

# Chapter 2

## On the PLS Algorithm for Multiple Regression (PLS1)

Yoshio Takane and Sébastien Loisel

**Abstract** Partial least squares (PLS) was first introduced by Wold in the mid 1960s as a heuristic algorithm to solve linear least squares (LS) problems. No optimality property of the algorithm was known then. Since then, however, a number of interesting properties have been established about the PLS algorithm for regression analysis (called PLS1). This paper shows that the PLS estimator for a specific dimensionality  $S$  is a kind of constrained LS estimator confined to a Krylov subspace of dimensionality  $S$ . Links to the Lanczos bidiagonalization and conjugate gradient methods are also discussed from a somewhat different perspective from previous authors.

**Keywords** Krylov subspace • NIPALS • PLS1 algorithm • Lanczos bidiagonalization • Conjugate gradients • Constrained principal component analysis (CPCA)

### 2.1 Introduction

Partial least squares (PLS) was first introduced by Wold (1966) as a heuristic algorithm for estimating parameters in multiple regression. Since then, it has been elaborated in many directions, including extensions to multivariate cases (Abdi 2007; de Jong 1993) and structural equation modeling (Lohmöller 1989; Wold 1982). In this paper, we focus on the original PLS algorithm for univariate regression (called PLS1), and show its optimality given the subspace in which the vector of regression coefficients is supposed to lie. Links to state-of-the-art algorithms for solving a system of linear simultaneous equations, such as the Lanczos bidiagonalization and the conjugate gradient methods, are also discussed

---

Y. Takane (✉)  
University of Victoria, Victoria, BC, Canada  
e-mail: [Yoshio.Takane@mcgill.ca](mailto:Yoshio.Takane@mcgill.ca)

S. Loisel  
Heriot-Watt University, Edinburgh, UK  
e-mail: [sloisel@gmail.com](mailto:sloisel@gmail.com)

from a somewhat different perspective from previous authors (Eldén 2004; Phatak and de Hoog 2002). We refer the reader to Rosipal and Krämer (2006) for more comprehensive accounts and reviews of new developments of PLS.

## 2.2 PLS1 as Constrained Least Squares Estimator

Consider a linear regression model

$$\mathbf{z} = \mathbf{G}\mathbf{b} + \mathbf{e}, \quad (2.1)$$

where  $\mathbf{z}$  is the  $N$ -component vector of observations on the criterion variable,  $\mathbf{G}$  is the  $N \times P$  matrix of predictor variables,  $\mathbf{b}$  is the  $P$ -component vector of regression coefficients, and  $\mathbf{e}$  is the  $N$ -component vector of disturbance terms. The ordinary LS (OLS) criterion is often used to estimate  $\mathbf{b}$  under the *iid* (independent and identically distributed) normal assumption on  $\mathbf{e}$ . This is a reasonable practice if  $N$  is large compared to  $P$ , and columns of  $\mathbf{G}$  are not highly collinear (i.e., as long as the matrix  $\mathbf{G}'\mathbf{G}$  is well-conditioned). However, if this condition is not satisfied, the use of OLS estimators (OLSE) is not recommended, because then these estimators tend to have large variances. Principal component regression (PCR) is often employed in such situations. In PCR, principal component analysis (PCA) is first applied to  $\mathbf{G}$  to find a low rank (say, rank  $S$ ) approximation, which is subsequently used as the set of new predictor variables in a linear regression analysis. One potential problem with PCR is that the low rank approximation of  $\mathbf{G}$  best accounts for  $\mathbf{G}$  but is not necessarily optimal for predicting  $\mathbf{z}$ . By contrast, PLS extracts components of  $\mathbf{G}$  that are good predictors of  $\mathbf{z}$ . For the case of univariate regression, the PLS algorithm (called PLS1) proceeds as follows:

### PLS1 Algorithm

Step 1. Column-wise center  $\mathbf{G}$  and  $\mathbf{z}$ , and set  $\mathbf{G}_0 = \mathbf{G}$ .

Step 2. Repeat the following substeps for  $i = 1, \dots, S$  ( $S \leq \text{rank}(\mathbf{G})$ ):

Step 2.1. Set  $\mathbf{w}_i = \mathbf{G}'_{i-1}\mathbf{z} / \|\mathbf{G}'_{i-1}\mathbf{z}\|$ , where  $\|\mathbf{G}'_{i-1}\mathbf{z}\| = (\mathbf{z}'\mathbf{G}_{i-1}\mathbf{G}'_{i-1}\mathbf{z})^{1/2}$ .

Step 2.2. Set  $\mathbf{t}_i = \mathbf{G}_{i-1}\mathbf{w}_i / \|\mathbf{G}_{i-1}\mathbf{w}_i\|$ .

Step 2.3. Set  $\mathbf{v}_i = \mathbf{G}'_{i-1}\mathbf{t}_i$ .

Step 2.4. Set  $\mathbf{G}_i = \mathbf{G}_{i-1} - \mathbf{t}_i\mathbf{v}'_i = \mathbf{Q}_{\mathbf{G}_{i-1}\mathbf{w}_i}\mathbf{G}_{i-1}$  (deflation),

where  $\mathbf{Q}_{\mathbf{G}_{i-1}\mathbf{w}_i} = \mathbf{I} - \mathbf{G}_{i-1}\mathbf{w}_i(\mathbf{w}'_i\mathbf{G}'_{i-1}\mathbf{G}_{i-1}\mathbf{w}_i)^{-1}\mathbf{w}'_i\mathbf{G}'_{i-1}$ , and where  $'$  denotes the transpose operation, and  $\|\cdot\|$  denotes the  $L_2$  norm of a vector (i.e.,  $\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}}$ , see, e.g., Takane (2014), for details); vectors  $\mathbf{w}_i$ ,  $\mathbf{t}_i$ , and  $\mathbf{v}_i$  are called (respectively) weights, scores, and loadings, and are collected in matrices  $\mathbf{W}_S$ ,  $\mathbf{T}_S$ , and  $\mathbf{V}_S$ . For a given  $S$ , the PLS estimator (PLSE) of  $\mathbf{b}$  is given by

$$\hat{\mathbf{b}}_{PLSE}^{(S)} = \mathbf{W}_S(\mathbf{V}'_S\mathbf{W}_S)^{-1}\mathbf{T}'_S\mathbf{z} \quad (2.2)$$

(see, e.g., Abdi 2007). The algorithm above assumes that  $S$  is known and, actually, the choice of its value is crucial for good performance of PLSE (a cross validation method is often used to choose the best value of  $S$ ). It has been demonstrated (Phatak and de Hoog 2002) that for a given value of  $S$ , the PLSE of  $\mathbf{b}$  has better predictability than the corresponding PCR estimator.

The PLSE of  $\mathbf{b}$  can be regarded as a special kind of constrained LS estimator (CLSE), in which  $\mathbf{b}$  is constrained to lie in the Krylov subspace of dimensionality  $S$  defined by

$$\mathcal{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z}) = \text{Sp}(\mathbf{K}_S), \quad (2.3)$$

where  $\text{Sp}(\mathbf{K}_S)$  is the space spanned by the column vectors of  $\mathbf{K}_S$ , and

$$\mathbf{K}_S = [\mathbf{G}'\mathbf{z}, (\mathbf{G}'\mathbf{G})\mathbf{G}'\mathbf{z}, \dots, (\mathbf{G}'\mathbf{G})^{S-1}\mathbf{G}'\mathbf{z}] \quad (2.4)$$

is called the Krylov matrix of order  $S$ . Because  $\text{Sp}(\mathbf{W}_S) = \mathcal{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$  (see Eldén 2004, proposition 3.1; Phatak and de Hoog 2002)  $\mathbf{b}$  can be re-parameterized as  $\mathbf{b} = \mathbf{W}_S\mathbf{a}$  for some  $\mathbf{a}$ . Then Eq. (2.1) can be rewritten as

$$\mathbf{z} = \mathbf{G}\mathbf{W}_S\mathbf{a} + \mathbf{e}. \quad (2.5)$$

The OLSE of  $\mathbf{a}$  is given by

$$\hat{\mathbf{a}} = (\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{G}'\mathbf{z}, \quad (2.6)$$

from which the CLSE of  $\mathbf{b}$  is found as

$$\hat{\mathbf{b}}_{CLSE}^{(S)} = \mathbf{W}_S\hat{\mathbf{a}} = \mathbf{W}_S(\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{G}'\mathbf{z}. \quad (2.7)$$

To show that (2.7) is indeed equivalent to (2.2), we need several well-known results in the PLS literature (Bro and Eldén 2009; de Jong 1993; Eldén 2004; Phatak and de Hoog 2002). First of all,  $\mathbf{W}_S$  is column-wise orthogonal, that is,

$$\mathbf{W}_S'\mathbf{W}_S = \mathbf{I}_S. \quad (2.8)$$

Secondly,  $\mathbf{T}_S$  is also column-wise orthogonal,

$$\mathbf{T}_S'\mathbf{T}_S = \mathbf{I}_S, \quad (2.9)$$

and

$$\mathbf{T}_S\mathbf{L}_S = \mathbf{G}\mathbf{W}_S, \quad (2.10)$$

where  $\mathbf{L}_S$  is an upper bidiagonal matrix. Relations (2.8), (2.9) and (2.10) imply that

$$\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S = \mathbf{L}_S'\mathbf{L}_S = \mathbf{H}_S, \quad (2.11)$$

where  $\mathbf{H}_S$  is tridiagonal. Thirdly,

$$\mathbf{V}'_S = \mathbf{T}'_S \mathbf{G}, \quad (2.12)$$

so that

$$\mathbf{L}_S = \mathbf{T}'_S \mathbf{G} \mathbf{W}_S = \mathbf{V}'_S \mathbf{W}_S. \quad (2.13)$$

Now it is straightforward to show that

$$\begin{aligned} \hat{\mathbf{b}}_{CLSE}^{(S)} &= \mathbf{W}_S (\mathbf{W}'_S \mathbf{G}' \mathbf{G} \mathbf{W}_S)^{-1} \mathbf{W}'_S \mathbf{G}' \mathbf{z} \\ &= \mathbf{W}_S \mathbf{H}_S^{-1} \mathbf{L}'_S \mathbf{T}'_S \mathbf{z} \\ &= \mathbf{W}_S (\mathbf{L}'_S \mathbf{L}_S)^{-1} \mathbf{L}'_S \mathbf{T}'_S \mathbf{z} \\ &= \mathbf{W}_S \mathbf{L}_S^{-1} \mathbf{T}'_S \mathbf{z} \\ &= \mathbf{W}_S (\mathbf{V}'_S \mathbf{W}_S)^{-1} \mathbf{T}'_S \mathbf{z} \\ &= \hat{\mathbf{b}}_{PLSE}^{(S)}, \end{aligned} \quad (2.14)$$

and this establishes the equivalence between Eqs. (2.7) and (2.2).

The PLSE of regression parameters reduces to the OLSE if  $S = \text{rank}(\mathbf{G})$  (when  $\text{rank}(\mathbf{G}) < P$ , we use  $\mathbf{G}^+$ , which is the Moore-Penrose inverse of  $\mathbf{G}$ , in lieu of  $(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}$  in the OLSE for regression coefficients).

### 2.3 Relations to the Lanczos Bidiagonalization Method

It has been pointed out (Eldén 2004) that PLS1 described above is equivalent to the following Lanczos bidiagonalization algorithm:

#### The Lanczos Bidiagonalization (LBD) Algorithm

- Step 1. Column-wise center  $\mathbf{G}$ , and compute  $\mathbf{u}_1 = \mathbf{G}'\mathbf{z}/\|\mathbf{G}'\mathbf{z}\|$  and  $\mathbf{q}_1 = \mathbf{G}\mathbf{u}_1/\delta_1$ , where  $\delta_1 = \|\mathbf{G}\mathbf{u}_1\|$ .  
 Step 2. For  $i = 2, \dots, S$  (this is the same  $S$  as in PLS1),

- (a) Compute  $\gamma_{i-1}\mathbf{u}_i = \mathbf{G}'\mathbf{q}_{i-1} - \delta_{i-1}\mathbf{u}_{i-1}$ .
- (b) Compute  $\delta_i\mathbf{q}_i = \mathbf{G}\mathbf{u}_i - \gamma_{i-1}\mathbf{q}_{i-1}$ .

Scalars  $\gamma_{i-1}$  and  $\delta_i$  ( $i = 2, \dots, S$ ) are the normalization factors to make  $\|\mathbf{u}_i\| = 1$  and  $\|\mathbf{q}_{i-1}\| = 1$ , respectively.

Let  $\mathbf{U}_S$  and  $\mathbf{Q}_S$  represent the collections of  $\mathbf{u}_i$  and  $\mathbf{q}_i$  for  $i = 1, \dots, S$ . It has been shown (Eldén 2004, Proposition 3.1) that these two matrices are essentially the same as  $\mathbf{W}_S$  and  $\mathbf{T}_S$ , respectively, obtained in PLS1. Here “essentially” means that these

two matrices are identical to  $\mathbf{W}_S$  and  $\mathbf{T}_S$  except that the even columns of  $\mathbf{U}_S$  and  $\mathbf{Q}_S$  are reflected (i.e., have their sign reversed). We show this explicitly for  $\mathbf{u}_2$  and  $\mathbf{q}_2$  (i.e.,  $\mathbf{u}_2 = -\mathbf{w}_2$  and  $\mathbf{q}_2 = -\mathbf{t}_2$ ). It is obvious from Step 1 of the two algorithms that

$$\mathbf{w}_1 = \mathbf{u}_1 \text{ and } \mathbf{t}_1 = \mathbf{q}_1. \quad (2.15)$$

Let  $\alpha_1 = \|\mathbf{G}'\mathbf{z}\|$ . Then

$$\begin{aligned} \mathbf{w}_2 &\propto \mathbf{G}'\mathbf{Q}_{G\mathbf{w}_1}\mathbf{z} \quad (\text{from Step 2.4 of the PLS1 algorithm}) \\ &= \mathbf{G}'\mathbf{z} - \mathbf{G}'\mathbf{G}\mathbf{w}_1(\mathbf{w}_1'\mathbf{G}'\mathbf{G}\mathbf{w}_1)^{-1}\mathbf{w}_1'\mathbf{G}'\mathbf{z} \\ &= \alpha_1(\mathbf{w}_1 - \mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1^2) \end{aligned} \quad (2.16)$$

$$\propto -\mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1 + \delta_1\mathbf{w}_1, \quad (2.17)$$

where  $\propto$  means “proportional.” To obtain the last expression, we multiplied Eq. (2.16) by  $\delta_1/\alpha_1$  ( $> 0$ ). This last expression is proportional to  $-\mathbf{u}_2$ , where  $\mathbf{u}_2 \propto \mathbf{G}'\mathbf{G}\mathbf{u}_1/\delta_1 - \delta_1\mathbf{u}_1$  from Step 2(a) of the Lanczos algorithm. This implies  $\mathbf{u}_2 = -\mathbf{w}_2$ , because both  $\mathbf{u}_2$  and  $\mathbf{w}_2$  are normalized.

Similarly, define  $\beta_1^2 = \mathbf{w}_1'(\mathbf{G}'\mathbf{G})^2\mathbf{w}_1$ . Then

$$\begin{aligned} \mathbf{t}_2 &\propto \mathbf{Q}_{G\mathbf{w}_1}\mathbf{G}\mathbf{G}'\mathbf{Q}_{G\mathbf{w}_1}\mathbf{z} \quad (\text{from Step 2.2 of the PLS1 algorithm}) \\ &= \alpha_1(\mathbf{G}\mathbf{w}_1 - \mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1^2 - \mathbf{G}\mathbf{w}_1 + \frac{\beta_1^2}{\delta_1^4}\mathbf{G}\mathbf{w}_1) \end{aligned} \quad (2.18)$$

$$\propto -\mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{w}_1 + \frac{\beta_1^2}{\delta_1^2}\mathbf{G}\mathbf{w}_1. \quad (2.19)$$

To obtain Eq. (2.19), we multiplied (2.18) by  $\delta_1^2/\alpha_1$  ( $> 0$ ). On the other hand, we have

$$\begin{aligned} \mathbf{q}_2 &\propto \frac{1}{\delta_1\gamma_1}(\mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{u}_1 - \delta_1^2\mathbf{G}\mathbf{u}_1 - \gamma_1^2\mathbf{G}\mathbf{u}_1) \quad (\text{from Step 2(b) of the Lanczos algorithm}) \\ &\propto \mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{u}_1 - (\delta_1^2 + \gamma_1^2)\mathbf{G}\mathbf{u}_1. \end{aligned} \quad (2.20)$$

To show that  $\mathbf{q}_2 \propto -\mathbf{t}_2$ , it remains to show that

$$\gamma^2 + \delta^2 = \beta_1^2/\delta_1^2. \quad (2.21)$$

From Step 2(a) of the Lanczos algorithm,

$$\begin{aligned} \gamma^2 &= (\mathbf{G}'\mathbf{G}\mathbf{u}_1/\delta_1 - \delta_1\mathbf{u}_1)'(\mathbf{G}'\mathbf{G}\mathbf{u}_1/\delta_1 - \delta_1\mathbf{u}_1) \\ &= \beta^2/\delta^2 - \delta^2, \end{aligned} \quad (2.22)$$

and so indeed (2.21) holds. Again, we have  $\mathbf{q}_2 = -\mathbf{t}_2$ , because both  $\mathbf{q}_2$  and  $\mathbf{t}_2$  are normalized.

The sign reversals of  $\mathbf{u}_2$  and  $\mathbf{q}_2$  yield  $\mathbf{u}_3$  and  $\mathbf{q}_3$  identical to  $\mathbf{w}_3$  and  $\mathbf{t}_3$ , respectively, by similar sign reversals, and  $\mathbf{u}_4$  and  $\mathbf{q}_4$  which are sign reversals of  $\mathbf{w}_4$  and  $\mathbf{t}_4$ , and so on. Thus, only even columns of  $\mathbf{U}_S$  and  $\mathbf{Q}_S$  are affected (i.e., have their sign reversed) relative to the corresponding columns of  $\mathbf{W}_S$  and  $\mathbf{T}_S$ , respectively. Of course, these sign reversals have no effect on estimates of regression parameters. The estimate of regression parameters by the Lanczos bidiagonalization method is given by

$$\hat{\mathbf{b}}_{LBD}^{(S)} = \mathbf{U}_S (\mathbf{L}_S^*)^{-1} \mathbf{Q}_S' \mathbf{z}, \quad (2.23)$$

where

$$\mathbf{L}_S^* = \mathbf{Q}_S' \mathbf{G} \mathbf{U}_S, \quad (2.24)$$

which is upper bidiagonal, as is  $\mathbf{L}_S$  (defined in Eq. (2.13)). matrix  $\mathbf{L}_S^*$  differs from matrix  $\mathbf{L}_S$  only in the sign of its super-diagonal elements. The matrices  $\mathbf{L}_S^{-1}$  and  $(\mathbf{L}_S^*)^{-1}$  are also upper bidiagonal, for which the super-diagonal elements are opposite in sign, while their diagonal elements remain the same. Thus

$$\begin{aligned} \mathbf{W}_S \mathbf{L}_S^{-1} \mathbf{T}_S' &= \sum_{i=1}^S (\ell_{i,i} \mathbf{w}_i \mathbf{t}_i' + \ell_{i,i+1} \mathbf{w}_i \mathbf{t}_{i+1}') \\ &= \sum_{i=1}^S (\ell_{i,i}^* \mathbf{u}_i \mathbf{q}_i' + \ell_{i,i+1}^* \mathbf{u}_i \mathbf{q}_{i+1}') \\ &= \mathbf{U}_S (\mathbf{L}_S^*)^{-1} \mathbf{Q}_S', \end{aligned} \quad (2.25)$$

where  $\ell_{i,j}$  and  $\ell_{i,j}^*$  are the  $ij$ -th element of (respectively)  $\mathbf{L}_S$  and  $\mathbf{L}_S^*$ . Note that

$$\ell_{i,i} = \ell_{i,i}^*, \quad \mathbf{w}_i \mathbf{t}_i' = \mathbf{u}_i \mathbf{q}_i', \quad \ell_{i,i+1} = -\ell_{i,i+1}^*, \quad \text{and} \quad \mathbf{w}_i \mathbf{t}_{i+1}' = -\mathbf{u}_i \mathbf{q}_{i+1}' \quad (2.26)$$

It is widely known (see, e.g., Saad 2003) that the matrix of orthogonal basis vectors generated by the Arnoldi orthogonalization of  $\mathbf{K}_S$  (Arnoldi 1951) is identical to  $\mathbf{U}_S$  obtained in the Lanczos algorithm. Starting from  $\mathbf{u}_1 = \mathbf{G}'\mathbf{z}/\|\mathbf{G}'\mathbf{z}\|$ , this orthogonalization method finds  $\mathbf{u}_{i+1}$  ( $i = 1, \dots, S-1$ ) by successively orthogonalizing  $\mathbf{G}'\mathbf{G}\mathbf{u}_i$  ( $i = 1, \dots, S-1$ ) to all previous  $\mathbf{u}_i$ 's by a procedure similar to the Gram-Schmidt orthogonalization method. This yields  $\mathbf{U}_S$  such that  $\mathbf{G}'\mathbf{G}\mathbf{U}_S = \mathbf{U}_S \mathbf{H}_S^*$ , or

$$\mathbf{U}_S' \mathbf{G}' \mathbf{G} \mathbf{U}_S = \mathbf{L}_S^* \mathbf{L}_S^* = \mathbf{H}_S^*, \quad (2.27)$$

where  $\mathbf{H}_S^*$  is tridiagonal as is  $\mathbf{H}_S$  defined in Eq. (2.11). The diagonal elements of this matrix are identical to those of  $\mathbf{H}_S$  while its sub- and super-diagonal elements have their sign reversed. Matrix  $\mathbf{H}_S^*$  is called the Lanczos tridiagonal matrix and it is useful to obtain eigenvalues of  $\mathbf{G}'\mathbf{G}$ .

## 2.4 Relations to the Conjugate Gradient Method

It has been pointed out (Phatak and de Hoog 2002) that the conjugate gradient (CG) algorithm (Hestenes and Stiefel 1951) for solving a system of linear simultaneous equations  $\mathbf{G}'\mathbf{G}\mathbf{b} = \mathbf{G}'\mathbf{y}$  gives solutions identical to  $\hat{\mathbf{b}}_{PLSE}^{(s)}$  [ $s = 1, \dots, \text{rank}(\mathbf{G})$ ], if the CG iteration starts from the initial solution  $\hat{\mathbf{b}}_{CG}^{(0)} \equiv \mathbf{b}_0 = \mathbf{0}$ . To verify their assertion, we look into the CG algorithm stated as follows:

### The Conjugate Gradient (CG) Algorithm

Step 1. Initialize  $\mathbf{b}_0 = \mathbf{0}$ . Then,  $\mathbf{r}_0 = \mathbf{G}'\mathbf{z} - \mathbf{G}'\mathbf{G}\mathbf{b}_0 = \mathbf{G}'\mathbf{z} = \mathbf{d}_0$ . (Vectors  $\mathbf{r}_0$  and  $\mathbf{d}_0$  are called initial residual and initial direction vectors, respectively.)

Step 2. For  $i = 0, \dots, s-1$ , compute:

- (a)  $a_i = \mathbf{d}_i'\mathbf{r}_i / \mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i = \|\mathbf{r}_i\|^2 / \mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i$ .
- (b)  $\mathbf{b}_{i+1} = \mathbf{b}_i + a_i\mathbf{d}_i$ .
- (c)  $\mathbf{r}_{i+1} = \mathbf{G}'\mathbf{z} - \mathbf{G}'\mathbf{G}\mathbf{b}_{i+1} = \mathbf{r}_i - a_i\mathbf{G}'\mathbf{G}\mathbf{d}_i = \mathbf{Q}_{d_i/G'G}'\mathbf{r}_i$ , where  $\mathbf{Q}_{d_i/G'G} = \mathbf{I} - \mathbf{d}_i(\mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i)^{-1}\mathbf{d}_i'\mathbf{G}'\mathbf{G}$  is the projector onto the space orthogonal to  $\text{Sp}(\mathbf{G}'\mathbf{G}\mathbf{d}_i)$  along  $\text{Sp}(\mathbf{d}_i)$  [its transpose, on the other hand, is the projector onto the space orthogonal  $\text{Sp}(\mathbf{d}_i)$  along  $\text{Sp}(\mathbf{G}'\mathbf{G}\mathbf{d}_i)$ ].
- (d)  $b_i = -\mathbf{r}_{i+1}'\mathbf{G}'\mathbf{G}\mathbf{d}_i / \mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i = \|\mathbf{r}_{i+1}\|^2 / \|\mathbf{r}_i\|^2$ .
- (e)  $\mathbf{d}_{i+1} = \mathbf{r}_{i+1} + b_i\mathbf{d}_i = \mathbf{Q}_{d_i/G'G}\mathbf{r}_{i+1}$ .

Let  $\mathbf{R}_j = [\mathbf{r}_0, \dots, \mathbf{r}_{j-1}]$  and  $\mathbf{D}_j = [\mathbf{d}_0, \dots, \mathbf{d}_{j-1}]$  for  $j \leq S$ . We first show that

$$\text{Sp}(\mathbf{R}_j) = \text{Sp}(\mathbf{D}_j) = \mathcal{H}_j(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z}) \quad (2.28)$$

by induction, where, as before,  $\text{Sp}(\mathbf{A})$  indicates the space spanned by the column vectors of matrix  $\mathbf{A}$ . It is obvious that  $\mathbf{r}_0 = \mathbf{d}_0 = \mathbf{G}'\mathbf{z}$ , so that  $\text{Sp}(\mathbf{R}_1) = \text{Sp}(\mathbf{D}_1) = \mathcal{H}_1(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ . From Step 2(c) of the CG algorithm, we have

$$\mathbf{r}_1 = \mathbf{Q}_{d_0/G'G}'\mathbf{r}_0 = \mathbf{r}_0 - \mathbf{G}'\mathbf{G}\mathbf{d}_0c_0 \quad (2.29)$$

for some scalar  $c_0$ , so that  $\mathbf{r}_1 \in \mathcal{H}_2(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$  because  $\mathbf{G}'\mathbf{G}\mathbf{d}_0 \in \mathcal{H}_2(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ . From Step 2(e), we also have

$$\mathbf{d}_1 = \mathbf{Q}_{d_0/G'G}\mathbf{r}_1 = \mathbf{r}_1 - \mathbf{d}_0c_0^* \quad (2.30)$$

for some  $c_0^*$ , so that  $\mathbf{d}_1 \in \mathcal{H}_2(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ . This shows that  $\text{Sp}(\mathbf{R}_2) = \text{Sp}(\mathbf{D}_2) = \mathcal{H}_2(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ . Similarly, we have  $\mathbf{r}_2 \in \mathcal{H}_3(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$  and  $\mathbf{d}_2 \in \mathcal{H}_3(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ , so that  $\text{Sp}(\mathbf{R}_3) = \text{Sp}(\mathbf{D}_3) = \mathcal{H}_3(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ , and so on.

The property of  $\mathbf{D}_j$  above implies that  $\text{Sp}(\mathbf{W}_S)$  is identical to  $\text{Sp}(\mathbf{D}_S)$ , which in turn implies that

$$\hat{\mathbf{b}}_{CG}^{(S)} = \mathbf{D}_S(\mathbf{D}_S'\mathbf{G}\mathbf{G}\mathbf{D}_S)^{-1}\mathbf{D}_S'\mathbf{G}\mathbf{z} \quad (2.31)$$

is identical to  $\hat{\mathbf{b}}_{CLSE}^{(S)}$  as defined in Eq. (2.7), which in turn is equal to  $\hat{\mathbf{b}}_{PLSE}^{(S)}$  defined in Eq. (2.2) (Phatak and de Hoog 2002) by virtue of Eq. (2.14). It remains to show that  $\hat{\mathbf{b}}_{CG}^{(S)}$  defined in (2.31) coincides with  $\mathbf{b}_S$  generated by the CG algorithm. By the  $\mathbf{G}'\mathbf{G}$ -conjugacy of  $\mathbf{d}_j$ 's (the orthogonality of  $\mathbf{d}_j$ 's with respect to  $\mathbf{G}'\mathbf{G}$ , i.e.,  $\mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0$  for any  $i \neq j$ , as will be shown later), Eq. (2.31) can be rewritten as

$$\hat{\mathbf{b}}_{CG}^{(S)} = \sum_{i=0}^{S-1} \mathbf{d}_i'(\mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i)^{-1}\mathbf{d}_i'\mathbf{G}'\mathbf{z}. \quad (2.32)$$

From Step 2(b) of the CG algorithm, on the other hand, we have

$$\mathbf{b}_1 = \mathbf{d}_0(\mathbf{d}_0'\mathbf{G}'\mathbf{G}\mathbf{d}_0)^{-1}\mathbf{d}_0'\mathbf{r}_0 = \mathbf{d}_0(\mathbf{d}_0'\mathbf{G}'\mathbf{G}\mathbf{d}_0)^{-1}\mathbf{d}_0'\mathbf{G}\mathbf{z} = \hat{\mathbf{b}}_{CG}^{(1)}, \quad (2.33)$$

and

$$\begin{aligned} \mathbf{b}_3 &= \hat{\mathbf{b}}_{CG}^{(1)} + \mathbf{d}_1(\mathbf{d}_1'\mathbf{G}'\mathbf{G}\mathbf{d}_1)^{-1}\mathbf{d}_1'\mathbf{r}_1, \\ &= \hat{\mathbf{b}}_{CG}^{(1)} + \mathbf{d}_1(\mathbf{d}_1'\mathbf{G}'\mathbf{G}\mathbf{d}_1)^{-1}\mathbf{d}_1'\mathbf{G}'\mathbf{z} = \hat{\mathbf{b}}_{CG}^{(2)}, \end{aligned} \quad (2.34)$$

since  $\mathbf{d}_1'\mathbf{r}_1 = \mathbf{d}_1'\mathbf{Q}'_{\mathbf{d}_0/\mathbf{G}'\mathbf{G}}\mathbf{r}_0 = \mathbf{d}_1'\mathbf{r}_0 = \mathbf{d}_1'\mathbf{G}\mathbf{z}$  (the second equality in the preceding equation holds again due to the  $\mathbf{G}'\mathbf{G}$ -conjugacy of  $\mathbf{d}_1$  and  $\mathbf{d}_0$ ). Similarly, we obtain

$$\begin{aligned} \mathbf{b}_3 &= \hat{\mathbf{b}}_{CG}^{(2)} + \mathbf{d}_2(\mathbf{d}_2'\mathbf{G}'\mathbf{G}\mathbf{d}_2)^{-1}\mathbf{d}_2'\mathbf{r}_2, \\ &= \hat{\mathbf{b}}_{CG}^{(2)} + \mathbf{d}_2(\mathbf{d}_2'\mathbf{G}'\mathbf{G}\mathbf{d}_2)^{-1}\mathbf{d}_2'\mathbf{G}'\mathbf{z} = \hat{\mathbf{b}}_{CG}^{(3)}, \end{aligned} \quad (2.35)$$

since  $\mathbf{d}_2'\mathbf{r}_2 = \mathbf{d}_2'\mathbf{Q}'_{\mathbf{d}_1/\mathbf{G}'\mathbf{G}}\mathbf{r}_1 = \mathbf{d}_2'\mathbf{r}_1 = \mathbf{d}_2'\mathbf{Q}'_{\mathbf{d}_0/\mathbf{G}'\mathbf{G}}\mathbf{r}_0 = \mathbf{d}_2'\mathbf{r}_0 = \mathbf{d}_2'\mathbf{G}\mathbf{z}$ . This extends to  $S$  larger than 3. This proves the claim made above that (2.31) is indeed identical to  $\mathbf{b}_S$  obtained from the CG iteration.

It is rather intricate to show the  $\mathbf{G}'\mathbf{G}$ -conjugacy of direction vectors (i.e.,  $\mathbf{d}_j'\mathbf{G}'\mathbf{G}\mathbf{d}_i = 0$  for  $j \neq i$ ), although it is widely known in the numerical linear algebra literature (Golub and van Loan 1989). The proofs given in Golub and van Loan (1989) are not very easy to follow, however. In what follows, we attempt to provide a step-by-step proof of this fact. Let  $\mathbf{R}_j$  and  $\mathbf{D}_j$  be as defined above. We temporarily assume that the columns of  $\mathbf{D}_j$  are already  $\mathbf{G}'\mathbf{G}$ -conjugate (i.e.,  $\mathbf{D}_j'\mathbf{G}'\mathbf{G}\mathbf{D}_j$  is diagonal). Later we show that such construction of  $\mathbf{D}_j$  is possible.

We first show that

$$\mathbf{d}'_{j-1}\mathbf{r}_j = 0. \quad (2.36)$$

From Step 2(c) of the CG algorithm, we have

$$\mathbf{d}'_{j-1}\mathbf{r}_j = \mathbf{d}'_{j-1}\mathbf{Q}'_{\mathbf{d}_{j-1}/\mathbf{G}'\mathbf{G}}\mathbf{r}_{j-1} = \mathbf{d}'_{j-1}(\mathbf{I} - \mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}'_{j-1}\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}'_{j-1})\mathbf{r}_{j-1} = 0, \quad (2.37)$$



as claimed above. We next show that

$$\mathbf{d}'_{j-2}\mathbf{r}_j = 0, \quad (2.38)$$

based on (2.36). From Step 2(c) of the algorithm, we have

$$\begin{aligned} \mathbf{d}'_{j-2}\mathbf{r}_j &= \mathbf{d}'_{j-2}\mathbf{Q}'_{d_{j-1}/G'}\mathbf{r}_{j-1} \\ &= \mathbf{d}'_{j-2}(\mathbf{I} - \mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}'_{j-1}\mathbf{G}\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}'_{j-1})\mathbf{r}_{j-1} \\ &= \mathbf{d}'_{j-2}\mathbf{r}_{j-1} = 0, \end{aligned} \quad (2.39)$$

as claimed. Note that  $\mathbf{d}'_{j-2}\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1} = 0$  by the assumption of the  $\mathbf{G}'\mathbf{G}$ -conjugacy (among the column vectors) of  $\mathbf{D}_j$ . The last equality in (2.39) holds due to (2.36). By repeating essentially the same process, we can prove that  $\mathbf{d}'_{j-k}\mathbf{r}_j = 0$  for  $k = 3, \dots, j$ , which implies

$$\mathbf{D}'_j\mathbf{r}_j = \mathbf{0}, \quad (2.40)$$

and

$$\mathbf{R}'_j\mathbf{r}_j = \mathbf{0}, \quad (2.41)$$

since  $\text{Sp}(\mathbf{D}_j) = \text{Sp}(\mathbf{R}_j) = \mathcal{K}_j(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ . These relations indicate that in the CG method, the residual vector  $\mathbf{r}_j$  is orthogonal to all previous search directions as well as all previous residual vectors.

We are now in a position to prove that

$$\mathbf{d}'_{j-1}\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0. \quad (2.42)$$

To do so, we first need to show that

$$\mathbf{d}'_j\mathbf{r}_j = \|\mathbf{r}_j\|^2, \quad (2.43)$$

and also that

$$\mathbf{d}'_j\mathbf{r}_{j-1} = \|\mathbf{r}_j\|^2. \quad (2.44)$$

For Eq. (2.43), we note that

$$\begin{aligned} \mathbf{d}'_j\mathbf{r}_j &= \mathbf{r}'_j\mathbf{Q}'_{d_{j-1}/G'}\mathbf{r}_j \quad (\text{by Step 2(e)}) \\ &= \|\mathbf{r}_j\|^2 - \mathbf{r}'_j\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}'_{j-1}\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}'_{j-1}\mathbf{r}_j = \|\mathbf{r}_j\|^2, \end{aligned} \quad (2.45)$$

due to Eq. (2.36). For Eq. (2.44), we have

$$\begin{aligned} \mathbf{d}'_j \mathbf{r}_{j-1} &= \mathbf{r}'_j \mathbf{r}_{j-1} + b_{j-1} \mathbf{d}'_{j-1} \mathbf{r}_{j-1} \quad (\text{by Step 2(e)}) \\ &= 0 + (||\mathbf{r}_j||^2 / ||\mathbf{r}_{j-1}||^2) ||\mathbf{r}_{j-1}||^2 = ||\mathbf{r}_j||^2. \end{aligned} \quad (2.46)$$

To show that (2.42) holds is now straightforward. We note that

$$\mathbf{r}'_j \mathbf{d}_j = \mathbf{r}'_{j-1} \mathbf{d}_j - a_{j-1} \mathbf{d}'_{j-1} \mathbf{G}' \mathbf{G} \mathbf{d}_j \quad (2.47)$$

by Step 2(c), and that  $\mathbf{r}'_j \mathbf{d}_j = \mathbf{r}'_{j-1} \mathbf{d}_j = ||\mathbf{r}_j||^2$  by Eqs. (2.43) and (2.44). Since  $a_{j-1} \neq 0$ , this implies that  $\mathbf{d}'_{j-1} \mathbf{G}' \mathbf{G} \mathbf{d}_j = 0$ . That is,  $\mathbf{d}_j$  is  $\mathbf{G}' \mathbf{G}$ -conjugate to the previous direction vector  $\mathbf{d}_{j-1}$ .

We can also show that  $\mathbf{d}_j$  is  $\mathbf{G}' \mathbf{G}$ -conjugate to all previous direction vectors despite the fact that at any specific iteration,  $\mathbf{d}_j$  is taken to be  $\mathbf{G}' \mathbf{G}$ -conjugate to only  $\mathbf{d}_{j-1}$ . We begin with

$$\mathbf{d}'_{j-2} \mathbf{G}' \mathbf{G} \mathbf{d}_j = 0. \quad (2.48)$$

We first note that

$$\begin{aligned} \mathbf{r}'_{j-2} \mathbf{d}_j &= \mathbf{r}'_{j-2} \mathbf{r}_j + b_{j-1} \mathbf{r}'_{j-2} \mathbf{d}_{j-1} \quad (\text{by Step 2(e)}) \\ &= 0 + (||\mathbf{r}_j||^2 / ||\mathbf{r}_{j-1}||^2) ||\mathbf{r}_{j-1}||^2 \quad (\text{by (2.44)}) \\ &= ||\mathbf{r}_j||^2. \end{aligned} \quad (2.49)$$

We also have

$$\mathbf{r}'_{j-1} \mathbf{d}_j = \mathbf{r}'_{j-2} \mathbf{d}_j - a_{j-2} \mathbf{d}'_{j-2} \mathbf{G}' \mathbf{G} \mathbf{d}_j \quad (2.50)$$

by Step 2(c). Since  $\mathbf{r}'_{j-1} \mathbf{d}_j = \mathbf{r}'_{j-2} \mathbf{d}_j = ||\mathbf{r}_j||^2$  and  $a_{j-2} \neq 0$ , this implies (2.48). We may follow a similar line of argument as above, and show that  $\mathbf{d}'_{j-k} \mathbf{G}' \mathbf{G} \mathbf{d}_j = 0$  for  $k = 3, \dots, j$ . This shows that  $\mathbf{D}'_j \mathbf{G}' \mathbf{G} \mathbf{d}_j = \mathbf{0}$ , as claimed.

In the proof above, it was assumed that the column vectors of  $\mathbf{D}_j$  were  $\mathbf{G}' \mathbf{G}$ -conjugate. It remains to show that such construction of  $\mathbf{D}_j$  is possible. We have  $\mathbf{D}'_1 \mathbf{r}_1 = \mathbf{d}'_0 \mathbf{r}_1 = 0$  by (2.36). This implies that  $\mathbf{R}'_1 \mathbf{r}_1 = 0$  (since  $\text{Sp}(\mathbf{D}_1) = \text{Sp}(\mathbf{R}_1)$ ), which in turn implies that  $\mathbf{D}'_1 \mathbf{G}' \mathbf{G} \mathbf{d}_1 = \mathbf{d}'_0 \mathbf{G}' \mathbf{G} \mathbf{d}_1 = 0$ . The columns of  $\mathbf{D}_2 = [\mathbf{d}_0, \mathbf{d}_1]$  are now shown to be  $\mathbf{G}' \mathbf{G}$ -conjugate. We repeat this process until we reach  $\mathbf{D}_j$  whose column vectors are all  $\mathbf{G}' \mathbf{G}$ -conjugate. This process also generates  $\mathbf{R}_j$  whose columns are mutually orthogonal. This means that all residual vectors are orthogonal in the CG method. The CG algorithm is also equivalent to the GMRES (Generalized Minimum Residual) method (Saad and Schultz 1986), when the latter is applied to the symmetric positive definite (*pd*) matrix  $\mathbf{G}' \mathbf{G}$ .

It may also be pointed out that  $\mathbf{R}_S$  is an un-normalized version of  $\mathbf{W}_S$  obtained in PLS1. This can be seen from the fact that the column vectors of both of these matrices are orthogonal to each other, and that  $\text{Sp}(\mathbf{W}_S) = \text{Sp}(\mathbf{R}_S) = \mathcal{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ . Although some columns of  $\mathbf{R}_S$  may be sign-reversed as are some columns of  $\mathbf{U}_S$  in the Lanczos method, it can be directly verified that this does not happen to  $\mathbf{r}_2$  (i.e.,  $\mathbf{r}_2/\|\mathbf{r}_2\| = \mathbf{w}_2$ ). So it is not likely to happen to other columns of  $\mathbf{R}_S$ .

## 2.5 Concluding Remarks

The PLS1 algorithm was initially invented as a heuristic technique to solve LS problems (Wold 1966). No optimality properties of the algorithm were known at that time, and for a long time it had been criticized for being somewhat ad-hoc. It was later shown, however, that it is equivalent to some of the most sophisticated numerical algorithms to date for solving systems of linear simultaneous equations, such as the Lanczos bidiagonalization and the conjugate gradient methods. It is amazing, and indeed admirable, that Herman Wold almost single-handedly reinvented the “wheel” in a totally different context.

## References

- Abdi, H.: Partial least squares regression. In: Salkind, N.J. (ed.) *Encyclopedia of Measurement and Statistics*, pp. 740–54. Sage, Thousand Oaks (2007)
- Arnoldi, W.E.: The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Q. Appl. Math.* **9**, 17–29 (1951)
- Bro, R., Eldén, L.: PLS works. *J. Chemom.* **23**, 69–71 (2009)
- de Jong, S.: SIMPLS: an alternative approach to partial least squares regression. *J. Chemom.* **18**, 251–263 (1993)
- Eldén, L.: Partial least-squares vs Lanczos bidiagonalization—I: analysis of a projection method for multiple regression. *Comput. Stat. Data Anal.* **46**, 11–31 (2004)
- Golub, G.H., van Loan, C.F.: *Matrix Computations*, 2nd edn. The Johns Hopkins University Press, Baltimore (1989)
- Hestenes, M., Stiefel, E.: Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur Stand.* **49**, 409–436 (1951)
- Lohmöller, J.B.: *Latent Variables Path-Modeling with Partial Least Squares*. Physica-Verlag, Heidelberg (1989)
- Phatak, A., de Hoog, F.: Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *J. Chemom.* **16**, 361–367 (2002)
- Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Saunders, C., et al. (eds.) *SLSFS 2005. LNCS 3940*, pp. 34–51. Springer, Berlin (2006)
- Saad, Y.: *Iterative Methods for Sparse Linear Systems*, 2nd edn. Society of Industrial and Applied Mathematics, Philadelphia (2003)
- Saad, Y., Schultz, M.H.: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Comput.* **7**, 856–869 (1986)
- Takane, Y.: *Constrained Principal Component Analysis and Related Techniques*. CRC Press, Boca Raton (2014)

- Wold, H.: Estimation of principal components and related models by iterative least squares. In: Krishnaiah, P.R. (ed.) *Multivariate Analysis*, pp. 391–420. Academic, New York (1966)
- Wold, H. (1982) Soft modeling: the basic design and some extensions. In: Jöreskog, K.G., Wold, H. (eds.) *Systems Under Indirect Observations, Part 2*, pp. 1–54. North-Holland, Amsterdam (1982)

The Multiple Facets of Partial Least Squares and  
Related Methods

PLS, Paris, France, 2014

Abdi, H.; Esposito Vinzi, V.; Russolillo, G.; Saporta, G.;  
Trinchera, L. (Eds.)

2016, XIV, 316 p. 68 illus., 39 illus. in color., Hardcover

ISBN: 978-3-319-40641-1