \text{ and } W_1 = w_1. \quad (13)$$

Note, that B_1 is the matrix of regression coefficients computed by the PLS algorithm when the number of components is equal to $a = 1$. It is now important to note that $w_1^T w_2 = 0$, i.e., w_1 is normal to the residual w_2. Hence, this residual w_2 after choosing $W_1 = w_1 = X^T Y$ is the 2nd weight vector used by the PLS algorithm. We now define a normal equation for the residual, i.e,

$$w_2 = X^T X B_2 \text{ where } B_2 = W_2(W_2^T X^T X W_2)^{-1} W_2^T w_2 \text{ and } W_2 = [w_1 \ w_2]. \quad (14)$$

The residual

$$w_3 = w_2 - X^T X B_2 \quad (15)$$

is taken as the 3rd weight vector in the PLS algorithm. We now define yet a new normal equation

$$w_3 = X^T X B_3 \text{ where } B_3 = W_3(W_3^T X^T X W_3)^{-1} W_3^T w_3 \text{ and } W_3 = [w_1 \ w_2 \ w_3]. \quad (16)$$

From this it is also simple to show that $w_2^T w_3 = 0$, because w_2 is normal to the residual w_3. The other weight vectors w_i for $i = 4, \ldots, a$ are defined similarly. The procedure for computing the weight vectors which is outlined above is presented in the following Theorem 4.1. We can now combine the above equations to give the following normal equation which give us an expression for the PLS estimate of the matrix of regression coefficients

$$X^T Y = X^T X \overbrace{(B_1 + B_2 + B_3 + \cdots + B_a)}^{B_{\text{PLS}}}. \quad (17)$$

This shows that the problem of computing the PLS solution can be reduced to computing the weight matrix W_a. We have the following theorem for computing the weight vectors, i.e., the columns in matrix W_a.

4.1. *Theorem (PLS1: weight vectors and LS solution)*

Given data matrices $X \in \mathbb{R}^{N \times r}$ and univariate $Y \in \mathbb{R}^N$. The weighting matrix $W_a \in \mathbb{R}^{r \times a}$ used by the PLS algorithm can be computed as follows. The first weighting vector w_1, i.e., the first column in matrix $W_a = [w_1 \cdots w_a]$ can be taken as

$$w_1 = X^T Y \quad (18)$$

when Y is univariate. The other weights $w_2, \ldots, w_a$ are computed recursively from w_1, $W_1 = w_1$ and $X^T X$ as follows. Compute for all $i = 1, \ldots, a-1$

$$w_{i+1} = w_i - X^T X B_i \text{ where } B_i = W_i(W_i^T X^T X W_i)^{-1} W_i^T w_i \quad (19)$$

where W_i increase by one column for each iteration, i.e.

$$W_i = [w_1 \cdots w_i]. \quad (20)$$

Finally, the PLS solution for the matrix of regression coefficients B is given by

$$B_{\text{PLS}} = \sum_{i=1}^{a} B_i \quad (21)$$

which is equivalent to

$$B_{\text{PLS}} = W_a(W_a^T X^T X W_a)^{-1} W_a^T w_1. \quad (22)$$

Theorem 4.1. states that the PLS solution B_{PLS} can be expressed in terms of a weighting matrix $W_a \in \mathbb{R}^{r\times a}$ where a is the number of components. The number of components are usually bounded by $1 \le a \le r$. We shall here note that when $a = r$ then W_a is square and non-singular because W_a is an orthogonal matrix, and that the PLS solution is equal to the ordinary LS estimate, i.e., $B_{\text{PLS}} = B_{\text{OLS}}$.

In Helland (1988) it was shown that the weight vector also can be computed as $w_{i+1} = w_1 - X^TXW_i(W_i^TX^TXW_i)^{-1}W_i^Tw_1$ where $w_1 = X^TY$. However, we can show that w_{i+1} can be computed from W_i and any of its columns, i.e., we have the following alternative equation which can be used instead of Equation (19)

$$w_{i+1} = w_j - X^TXH_iw_j \forall j = 1,\ldots,i \tag{23}$$

where

$$H_i = W_i(W_i^TX^TXW_i)^{-1}W_i^T \tag{24}$$

We shall here note that the matrix product X^TXH_i is an oblique projection. The algorithm for computing the weighting matrix W_i in Theorem 4.1. can be viewed as an orthogonalization process, e.g., Gram-Smith orthogonalization, Golub (1983). The weight vector w_i computed after the ith iteration is orthogonal to the previous weight vectors $w_1,\ldots,w_{i-1}$. This means that $W_i^Tw_i = [0\,0\ldots w_i^Tw_i]^T$. The orthogonalization process in Theorem 4.1. is not unique. For instance, define a non-singular scaling or transformation matrix $T \in \mathbb{R}^{a\times a}$. It is then evident that any weighting matrix defined as $W_a := W_aT$ gives the same PLS solution. This can be proved by substituting W_aT for W_a in Equation (22).

In the literature the PLS algorithm for multivariate Y data is denoted PLS2. In this case we have the following result.

4.2. *Theorem (PLS2: weight vectors and LS solution)*

Given data matrices $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^{N\times m}$. The weighting matrix $W_a \in \mathbb{R}^{r\times a}$ used by the PLS algorithm can be computed as follows. The first weighting vector w_1, i.e., the first column in matrix $W_a = [w_1\ldots w_a]$ can be taken as

$$w_1 := u_1 \text{ where } USV^T := X^TY \text{ and } U = [u_1\ldots u_m] \tag{25}$$

i.e., w_1 can be chosen as the left singular vector which corresponds to the largest singular value of matrix X^TY.

The other weights $w_2,\ldots w_a$ are computed recursively from $W_1 = w_1$, $(X^TY)_1 := X^TY$ and X^TX as follows. Compute for all $i = 1,\ldots a-1$

$$(X^TY)_{i+1} = (I_r - X^TXW_i(W_i^TX^TXW_i)^{-1}W_i^T)(X^TY)_i \tag{26}$$

and

$$w_{i+1} := u_1 \text{ where } USV^T := (X^TY)_{i+1} \text{ and } U = [u_1\ldots u_m] \tag{27}$$

where W_i increase by one column for each iteration, i.e.

$$W_i = [w_1\ldots w_i]. \tag{28}$$

Finally, the PLS solution for the matrix of regression coefficients B is given by

$$B_{\text{PLS}} = W_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY. \tag{29}$$

We will here present some alternative formulations for the problem of computing the PLS weighting vectors. The weight vectors in Theorem 4.1. can equivalently be computed by the following process

$$X_{i+1} = X_i - \frac{X_i w_i w_i^T X_i^T X_i}{w_i^T X_i^T X_i w_i} \tag{30}$$

$$w_{i+1} = X_{i+1}^T Y \tag{31}$$

with $w_1 = X^T Y$ and $X_1 = X$.

The following formulation can also be used in the univariate case ($m = 1$).

$$w_{i+1} = w_i - X^T X w_i \frac{w_i^T w_i}{w_i^T X^T X w_i} \tag{32}$$

where $w_1 = X^T Y$. Note however that the weight vectors computed from this last process may differ from that presented in Theorem 4.1. by a different scaling. This last formulation will be used in the analysis of the PLS algorithm.

The PLS algorithm can be implemented with different formulations of the orthogonalization process, as pointed out above. However, it is important that these weight vectors span the same subspace. The subspace spanned by these weight vectors will be pointed out in the next section.

5. Relationship between PLS and a controllability matrix

It is now important to recognize a relationship between the weight matrix W_a and a so called Krylov matrix. It is known that the problem of computing many orthogonal decompositions has an equivalent problem of computing subspaces for a Krylov matrix. Correspondence with Krylov matrices and orthogonal decompositions are pointed out in Golub (1983). In the control literature the Krylov matrix is known as the controllability matrix.

We will later present the relationship between the PLS solution and the problem of computing the subspace spanned by a controllability matrix. First let us illustrate how the ordinary LS solution is related to a controllability matrix of the pair $(X^T X, X^T Y)$. We have the following proposition.

5.1. *Proposition*

The ordinary LS solution B_{OLS} can be expressed in terms of the controllability matrix of the pair $(X^T X, X^T Y)$ and the coefficients of the characteristic polynomial $\det(\lambda I_r - X^T X) = \lambda^r + p_2 \lambda^{r-1} + \cdots + p_r \lambda + p_{r+1}$. *Assume that $X^T X$ is non-singular, then*

$$B_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y = K_r p \tag{33}$$

where K_r is the controllability matrix for the pair $(X^T X, X^T Y)$, defined as

$$K_r = [X^T Y \quad X^T X X^T Y \quad (X^T X)^2 X^T Y \cdots (X^T X)^{r-1} X^T Y], \tag{34}$$

and p is a vector formed from the coefficients of the characteristic polynomial.

5.1. *Proof*

From the Cayley-Hamilton Theorem we have that $X^T X$ satisfies its own characteristic polynomial, i.e.,

$$(X^T X)^r + p_2 (X^T X)^{r-1} + \cdots + p_r X^T X + p_{r+1} = 0 \tag{35}$$

where $p_2, \ldots, p_{r+1}$ are the coefficients of the characteristic polynomial $\det(\lambda I_r - X^T X)$. *This can be used to form the matrix inverse*

$$(X^TX)^{-1} = \frac{1}{p_{r+1}}(p_r I_r + p_{r-1}X^TX + \cdots p_2(X^TX)^{r-2} + (X^TX)^{r-1}) \tag{36}$$

which is derived by post-multiplying (35) and then solving for the inverse. Substituting (36) in the LS solution gives Equation (33) where

$$p = \frac{1}{p_{r+1}}[p_r\ p_{r-1} \cdots p_2\ 1]^T \tag{37}$$

and the proposition follows. □

A consequence of Proposition 5.1. is that the ordinary LS solution can be expressed as a linear combination of the columns in the controllability matrix. The coefficient p_{r+1} in the characteristic polynomial can be computed as $p_{r+1} = \det(X^TX) = \lambda_1\lambda_2 \cdots \lambda_r$. Assume that X^TX is singular (rank deficient) or nearly rank deficient, then, $p_{r+1} = 0$ or approximately zero. The problem of computing the vector p given by Equation (37) may in this case be ill-conditioned. This illustrates the problem with the OLS solution when X^TX is nearly rank deficient. We can instead look for a regularized solution in the subspace spanned by the (reduced) controllability matrix K_a where $1 \leq a \leq r$.

In fact, we will below show that the column space of the weighting matrix W_a computed by the PLS algorithm and the column space of the reduced controllability matrix K_a coincide.

5.2. *Proposition*

The weighting matrix W_a which results from the PLS algorithm is related to the controllability (Krylov) matrix K_a of the pair (X^TX, X^Ty).

The weight matrix W_a is given by the following QR decomposition

$$K_a = W_aR_1 \tag{38}$$

where $K_a \in \mathbb{R}^{r\times a}$ is the controllability matrix and $R_1 \in \mathbb{R}^{a\times a}$ is an upper triangular matrix.

The weight vectors w_i is a linear combination of the columns in the controllability, i.e.

$$W_a = K_aR_1^{-1} \tag{39}$$

where $R_1^{-1}$is upper triangular.

The following are equivalent. W_a is an orthogonal/orthonormal basis for the column space of K_a. The columns of W_a span the same space as the columns of K_a.

5.2. *Proof*

This proposition can be proved from the weight vectors as computed in Theorem 4.1. and the controllability matrix K_a. We simply have to prove that $R_1 = W_a^TK_a$ is upper triangular or that $W_a = K_aR_1^{-1}$. A proof is presented in Appendix B.

Note that Helland (1988) has pointed out that the space spanned by the columns in the PLS weighting matrix W_a and the space spanned by the Krylov sequence $X^TY, \ldots, (X^TX)^{a-1}X^TY$ is the same. □

Define now the QR decomposition of the controllability matrix as

$$K_a = Q_aR \tag{40}$$

where $Q_a \in \mathbb{R}^{r\times a}$ is orthogonal and $R \in \mathbb{R}^{a\times a}$ is upper triangular. A QR decomposition of the relationship (38) is then given by

$$W_a = Q_a R_2 \tag{41}$$

where $R_2 = RR_1^{-1}$ (usually diagonal and $R_2 = I$) also is upper triangular.

This implies that the weighting matrix W_a, computed by any PLS implementation, irrespective of scaling, etc., have the same column space as Q_a. Furthermore, this column space can be computed from the QR decomposition of the controllability matrix K_a. An orthogonal PLS weighting matrix is then defined as $W_a := Q_a$. These important results are presented in Theorem 5.1.

5.1. *Theorem (PLS: a QR decomposition of a Controllability matrix)*

Given data matrices $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^{N\times m}$. Define the (reduced) controllability (Krylov) matrix from X, Y and the number of components $1 \le a \le r$ as follows

$$K_a = [X^TY \quad X^TXX^TY \quad (X^TX)^2X^TY \cdots (X^TX)^{a-1}X^TY]. \tag{42}$$

The column space of the weighting matrix W_a and the controllability matrix K_a coincide. The QR decomposition is a numerically stable method for computing the column space. We have

$$K_a = Q_a R \tag{43}$$

where $R \in \mathbb{R}^{a\times a}$ is upper triangular and $Q \in \mathbb{R}^{r\times a}$ is orthogonal.

A Q-orthogonal PLS solution is then given by

$$B_{\mathrm{QPLS}} = Q_a(Q_a^TX^TXQ_a)^{-1}Q_a^TX^TY. \tag{44}$$

Furthermore, for univariate Y, i.e., when $m = 1$, then the orthogonal weighting matrix W_a which results from the PLS algorithm is identical to Q_a, up to within sign differences., i.e., the PLS weighting matrix is given by

$$W_a = Q_a. \tag{45}$$

Hence, when $m = 1$ the PLS solution is given by

$$B_{\mathrm{PLS}} = B_{\mathrm{QPLS}}. \tag{46}$$

We have defined the LS solution defined in Theorem 5.1. for the Q-orthogonal PLS solution (QPLS). The reason for this is that this solution differs from PLS (more precisely PLS2) when Y is multivariate (multivariable), i.e., when $m > 1$. However, the QPLS solution is identical to the PLS solution when $m = 1$, i.e., when Y is univariate.

Theorem 5.1. states that the weighting matrix W_a can be computed directly from a single QR decomposition of one single data matrix. This data matrix is the controllability (Krylov) matrix which is defined in terms of X and Y.

Note also that putting $W_a = K_a$ also gives the same PLS solution. This can be proved by substituting $W_a = K_aR_1^{-1}$ into the solution and using the assumption that R_1 is non-singular. We have the following proposition with proof.

5.3. *Proposition (PLS1: a non-iterative solution)*

Given data matrices $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^N$. The PLS solution is given by

$$B_{\mathrm{PLS}} = K_a p^* \tag{47}$$

where $K_a \in \mathbb{R}^{r\times a}$ is the controllability matrix for the pair (X^TX, X^TY) and the polynomial coefficient vector $p^* \in \mathbb{R}^a$ is determined as the LS optimal solution to

$$p^* = \arg\min_p \|V(p)\|_F^2, \tag{48}$$

$$V(p) = \|Y - X\overbrace{K_a p}^{B_{\mathrm{PLS}}}\|_F^2, \tag{49}$$

hence,

$$p^* = (K_a^T X^T X K_a)^{-1} K_a^T X^T Y \tag{50}$$

which gives the PLS solution

$$B_{\mathrm{PLS}} = K_a (K_a^T X^T X K_a)^{-1} K_a^T X^T Y \tag{51}$$

where we have assumed that $(W_a^T X^T X W_a)^{-1}$ *is non-singular for some* $1 \leq a \leq r$. *Furthermore, the minimum is*

$$V(p^*) = trace(Y^T Y) - trace(Y^T X K_a (K_a^T X^T X K_a)^{-1} K_a^T X^T Y) \tag{52}$$

5.3. *Proof*

A truncated Cayley-Hamilton polynomial approximation of the matrix inverse in Equation (36) is defined as

$$(X^T X)^{-1} := p_1 I_r + p_2 X^T X + p_3 (X^T X)^2 + \cdots + p_a (X^T X)^{a-1} \tag{53}$$

when $1 \leq a \leq r$, *which substituted into the OLS solution* $(X^T X)^{-1} X^T Y$ *gives the truncated solution*

$$B(p) = K_a p \tag{54}$$

where K_a *is the controllability matrix and p is the coefficient vector. Instead of putting the vector p equal to the coefficients in the truncated characteristic polynomial, the vector p is taken as the LS optimal solution to the squared Frobenius norm of the prediction error. Hence,*

$$p^* = \arg\min_p V(p), \tag{55}$$

where the PE criterion for the coefficient vector is given by

$$\begin{aligned} V(p) &= \|Y - X\overbrace{K_a p}^{B_{(p)}}\|_F^2 \\ &= trace(Y^T Y) - 2trace(p^T K_a^T X^T Y) + trace(p^T K_a^T X^T X K_a p) \end{aligned} \tag{56}$$

Putting the gradient

$$\frac{dV(p)}{dp} = -2K_a^T X^T Y + 2K_a^T X^T X K_a p \tag{57}$$

equal to zero gives the optimal solution (50) which substituted into (47) gives (51). Furthermore, the minimum value (52), can be found by substituting the optimal truncated polynomial coefficients p into (56).* □

Proposition 5.3. and Theorem 5.1. is believed to be important for its simple and non-iterative interpretation and implementation of the PLS algorithm. The problem of computing the PLS solution to the LS problem is in the literature presented as an iterative algorithm, or piecewise linear regression algorithm.

6. Multivariate extensions

We will in this section propose a new latent variable regression method for multivariate *Y* data. The solution reduces to the PLS1 solution for univariate *Y* data. The

new method is an extension of PLS1 to incorporate multivariate Y data. The method is found to be optimal compared with PLS2.

Consider the OLS solution substituted into the model, i.e.

$$Y = \overbrace{X(X^TX)^{-1}X^TY}^{B_{\text{OLS}}} + E \tag{58}$$

where E is the prediction error. Let us, instead of using the inverse $(X^TX)^{-1}$ as in the OLS solution, use a truncated Cayley-Hamilton series approximation for the inverse, i.e.,

$$(X^TX)^{-1} := p_1I_r + p_2X^TX + p_3(X^TX)^2 + \cdots + p_a(X^TX)^{a-1} \tag{59}$$

where a is the number of components which we will restrict to be bounded by $1 \leq a \leq r$. Hence, we have the following prediction error

$$E = Y - \overbrace{X(p_1I_r + p_2X^TX + p_3(X^TX)^2 + \cdots + p_a(X^TX)^{a-1})X^TY}^{B_{\text{CPLS}}} \tag{60}$$

which can be expressed as

$$E = Y - \overbrace{X[X^TY \quad (X^TX)X^TY \quad \cdots \quad (X^TX)^{a-1}X^TY]}^{K_a} \begin{bmatrix} p_1I_m \\ p_1I_m \\ \vdots \\ p_aI_m \end{bmatrix} \tag{61}$$

Let us now find the coefficients $p_1, p_2, \ldots, p_a$ that minimize a norm of the prediction error and use these optimal coefficients in the expression for the truncated LS solution. Define this solution

$$B_{\text{CPLS}} = \overbrace{[X^TY \quad (X^TX)X^TY \quad \cdots \quad (X^TX)^{a-1}X^TY]}^{K_a} \begin{bmatrix} p_1I_m \\ p_1I_m \\ \vdots \\ p_aI_m \end{bmatrix} \tag{62}$$

for the truncated Cayley-Hamilton PLS solution, or Controllability PLS solution. We have the following theorem

6.1. *Theorem (Controllability PLS solution)*

Given data matrices $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^{N\times m}$ and a number of components $1 \leq a \leq r$. The optimal solution is

$$\begin{aligned} B_{\text{CPLS}} &= (p_1I_r + p_2X^TX + p_3(X^TX)^2 + \cdots + p_a(X^TX)^{a-1})X^TY \\ &= \sum_{i=1}^{a} p_i(X^TX)^{i-1}X^TY \end{aligned} \tag{63}$$

where a vector with the polynomial coefficients

$$p^* = [p_1 \quad p_2 \quad \cdots \quad p_a]^T \in \mathbb{R}^a \tag{64}$$

is found from the solution to the LS problem

$$p^* = \arg\min_p \|cs(Y) - X_pp\|_F. \tag{65}$$

The minimizing solution is given by

$$p^* = (X_p^T X_p)^{-1} X_p cs(Y) \tag{66}$$

where

$$X_p = [cs(XX^TY) \quad cs(XX^TXX^TY) \quad \cdots \quad cs(X(X^TX)^{a-1}X^TY)] \in \mathbb{R}^{Nm \times a} \tag{67}$$

and where $cs(\cdot)$ is the column string (vector) operator.

6.1. *Proof*

The prediction error, Equation (60), can be written as

$$cs(E) = cs(Y) - \overbrace{[cs(XX^TY) \quad cs(XX^TXX^TY) \quad \cdots \quad cs(X(X^TX)^{a-1}X^TY)]}^{X_p} p \tag{68}$$

where p is defined in (64). Using that $V(p) = \|E\|_F = \|cs(E)\|_F$ where E is the prediction error (i.e. a real matrix), gives the LS optimal solution (66) by putting the gradient $dV(p)/dp = 0$. See also Appendix A for an alternative proof. □

The above method denoted CPLS is clearly a latent variable method for multivariate Y data. All variables in Y are used to identify a common vector $p \in \mathbb{R}^a$ of latent variables.

Note also that the CPLS algorithm gives the same solution as the univariate PLS algorithm applied to the model (5). See Appendix A.

7. Optimal weights

From the discussion in this work we have shown that the PLS estimate B_{PLS} can be expressed in terms of X, Y and a weighting matrix $W_a \in \mathbb{R}^{r \times a}$, and that this weighting matrix is a function of a polynomial coefficients.

It makes sense that different LS regression methods are different because they are using different weighting matrices. This means in other words that different weighting matrices give different least squares regression methods.

We will in this section show that there exist an optimal weighting matrix, i.e., a weighting matrix W_a which minimizes the squared Frobenius matrix norm of the residual $Y - XB(W_a)$. We will also show that there exist a minimum number a of the columns in the weighting matrix.

The resulting optimal LS solution is, not surprisingly, identical to the OLS solution. However, this result is believed to be of interest and will be used in the next section in order to develop a regularized estimator for the PLS weighting matrix.

7.1. *Theorem (The estimate of the regression matrix)*

Assume that $Y \in \mathbb{R}^{N \times m}$ and $X \in \mathbb{R}^{N \times r}$ are the known data matrices. Given a weighting matrix $W_a \in \mathbb{R}^{r \times a}$ where a is the number of components which is bounded by $1 \le a \le r$. The solution $B(W_a)$ of the matrix of regression coefficients B given by

$$B(W_a) = W_a(W_a^T X^T X W_a)^{-1} W_a^T X^T Y \in \mathbb{R}^{r \times m}, \tag{69}$$

where we have assumed that $W_a^T X^T X W_a \in \mathbb{R}^{a \times a}$ is nonsingular, satisfy the weighted normal equation

$$W_a^T X^T Y = W_a^T X^T X B(W_a) \tag{70}$$

7.1. *Proof*

Theorem 7.1. can be proved by substituting the LS solution $B(W_a)$ defined in (69) into the weighted normal equation (70). □

Note the obvious that when W_a is equal to the identity matrix and X^TX is non-singular then $B(W_a)$ is identical to the ordinary least squares estimate.

We will now search for the weighting matrix W_a which is optimal in the sense that it minimizes the Frobenius norm of the residual. Assume first for simplicity that W_a is equal to a vector $w \in \mathbb{R}^r$. The general case will be discussed and presented later. The squared Frobenius norm of the residual is in this case given by

$$V(w) = \|Y - XB_w\|_F^2 = Y^TY - \frac{Y^TXww^TX^TY}{w^TX^TXw} = Y^TY - \frac{w^TX^TYY^TXw}{w^TX^TXw}. \tag{71}$$

where we also for the sake of simplicity have assumed that Y is a vector.[1]

The minimizing weight vector w can be found by putting the gradient of $V(w)$ with respect to w equal to zero. The gradient is given by

$$\frac{dV(w)}{dw} = -\frac{2X^TYY^TXw(w^TX^TXw) - w^TX^TYY^TXw(2X^TXw)}{(w^TX^TXw)^2} \tag{72}$$

Putting the gradient equal to zero gives

$$X^TYY^TXw = \frac{w^TX^TYY^TXw}{w^TX^TXw}X^TXw \tag{73}$$

This is a generalized eigenvalue problem, i.e.,

$$\lambda_1 = \frac{w^TX^TYY^TXw}{w^TX^TXw}$$

is a generalized eigenvalue of the square matrices X^TYY^TX and X^TX and w is the corresponding generalized eigenvector.

From this we have that a solution in general can be computed by a generalized eigenvalue problem as stated in the following theorem.

7.2. *Theorem (generalized eigenvalue problem)*

The optimal weighting matrix $W_a \in \mathbb{R}^{r\times a}$ where the number of components is bounded by $1 \le a \le r$, which minimize the PE (defined here as the squared Frobenius matrix norm)

$$V(W_a) = \|Y - XB(W_a)\|_F^2 = trace(Y^TY) - trace(Y^TXW_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY). \tag{74}$$

can be computed by the following generalized eigenvalue problem

$$X^TYY^TXW_a = X^TXW_a\Lambda_a \tag{75}$$

where

$$\Lambda_a = (W_a^TX^TXW_a)^{-1}W_a^TX^TYY^TXW_a \in \mathbb{R}^{a\times a} \tag{76}$$

is a diagonal matrix with the generalized eigenvalues on the diagonal, and where W_a

[1]Note that if Y is a matrix then the matrix model $Y = XB + E$ can be written as a vector model.

is the corresponding generalized eigenvector matrix. Furthermore, the minimum value of the PE

$$V(W_a) = \|Y - XB(W_a)\|_F^2 = trace(Y^TY) - trace(\Lambda_a). \tag{77}$$

7.2. *Proof*

We will prove the Theorem from an expression of the covariance matrix of X^TY. *Substituting the LS solution* $B(W_a)$ *into the model gives*

$$Y = XW_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY. \tag{78}$$

Pre-multiplication with X^T *gives the normal equation*

$$X^TY = X^TXW_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY \tag{79}$$

and post-multiplication with Y^TXW_a *gives*

$$X^TYY^TXW_a = X^TXW_a\overbrace{(W_a^TX^TXW_a)^{-1}W_a^TX^TYY^TXW_a}^{\Lambda_a} \tag{80}$$

which is equivalent to the following generalized eigenvalue problem

$$X^TYY^TXW_a = X^TXW_a\Lambda_a \tag{81}$$

where W_a *is the generalized eigenvector matrix of the square matrices* X^TYY^TX *and* X^TX *and*

$$\Lambda_a = (W_a^TX^TXW_a)^{-1}W_a^TX^TYY^TW_a \tag{82}$$

is the corresponding generalized eigenvalue matrix.

Note that the above is equivalent to formulate the correlation matrix of X^TY *given by the normal equation, i.e.*

$$X^TY(X^TY)^T = X^TXW_a(W_a^TX^TXW_a)^{-1}W_a^TX^TYY^TXW_a(W_a^TX^TXW_a)^{-1}W_a^TX^TX \tag{83}$$

Post-multiplying with W_a *gives Equations (81) and (82).*

The minimum value can be found as follows:

$$\begin{aligned} V(W_a) &= \|Y - XB(W_a)\|_F^2 \\ &= trace(Y^TY) - trace(Y^TXW_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY) \\ &= trace(Y^TY) - trace(\underbrace{W_a^TX^TYY^TXW_a}_{X^TXW_a\Lambda_a}(W_a^TX^TXW_a)^{-1}) \end{aligned} \tag{84}$$

Substituting for the stationary condition Equation (75) gives

$$V(W_a) = \|Y - XB(W_a)\|_F^2 = trace(Y^TY) - trace(\Lambda_a). \tag{85}$$

We have here used that $trace(AB) = trace(A^TB^T)$ *for two matrices A and B with compatible dimensions.* □

The generalized eigenproblem in Theorem 7.2. can be solved by the QZ algorithm (Golub 1983). The weighting matrix W_a can be computed in MATLAB as $[Aa, Bb, q, Z, V] = qz(X^TYY^TX, X^TX)$ and putting $W_a = V(:, 1:a)$.

Note that W and Λ also can be computed by the MATLAB function eig(·,·), i.e., $[W, \Lambda] = eig(X^TYY^TX, X^TX)$. The weight matrix corresponding to the first a generalized eigenvalues is then given by $W_a := W(:, 1:a)$. Note that it is possible to compute only

the a first generalized eigenvectors. However, we recommend to use the MATLAB function qz$(\cdot,\cdot)$ instead of using the function eig.

Investigations of the above results indicated that the resulting LS optimal solution is the same for all $m \leq a \leq r$, and that this solution is the same as the OLS solution. A question is whether the minimum number of components is $a = m$ or not. In the case when X^TX is non singular the above corresponds to take the weights from the columns space of the OLS solution $(X^TX)^{-1}X^TY$. We will in the next section use the results presented in this section to develop a regularized estimator for the PLS weights.

8. An estimator for the PLS weights

The number of parameters in the PLS weighting matrix W_a is ra but it is rm parameters in the PLS solution B_{PLS}. Assume the existence of a parameter estimator for the PLS algorithm. It makes sense that in order for this parameter estimator to have a unique optimum, it must be a function of at least rm parameters, and not a function of all ra unknown parameters in W_a. We have here assumed $1 \leq m \leq a$.

In order to formulate the PLS algorithm as an estimator we must find the relationship between the PLS solution and rm unknown parameters. This relationship is presented in the following theorem.

8.1. *Theorem (The number of unique PLS parameters)*

Assume that a weighting matrix W_a with $m \leq a \leq r$ for the PLS solution B_{PLS} is given. The PLS solution can be expressed in terms of $X \in \mathbb{R}^{N\times r}$, $Y \in \mathbb{R}^{N\times m}$ and a weighting matrix $w \in \mathbb{R}^{r\times m}$ with only rm parameters as follows

$$B_{\text{PLS}} = w(w^TX^TXw)^{-1}w^TX^TY \tag{86}$$

where the weighting matrix w is the eigenvectors of $W_a(W_a^TX^TXW_a)^{-1}W_a^TX^TYY^TX$ corresponding to the m largest eigenvalues, i.e., w is a solution to the following eigenvalue problem

$$W_a(W_a^TX^TXW_a)^{-1}W_a^TX^TYY^TXw = w\lambda \tag{87}$$

where

$$\lambda = (w^TX^TXw)^{-1}w^TX^TYY^TXw \in \mathbb{R}^{m\times m} \tag{88}$$

8.1. *Proof*

It results from putting the two expressions for the same solution equal to each other. We have

$$\overbrace{W_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY}^{B_{\text{PLS}}(W_a)} = \overbrace{w(w^TX^TXw)^{-1}w^TX^TY}^{B_{\text{PLS}}(w)}. \tag{89}$$

Post-multiplication with Y^TXw gives an eigenvalue problem $Zw = \lambda w$, i.e.,

$$\overbrace{W_a(W_a^TX^TXW_a)^{-1}W_a^TX^TYY^TXw}^{Z} = \overbrace{w(w^TX^TXw)^{-1}w^TX^TYY^TXw}^{\lambda}. \tag{90}$$

□

We can now present the PLS algorithm as an estimator.

8.2. *Theorem (PLS optimization criterion)*

The PLS estimate B_{PLS} of the matrix of regression coefficients B can be expressed in terms of $X \in \mathbb{R}^{N\times r}$, $Y \in \mathbb{R}^{N\times m}$ and an estimate $\hat{w}$ of a single weight vector $w \in \mathbb{R}^r$. The PLS estimate is given by

$$B_{\mathrm{PLS}} = \hat{w}(\hat{w}^T X^T X \hat{w})^{-1} \hat{w}^T X^T Y \tag{91}$$

where

$$\hat{w} = \arg\min_w V(w) \tag{92}$$

where (we for simplicity consider only $m = 1$)

$$V(w) = trace(Y^T Y) - \lambda \tag{93}$$

where

$$\lambda = \frac{w^T (X^T Y - z)(Y^T X - z^T) w}{w^T X^T X w}. \tag{94}$$

and for PLS we choose

$$z = w_{a+1} = X^T Y - X^T X H_a X^T Y, H_a = K_a (K_a^T X^T X K_a)^{-1} K_a^T, \tag{95}$$

where a is the number of components and K_a is the controllability matrix for the pair $(X^T X, X^T Y)$. The vector w_{a+1} can also be computed from Theorem 4.1. Furthermore, this can be written as

$$V(w) = trace(Y^T Y) - \frac{w^T X^T Y Y^T X w}{w^T X^T X w} + \frac{w^T (2 X^T Y z^T - z z^T) w}{w^T X^T X w} \tag{96}$$

and

$$V(w) = \|Y - XB(w)\|_F^2 + \frac{w^T (2 X^T Y z^T - z z^T) w}{w^T X^T X w} \tag{97}$$

where

$$B(w) = w(w^T X^T X w)^{-1} w^T X^T Y \tag{98}$$

Theorem 8.2. is important from a statistical point of view. It implies that PLS is a regularized prediction error estimator. It implies that it is only one single weight vector w which has to be estimated. The theorem also defines a class of regularized LS estimators, i.e., one estimator for each choice of vector $z \in \mathbb{R}^r$. Note that $z = 0$ or $z = X^T(Y - XB_{\mathrm{OLS}})$ gives the ordinary LS estimator and that $z = X^T(Y - B_{\mathrm{PCR}})$ gives the PCR estimator. The vector z can be viewed as regularization parameters which attract the parameter estimator to a point in the parameter space.

The solution to the optimization problem can be found from a generalized eigenvalue problem. The solution is presented in the next theorem.

8.3. *Theorem (PLS as a generalized eigenvalue problem)*

$$(X^T Y - z)(Y^T X - z^T) w = X^T X w \lambda \tag{99}$$

where $w \in \mathbb{R}^r$ is the generalized eigenvector corresponding to the generalized eigenvalue

$$\lambda = \frac{w^T (X^T Y - z)(Y^T X - z^T) w}{w^T X^T X w} \tag{100}$$

where

$$z = w_{a+1} \tag{101}$$

Finally, the PLS estimate of the matrix of regression coefficients B can be computed from the generalized eigenvector w, X and Y as follows

$$B_{\mathrm{PLS}} = w(w^T X^T X w)^{-1} w^T X^T Y \tag{102}$$

8.2. *Proof*

We have that the residual of the normal Equation is

$$z = X^T Y - X^T X W_a (W_a^T X^T X W_a)^{-1} W_a^T X^T Y \tag{103}$$

where z is the residual of the normal equation, e.g., $z = w_{a+1}$. We have shown that W_a can be replaced by a weight matrix W_m when $m \leq a$. This gives

$$X^T Y - z = X^T X W_m (W_m^T X^T X W_m)^{-1} W_m^T X^T Y. \tag{104}$$

The covariance matrix of $X^T Y - z$ post-multiplied by W_m is expressed as

$$(X^T Y - z)(X^T Y - z)^T W_m = X^T X W_m \overbrace{(W_m^T X^T X W_m)^{-1} W_m^T X^T Y Y^T X W_m}^{\Lambda_m} \tag{105}$$

which is a generalized eigenvalue problem for W_m and Λ_m.

Consider the following regularized PE criterion

$$V(W_m) = \|Y - XB(W_m)\|_F^2 + \mathrm{trace}(W_m^T (2X^T Y - z) z^T W_m (W_m^T X^T X W_m)^{-1}). \tag{106}$$

This can be written as

$$\begin{aligned} V(W_m) &= \mathrm{trace}(Y^T Y) - \mathrm{trace}(\overbrace{W_m^T (X^T Y - z)(X^T Y - z)^T W_m}^{X^T T X W_m \Lambda_m} (W_m^T X^T X W_m)^{-1} \\ &= \mathrm{trace}(Y^T Y) - \mathrm{trace}(\Lambda_m) \end{aligned} \tag{107}$$

Note that the second term in the PE is equal to zero if the weighting matrix W_m is orthogonal to the residual z. Hence, the estimator attracts weighting matrices such that $z^T W_m = 0$.

9. Examples

9.1. *Example*

Consider the following example from Hansen (1992)

$$\overbrace{\begin{bmatrix} 0{\cdot}27 \\ 0{\cdot}25 \\ 3{\cdot}33 \end{bmatrix}}^{Y} = \overbrace{\begin{bmatrix} 0{\cdot}16 & 0{\cdot}10 \\ 0{\cdot}17 & 0{\cdot}11 \\ 2{\cdot}02 & 1{\cdot}29 \end{bmatrix}}^{X} \overbrace{\begin{bmatrix} 1{\cdot}00 \\ 1{\cdot}00 \end{bmatrix}}^{B} + \overbrace{\begin{bmatrix} 0{\cdot}01 \\ -0{\cdot}03 \\ 0{\cdot}02 \end{bmatrix}}^{E}. \tag{108}$$

The problem addressed is to find the best estimate of B from given data matrices X and Y and knowledge of the model structure (3).

$$B_{\mathrm{OLS}} = \begin{bmatrix} 7{\cdot}01 \\ -8{\cdot}40 \end{bmatrix}, \|B_{\mathrm{OLS}}\|_F = 10{\cdot}94, \|Y - XB_{\mathrm{LS}}\|_F = 0{\cdot}02. \tag{109}$$

$$B_{\mathrm{PLS}} = \begin{bmatrix} 1{\cdot}1703 \\ 0{\cdot}7473 \end{bmatrix}, \|B_{\mathrm{PLS}}\|_F = 1{\cdot}3885, \|Y - XB_{\mathrm{PLS}}\|_F = 0{\cdot}0322. \tag{110}$$

$$B_{\mathrm{TTLS}} = \begin{bmatrix} 1{\cdot}1703 \\ 0{\cdot}7473 \end{bmatrix}, \|B_{\mathrm{TTLS}}\|_F = 1{\cdot}3885, \|Y - XB_{\mathrm{TTLS}}\|_F = 0{\cdot}0322. \tag{111}$$

A major difficulty with the above ordinary least squares solution B_{OLS} in (109) is that its norm is significantly greater than the norm of the exact solution, which is $\|B\|_F = \sqrt{2}$. One component ($a = 1$) was specified for the PLS and TTLS algorithms. See, e.g., Fierro et al. (1997) for a description of regularization and the Truncated Total Least Squares (TTLS) solution. The PLS and TTLS solutions are almost similar for this example. The effect of the latent variable ($a = 1$) solutions is that regularization is introduced in order to stabilize the solution.

9.2. *Example*

Assume that data matrix $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^N$ are given and that we want to compute PLS estimate of the regression matrix B_{PLS} by using $a = 2$ components. We have shown that the solution to this problem is to first compute a weighting matrix

$$W = [w_1 \quad w_2] \in \mathbb{R}^{r\times 2}, \tag{112}$$

where the two columns in the weighting matrix can be computed as

$$w_1 = X^T Y, \tag{113}$$

$$w_2 = w_1 - \frac{X^T X w_1 w_1^T w_1}{w_1^T X^T X w_1}. \tag{114}$$

Note that $w_1^T w_2 = 0$. It is often convenient to scale the columns so that $W^T W = I$. In this case we have

$$\begin{aligned} \tilde{w}_1 &= X^T Y, & w_1 &= \frac{\tilde{w}_1}{(\tilde{w}_1^T \tilde{w}_1)^{\frac{1}{2}}} \\ \tilde{w}_2 &= w_1 - \frac{X^T X w_1}{w_1^T X^T X w_1}, & w_2 &= \frac{\tilde{w}_2}{(\tilde{w}_2^T \tilde{w}_2)^{\frac{1}{2}}} \end{aligned} \tag{115}$$

This example shows that there is only r unknown parameters in the weighting matrix W, i.e., the parameters in w_1. The PLS algorithm computes w_1 as the solution to the problem of minimizing $\|Y - Xw_1\|_F$ (when $\|w_1\|_F = 1$).

9.3. *Example*

Given data matrices X, Y and the weighting matrix W_2. We will in this example illustrate that the columns of W_2 can be expressed as linear combinations of the columns in the controllability matrix K_2 of the pair $(X^T X, X^T Y)$. The weighting vectors w_1 and w_2, defined in Example 9.2, can be written as

$$[w_1 \quad w_2] = [X^T Y \quad X^T X X^T Y] \begin{bmatrix} 1 & 1 \\ 0 & -\frac{1}{\lambda} \end{bmatrix} \tag{116}$$

where

$$\lambda = \frac{w_1^T X^T X w_1}{w_1^T w_1}$$

is an eigenvalue of $X^T X$.

9.4. *Example (Real world data from a pulp and paper mill)*

A refiner experiment was designed in order to investigate the relationship between refiner manipulable variables and the freeness of the pulp. The freeness is one of the

main variables which are frequently used as a measure of the quality of the pulp. The four input variables used in the experiment are the refiner plate gap u_1 [mm], the flow of dilution water u_2 [kg/s], the refiner casing pressure u_3 [bar] and the dosage screw speed u_4 [1000 Kg/h]. The sampling rate for the experiment was one hour. $N = 16$ samples of the freeness was measured in the blow-line and in the latency chest. The freeness in the blow-line y_1 was analyzed in the laboratory from samples which were taken each hour. The freeness in the latency chest y_2 was measured by a Pulp Expert analysator, also with one hour sampling rate.

The data is organized into X and Y as follows.

$$X = \begin{bmatrix} 9{\cdot}3 & 0{\cdot}54 & 4{\cdot}5 & 13{\cdot}0 \\ 8{\cdot}3 & 0{\cdot}64 & 4{\cdot}0 & 13{\cdot}0 \\ 9{\cdot}3 & 0{\cdot}54 & 4{\cdot}0 & 13{\cdot}0 \\ 8{\cdot}3 & 0{\cdot}64 & 4{\cdot}5 & 13{\cdot}0 \\ 8{\cdot}3 & 0{\cdot}54 & 4{\cdot}5 & 13{\cdot}0 \\ 9{\cdot}3 & 0{\cdot}64 & 4{\cdot}5 & 13{\cdot}0 \\ 8{\cdot}3 & 0{\cdot}54 & 4{\cdot}0 & 13{\cdot}0 \\ 9{\cdot}3 & 0{\cdot}64 & 4{\cdot}0 & 13{\cdot}0 \\ 7{\cdot}0 & 0{\cdot}70 & 4{\cdot}5 & 11{\cdot}0 \\ 8{\cdot}0 & 0{\cdot}60 & 4{\cdot}0 & 11{\cdot}0 \\ 8{\cdot}0 & 0{\cdot}70 & 4{\cdot}5 & 11{\cdot}0 \\ 8{\cdot}0 & 0{\cdot}70 & 4{\cdot}0 & 11{\cdot}0 \\ 7{\cdot}0 & 0{\cdot}60 & 4{\cdot}0 & 11{\cdot}0 \\ 8{\cdot}0 & 0{\cdot}60 & 4{\cdot}5 & 11{\cdot}0 \\ 7{\cdot}0 & 0{\cdot}70 & 4{\cdot}0 & 11{\cdot}0 \\ 7{\cdot}0 & 0{\cdot}60 & 4{\cdot}5 & 11{\cdot}0 \end{bmatrix}, \quad Y = \begin{bmatrix} 181 & 167 \\ 241 & 206 \\ 161 & 172 \\ 230 & 198 \\ 154 & 157 \\ 231 & 209 \\ 154 & 145 \\ 203 & 220 \\ 216 & 185 \\ 135 & 152 \\ 257 & 223 \\ 185 & 208 \\ 102 & 131 \\ 156 & 155 \\ 204 & 182 \end{bmatrix} \tag{117}$$

The X and Y data was centered (sample mean removed from each variable) prior to identification. The data was first used to compare the multivariate algorithms CPLS, PLS, SIMPLS and PCR. The results are illustrated in Table 1.

Table 1. Comparison of the multivariate regression method CPLS against PLS, SIMPLS (see Section 10.5) and PCR. The norm $\|Y - XB_M\|_F$ where B_M is the solution from the particular Method, is taken as our PE criterion and is presented in the table

a	CPLS	PLS	SIMPLS	PCR
1	194·798	195·103	195·103	196·027
2	185·171	186·621	186·714	193·759
3	174·322	176·327	178·369	188·108
4	68·795	68·795	68·795	68·795

This example clearly illustrates the optimality (minimizing PE) of CPLS compared to PLS, SIMPLS and PCR.

Assume now that we are only interested in a good model for the freeness y_1 in the blow-line. The model predictions will in this case be improved by including y_2 in the X data matrix, i.e., as an additional regressor.

Table 2 shows that the prediction of y_1 is improved by incorporating y_2 as a regressor. This is quite expected since the regressor y_2 is an indirect measure of the response (output) y_1.

We also note that the Truncated Total Least Squares (TTLS) method gives larger

Table 2. Comparison of the univariate regression methods PLS, PCR and TSVD. u_1, u_2, u_3 u_4 and y_2 are used as regressors, i.e., in order to define the X data matrix. y_1 is used as the response variable, i.e., in order to define Y. The norm $\|Y - XB_M\|_F$ where B_M is the solution from the particular Method, is taken as our PE criterion and is presented in the table

a	PLS	PCR	TTLS
1	72·93	75·70	76·85
2	72·76	72·77	76·52
3	69·73	72·36	244·6
4	64·47	64·49	141·66
5	57·31	57·31	124·93

PE compared to PLS and PCR. This is also quite expected since TTLS are minimizing an objective function $\|X - Z\|_F^2 + \|Y - ZB_{TTLS}\|_F^2$, which is a solution to the *errors-in-variables* regression problem where not only Y is subject to errors but also X is assumed to be subject to errors. Note that *PLS* and *PCR* gives biased solutions for B in case of an *errors-in-variables* model.

The reliability of the different models should be investigated further by model validation. This work is in progress.

10. Discussion

10.1. *Weights W_a from the SVD of the controllability matrix K_a*

In Burnham *et al.* (1996) an Undeflated PLS like solution (UPLS) was proposed in order to illustrate the need for the deflation process in PLS. It was proposed that the weighting matrix W_a should be taken as the a left singular vectors of X^TY. We have in this paper proved that the PLS solution in general is related to the controllability matrix K_a of the pair (X^TX, X^TY). In the univariate case we have $B = K_a p^*$ (Theorem 5.1. and in the multivariate case

$$B_{CPLS} = \overbrace{[X^TY \quad (X^TX)X^TY \quad \cdots \quad (X^TX)^{a-1}X^TY]}^{K_a} \begin{bmatrix} p_1 I_m \\ p_1 I_m \\ \vdots \\ p_a I_m \end{bmatrix}$$

(Theorem 6.1.). A more general alternative to UPLS is then to take the weighting matrix W_a equal to the a first left singular vectors of K_a, i.e., $W_a = U(:, 1:a)$ where $USV^T = K_a$.

Another choice is to choose W_a equal to a controllability matrix of the pair (X^TX, w_1) where w_1 is equal to the first singular vector of X^TY. We have found that this basis (W_a from SVD of K_a) for multivariate Y data, in some cases gives smaller prediction errors compared to the multivariate CPLS solution in Theorem 6.1., however, in most cases it gave larger PE. Note that CPLS is the minimizing solution to a well defined prediction error, but that the above solutions have diffuse statistical properties. We mention this as a comment to the UPLS solution, and will not elaborate this further.

10.2. *Prediction*

In chemometrics one is often only concerned with the prediction properties of the model. One of the main points for using PLS instead of PCR (truncated SVD solution) is that PLS usually gives a smaller prediction error compared to PLS, for the same number of components. This is also illustrated in Example 9.3. The reason for this is that PCR uses only information in X in order to construct the pseudo inverse, but as

shown in this paper, the parameters in the approximate inverse used by PLS is taken as the minimizing parameters of the prediction error.

10.3. *Bias on parameter estimates*

Like PCR, PLS gives bias free estimates in case of measurements noise only (noise on Y), assuming that the rank of X actually is $a \leq r$ and that the same number of components is used in the two algorithms.

PLS may give a bias on the parameter estimates in case of an errors-in-variables model, i.e., in the case when X is corrupted with measurements noise. Note that also OLS and PCR gives bias in this case. An interesting solution to the errors-in-variables problem is the Truncated Total Least Squares (TTLS) solution of De Moor *et al.* (1996, Fierro *et al.* (1997) and Hansen (1992).

10.4. *Bias and variance*

Based on our simulation experiments, we believe that PLS is a valuable tool in order to stabilize the solution in case of a rank deficient or nearly rank deficient data matrix X. The problem of choosing the number of components $1 \leq a \leq r$ is in general a trade off between bias and variance, and model validation. The number of components a used to compute the PLS solution is a regularization parameter. The bias and variance properties of the PLS solution should be investigated further. However, we will refer to Johansen (1997) for a discussion of bias and variance because of regularization in system identification.

10.5. *SIMPLS*

We are aware of the variant of PLS which is denoted SIMple PLS discussed in ter Braak and de Jong (1998). SIMPLS gives the same solution as PLS for univariate Y data, but gives in general different solutions for multivariate Y data. This is illustrated in Example 9.3. Like PLS, the first weight vector w_1 in SIMPLS can be taken as the left singular vector of X^TY, i.e., $w_1 = U(:,1)$ where $USV^T = X^TY$. The next weight vectors are computed iteratively as follows. Put $w_i = w_1$ and for all $i = 2, \ldots, a$ construct a projection matrix $P_i = X^TXw_i/(w_i^TX^TXw_i)$. The weight vector w_i can be taken as the first left singular vector of $(I_r - P_i)X^TY$, i.e., $w_i = U(:,1)$ where $(I_r - P_i)X^TY = USV^T$. As also pointed out by ter Braak and de Jong (1998), SIMPLS may in some cases give a smaller PE than PLS2 (for multivariate Y data and the same number of components). On our Example 9.3. SIMPLS gives equal or larger PE compared to PLS. However, the CPLS solution which is presented in this work gave smaller PE than both PLS and SIMPLS. Note that a well defined PE criterion is defined for the CPLS solution, but such a PE criterion does not exist for PLS2 and SIMPLS.

11. Conclusions

The PLS solution for univariate Y data is equivalent to using a truncated Cayley-Hamilton series approximation to the matrix inverse $(X^TX)^{-1}$ in the OLS solution. This implies that the PLS solution can be written as $B_{\mathrm{PLS}} = K_a p^*$ where K_a is the controllability matrix for the matrix pair (X^TX, X^TY). Furthermore, the polynomial coefficients (in vector $p^* \in \mathbb{R}^a$), are determined as the LS optimal solution to the squared Frobenius norm of the prediction error, i.e., $p^* = \arg\min_p \|Y - XK_a p\|_F^2$. Furthermore, this implies that the controllability matrix K_a is a valid weighting matrix for the PLS solution. Hence, the PLS solution for univariate Y can be computed directly

as $B_{\mathrm{PLS}} = K_a(K_a^T X^T X K_a)^{-1} K_a^T X^T Y$. We have proved that the PLS solution for univariate Y data is non-iterative. Hence, there is no need for any deflation (rank one reduction) process for computing the PLS solution.

The optimal polynomial coefficient vector p^* may be a function of both Y as well as the X matrix, i.e., it results in the minimal PE. This is probably the reason why PLS often gives a smaller PE than the corresponding PE by using a PCR solution, assuming the same number of components. In PCR the approximate inverse of $X^T X$ is constructed from information in X only.

The usual algorithm for computing the PLS weighting matrix W_a which is presented in the literature is equivalent to computing an (orthogonal matrix with orthonormal columns) basis for the column space of the controllability (Krylov) matrix. This basis is equivalent to the Q-orthogonal matrix Q_a from the QR decomposition of the controllability matrix., i.e., a Gram-Schmidt procedure can be used to compute orthogonal Q_a that satisfy $K_a = Q_a R$ where R is upper triangular. Furthermore, an orthogonal PLS weighting matrix is $W_a := Q_a$, and the solution can equivalently be computed as $B_{\mathrm{PLS}} = Q_a(Q_a^T X^T X Q_a)^{-1} Q_a^T X^T Y$.

A QR updating technique (one column at a time) can be used to compute the QR decomposition of K_a and thereby avoiding explicit formulation of the controllability matrix K_a. The problem of computing an orthogonal basis for the controllability subspace may be better conditioned compared to explicitly forming the controllability matrix. The problem of forming the controllability matrix may be ill-conditioned due to rounding off errors when computing powers of $X^T X$. The so-called Arnoldi's method to construct the basis for the Krylov subspace should be considered.

The PLS solution is not optimal for multivariate Y data. This is shown by counter-example. An optimal latent variable LS solution B_{CPLS} is presented in the paper. This optimal solution follows from an extension of the Cayley-Hamilton series approach that we derived the PLS algorithm for univariate data to incorporate multivariate data. The optimality is illustrated by real world data from the pulp and paper industry.

A. Appendix—proof of Theorem 6.1.

The expression for the PE, Equation (60), gives

$$\mathrm{cs}(E) = \mathrm{cs}(Y) - (I_m \otimes X)\mathrm{cs}(K_a(p)) \tag{118}$$

where we have used that $\mathrm{cs}(AXB) = (B^T \otimes A)\mathrm{cs}(X)$ for the column string (vector) operation of the product of the triple matrices (A, X, B) with compatible dimensions, see, e.g., Vetter (1973). Furthermore, Equation (118) can be written as

$$\mathrm{cs}(E) = \mathrm{cs}(Y) - (I_m \otimes X)\mathrm{bcs}(K_a)p \tag{119}$$

where $p \in \mathbb{R}^a$, $(I_m \otimes X) \in \mathbb{R}^{NM \times mr}$ and where we have defined (and introduced)

$$\mathrm{bcs}(K_a) = [\mathrm{cs}(X^T Y) \quad \mathrm{cs}(X^T x x^T y) \quad \cdots \quad \mathrm{cs}((X^T X)^{a-1} X^T Y)] \in \mathbb{R}^{rm \times a} \tag{120}$$

as a block column string operator.

Equation (119) can be solved for p in a LS optimal sense by minimizing $V(p) = \|\mathrm{cs}(E)\|_F^2$ with respect to p. This gives the optimal parameter vector

$$p^* = M^{\dagger}\mathrm{cs}(Y) \tag{121}$$

where we have defined

$$M = (I_m \otimes X)\text{bcs}(K_a) \in \mathbb{R}^{Nm \times a} \tag{122}$$

and where $M^{\dagger} = (M^T M)^{-1} M^T$ is the More-Penrose pseudo-inverse of the matrix M. □

B. Appendix—proof of Proposition 5.2.

We want to prove that $W_a = K_a R_1^{-1}$ is upper triangular.

From Theorem 4.1. we have that

$$w_1 = X^T Y \tag{123}$$

$$w_{i+1} = w_i - X^T X W_i c_i \in i = 1, \ldots, a-1 \tag{124}$$

where it is important to note that

$$c_i = (W_i^T X^T X W_i)^{-1} W_i^T w_i \in \mathbb{R}^a \tag{125}$$

is a vector. This implies directly that w_{i+1} is a linear combination of the sequence $w_i, X^T X w_i, X^T X w_2, \ldots, X^T X w_i$.

From this we can prove that w_i is a linear combination of the sequence $w_i, X^T X w_1, (X^T X)^2 w_1, \ldots, (X^T X)^{i-1} w_1$ as follows.

From the above we have that w_i is a linear combination of the sequence $w_{i-1}, X^T X w_1, X^T X w_2, \ldots, X^T X w_{i-1}$. Substituting for $w_2, \ldots, w_{i-1}$ into this sequence, by noting that w_2 is a linear combination of w_1 and $X^T X w_1$, w_3 is a linear combination of $w_2, X^T X w_1$ and $X^T X w_2$, and so on, proves that w_i is a linear combination of the columns in the controllability matrix K_i of the pair $(X^T X, w_1)$. By induction, this must also hold for $i = a$.

The fact that $W_a = K_a R_1^{-1}$ is upper triangular follows from the fact, as proved above, that each column w_i in W_a only is a linear combination of columns 1 to i in the controllability matrix.

We will illustrate the proof for $i = 1, 2, 3$ in the following.

$i = 1$

$$w_2 = w_1 - c_1 X^T X w_1 \quad \text{where} \quad c_1 = \frac{w_1^T w_1}{w_1^T X^T X w_1} \tag{126}$$

which is a linear combination of $X^T Y$ and $X^T X X^T Y$.

$i = 2$

$$w_3 = w_2 - X^T X \overbrace{[w_1 \quad w_2]}^{W_2} \overbrace{\begin{bmatrix} c_{21} \\ c_{22} \end{bmatrix}}^{c_2} \tag{127}$$

where

$$c_2 = (W_2^T X^T X W_2)^{-1} W_2^T w_2 \tag{128}$$

which can be written as

$$w_3 = \overbrace{[w_1 \quad X^T X w_1 \quad (X^T X)^2 w_1]}^{K_3} \begin{bmatrix} 1 \\ -c_1 + c_{21} + c_{22} \\ c_1 c_{22} \end{bmatrix} \tag{129}$$

Hence,

$$[w_1 \quad w_2 \quad w_3] = [w_1 \quad X^TXw_1 \quad (X^TX)^2w_1] \begin{bmatrix} 1 & 1 & 1 \\ 0 & -c_1 & -(c_1 + c_{21} + c_{22}) \\ 0 & 0 & c_1c_{22} \end{bmatrix} \tag{130}$$

References

Burnham, A. J., Viveros, R. and MacGregor, J. F. (1996). Frameworks for Latent Variable Multivariate Regression. *Journal of Chemometrics*, Vol. 10, pp. 31–45.

DeMoor, B. and David, J. (1996). Total least squares and the algebraic Riccati equation. Katholieke Universiteit Leuven, B-3001 Leuven, Belgium. Internal report.

Di Ruscio, D. (1997). On subspace identification of the extended observability matrix. In: Proceedings of the *1997 IEEE Conference on Decision and Control*, San Diego, California, December 10–12.

Fierro, R. D., Golub, G. H., Hansen, P. C. and O'Leary, D. P. (1997). Regularization by truncated total least squares. *Siam Journal on Scientific Computing*, Vol. 18, No. 4, pp. 1223–1241.

Frank, L. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tool. *Technometrics*, Vol. 35, No. 2, pp. 109–135.

Hansen, P. C. (1992). Regularization Tools. A Matlab Package for Analysis and Solution of Discrete Ill-Posed Problems. Danish Computing Centre for Research and Education, DK-2800 Lyngby, Denmark.

Helland, I. S. (1988). On the Structure of Partial Least Squares Regression. *Commun. in Stat. Simulation and Computation*, Vol. 17, No. 2, pp. 581–607.

Höskuldsson, A. (1996). *Prediction Methods in Science and Technology*, COLOURSCAN Warsaw, ISBN 87-985941-0-9.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, Vol. 2, pp. 211–228.

Johansen, T. A. (1997). On Tikhonov Regularization, Bias and Variance In Nonlinear System Identification. *Automatica*, Vol. 33, No. 3, pp. 441–446.

Manne, R. (1987). Analysis of Two Partial-Least-Squares Algorithms for Multivariate Calibration. *Chemometrics and Intelligent Laboratory Systems*, Vol. 2, pp. 187–197.

Martens, H. and Næs, T. (1989). *Multivariate Calibration*, John Wiley and Sons Ltd.

ter Braak, C. J. and de Jong, S. (1998). The objective function of Partial Least Squares. *Journal of Chemometrics*, Vol. 12, pp. 41–54.

Martens, H. and Næs, T. (1985). Comparison of Prediction Methods for Multicolinear Data. *Commun. in Stat. Simulation and Computation*, Vol. 14, No. 3, pp. 544–576.

Wold, H. (1966). Non-linear estimation by iterative least squares procedures. *Research papers in statistics*, Ed. F. David. Wiley, New York, pp. 411–444.

Wold, H. (1975). Soft modeling by latent variables: the Non-linear Iterative Partial Least Squares Approach. In: *In Perspectives in Probability and Statistics*, Editor J. Gani. London, Academic Press.

Wold, H. (1985). Partial Least Squares. In: *Encyclopedia of statistics sciences*, Editors S. Kotz and N. L. Johnson. Wiley, Vol. 6, pp. 581–591.

Lorber, A., Lawrence, E. W. and Kowalski, B. R. (1987). A Theoretical Foundation for the PLS Algorithm. *Journal of Chemometrics*, Vol. 1, pp. 19–31.

Vetter, W. J. (1973). Matrix calculus operations and Taylor expansions. Siam review, Vol. 15, No. 2, pp. 352–369.