PERGAMON

Automatica 36 (2000) 831–850

www.elsevier.com/locate/automatica

A weighted view on the partial least-squares algorithm☆

David Di Ruscio*

Department of Process Automation, Telemark Institute of Technology, N-3914 Porsgrunn, Norway

Received 9 March 1998; revised 26 January 1999; received in final form 17 September 1999

Abstract

In this paper it is shown that the Partial Least-Squares (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 $(X^{\mathrm{T}}X)^{-1} \in \mathbb{R}^{r\times r}$ which is used to compute the least-squares (LS) solution. Furthermore, the a coefficients in this polynomial are computed as the optimal LS solution (minimizing parameters) to the prediction error. The resulting solution is non-iterative. The solution can be expressed in terms of a matrix inverse and is given by $B_{\mathrm{PLS}} = K_a(K_a^{\mathrm{T}}X^{\mathrm{T}}XK_a)^{-1}K_a^{\mathrm{T}}X^{\mathrm{T}}Y$ where $K_a \in \mathbb{R}^{r\times a}$ is the controllability (Krylov) matrix for the pair $(X^{\mathrm{T}}X, X^{\mathrm{T}}Y)$. The iterative PLS algorithm for computing the orthogonal weighting matrix W_a as presented in the literature, is shown here 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 equivalently be computed as $B_{\mathrm{PLS}} = W_a(W_a^{\mathrm{T}}X^{\mathrm{T}}XW_a)^{-1}W_a^{\mathrm{T}}X^{\mathrm{T}}Y$, 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. The free parameters in this solution is found as the minimizing solution of a prediction error criterion.

Keywords: Partial least squares; Prediction error methods; Controllability matrix; Regularization

1. Introduction

The Partial Least-Squares (PLS) algorithm and its solution has received great attention and is widely used in chemometrics, which has been 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 computing a solution B_{PLS} for the regression coefficients B in a linear model $Y = XB + E$ from known data matrices X and Y. One of the main purpose of using the PLS algorithm is to handle multicollinearity problems, i.e. problems where there are (approximate) linear dependencies between the columns of X which results in a (nearly) rank deficient data matrix X. An unbiased LS solution may in such situations have large variances and may therefore not be a reliable solution. The PLS algorithm is a tool to introduce a (small) bias and thereby reduce the variance. The PLS algorithm is analyzed and reviewed in some detail in among others, Næs and Martens (1985), Manne (1987), Lorber, Lawrence and Kowalski (1987), Helland (1988), Höskuldsson (1988,1996), Frank and Friedman (1993), Phatak (1993), Burnham, Viveros and MacGregor (1996), de Jong and Phatak (1997), Phatak and de Jong (1997), and ter Braak and de Jong (1998).

While PLS has been used in many applications in chemometrics, there have been few applications to system parameter identification. PLS has 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 basis of most subspace identification algorithms.

PLS is presented in the literature as an iterative algorithm, i.e. partial or piece-wise linear regression. One of the main contributions in this paper is to give a new 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 optimal solution to a prediction error minimization problem. This is believed

☆This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor B. Ninness under the direction of Editor T. Söderström.

* Tel.: + 47-35-57-51-68; fax: + 47-35-57-52-50.

E-mail address: david.di.ruscio@hit.no (D.D. Ruscio)

0005-1098/00/$ - see front matter
PII: S0005-1098(99)00210-1

to be of interest to researchers working with system identification in general, as well as to chemometricians.

We will try to give a simple description. We believe that this can only be done by introducing as few definitions and variables as possible. In the PLS literature, the algorithm and its solution are usually presented in terms of the 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 shows that there exists a very simple and non-iterative algorithm for computing the PLS solution. It will 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. However, for the sake of completeness, a discussion of the relationship between the weight vectors and the score vectors and loading vectors, which are usually defined in connection with the PLS algorithm, are presented. Further details can be found elsewhere.

The rest of this paper is organized as follows. Some basic system definitions are presented in Section 2.1. A basic preliminary result concerning the latent variable LS solution is presented in Section 2.2. The PLS algorithm is reviewed and some new results are presented in Section 3.1. The main contributions concerning the interpretation of the PLS solution are presented in Sections 3.2 and 4. Some additional results concerning LS and PLS are presented in Section 5. Some discussions follow in Section 6. Two real-world examples from the pulp and paper industry are presented in Section 7 and some conclusions follow in Section 8.

2. System definitions and preliminary results

2.1. System definitions

Define $y_k \in \mathbb{R}^m$ as the 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 (or regressors) 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:

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

where e_k is a vector of white noise with covariance matrix $\mathrm{E}(e_k e_k^{\mathrm{T}})$ and k is the observation index. With N observations $k = 1, \ldots, N$ we define an output data matrix $Y \in \mathbb{R}^{N \times m}$ and an input data matrix $X \in \mathbb{R}^{N \times r}$ as follows:

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

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

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

The linear relationship between the output (response) and the input data (or regressors) is an important assumption and condition for the PLS as well as any LS algorithm to work. In this work we will analyze systems with multiple output variables in the data matrix Y. This is often referred to a multivariate (or multivariable) system.

If we are only interested in the matrix of regression coefficients B, and that the LS solution is linear in Y, i.e. computed as $(X^{\mathrm{T}}X)^{\dagger}X^{\mathrm{T}}Y$ where $(X^{\mathrm{T}}X)^{\dagger}$ denotes a pseudo-inverse of $X^{\mathrm{T}}X$, and that this matrix is independent of Y, then one should note that (for steady-state systems) it suffices 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 output LS problems, i.e. each column in B is estimated from a separate univariate LS problem. However, this is in general not true if $(X^{\mathrm{T}}X)^{\dagger}$ is computed by the use of both X and Y, i.e. if the LS solution is non-linear in Y.

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

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

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

where $\mathrm{vec}(\cdot)$ is the column string (vector) operator and $\otimes$ is the Kronecker product. $\mathrm{vec}(Y) \in \mathbb{R}^{Nm}$ is a column vector constructed from Y by stacking each column of Y onto another. We also have $(I_m \otimes X) \in \mathbb{R}^{Nm \times rm}$ and $\mathrm{vec}(B) \in \mathbb{R}^{rm}$. Note that (6) can be constructed directly from (1) by first writing (1) as

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

and then combine all N equations ($k = 1, \ldots, N$) into a matrix equation of the form (3). Note that the variance of the noise terms in the univariate models (5) and (6) is related to the covariance matrix of the noise term in the models (1) and (3) as

$$\lim_{N\to\infty} \frac{\mathrm{vec}(E)^{\mathrm{T}}\mathrm{vec}(E)}{mN} = \lim_{N\to\infty} \frac{\mathrm{vec}(E^{\mathrm{T}})^{\mathrm{T}}\mathrm{vec}(E^{\mathrm{T}})}{mN} = \frac{1}{m}\mathrm{trace}(\mathrm{E}(e_k e_k^{\mathrm{T}})).$$

This can be proved by using that the noise term in (3) has the asymptotic covariance $\mathrm{E}(e_k e_k^{\mathrm{T}}) = \lim_{N\to\infty}(1/N)E^{\mathrm{T}}E$.

However, for the sake of completeness we will, in general, consider 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 a matrix Y^{T} onto the row space of X^{T}, or equivalently, the projection of the column space of Y onto the column space of X). For this problem it is convenient with a multivariate description. 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 this definition. The following definition is frequently used throughout the paper. The squared Frobenius norm of a matrix $A \in \mathbb{R}^{m\times n}$ is equal to the trace of the product $A^{\mathrm{T}}A$, and defined as follows:

$$\|A\|_{\mathrm{F}}^2 = \mathrm{trace}(A^{\mathrm{T}}A) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2.$$

2.2. Preliminary results

In this paper we will consider Least-Squares solutions which may be regularized approximations to the Ordinary Least-Squares (OLS) solution, as defined below.

Definition 2.1. Consider a Least-Squares solution of the form

$$B_M = W_a p^* \tag{8}$$

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 $p^* \in \mathbb{R}^{a\times m}$ is the LS solution to

$$p^* = \arg\min_p \|Y - X\overbrace{W_a p}^{B_M(p)}\|_F^2, \tag{9}$$

where $p \in \mathbb{R}^{a\times m}$. Furthermore, p^* and the LS solution B_M corresponding to the particular weighting matrix W_a, are given by

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

and

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

where we assume that $(W_a^{\mathrm{T}}X^{\mathrm{T}}XW_a)^{-1}$ is non-singular for some $1 \le a \le r$. The resulting prediction of Y is defined as

$$Y_M = XW_a p^*, \tag{12}$$

where p^* is given by (11).

Note that any square non-singular matrix W_r gives the OLS solution $B_{\mathrm{OLS}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$. Hence, $M = \mathrm{OLS}$ in Eq. (10). One should also note that any weighting matrix $W_m \in \mathbb{R}^{r\times m}$ with the same column (range) space as the solution B_{OLS} also gives the OLS solution. This can be proved by letting $W_m = B_{\mathrm{OLS}}R$, with $R \in \mathbb{R}^{m\times m}$ non-singular, in the solution (10). Furthermore, choosing $W_a = V_1$ where $V_1 \in \mathbb{R}^{r\times a}$ are the first a columns in the right singular vector matrix V from the SVD,

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

where $U_1 \in \mathbb{R}^{N\times a}$ and $S_1 \in \mathbb{R}^{a\times a}$ is non-singular, gives the Principal Component Regression (PCR) solution (truncated SVD solution), $B_{\mathrm{PCR}} = V_1 S_1^{-1} U_1^{\mathrm{T}} Y$. This can be proved by letting $W_a := V_1$ and $X := U_1 S_1 V_1^{\mathrm{T}}$ in solution (10). PCR is frequently used when the X data are multicollinear, i.e. when the columns in X are linearly or nearly linearly dependent. In this paper we will show that the PLS solution can be defined similarly. The key is to understand how the PLS algorithm defines W_a and why the parameterization $W_a p$ of the solution makes sense. Note also that W_a can be interpreted as a column weighting matrix for X, i.e. a column weighting for X in the LS problem (9) and a column weighting for X in prediction (12). Furthermore, from (8) we have that the columns in B_M are contained in the column space of W_a. Hence, $R(B_M) \subseteq R(W_a)$, or simply $B_M \in R(W_a)$ in the univariate case. The prediction Y_M given by (11) and (12), is the orthogonal projection of the column space of Y onto the column space of XW_a, i.e. onto $R(XW_a)$. Hence, $R(Y_M) \subseteq R(XW_a)$.

3. The PLS solution

3.1. The weights used by PLS

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

The normal equations are of central importance in LS problems and its solutions. Therefore it makes sense to study the PLS algorithm with the normal equations as a starting point. The normal equations $X^TY = X^TXB(W_a)$ substituted for a LS solution $B(W_a) = W_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY$ yields

$$X^TY = X^TXW_a(W_a^TX^TXW_a)^{-1}W_a^TX^TY. \tag{13}$$

The first weight vector w_1 in the PLS weighting matrix W_a can be taken directly as the correlation $w_1 = X^TY$ when Y is a vector. When Y is a matrix then w_1 can be taken as the left singular vector of X^TY 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^TYY^TX. Power iteration is a convenient tool for this computation. In the following discussion we assume univariate $Y \in \mathbb{R}^N$. The extension to the multivariate case will be clarified later.

The PLS algorithm was probably derived in a rather ad-hoc manner (Helland, 1988). Having this in mind, it is not unusual to choose a weight vector $w_1 = X^TY$. For the sake of convenience w_1 is often scaled, e.g. the choice $w_1 = X^TY/\|X^TY\|_F$ gives an orthonormal weight vector, i.e. $w_1^Tw_1 = 1$. However, as also pointed out in Helland (1988), this scaling is not necessary. In order not to complicate the discussion we chose not to use scaled weight vectors. Substituting this and $W_1 = w_1$ into the normal equations (13) gives us a residual

$$\begin{aligned} w_2 &= w_1 - X^TXB_1, \\ &\quad \text{where } B_1 = W_1(W_1^TX^TXW_1)^{-1}W_1^Tw_1 \\ &\quad \text{and } W_1 = w_1. \end{aligned} \tag{14}$$

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 $W_1^Tw_2 = w_1^Tw_2 = 0$, i.e. w_1 is normal to the residual w_2. Hence, this residual w_2, after choosing $W_1 = w_1 = X^TY$, is the second weight vector used by the PLS algorithm. We now define the normal equations for the residual, w_2, i.e.

$$\begin{aligned} w_2 &= X^TXB_2, \\ &\quad \text{where } B_2 = W_2(W_2^TX^TXW_2)^{-1}W_2^Tw_2 \\ &\quad \text{and } W_2 = [w_1 \quad w_2]. \end{aligned} \tag{15}$$

The residual w_3, defined as

$$w_3 = w_2 - X^TXB_2 \tag{16}$$

is taken as the third weight vector in the PLS algorithm. We define yet a new set of normal equations

$$\begin{aligned} w_3 &= X^TXB_3, \\ &\quad \text{where } B_3 = W_3(W_3^TX^TXW_3)^{-1}W_3^Tw_3 \\ &\quad \text{and } W_3 = [w_1 \quad w_2 \quad w_3]. \end{aligned} \tag{17}$$

From this it is also simple to show that $W_2^Tw_3 = 0$ (premultiplying (16) with W_2^T). This gives $w_1^Tw_3 = 0$ and $w_2^Tw_3 = 0$, 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 Theorem 3.1. We can now combine the above equations to give the following normal equations which give us an expression for the PLS estimate of the matrix of regression coefficients

$$X^TY = X^TX\overbrace{(B_1 + B_2 + B_3 + \cdots + B_a)}^{B_{\mathrm{PLS}}}. \tag{18}$$

This shows that the problem of computing the PLS solution can be reduced to computing the weight matrix W_a. The procedure for computing the weight vectors, and the PLS solution B_{PLS} is presented in the following Theorem 3.1.

Theorem 3.1 (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 weight vector w_1, i.e., the first column in matrix $W_a = [w_1, \ldots, w_a]$ can be taken as*

$$w_1 = X^TY. \tag{19}$$

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

$$\begin{aligned} w_{i+1} &= w_i - X^TXB_i \\ &\quad \text{where } B_i = W_i(W_i^TX^TXW_i)^{-1}W_i^Tw_i, \end{aligned} \tag{20}$$

where W_i increases by one column at each iteration, i.e.

$$W_i = [w_1 \quad \ldots \quad w_i], \tag{21}$$

and $W_i^Tw_{i+1} = 0$. Finally, the PLS solution for the matrix of regression coefficients B is given by

$$B_{\mathrm{PLS}} = \sum_{i=1}^{a} B_i \tag{22}$$

which is equivalent to

$$B_{\mathrm{PLS}} = W_a(W_a^TX^TXW_a)^{-1}W_a^Tw_1. \tag{23}$$

Proof. See Appendix B. □

Theorem 3.1 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 nonsingular because W_a is an orthogonal matrix, and the PLS solution is equal to the ordinary LS estimate, i.e. $B_{\mathrm{PLS}} = B_{\mathrm{OLS}}$.

In Helland (1988) it was shown that the weight vector can also be computed as $w_{i+1} = w_1 - X^T X W_i (W_i^T X^T X W_i)^{-1} W_i^T w_1$ where $w_1 = X^T Y$. This is different from the iterations in Theorem 3.1. However, we can show that w_{i+1} can be computed from W_i and any of its columns w_j, i.e. we have the following alternative equation which can be used instead of Eq. (20)

$$w_{i+1} = w_j - X^T X H_i w_j \quad \forall j = 1, \ldots, i, \tag{24}$$

where

$$H_i = W_i (W_i^T X^T X W_i)^{-1} W_i^T. \tag{25}$$

The reader should note that the matrix product $X^T X H_i$ is an oblique projection. See e.g., Phatak and de Jong (1997) for a discussion of oblique projections and PLS. The algorithm for computing the weighting matrix W_i in Theorem 3.1 can be viewed as an orthogonalization process, e.g., Gram–Smith orthogonalization, Golub and Van Loan (1986). 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 $W_i^T w_i = [0 \ \ldots \ 0 \ \ w_i^T w_i]^T$. The orthogonalization process in Theorem 3.1 is not unique. For instance, define a non-singular scaling or transformation matrix $D \in \mathbb{R}^{a \times a}$. It is then evident that any weighting matrix defined as $W_a := W_a D$ gives the same PLS solution. This can be proved by substituting $W_a D$ for W_a in Eq. (23).

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

Theorem 3.2 (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*

$$w_1 := u_1, \quad \textit{where } USV^T := X^T Y \textit{ and } U = [u_1 \ \ldots \ u_m], \tag{26}$$

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

The other weight vectors $w_2, \ldots, w_a$ are computed recursively from $W_1 = w_1, (X^T Y)_1 = X^T Y$ and $X^T X$ as follows. Compute for all $i = 1, \ldots, a-1$

$$(X^T Y)_{i+1} = (I_r - X^T X W_i (W_i^T X^T X W_i)^{-1} W_i^T)(X^T Y)_i \tag{27}$$

and

$$w_{i+1} := u_1, \quad \textit{where } USV^T := (X^T Y)_{i+1} \textit{ and } U = [u_1 \ \ldots \ u_m], \tag{28}$$

where W_i increases by one column at each iteration, i.e.

$$W_i = [w_1 \ \ \ldots \ \ w_i]. \tag{29}$$

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

$$B_{\mathrm{PLS}} = W_a (W_a^T X^T X W_a)^{-1} W_a^T X^T Y. \tag{30}$$

Proof. See Appendix A. □

The resulting PLS2 solution is equivalent to the solution of the PLS2 kernal algorithm in Lindgren, Geladi and Wold (1993), de Jong and ter Braak (1994) and the PLS2 solution in Höskuldsson (1988,1996). Here we will present some alternative formulations for the problem of computing the PLS weighting vectors. The weight vectors (in Theorem 3.1) can equivalently be computed by the following process (which is standard in the PLS literature)

$$X_{i+1} = X_i - \frac{X_i w_i w_i^T X_i^T}{w_i^T X_i^T X_i w_i} X_i, \tag{31}$$

with $w_1 = X^T Y$, $X_1 = X$ and $w_{i+1} = X_{i+1}^T Y$ in Theorem 3.1. Furthermore, the weight vectors in Theorem 3.2 can equivalently be taken as the left singular vectors of $X^T Y$ and $X_{i+1}^T Y \ \forall i = 1, \ldots, a-1$ where $X_1 = X$ and X_{i+1} is defined in (31). See Appendix A for further details. 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 3.1 by a different scaling.

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 further in the next section.

3.2. Relationship between PLS and a controllability matrix

It is 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 have an equivalent problem of computing subspaces for a Krylov matrix. Correspondence with Krylov matrices and orthogonal decompositions are pointed out in Golub and Van Loan (1986). In the control literature the Krylov matrix is known as the controllability matrix. Krylov subspaces and PLS is discussed in Helland (1988). We have the following definition.

Definition 3.1 (Controllability (Krylov) matrix). Given matrices $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^{N\times m}$. The controllability (Krylov) matrix $K_r \in \mathbb{R}^{r\times rm}$ for the pair (X^TX, X^TY) is defined by

$$K_r = [X^TY \quad X^TXX^TY \quad (X^TX)^2X^TY \quad \ldots \quad (X^TX)^{r-1}X^TY]. \tag{33}$$

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

Proposition 3.1. *Given matrices $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^{N}$. The ordinary LS solution B_{OLS} can be expressed in terms of the controllability matrix of the pair (X^TX, X^TY) and the coefficients of the characteristic polynomial $\det(\lambda I_r - X^TX) = \lambda^r + p_2\lambda^{r-1} + \cdots + p_r\lambda + p_{r+1}$. Assume that X^TX is non-singular, then*

$$B_{\mathrm{OLS}} = (X^TX)^{-1}X^TY = K_r p, \tag{34}$$

where $K_r \in \mathbb{R}^{r\times r}$ is the controllability matrix for the pair (X^TX, X^TY) as defined in (33) and $p \in \mathbb{R}^r$ is a vector formed from the coefficients of the characteristic polynomial.

Proof. From the Cayley–Hamilton Theorem we have that X^TX satisfies its own characteristic equation, i.e.

$$(X^TX)^r + p_2(X^TX)^{r-1} + \cdots + p_rX^TX + p_{r+1}I_r = 0, \tag{35}$$

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

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

which is derived by post-multiplying (or equivalently, pre-multiplying) (35) with $(X^TX)^{-1}$ and then solving for the inverse. Substituting (36) into the LS solution gives Eq. (34) where

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

and the proposition follows. □

A consequence of Proposition 3.1 is that the ordinary LS solution for univariate Y data can be expressed as a linear combination of the columns in the controllability matrix (the multivariate case will be discussed in the next Section 4). The coefficient p_{r+1} in the characteristic polynomial can be computed as $p_{r+1} = \det(X^TX) = ?\lambda_1\lambda_2\cdots\lambda_r$. If 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 Eq. (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 $K_a \in \mathbb{R}^{r\times a}$, where $1 \le a \le r$. The matrix K_a is in general (i.e., for $m \ge 1$) defined as follows.

Definition 3.2 (Reduced controllability (Krylov) matrix). Given data matrices $X \in \mathbb{R}^{N\times r}$ and $Y \in \mathbb{R}^{N\times m}$, the reduced controllability (Krylov) matrix $K_a \in \mathbb{R}^{r\times am}$ for the pair (X^TX, X^TY) is defined by

$$K_a = [X^TY \quad X^TXX^TY \quad (X^TX)^2X^TY \quad \ldots \quad (X^TX)^{a-1}X^TY], \tag{38}$$

where $1 \le a \le r$.

Consider the univariate case. The number of columns, a, in the reduced controllability matrix K_a can in principle be taken as the (effective) rank of the Krylov matrix K_r, i.e., $a = \mathrm{rank}(K_r)$. In fact, we will now show that the column space of the weighting matrix W_a computed by the PLS1 algorithm and the column space of the reduced controllability matrix K_a coincide.

Proposition 3.2. *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*

$$K_a = W_aR_1, \tag{39}$$

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, $i = 1, \ldots, a$, are a linear combination of the columns of the controllability matrix, i.e.

$$W_a = K_aR_1^{-1}, \tag{40}$$

where R_1^{-1} is upper triangular. Furthermore, 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.

Proof. This result follows from that in Helland (1988) where it is 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.

This proposition can be proved from the weight vectors as computed in Theorem 3.1 and the controllability matrix K_a. We simply have to prove that $R_1 = W_a^T K_a$ is upper triangular or that $W_a = K_a R_1^{-1}$. A proof is presented in Appendix C. □

Define now the QR decomposition of the controllability matrix as

$$K_a = Q_a R, \tag{41}$$

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 (39) is then given by

$$W_a = Q_a R_2, \tag{42}$$

where $R_2 = R R_1^{-1}$ (usually diagonal and $R_2 = I$) is also upper triangular.

This implies that the weighting matrix W_a, computed by any PLS implementation, irrespective of scaling, etc., has 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 $W_a := Q_a$. This important result is presented in Theorem 3.3.

Proposition 3.3 (PLS: a QR decomposition of a controllability matrix). *Given data matrices $X \in \mathbb{R}^{N \times r}$ and $Y \in \mathbb{R}^N$, define the reduced controllability (Krylov) matrix K_a from X, Y and the number of components $1 \le a \le r$ as in* (38). *The column space of the weighting matrix W_a and the controllability (Krylov) 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 Controllability based PLS solution is then given by

$$B_{\mathrm{QPLS}} = Q_a (Q_a^T X^T X Q_a)^{-1} Q_a^T X^T Y. \tag{44}$$

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

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

and hence when $m = 1$

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

Proof. This result follows from Theorem 3.1, Proposition 3.2 and (42). □

We have defined the LS solution defined in Theorem 3.3 for the QR-based PLS solution (QPLS). The reason for this is that the solution differs from PLS when Y is multivariate, i.e. when $m > 1$. Theorem 3.3 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. Furthermore, the matrix $Q_a X^T X Q_a$ is tridiagonal since $Q_a = K_a R^{-1}$ is an (orthogonal) basis for $R(K_a)$ (Parlett, 1998, Section 12.7) and Golub and Van Loan (1986, Sections 7.4 and 9.1). Note also that letting $W_a := K_a$ gives the same PLS1 solution. This can be proved by substituting $W_a = K_a R_1^{-1}$ into solution 23 and using the assumption that R_1 is non-singular. We have the following proposition.

Proposition 3.4 (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 reduced controllability matrix for the pair $(X^T X, X^T Y)$ defined in (38) *and the polynomial coefficient vector $p^* \in \mathbb{R}^a$ is determined as the LS solution to*

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

where

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

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 $(K_a^T X^T X K_a)^{-1}$ is non-singular for some $1 \le a \le r$. The PLS prediction of Y is given by

$$Y_{\mathrm{PLS}} = X K_a p^*, \tag{52}$$

where p^* is given by (50). *Furthermore, the minimum is*

$$V(p^*) = \mathrm{trace}(Y^T Y) - \mathrm{trace}(Y^T X K_a (K_a^T X^T X K_a)^{-1} K_a^T X^T Y). \tag{53}$$

Proof. A truncated Cayley–Hamilton polynomial approximation of the matrix inverse in Eq. (36) is defined as

$$(X^T X)^{-1} := p_1 I_r + p_2 X^T X + p_3 (X^T X)^2 + \cdots + p_a (X^T X)^{a-1} \tag{54}$$

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

$$B(p) = K_a p, \tag{55}$$

where K_a is the controllability matrix and $p \in \mathbb{R}^a$ 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 solution to the squared Frobenius norm of the prediction error. Hence,

$$p^* = \arg\min_{p} V(p), \tag{56}$$

where the PE criterion for the coefficient vector is given by

$$\begin{aligned} V(p) &= \|Y - X\overbrace{K_a p}^{B(p)}\|_F^2 \\ &= \operatorname{trace}(Y^T Y) - 2\operatorname{trace}(p^T K_a^T X^T Y) \\ &\quad + \operatorname{trace}(p^T K_a^T X^T X K_a p). \end{aligned} \tag{57}$$

Letting the gradient

$$\frac{dV(p)}{dp} = -2K_a^T X^T Y + 2K_a^T X^T X K_a p \tag{58}$$

equal to zero gives the optimal solution (50) which, when substituted into (47) gives (51). Furthermore, the minimum value (53), can be found by substituting the optimal truncated polynomial coefficients p^* into (57). □

Proposition 3.4 and Theorem 3.3 are believed to be important for their simple and non-iterative interpretation and implementation of the PLS algorithm. The problem of computing the PLS solution to the LS problem is presented in the literature as an iterative algorithm, or as a piecewise linear regression algorithm. The explicit formulation (51) of the solution is presented in Helland (1988) but the prediction error interpretation of the solution is new.

The PLS algorithm is in the literature usually presented in terms of a score vector matrix $T \in \mathbb{R}^{N\times a}$, a loading matrix $C \in \mathbb{R}^{m\times a}$ for Y, a loading matrix $P \in \mathbb{R}^{r\times a}$ for X, in addition to the weighting matrix W_a. This notation is similar as in Helland (1988) and Lindgren et al. (1993). Furthermore, the a columns in T represents the latent variables. Y is decomposed as $Y = TC^T + \mathscr{E}$ where $C = (T^T T)^{-1} T^T Y$ and $\mathscr{E}$ is the prediction error. X is decomposed as $X = TP^T + \mathscr{E}_X$ where $P = (T^T T)^{-1} T^T X$ and $\mathscr{E}_X$ is a residual. One should note that these definitions of the loading matrices ensures that the score vector matrix is normal to the prediction error and the residual, i.e. $T^T \mathscr{E} = 0$ and $T^T \mathscr{E}_X = 0$. Furthermore, the PLS solution can be expressed as $B_{\text{PLS}} = W_a (P^T W_a)^{-1} C^T$ (Manne, 1987; Helland, 1988). This is an alternative to (23), (44) or (51). It follows from Proposition 3.4 that the PLS1 algorithm decomposes Y as $Y = Y_M + \mathscr{E}$ where the prediction is given by $Y_M = XK_a(K_a^T X^T X K_a)^{-1} K_a^T X^T Y$ and where $\mathscr{E}$ is the prediction error. Comparing the column space of this and the column space of the prediction $Y_M = TC^T$ we have that the PLS score vector matrix T is related to XK_a as $T = XK_a D$ for some non-singular matrix $D \in R^{a\times a}$. Choosing $D = I_a$ gives a score matrix $T := XK_a$. This shows that X can be decomposed as $X = (XK_a(K_a^T X^T X K_a)^{-1} K_a^T X^T X + \mathscr{E}_X$.

From this it is clear that the columns in XK_a are a basis for the score vector matrix. See de Jong (1993). Consider now the QR decomposition

$$\tilde{Q}\tilde{R} = XK_a, \tag{59}$$

which gives an orthonormal basis for the range of XK_a. Hence, $\tilde{Q}$ is an orthogonal (with orthonormal columns) score vector matrix and we can let $T := \tilde{Q}$. See, e.g. Martens and Næs (1989) for a PLS1 algorithm with orthogonal scores. Substituting (59) and the QR decomposition (41) into the solution (51) gives

$$B_{\text{PLS}} = Q_a(\tilde{Q}^T X Q_a)^{-1} \tilde{Q}^T Y, \tag{60}$$

where $\tilde{Q} X Q_a$ is (upper) bidiagonal (Golub and Van Loan, 1986, Sections 6.5 and 9.3; Manne, 1987). Hence, the loadings can be defined as $P^T = \tilde{Q} X$ and $C^T = \tilde{Q} Y$. The PLS1 solution turns out (Wold, Ruhe, Wold & Dunn, 1984) to be similar to the bidiagonalization LS algorithm in Paige and Saunders (1982). Note that (59) can be changed to $\tilde{Q}\tilde{R} = XW_a$ in the PLS2 algorithm. Substituting this into (30) gives $B_{\text{PLS}} = W_a(\tilde{Q}^T X W_a)^{-1} \tilde{Q}^T Y$ where $\tilde{Q}^T X W_a = \tilde{R}$ is upper triangular.

It is interesting to recognize the relationship between the PLS1 solution in (44) and the Lanczos method for tridiagonalizing a symmetric matrix ($Q_a^T X^T X Q_a$ tridiagonal). See Golub and Van Loan (1986) and in particular Algorithm 9.3.1 where Lanczos tridiagonalization is used to iteratively solve LS problems. A truncated version of this iterative LS algorithm results in a PLS1 algorithm. Furthermore, this algorithm is similar to the method of conjugate gradients, Algorithm 10.2.13 in Golub and Van Loan (1986) (a truncated version of this algorithm gives the PLS1 solution).

One should note that it is possible to modify the solution in Proposition 3.4 in order to incorporate a possible known row weighting matrix $Z \in \mathbb{R}^{N\times N}$, by letting, e.g. $p^* = \arg\min_p \|Z^{1/2}(Y - XK_a p)\|_F^2$ which gives the non-iterative PLS solution with row weighting $B_{\text{PLS}} = K_a(K_a^T X^T Z X K_a)^{-1} K_a^T X^T Z Y$. This is equivalent to the Best Linear Unbiased Estimator (BLUE), (see, e.g. Söderström and Stoica (1989) for further details) i.e. $B_{\text{BLUE}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$ when $a = r$, K_r non-singular and $Z = \Sigma^{-1}$ where $\Sigma = \mathrm{E}(EE^T) > 0$.

4. Multivariate extensions

In this section we will 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.

$$Y = X\overbrace{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y}^{B_{\mathrm{OLS}}} + \mathscr{E}, \tag{61}$$

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

$$(X^{\mathrm{T}}X)^{-1} := p_1 I_r + p_2 X^{\mathrm{T}}X + p_3(X^{\mathrm{T}}X)^2 + \cdots + p_a(X^{\mathrm{T}}X)^{a-1}, \tag{62}$$

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

$$\mathscr{E} = Y - X \overbrace{(p_1 I_r + p_2 X^{\mathrm{T}}X + p_3(X^{\mathrm{T}}X)^2 + \cdots + p_a(X^{\mathrm{T}}X)^{a-1})X^{\mathrm{T}}Y}^{B_{\mathrm{CPLS}}(p)}, \tag{63}$$

which can be expressed as

$$\mathscr{E} = Y - X\overbrace{\overbrace{[X^{\mathrm{T}}Y \quad (X^{\mathrm{T}}X)X^{\mathrm{T}}Y \quad \ldots \quad (X^{\mathrm{T}}X)^{a-1}X^{\mathrm{T}}Y]}^{K_a}\begin{bmatrix} p_1 I_m \\ p_2 I_m \\ \vdots \\ p_a I_m \end{bmatrix}}^{B_{\mathrm{CPLS}}(p)}. \tag{64}$$

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

Theorem 4.1 (CPLS: 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* $1 \le a \le r$, *the optimal solution is*

$$\begin{aligned} B_{\mathrm{CPLS}} &= \overbrace{[X^{\mathrm{T}}Y \quad (X^{\mathrm{T}}X)X^{\mathrm{T}}Y \quad \ldots \quad (X^{\mathrm{T}}X)^{a-1}X^{\mathrm{T}}Y]}^{K_a} \\ &\quad \times \begin{bmatrix} p_1 I_m \\ p_2 I_m \\ \vdots \\ p_a I_m \end{bmatrix} \\ &= (p_1 I_r + p_2 X^{\mathrm{T}}X + p_3(X^{\mathrm{T}}X)^2 + \cdots \\ &\quad + p_a(X^{\mathrm{T}}X)^{a-1})X^{\mathrm{T}}Y \\ &= \sum_{i=1}^{a} p_i (X^{\mathrm{T}}X)^{i-1}X^{\mathrm{T}}Y, \end{aligned} \tag{65}$$

where the vector of polynomial coefficients

$$p^* = [p_1 \quad p_2 \quad \ldots \quad p_a]^{\mathrm{T}} \in \mathbb{R}^a \tag{66}$$

is found from the solution to the LS problem

$$p^* = \arg\min_p \|\mathrm{vec}(Y) - X_p p\|_{\mathrm{F}}^2. \tag{67}$$

The minimizing solution is given by

$$p^* = (X_p^{\mathrm{T}} X_p)^{-1} X_p \mathrm{vec}(Y), \tag{68}$$

where

$$X_p = [\mathrm{vec}(XX^{\mathrm{T}}Y) \quad \mathrm{vec}(XX^{\mathrm{T}}XX^{\mathrm{T}}Y) \quad \ldots \quad \mathrm{vec}(X(X^{\mathrm{T}}X)^{a-1}X^{\mathrm{T}}Y)] \in \mathbb{R}^{Nm\times a}. \tag{69}$$

Proof. The prediction error, Eq. (63), can be written as

$$\mathrm{vec}(\mathscr{E}) = \mathrm{vec}(Y) - \overbrace{[\mathrm{vec}(XX^{\mathrm{T}}Y) \quad \mathrm{vec}(XX^{\mathrm{T}}XX^{\mathrm{T}}Y) \quad \ldots \quad \mathrm{vec}(X(X^{\mathrm{T}}X)^{a-1}X^{\mathrm{T}}Y)]}^{X_p}\, p, \tag{70}$$

where p is defined in (66). Using that $V(p) = \|\mathscr{E}\|_{\mathrm{F}}^2 = \|\mathrm{vec}(\mathscr{E})\|_{\mathrm{F}}^2$ where $\mathscr{E}$ is the prediction error (i.e. a real matrix), gives the optimal LS solution (68) by letting the gradient $\mathrm{d}V(p)/\mathrm{d}p = 0$. See also Appendix D 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. The CPLS solution for multivariate Y data can be expressed as a linear combination of the $r \times m$ block columns in the reduced controllability matrix $K_a \in \mathbb{R}^{r\times ma}$. The CPLS solution is identical to the PLS1 solution for univariate data. The solution for univariate data can be expressed as a linear combination of the columns in the controllability matrix $K_a \in \mathbb{R}^{r\times a}$. Note also that the CPLS algorithm gives the same solution as the univariate PLS1 algorithm applied to the model (5). In order to give further insight into the CPLS solution in Theorem 4.1 and to present an alternative method for defining the coefficient vector p we have the following proposition.

Proposition 4.1 (Coefficient vector in CPLS solution). *The coefficient vector* $p \in \mathbb{R}^a$ *in Theorem* 4.1 *can be defined by the linear equation*

$$\mathscr{H} p = f, \tag{71}$$

where the matrix $\mathscr{H} \in \mathbb{R}^{a \times a}$ and the vector $f \in \mathbb{R}^a$ are given by

$$\mathscr{H} = \begin{bmatrix} \mathrm{trace}(Y^T X X^T X X^T Y) & \cdots & \mathrm{trace}(Y^T X (X^T X)^a X^T Y) \\ \vdots & \ddots & \vdots \\ \mathrm{trace}(Y^T X (X^T X)^a X^T Y) & \cdots & \mathrm{trace}(Y^T X (X^T X)^{2a-1} X^T Y) \end{bmatrix}, \quad (72)$$

and

$$f = \begin{bmatrix} \mathrm{trace}(Y^T X X^T Y) \\ \vdots \\ \mathrm{trace}(Y^T X (X^T X)^{a-1} X^T Y) \end{bmatrix}. \quad (73)$$

Furthermore, when $\mathscr{H}$ is non-singular we have the solution $p^* = \mathscr{H}^{-1} f$.

Proof. The squared Frobenius norm of the prediction error (63) can be written as

$$\begin{aligned} V(p) &= \| Y - X(p_1 I_r + p_2 X^T X + p_3 (X^T X)^2 + \cdots \\ &\quad + p_a (X^T X)^{a-1}) X^T Y \|_F^2 \\ &= \mathrm{trace}(Y^T Y) - 2 f^T p + p^T \mathscr{H} p, \end{aligned} \quad (74)$$

where $\mathscr{H}$ and f are defined in (72) and (73), respectively. Letting the gradient $\mathrm{d}V(p)/\mathrm{d}p = 0$ gives the condition (71). Furthermore, the optimal solution is $p^* = \mathscr{H}^{-1} f$ when the Hessian matrix $\mathrm{d}^2 V(p)/\mathrm{d}p^2 = H$ is non-singular. □

The reader should note that, in the univariate case, Proposition 4.1 reduces to $(K_a^T X^T X K_a) p = K_a^T X^T Y$ where K_a is the reduced controllability matrix as defined in (38). This results in a coefficient vector which is identical to the PLS1 coefficient vector, p, in Eq. (50). This shows that, for univariate data, the CPLS solution reduces to the PLS1 solution.

5. Generalized eigenvalue problem and LS solutions

5.1. Optimal weights

From the previous discussion we have shown that the PLS estimate B_{PLS} can be expressed in terms of X, Y and a weighting matrix $W_a \in \mathbb{R}^{r \times a}$, which is a function of a set of polynomial coefficients. Different LS regression methods use different weighting matrices, thus leading to different least-squares regression methods. We will now show that there exists 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 exists a minimum number a of columns in the weighting matrix. The resulting optimal LS solution is, 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.

Theorem 5.1 (The estimate of the matrix of regression coefficients). *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 is given by*

$$B(W_a) = W_a (W_a^T X^T X W_a)^{-1} W_a^T X^T Y \in \mathbb{R}^{r \times m}, \quad (75)$$

where we have assumed that $W_a^T X^T X W_a \in \mathbb{R}^{a \times a}$ is non-singular, and satisfies the weighted normal equations

$$W_a^T X^T Y = W_a^T X^T X B(W_a). \quad (76)$$

Proof. Theorem 5.1 can be proved by substituting the LS solution $B(W_a)$ defined in (75) into the weighted normal equations (76). □

It is obvious that when W_a is equal to the identity matrix and $X^T X$ 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 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

$$\begin{aligned} V(w) = \| Y - XB(w) \|_F^2 &= Y^T Y - \frac{Y^T X w w^T X^T Y}{w^T X^T X w} \\ &= Y^T Y - \frac{w^T X^T Y Y^T X w}{w^T X^T X w}, \end{aligned} \quad (77)$$

where $B(w) = w(w^T X^T X w)^{-1} w^T X^T Y$. For the sake of simplicity we have also 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

$$\frac{\mathrm{d}V(w)}{\mathrm{d}w} = - \frac{2 X^T Y Y^T X w (w^T X^T X w) - w^T X^T Y Y^T X w (2 X^T X w)}{(w^T X^T X w)^2}. \quad (78)$$

Letting the gradient equal to zero gives

$$X^T Y Y^T X w = \frac{w^T X^T Y Y^T X w}{w^T X^T X w} X^T X w. \quad (79)$$

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

This is a generalized eigenvalue problem, i.e. $\lambda_1 = w^T X^T Y Y^T X w / w^T X^T X w$ is a generalized eigenvalue of the square matrices $X^T Y Y^T X$ and $X^T X$ 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.

Theorem 5.2 (Generalized eigenvalue problem). *The optimal weighting matrix $W_a \in \mathbb{R}^{r \times a}$ where the number of components is bounded by $1 \leq a \leq r$, which minimizes the PE (defined here as the squared Frobenius matrix norm)*

$$V(W_a) = \|Y - XB(W_a)\|_F^2 = \mathrm{trace}(Y^T Y) - \mathrm{trace}(Y^T X W_a (W_a^T X^T X W_a)^{-1} W_a^T X^T Y) \quad (80)$$

can be computed by the following generalized eigenvalue problem

$$X^T Y Y^T X W_a = X^T X W_a \Lambda_a, \quad (81)$$

where

$$\Lambda_a = (W_a^T X^T X W_a)^{-1} W_a^T X^T Y Y^T X W_a \in \mathbb{R}^{a \times a} \quad (82)$$

is a diagonal matrix with the generalized eigenvalues on the diagonal, and where W_a is the corresponding generalized eigenvector matrix. Furthermore, the minimum value of the PE

$$V(W_a) = \|Y - XB(W_a)\|_F^2 = \mathrm{trace}(Y^T Y) - \mathrm{trace}(\Lambda_a). \quad (83)$$

Proof. We will prove the Theorem from an expression of the covariance matrix of $X^T Y$. Using the LS solution $B(W_a)$, gives the normal equations

$$X^T Y = X^T X W_a (W_a^T X^T X W_a)^{-1} W_a^T X^T Y. \quad (84)$$

Post-multiplication with $Y^T X W_a$ gives

$$X^T Y Y^T X W_a = X^T X W_a \overbrace{(W_a^T X^T X W_a)^{-1} W_a^T X^T Y Y^T X W_a}^{\Lambda_a}, \quad (85)$$

which is equivalent to the following generalized eigenvalue problem:

$$X^T Y Y^T X W_a = X^T X W_a \Lambda_a, \quad (86)$$

where W_a is the generalized eigenvector matrix of the square matrices $X^T Y Y^T X$ and $X^T X$ and

$$\Lambda_a = (W_a^T X^T X W_a)^{-1} W_a^T X^T Y Y^T X W_a \quad (87)$$

is the corresponding generalized eigenvalue matrix. Note that the above is equivalent to formulating the correlation matrix of $X^T Y$ given by the normal equation, i.e.

$$X^T Y (X^T Y)^T = X^T X W_a (W_a^T X^T X W_a)^{-1} \times W_a^T X^T Y Y^T X W_a (W_a^T X^T X W_a)^{-1} W_a^T X^T X. \quad (88)$$

Post-multiplying with W_a gives Eqs. (86) and (87). The minimum value can be found as follows:

$$V(W_a) = \|Y - XB(W_a)\|_F^2 = \mathrm{trace}(Y^T Y) - \mathrm{trace}(Y^T X W_a (W_a^T X^T X W_a)^{-1} W_a^T X^T Y) = \mathrm{trace}(Y^T Y) - \mathrm{trace}(W_a^T \underbrace{X^T Y Y^T X W_a}_{X^T X W_a \Lambda_a} (W_a^T X^T X W_a)^{-1}). \quad (89)$$

Substituting for the stationary condition Eq. (81) gives

$$V(W_a) = \|Y - XB(W_a)\|_F^2 = \mathrm{trace}(Y^T Y) - \mathrm{trace}(\Lambda_a). \quad \square \quad (90)$$

The generalized eigenproblem in Theorem 5.2 can be solved by the QZ algorithm (Golub, 1983). The weighting matrix W_a can be computed in MATLAB as $[Aa, Bb, q, Z, V] = \mathrm{qz}(X^T Y Y^T X, X^T X)$ and putting $W_a = V(:, 1:a)$. Note that W and Λ can also be computed by the MATLAB function eig$(\cdot,\cdot)$, i.e. $[W, \Lambda] = \mathrm{eig}(X^T Y Y^T X, X^T X)$. 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$(\cdot,\cdot)$. Investigations of the above result indicate that the resulting optimal LS solution is the same for all $m \leq a \leq r$, and that this solution is the same as the OLS solution. The question is whether the minimum number of components is $a = m$ or not. In the case when $X^T X$ is non-singular the above corresponds to taking the weights from the column space of the OLS solution $(X^T X)^{-1} X^T Y$. In the next section we will use the results presented in this section to develop a regularized estimator for the PLS weights.

5.2. An estimator for the PLS weights

The number of parameters in the PLS weighting matrix W_a is ra but there are 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, where we assumed that $1 \leq m \leq a$. In order to formulate the PLS algorithm as an estimator we must find the relationship between the PLS solution and the rm unknown parameters. This relationship is presented in the following theorem.

Theorem 5.3 (The number of unique PLS parameters). *Assume that a weighting matrix W_a with*

$m \le a \le r$ 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:

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

where the weighting matrix w is composed of the eigenvectors of $W_a(W_a^{\mathrm{T}}X^{\mathrm{T}}XW_a)^{-1}W_a^{\mathrm{T}}X^{\mathrm{T}}YY^{\mathrm{T}}X$ corresponding to the m largest eigenvalues, i.e., w is a solution to the following eigenvalue problem:

$$W_a(W_a^{\mathrm{T}}X^{\mathrm{T}}XW_a)^{-1}W_a^{\mathrm{T}}X^{\mathrm{T}}YY^{\mathrm{T}}Xw = w\lambda, \tag{92}$$

where

$$\lambda = (w^{\mathrm{T}}X^{\mathrm{T}}Xw)^{-1}w^{\mathrm{T}}X^{\mathrm{T}}YY^{\mathrm{T}}Xw \in \mathbb{R}^{m\times m}. \tag{93}$$

Proof. Assume first that there exists an equivalent weighting matrix w. Putting the two expressions for the same solution equal to each other gives

$$\overbrace{W_a(W_a^{\mathrm{T}}X^{\mathrm{T}}XW_a)^{-1}W_a^{\mathrm{T}}X^{\mathrm{T}}}^{B_{\mathrm{PLS}}(W_a)} = \overbrace{w(w^{\mathrm{T}}X^{\mathrm{T}}Xw)^{-1}w^{\mathrm{T}}X^{\mathrm{T}}Y}^{B_{\mathrm{PLS}}(w)}. \tag{94}$$

Post-multiplication with $Y^{\mathrm{T}}Xw$ gives an eigenvalue problem $Zw = \lambda w$, i.e.,

$$\overbrace{W_a(W_a^{\mathrm{T}}X^{\mathrm{T}}XW_a)^{-1}W_a^{\mathrm{T}}X^{\mathrm{T}}YY^{\mathrm{T}}X}^{Z}w = w\overbrace{(w^{\mathrm{T}}X^{\mathrm{T}}Xw)^{-1}w^{\mathrm{T}}X^{\mathrm{T}}YY^{\mathrm{T}}Xw}^{\lambda}. \tag{95}$$

A basis for the weighting matrix w can be taken from the column space of the solution $B_{\mathrm{PLS}} = W_a(W_a^{\mathrm{T}}X^{\mathrm{T}}XW_a)^{-1}W_a^{\mathrm{T}}X^{\mathrm{T}}Y$, i.e. $R(w) \subseteq R(B_{\mathrm{PLS}})$. This gives a solution of the form $B_{\mathrm{PLS}}(p) = wp$ where $p \in \mathbb{R}^{m\times m}$. Solving for p in a LS optimal sense (as in Definition 2.1) gives (91). Hence, there exists an equivalent weighting matrix $w \in \mathbb{R}^{r\times m}$. □

We can now present the PLS algorithm as an estimator. The following result is presented for the univariate case. The extention to the multivariate case is clarified later.

Theorem 5.4 (PLS1 optimization criterion). *The PLS estimate B_{PLS} of the matrix of regression coefficients B can be expressed in terms of $X \in \mathbb{R}^{N\times r}$, $Y \in \mathbb{R}^{N}$, and an estimate $\hat{w}$ of a single weight vector $w \in \mathbb{R}^{r}$. The PLS estimate is given by*

$$B_{\mathrm{PLS}} = \hat{w}(\hat{w}^{\mathrm{T}}X^{\mathrm{T}}X\hat{w})^{-1}\hat{w}^{\mathrm{T}}X^{\mathrm{T}}Y, \tag{96}$$

where

$$\hat{w} = \arg\min_{w} V(w), \tag{97}$$

where

$$V(w) = \mathrm{trace}(Y^{\mathrm{T}}Y) - \lambda, \tag{98}$$

where

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

and for PLS we choose

$$z = w_{a+1} = X^{\mathrm{T}}Y - X^{\mathrm{T}}XH_aX^{\mathrm{T}}Y,$$
$$H_a = K_a(K_a^{\mathrm{T}}X^{\mathrm{T}}XK_a)^{-1}K_a^{\mathrm{T}}, \tag{100}$$

where a is the number of components and K_a is the controllability matrix for the pair $(X^{\mathrm{T}}X, X^{\mathrm{T}}Y)$. The vector w_{a+1} can also be computed from Theorem 3.1. *Furthermore, this can be written as*

$$V(w) = \mathrm{trace}(Y^{\mathrm{T}}Y) - \frac{w^{\mathrm{T}}X^{\mathrm{T}}YY^{\mathrm{T}}Xw}{w^{\mathrm{T}}X^{\mathrm{T}}Xw} + \frac{w^{\mathrm{T}}(2X^{\mathrm{T}}Yz^{\mathrm{T}} - zz^{\mathrm{T}})w}{w^{\mathrm{T}}X^{\mathrm{T}}Xw}, \tag{101}$$

and

$$V(w) = \|Y - XB(w)\|_F^2 + \frac{w^{\mathrm{T}}(2X^{\mathrm{T}}Yz^{\mathrm{T}} - zz^{\mathrm{T}})w}{w^{\mathrm{T}}X^{\mathrm{T}}Xw}, \tag{102}$$

where

$$B(w) = w(w^{\mathrm{T}}X^{\mathrm{T}}Xw)^{-1}w^{\mathrm{T}}X^{\mathrm{T}}Y. \tag{103}$$

Proof. Note that the second term in the PE criterion is equal to zero if the weight w is orthogonal to the residual z. Hence, the estimator attracts weighting matrices such that $z^{\mathrm{T}}w = 0$. For the rest of the proof, see Theorems 5.3 and 5.5 and the comments at the end of this section.

Theorem 5.4 is important from a statistical point of view. It implies that PLS is a regularized prediction error estimator. It implies that it is only a 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 $z = X^{\mathrm{T}}(Y - XB_{\mathrm{OLS}})$ gives the ordinary LS estimator and that $z = X^{\mathrm{T}}(Y - XB_{\mathrm{PCR}})$ gives the PCR estimator. The vector z can be viewed as regularization parameters which attracts the parameter estimator to a point in the parameter space. The solution to the optimization problem can be found from a generalized eigenvalue problem and presented in the next theorem.

Theorem 5.5 (PLS as a generalized eigenvalue problem).

$$(X^{\mathrm{T}}Y - z)(Y^{\mathrm{T}}X - z^{\mathrm{T}})w = X^{\mathrm{T}}Xw\lambda, \tag{104}$$

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

$$\lambda = \frac{w^{\mathrm{T}}(X^{\mathrm{T}}Y - z)(Y^{\mathrm{T}}X - z^{\mathrm{T}})w}{w^{\mathrm{T}}X^{\mathrm{T}}Xw}, \tag{105}$$

where

$$z = w_{a+1}. \tag{106}$$

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

$$B_{\mathrm{PLS}} = w(w^{\mathrm{T}}X^{\mathrm{T}}Xw)^{-1}w^{\mathrm{T}}X^{\mathrm{T}}Y. \tag{107}$$

Proof. We have that the residual of the normal equations are

$$z = X^{\mathrm{T}}Y - X^{\mathrm{T}}XW_a(W_a^{\mathrm{T}}X^{\mathrm{T}}XW_a)^{-1}W_a^{\mathrm{T}}X^{\mathrm{T}}Y, \tag{108}$$

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

$$X^{\mathrm{T}}Y - z = X^{\mathrm{T}}XW_m(W_m^{\mathrm{T}}X^{\mathrm{T}}XW_m)^{-1}W_m^{\mathrm{T}}X^{\mathrm{T}}Y. \tag{109}$$

The covariance matrix of $X^{\mathrm{T}}Y - z$, post-multiplied by W_m, is expressed as

$$(X^{\mathrm{T}}Y - z)(X^{\mathrm{T}}Y - z)^{\mathrm{T}}W_m = X^{\mathrm{T}}XW_m \overbrace{(W_m^{\mathrm{T}}X^{\mathrm{T}}XW_m)^{-1}W_m^{\mathrm{T}}X^{\mathrm{T}}YY^{\mathrm{T}}XW_m}^{\Lambda_m}, \tag{110}$$

which is a generalized eigenvalue problem for W_m and Λ_m. □

Consider the following regularized PE criterion:

$$V(W_m) = \|Y - XB(W_m)\|_{\mathrm{F}}^2 + \mathrm{trace}(W_m^{\mathrm{T}}(2X^{\mathrm{T}}Y - z)z^{\mathrm{T}}W_m(W_m^{\mathrm{T}}X^{\mathrm{T}}XW_m)^{-1}), \tag{111}$$

which can be written as

$$V(W_m) = \mathrm{trace}(Y^{\mathrm{T}}Y) - \mathrm{trace}(W_m^{\mathrm{T}}\overbrace{(X^{\mathrm{T}}Y - z)(X^{\mathrm{T}}Y - z)^{\mathrm{T}}W_m}^{X^{\mathrm{T}}XW_m\Lambda_m}(W_m^{\mathrm{T}}X^{\mathrm{T}}XW_m)^{-1})$$

$$= \mathrm{trace}(Y^{\mathrm{T}}Y) - \mathrm{trace}(\Lambda_m). \tag{112}$$

For univariate data, this reduces to the results in Theorem 5.4. 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 $z^{\mathrm{T}}W_m = 0$.

6. Discussion

6.1. 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 first a left singular vectors of $X^{\mathrm{T}}Y$. We have in this paper proved that the PLS solution in general is related to the controllability matrix K_a of the pair $(X^{\mathrm{T}}X, X^{\mathrm{T}}Y)$. In the univariate case we have $B_{\mathrm{PLS}} = K_a p^*$ (Theorem 3.3) and in the multivariate case

$$B_{\mathrm{CPLS}} = [\overbrace{X^{\mathrm{T}}Y \quad (X^{\mathrm{T}}X)X^{\mathrm{T}}Y \quad \dots \quad (X^{\mathrm{T}}X)^{a-1}X^{\mathrm{T}}Y}^{K_a}] \times \begin{bmatrix} p_1 I_m \\ p_2 I_m \\ \vdots \\ p_a I_m \end{bmatrix}$$

as presented in Theorem 4.1. A more general alternative to UPLS is then to take the weighting matrix W_a equal to the first a left singular vectors of K_a, i.e. $W_a = U(:, 1:a)$ where $USV^{\mathrm{T}} = K_a$.

Another choice is to choose W_a equal to a controllability matrix of the pair $(X^{\mathrm{T}}X, w_1)$ where w_1 is equal to the first singular vector of $X^{\mathrm{T}}Y$. 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 4.1. However, note that CPLS is the minimizing solution to a well defined prediction error, but the above solution has diffuse statistical properties. We mention this as a comment to the UPLS solution, but we will not elaborate this further.

6.2. Prediction, bias and variance

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 PCR, for the same number of components. This is also illustrated in Examples 7.2 and 7.3. 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 PLS1 are taken as the minimizing parameters of the prediction error. One should note that PLS2 is usually not optimal on the identification data, i.e. not optimal with respect to minimizing (a norm) of the prediction error. However, as claimed in the PLS literature, PLS2 may be good for predicting validation (independent output) data.

Like PCR, PLS gives bias free estimates in case of measurement noise only (noise on Y), assuming that the rank of X actually is $a \le r$ and that a sufficient number of components is used in the two algorithms, i.e. $a = \mathrm{rank}(X)$ components are used in PCR and $a = \mathrm{rank}(K_r)$ components are used in PLS. In order to illustrate the difference, note that if X is orthogonal, then only one ($a=1$) component is needed in PLS1 but that $a=r$ components has to be used in PCR (Frank and Friedman, 1993). For PLS2 one has to use $a=m$ components in order for the solution to be identical with the OLS solution.

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 also that OLS and PCR gives bias in this case. An interesting solution to the errors-in-variables problem is the Total Least Squares (TLS), (Van Huffel and Vandewalle, 1991), and the Truncated Total Least Squares (TTLS) solution, (De Moor & David, 1996; Fierro, Golub, Hansen & O'Leary, 1997; and Hansen, 1992). The TTLS solution can be computed as $B_{\mathrm{TTLS}} = -V_{12}V_{22}^{\dagger}$ where $V_{12}\in\mathbb{R}^{r\times r+m-a}$ and $V_{22}\in\mathbb{R}^{m\times r+m-a}$ are taken from the SVD of the compond matrix

$$[X \quad Y] = USV^{\mathrm{T}} = [U_1 \quad U_2]\begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}^{\mathrm{T}}.$$

In MATLAB notation, $V_{12} := V(1:r, a+1:r+m)$ and $V_{22} := V(r+1:r+m, a+1:r+m)$. This is the solution to the problem of minimizing $\|[X \quad Y] - [X_{\mathrm{TTLS}} \quad Y_{\mathrm{TTLS}}]\|_{\mathrm{F}}^2 = \|X - X_{\mathrm{TTLS}}\|_{\mathrm{F}}^2 + \|Y - Y_{\mathrm{TTLS}}\|_{\mathrm{F}}^2$ with respect to X_{TTLS} and Y_{TTLS} where $Y_{\mathrm{TTLS}} = X_{\mathrm{TTLS}}B_{\mathrm{TTLS}}$ is the TTLS prediction.

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 $1 \le a \le r$ is in general a trade-off between bias and variance, and model validation, e.g. cross 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 when using regularization in system identification. The exact statistical properties like mean and variance of the PLS solution is hard to derive due to the fact that B_{PLS} is non-linear in Y when $1 \le a < r$. Approximations based on 1. order derivatives are presented in Pathak (1993).

6.3. SIMPLS

We are aware of the variant of PLS which is denoted by SIMple PLS presented in de Jong (1993) and discussed further in ter Braak and de Jong (1998). SIMPLS gives the same solution as PLS for univariate Y data, but in general gives different solutions for multivariate Y data. This is illustrated in Examples 7.2 and 7.3. Like PLS, the first weight vector w_1 in SIMPLS can be taken as the left singular vector of $X^{\mathrm{T}}Y$, i.e. $w_1 = U(:,1)$ where $USV^{\mathrm{T}} = X^{\mathrm{T}}Y$. The next weight vectors are computed iteratively as follows. Let $w_i = w_1$ and for all $i = 2, \dots, a$ construct a projection matrix $P_i = X^{\mathrm{T}}Xw_i/(w_i^{\mathrm{T}}X^{\mathrm{T}}Xw_i)$. The weight vector w_i can be taken as the first left singular vector of $(I_r - P_i)X^{\mathrm{T}}Y$, i.e. $w_i = U(:,1)$ where $(I_r - P_i)X^{\mathrm{T}}Y = USV^{\mathrm{T}}$. 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). In our Example 7.2 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.

7. Examples

Example 7.1. Consider the following example from Hansen (1992)

$$\overbrace{\begin{bmatrix} 0.27 \\ 0.25 \\ 3.33 \end{bmatrix}}^{Y} = \overbrace{\begin{bmatrix} 0.16 & 0.10 \\ 0.17 & 0.11 \\ 2.02 & 1.29 \end{bmatrix}}^{X} \overbrace{\begin{bmatrix} 1.00 \\ 1.00 \end{bmatrix}}^{B} + \overbrace{\begin{bmatrix} 0.01 \\ -0.03 \\ 0.02 \end{bmatrix}}^{E}. \tag{113}$$

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

$$B_{\mathrm{OLS}} = \begin{bmatrix} 7.01 \\ -8.40 \end{bmatrix}, \quad \|B_{\mathrm{OLS}}\|_{\mathrm{F}} = 10.94,$$

$$\|Y - XB_{\mathrm{OLS}}\|_{\mathrm{F}} = 0.02. \tag{114}$$

$$B_{\mathrm{PLS}} = \begin{bmatrix} 1.1703 \\ 0.7473 \end{bmatrix}, \quad \|B_{\mathrm{PLS}}\|_{\mathrm{F}} = 1.3885,$$

$$\|Y - XB_{\mathrm{PLS}}\|_{\mathrm{F}} = 0.0322. \tag{115}$$

$$B_{\mathrm{TTLS}} = \begin{bmatrix} 1.1703 \\ 0.7473 \end{bmatrix}, \quad \|B_{\mathrm{TTLS}}\|_{\mathrm{F}} = 1.3885,$$

$$\|Y - XB_{\mathrm{TTLS}}\|_{\mathrm{F}} = 0.0322. \tag{116}$$

A major difficulty with the above ordinary least squares solution B_{OLS} in (114) is that its norm is significantly greater than the norm of the exact solution, which is $\|B\|_{\mathrm{F}} = \sqrt{2}$. 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 ($a=1$) solution is that regularization is introduced in order to stabilize the solution.

Example 7.2 (Real world data from a pulp and paper mill I). A refiner experiment at Union Co, Skien, Norway, 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 with one hour sampling rate. The data is organized into data matrices $X \in \mathbb{R}^{16\times 4}$ and $Y \in \mathbb{R}^{16\times 2}$ as follows.

$$X = \begin{bmatrix} 9.3 & 0.54 & 4.5 & 13.0 \\ 8.3 & 0.64 & 4.0 & 13.0 \\ 9.3 & 0.54 & 4.0 & 13.0 \\ 8.3 & 0.64 & 4.5 & 13.0 \\ 8.3 & 0.54 & 4.5 & 13.0 \\ 9.3 & 0.64 & 4.5 & 13.0 \\ 8.3 & 0.54 & 4.0 & 13.0 \\ 9.3 & 0.64 & 4.0 & 13.0 \\ 7.0 & 0.70 & 4.5 & 11.0 \\ 8.0 & 0.60 & 4.0 & 11.0 \\ 8.0 & 0.70 & 4.5 & 11.0 \\ 8.0 & 0.70 & 4.0 & 11.0 \\ 7.0 & 0.60 & 4.0 & 11.0 \\ 8.0 & 0.60 & 4.5 & 11.0 \\ 7.0 & 0.70 & 4.0 & 11.0 \\ 7.0 & 0.60 & 4.5 & 11.0 \end{bmatrix}, \quad Y = \begin{bmatrix} 181 & 167 \\ 241 & 206 \\ 161 & 172 \\ 230 & 198 \\ 154 & 157 \\ 231 & 209 \\ 154 & 145 \\ 203 & 220 \\ 216 & 185 \\ 135 & 152 \\ 257 & 223 \\ 185 & 208 \\ 102 & 131 \\ 156 & 155 \\ 204 & 182 \\ 141 & 164 \end{bmatrix}, \tag{117}$$

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

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

Assume now that we are only interested in a god 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 an 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 PE compared to PLS and PCR. This is also quite expected since TTLS are minimizing an objective function $\|X - Z\|_F^2 + \|Y - ZB_{\mathrm{TTLS}}\|_F^2$, 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.

Table 1
Comparison of the multivariate regression method CPLS against PLS, SIMPLS (see Section 6.3) and PCR[a]

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

[a] 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.

Table 2
Comparison of the univariate regression methods PLS, PCR and TSVD[a]

a	PLS	PCR	TTLS
1	79.13	79.14	84.18
2	74.44	78.83	220.2
3	66.42	71.02	124.6
4	64.43	64.51	137.7
5	57.31	57.31	124.7

[a] 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.

Example 7.3 (Real world data from a pulp and paper mill II). The variables tensile, y_1, and tear, y_2, are important for describing the quality of the paper. These variables are usually measured in the laboratory. It is of interest to predict these variables from the X data measured from a Pulp Expert (PEX) online analysator. The (input) data measured by the PEX are the freeness, x_1, the fiber length distribution (x_2, x_3, x_4 and x_5) and the shive contents, x_6, of the pulp. The length distribution is classified according to the Bauer 30, 100, 200 and -200 fractions. The data (which are from Union Co, Skien, Norway) are ordered into X and Y matrices as presented in Appendix E. The refiner manipulable variables (earlier in the process) were perturbed in order to ensure sufficient variability in the X and Y data. Furthermore, when the length distribution is exactly measured we have a linear dependency, $x_2 + x_3 + x_4 + x_5 = 100$. It is also a common belief in the pulp and paper industry that the

Table 3
Comparison of the multivariate regression methods CPLS, PLS2, SIMPLS, PCR and PLS1 (for each output) on the identification data[a]

a	CPLS	PLS2	SIMPLS	PCR	PLS1
1	1.0038	1.0266	1.0266	1.0422	1.0029
2	0.9627	0.9923	0.9922	1.0098	0.9416
3	0.9381	0.9343	0.9355	0.9895	0.9268
4	0.9217	0.9236	0.9235	0.9708	0.9217
5	0.9216	0.9216	0.9216	0.9216	0.9216

[a]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.

Table 4
Comparison of the multivariate regression methods CPLS, PLS2, SIMPLS, PCR and PLS1 (for each output) on the validation data[a]

a	CPLS	PLS2	SIMPLS	PCR	PLS1
1	0.3837	0.3774	0.3774	0.3782	0.3874
2	0.3956	0.3963	0.3963	0.4014	0.3556
3	0.3395	0.3438	0.3444	0.3974	0.3427
4	0.3453	0.3465	0.3465	0.3681	0.3455
5	0.3457	0.3457	0.3457	0.3457	0.3457

[a]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.

freeness, x_1, can be described by the length distribution, the shive content and the flexibility of the fibers (not a measured variable). Hence, from this aprior knowledge, the effective rank of X is believed to be close to four. The data were both centered and scaled for unit variance prior to identification. Hence, the sample mean were first removed from the data. Then the columns in the centered data were divided by the Frobenius norm of the respective columns. The observations used for identification were taken from row number 5 to row number 34 in the data matrices, i.e. $N = 30$ observations. The rest were used for validation, i.e. 8 observations. The results (norm of the PEs) based on the identification data are presented in Table 3. We can see that CPLS is optimal compared to the other multivariate methods. PLS2 and SIMPLS gave almost similar results. PCR gave the largest PEs. However, the strategy by modeling each output at a time with PLS1 gave the smallest PEs on the identification data. The results from the validation are presented in Table 4. Successive use of PLS1 gave worse results than the multivariate CPLS method for prediction on the validation data (eccept for $a = 2$). All methods gave a minimum for $a = 3$ components and CPLS produced the model with the smallest PEs. However, the methods produced very simila(r models. We can conclude that PLS2 is not necessarily optimal for prediction on validation data. The tensile, y_1, were well described by the model. The tear, y_2, were also reasonable described. The resulting $a = 3$ component model is promising and inspires for more work on model validation and online implementation.

8. 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 $B_{PLS} = K_a p^*$ 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 optimal LS 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 $B_{PLS} = K_a(K_a^T X^T X K_a)^{-1} K_a^T X^T Y$. 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 for 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^TX is constructed from information in X only.

The usual algorithm for computing the PLS weighting matrix W_a presented in the literature is equivalent to computing an orthogonal basis matrix (with orthonormal columns) 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 $B_{PLS} = Q_a(Q_a^T X^T X Q_a)^{-1} Q_a^T X^T Y$.

A QR updating technique (one column at a time) can be used to compute the QR decomposition of K_a, 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 round off errors when computing powers of X^TX. 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 a counterexample. An optimal latent variable LS solution B_{CPLS} has been presented in the paper. This optimal solution follows from an extension of the non-iterative Cayley–Hamilton series approach that we derived for the PLS1 algorithm to account for multivariate data. The optimality was illustrated by real world data from the pulp and paper industry.

Appendix A. Proof and implementation of Theorem 3.2

A procedure for updating the inverse of the matrix $W_i X^{\mathrm{T}} X W_i$ is needed in order to efficiently implement the PLS2 iteration algorithm in Theorem 3.2. Assume that the QR decomposition of the matrix XW_i, i.e. $T_i R_i = XW_i$, can be computed in parallel and in the same iteration loop as the weights are computed. Here, $T_i = [t_1 \ \dots \ t_i] \in \mathbb{R}^{N \times i}$ is orthogonal and $R_i \in \mathbb{R}^{i \times i}$ is upper triangular. Substituting this into (27) gives

$$(X^{\mathrm{T}} Y)_{i+1} = (X^{\mathrm{T}} Y)_i - X^{\mathrm{T}} T_i (W_i^{\mathrm{T}} X^{\mathrm{T}} T_i)^{-1} W_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_i, \tag{A.1}$$

where $W_i^{\mathrm{T}} X^{\mathrm{T}} T_i = R_i^{\mathrm{T}}$ is lower triangular. The weight vectors, $w_j \ \forall j = 1, \dots, i-1$, are normal to the residuals $(X^{\mathrm{T}} Y)_i$. This property follows by premultiplying (27) with W_i^{T} which gives $W_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_{i+1} = 0_{i \times m}$. This, gives that

$$W_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_i = \begin{bmatrix} 0_{i-1 \times m} \\ w_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_i \end{bmatrix}.$$

Hence, it is only the lower left element in R_i^{T} which is needed, and has to be inverted. This element is given by $r_{ii} = w_i^{\mathrm{T}} X^{\mathrm{T}} t_i$, and we have

$$(W_i^{\mathrm{T}} X^{\mathrm{T}} T_i)^{-1} W_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_i = \begin{bmatrix} 0_{i-1 \times m} \\ \dfrac{w_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_i}{w_i^{\mathrm{T}} X^{\mathrm{T}} t_i} \end{bmatrix}$$

and

$$T_i (W_i^{\mathrm{T}} X^{\mathrm{T}} T_i)^{-1} W_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_i = \frac{t_i w_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_i}{w_i^{\mathrm{T}} X^{\mathrm{T}} t_i}.$$

This gives the residual update equation

$$(X^{\mathrm{T}} Y)_{i+1} = (X^{\mathrm{T}} Y)_i - \frac{X^{\mathrm{T}} t_i w_i^{\mathrm{T}} (X^{\mathrm{T}} Y)_i}{w_i^{\mathrm{T}} X^{\mathrm{T}} t_i}. \tag{A.2}$$

The orthogonal (score) vector t_i can be computed by

$$t_i = X_i w_i, \qquad t_i := \frac{t_i}{(t_i^{\mathrm{T}} t_i)^{1/2}}, \tag{A.3}$$

where $X_1 = X$ and where X_{i+1} is computed by projecting the column space of X_i onto the orthogonal complement of the column space of t_i (Gram–Schmidt orthogonalization), i.e.

$$X_{i+1} = X_i - \frac{t_i t_i^{\mathrm{T}}}{t_i^{\mathrm{T}} t_i} X_i. \tag{A.4}$$

The definition (A.3) ensures that t_i is normalized to give $t_i^{\mathrm{T}} t_i = 1$. However, (A.2) shows that the update equation is independent of score vector, t_i, and weight vector, w_i, scalings. The update Eq. (A.2) (with (A.3) and (A.4)) is equivalent to (27).

From the rank one reduction (deflation) process in (31) we have

$$X_{i+1}^{\mathrm{T}} Y = X_i^{\mathrm{T}} Y - \frac{X_i^{\mathrm{T}} X_i w_i w_i^{\mathrm{T}} X_i^{\mathrm{T}} Y}{w_i^{\mathrm{T}} X_i^{\mathrm{T}} X_i w_i}. \tag{A.5}$$

Using (A.4) shows that $X_i^{\mathrm{T}} X_i = X^{\mathrm{T}} X_i$ because $I_N - t_i t_i^{\mathrm{T}} / t_i^{\mathrm{T}} t_i$ is a projection matrix. Hence, (A.5) is equivalent to the above formulation (A.3) of the iterations in Theorem 3.2. □

Appendix B. Proof and implementation of Theorem 3.1

The proof is divided into three parts.

Part 1 (*Equivalence condition*). In Helland (1988) it is proved that the columns in the weighting matrix used by the PLS1 algorithm (see, e.g. Wold, 1985; Næs and Martens, 1985) span the same space as the Krylov sequence $\{X^{\mathrm{T}} Y, X^{\mathrm{T}} X X^{\mathrm{T}} Y, \dots, (X^{\mathrm{T}} X)^{a-1} X^{\mathrm{T}} Y\}$.

Part 2 (*Subspace spanned by* W_a). In Appendix C it is proved that the columns in the weighting matrix W_a as defined in Theorem 3.1 span the same space as the Krylov sequence $\{X^{\mathrm{T}} Y, X^{\mathrm{T}} X X^{\mathrm{T}} Y, \dots, (X^{\mathrm{T}} X)^{a-1} X^{\mathrm{T}} Y\}$.

Part 3: Since the columns in the weighting matrix W_a provided by Theorem 3.1 (as in Part 2) span the same space as the columns in the weighting matrix used by the PLS1 algorithm in Helland (1988) (as in Part 1) the theorem is proved. □

An alternative to (20) in Theorem 3.1 can be derived as follows. The matrix $W_i^{\mathrm{T}} X^{\mathrm{T}} X W_i$ in (20) is tridiagonal since W_i is a basis for the Krylov matrix K_i. Let $T_i R_i$ be the QR decomposition of XW_i. Then, $W_i^{\mathrm{T}} X^{\mathrm{T}} T_i$ is lower bidiagonal. Following the lines in Appendix A we have that $T_i (W_i^{\mathrm{T}} X^{\mathrm{T}} T_i)^{-1} W_i^{\mathrm{T}} w_i = t_i w_i^{\mathrm{T}} w_i / w_i^{\mathrm{T}} X^{\mathrm{T}} t_i$. Hence, the update Eq. (20) is equivalent to

$$w_{i+1} = w_i - \frac{X^{\mathrm{T}} t_i w_i^{\mathrm{T}}}{w_i^{\mathrm{T}} X^{\mathrm{T}} t_i} w_i, \tag{B.1}$$

where the (score) vectors, t_i, is defined by (A.3) and (A.4).

Appendix C. Proof of Proposition 3.2

We want to prove that $W_a = K_a R_1^{-1}$ where R_1^{-1} is upper triangular. From Theorem 3.1 we have that

$$w_1 = X^{\mathrm{T}} Y \tag{C.1}$$

$$w_{i+1} = w_i - X^{\mathrm{T}} X W_i c_i \in \; i = 1, \ldots, a-1 \tag{C.2}$$

where it is important to note that

$$c_i = (W_i^{\mathrm{T}} X^{\mathrm{T}} X W_i)^{-1} W_i^{\mathrm{T}} w_i \in \mathbb{R}^a \tag{C.3}$$

is a vector. This implies directly that w_{i+1} is a linear combination of the sequence w_i, $X^{\mathrm{T}} X w_1$, $X^{\mathrm{T}} X w_2$, $\ldots, X^{\mathrm{T}} X w_i$.

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

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

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

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

$i = 1$

$$w_2 = w_1 - c_1 X^{\mathrm{T}} X w_1 \quad \text{where } c_1 = \frac{w_1^{\mathrm{T}} w_1}{w_1^{\mathrm{T}} X^{\mathrm{T}} X w_1}, \tag{C.4}$$

which is a linear combination of $X^{\mathrm{T}} Y$ and $X^{\mathrm{T}} X X^{\mathrm{T}} Y$.

$i = 2$

$$w_3 = w_2 - X^{\mathrm{T}} X \overbrace{[w_1, w_2]}^{W_2} \overbrace{\begin{bmatrix} c_{21} \\ c_{22} \end{bmatrix}}^{c_2}, \tag{C.5}$$

where

$$c_2 = (W_2^{\mathrm{T}} X^{\mathrm{T}} X W_2)^{-1} W_2^{\mathrm{T}} w_2, \tag{C.6}$$

which can be written as

$$w_3 = \overbrace{[w_1 \quad X^{\mathrm{T}} X w_1 \quad (X^{\mathrm{T}} X)^2 w_1]}^{K_3} \times \begin{bmatrix} 1 \\ -(c_1 + c_{21} + c_{22}) \\ c_1 c_{22} \end{bmatrix}. \tag{C.7}$$

Hence,

$$\overbrace{[w_1 \quad w_2 \quad w_3]}^{W_3} = \overbrace{[w_1 \quad X^{\mathrm{T}} X w_1 \quad (X^{\mathrm{T}} X)^2 w_1]}^{K_3} \times \overbrace{\begin{bmatrix} 1 & 1 & 1 \\ 0 & -c_1 & -(c_1 + c_{21} + c_{22}) \\ 0 & 0 & c_1 c_{22} \end{bmatrix}}^{R_1^{-1}} \tag{C.8}$$

and the proof is complete. □

Appendix D. Proof of Theorem 4.1

The expression for the PE, Eq. (63), gives

$$\mathrm{vec}(\mathscr{E}) = \mathrm{vec}(Y) - (I_m \otimes X)\mathrm{vec}(K_a(p)), \tag{D.1}$$

where we have used that $\mathrm{vec}(AXB) = (B^{\mathrm{T}} \otimes A)\mathrm{vec}(X)$ 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, Eq. (D.1) can be written as

$$\mathrm{vec}(\mathscr{E}) = \mathrm{vec}(Y) - (I_m \otimes X)\mathrm{bcs}(K_a)p, \tag{D.2}$$

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

$$\mathrm{bcs}(K_a) = [\mathrm{vec}(X^{\mathrm{T}} Y) \quad \mathrm{vec}(X^{\mathrm{T}} X X^{\mathrm{T}} Y) \quad \ldots \quad \mathrm{vec}((X^{\mathrm{T}} X)^{a-1} X^{\mathrm{T}} Y)] \in \mathbb{R}^{rm \times a} \tag{D.3}$$

as a block column string operator. Eq. (D.2) can be solved for p in a LS optimal sense by minimizing $V(p) = \|\mathrm{vec}(\mathscr{E})\|_{\mathrm{F}}^2$ with respect to p. This gives the optimal parameter vector

$$p^* = M^{\dagger} \; \mathrm{vec}(Y), \tag{D.4}$$

where we have defined

$$M = (I_m \otimes X)\mathrm{bcs}(K_a) \in \mathbb{R}^{Nm \times a} \tag{D.5}$$

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

Appendix E. Data for Example 7.3

$$X = \begin{bmatrix} 167.00 & 37.90 & 26.90 & 5.80 & 29.40 & 1.62 \\ 206.00 & 38.60 & 27.30 & 5.50 & 28.60 & 1.48 \\ 172.00 & 37.80 & 27.80 & 5.80 & 28.60 & 1.41 \\ 198.00 & 40.80 & 28.00 & 5.70 & 25.50 & 1.68 \\ 157.00 & 38.60 & 27.50 & 6.00 & 27.90 & 1.20 \\ 209.00 & 39.60 & 27.50 & 5.70 & 27.20 & 1.68 \\ 145.00 & 37.60 & 27.70 & 6.10 & 28.60 & 1.27 \\ 220.00 & 41.00 & 27.60 & 5.60 & 25.80 & 1.87 \\ 185.00 & 39.30 & 27.70 & 5.80 & 27.20 & 1.42 \\ 152.00 & 38.50 & 27.90 & 6.20 & 27.40 & 1.55 \\ 223.00 & 37.80 & 27.30 & 5.90 & 29.00 & 2.06 \\ 208.00 & 39.40 & 28.50 & 6.40 & 25.70 & 1.72 \\ 131.00 & 36.70 & 27.40 & 6.40 & 29.50 & 1.40 \\ 155.00 & 36.70 & 27.10 & 6.50 & 29.70 & 1.31 \\ 182.00 & 36.70 & 25.80 & 5.80 & 31.70 & 1.32 \\ 164.00 & 38.50 & 26.50 & 6.30 & 28.70 & 1.41 \\ 171.00 & 36.60 & 29.80 & 6.50 & 27.10 & 0.94 \\ 177.00 & 35.80 & 29.70 & 6.10 & 28.40 & 1.16 \\ 123.00 & 33.20 & 29.90 & 6.70 & 30.20 & 0.69 \\ 119.00 & 34.50 & 29.60 & 6.80 & 29.10 & 1.17 \\ 140.00 & 32.90 & 28.10 & 6.40 & 32.60 & 0.90 \\ 166.00 & 38.00 & 28.50 & 6.10 & 27.40 & 1.36 \\ 144.00 & 38.10 & 27.00 & 6.20 & 28.70 & 0.98 \\ 194.00 & 38.50 & 26.90 & 6.00 & 28.60 & 1.37 \\ 132.00 & 36.20 & 27.30 & 6.30 & 30.20 & 1.18 \\ 171.00 & 38.10 & 27.40 & 6.50 & 28.00 & 1.26 \\ 139.00 & 36.30 & 27.40 & 6.50 & 29.80 & 0.92 \\ 173.00 & 37.80 & 28.40 & 6.70 & 27.10 & 1.25 \\ 131.00 & 36.80 & 28.40 & 6.80 & 28.00 & 0.96 \\ 170.00 & 38.20 & 28.30 & 6.30 & 27.20 & 1.36 \\ 188.00 & 38.80 & 28.00 & 6.60 & 26.60 & 1.18 \\ 151.00 & 36.30 & 28.20 & 6.70 & 28.80 & 1.06 \\ 201.00 & 39.70 & 28.90 & 6.60 & 24.80 & 1.32 \\ 166.00 & 37.60 & 28.60 & 6.50 & 27.30 & 1.18 \\ 133.00 & 36.20 & 28.00 & 6.90 & 28.90 & 0.84 \\ 166.00 & 36.10 & 28.50 & 6.60 & 28.80 & 1.08 \\ 122.00 & 35.50 & 29.90 & 7.10 & 27.50 & 0.77 \\ 133.00 & 34.50 & 29.20 & 6.80 & 29.50 & 0.72 \end{bmatrix}, \quad Y = \begin{bmatrix} 34.10 & 7.60 \\ 33.00 & 7.37 \\ 35.10 & 7.57 \\ 33.40 & 7.69 \\ 38.00 & 7.64 \\ 31.70 & 7.73 \\ 39.00 & 7.37 \\ 29.50 & 7.35 \\ 32.60 & 7.52 \\ 35.10 & 7.33 \\ 29.30 & 7.49 \\ 32.30 & 7.25 \\ 35.10 & 7.81 \\ 34.00 & 7.47 \\ 32.60 & 7.68 \\ 33.70 & 7.36 \\ 34.60 & 7.12 \\ 35.30 & 7.15 \\ 39.40 & 6.77 \\ 38.60 & 6.88 \\ 38.70 & 7.11 \\ 37.60 & 7.47 \\ 37.50 & 7.07 \\ 32.60 & 7.24 \\ 38.30 & 7.16 \\ 35.90 & 6.62 \\ 38.70 & 7.21 \\ 35.10 & 7.35 \\ 39.70 & 7.23 \\ 37.30 & 7.27 \\ 31.20 & 6.97 \\ 36.20 & 7.86 \\ 31.40 & 7.31 \\ 34.60 & 7.85 \\ 39.90 & 7.19 \\ 34.00 & 7.37 \\ 40.00 & 7.16 \\ 37.90 & 7.52 \end{bmatrix}.$$

References

Burnham, A. J., Viveros, R., & MacGregor, J. F. (1996). Framevorks for latent variable multivariate regression. *Journal of Chemometrics*, *10*, 31–45.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*, 251–263.

de Jong, S., & Phatak, A. (1997). Partial least squares regression. In: Van Huffel (Ed.), *Recent advances in total least squares techniques and errors-in-variables modeling* (pp. 25–36). SIAM Proceedings series, Philadelphia: SIAM.

de Jong, S., & ter Braak, C. J. F. (1994). Comments on the PLS kernal algorithm. *Journal of Chemometrics*, *8*, 169–174.

De Moor, B., & 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., & O'Leary, D. P. (1997). Regularization by truncated total least squares. *SIAM Journal on Scientific Computing*, *18*(4), 1223–1241.

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135.

Golub, G. H., & Van Loan, C. F. (1986). *Matrix computations.* London: North Oxford Academic.

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. *Communications in Statistics, Simulation and Computation*, *17*(2), 581–607.

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

Höskuldsson, A. (1996). *Prediction methods in science and technology.* COLOURSCAN Warsaw, ISBN 87-985941-0-9.

Johansen, T. A. (1997). On Tikhonov regularization, bias and variance in nonlinear system identification. *Automatica*, *33*(3), 441–446.

Lindgren, F., Geladi, P., & Wold, S. (1993). The kernal algorithm for PLS. *Journal of Chemometrics*, *7*, 45–59.

Lorber, A., Lawrence, E. W., & Kowalski, B. R. (1987). A theoretical foundation for the PLS algorithm. *Journal of Chemometrics*, *1*, 19–31.

Manne, R. (1987). Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, *2*, 187–197.

Martens, H., & Næs, T. (1989). *Multivariate calibration*, New York: Wiley.

Næs, T., & Martens, H. (1985). Comparison of prediction methods for multicolinear data. *Communications in Statistics Simulation and Computation*, *14*(3), 544–576.

Paige, C., & Saunders, M. (1982). A bidiagonalization algorithm for sparse linear equations and least squares problems. *ACM Transactions on Mathematical Software*, *8*, 43–71.

Parlett, B. N. (1998). *The symmetric eigenvalue problem.* Philadelphia, PA: SIAM.

Phatak, A. (1993). *Evaluation of some multivariate methods and their applications in chemical engineering*. Ph.D. thesis, University of Waterloo.

Phatak, A., & de Jong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics*, *11*, 311–338.

Söderström, T., & Stoica, P. (1989). *System identification.* Englewood Cliffs, NJ: Prentice-Hall.

ter Braak, C. J., & de Jong, S. (1998). The objective function of partial least squares. *Journal of Chemometrics*, *12*, 41–54.

Van Huffel, S., & Vandewalle, J. (1991). *The total least squares problem*: *Computational aspects and analysis*. Philadelphia: SIAM.

Vetter, W. J. (1973). Matrix calculus operations and Taylor expansions. *SIAM Review*, *15*(2), 352–369.

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

Wold, H. (1975). Soft modeling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In: M.S. Bartkett, & J. Gani (Eds.), *Perspectives in probability and statistics*, New York: Academic Press.

Wold, H. (1985). Partial least squares. In: S. Kotz, & N. L. Johnson (Eds.), *Encyclopedia of statistics sciences*, vol. 6 (pp. 581–591). New York: Wiley.

Wold, H., Ruhe, A., Wold, H., & Dunn, W. (1984). The collinearity problem in regression. The PLS approach to generalized inverses. *SIAM Journal of Science Statistics and Computers*, *5*, 735–743.

David Di Ruscio was born in Kongsvinger, Norway in 1962. He received the M.Sc. degree in Electrical and Computer Engineering from the Norwegian University of Science and Technology (NTNU), Trondheim in 1986, and the Ph.D. degree in Engineering Cybernetics from NTNU in 1993. In the period 1986–1989 Di Ruscio was a research assistant at the Department of Engineering Cybernetics. He positioned a post doctoral scholarship from Norske Skog (a pulp and paper industry in Norway), at the same department, in the period 1993–1995. Since 1995 Di Ruscio has been an Associate professor in Process Control at Telemark Institute of Technology, Porsgrunn where he is teaching advanced process control and system identification. His research interests include system identification, subspace methods, numerical methods, process and optimal control, and application to industry.