Chapter 1
On the PLS algorithm for multiple regression (PLS1)

Yoshio Takane and Sèbastien Loisel

Abstract Partial least squares (PLS) was first introduced by Wold in the mid 1960's 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.

Key words: Krylov supspace, NIPALS, PLS1 algorithm, Lanczos bidiagonalization, conjugate gradients

1.1 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 [1, 4] and structural equation modeling [8, 15]. 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

Yoshio Takane
University of Victoria, Victoria, Bristish Columbia Canada e-mail: Yoshio.Takane@mcgill.ca

Sèbastien Loisel
Heriot-Watt University, Edinburgh, UK. e-mail: sloisel@gmail.com

discussed from a somewhat different perspective from previous authors [5, 9]. We refer the reader to [10] for more comprehensive accounts and reviews of new developments of PLS.

1.2 PLS1 as Constrained Least Squares Estimator

Consider a linear regression model

$$\mathbf{z} = \mathbf{G}\mathbf{b} + \mathbf{e}, \tag{1.1}$$

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 $i = 1, \cdots, S$ ($S \leq \text{rank}(\mathbf{G})$):

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 $\mathbf{G}_i = \mathbf{G}_{i-1} - \mathbf{t}_i\mathbf{v}'_i = \mathbf{Q}_{G_{i-1}w_i}\mathbf{G}_{i-1}$ (deflation),

where $\mathbf{Q}_{G_{i-1}w_i} = \mathbf{I} - \mathbf{G}_{i-1}\mathbf{w}_i(\mathbf{w}'_i\mathbf{G}'_{i-1}\mathbf{G}_{i-1}\mathbf{w}_i)^{-1}\mathbf{w}'_i\mathbf{G}'_{i-1}$, and where $'$ denotes the transpose operation, and $||.||$ denotes the L_2 norm of a vector (i.e., $||\mathbf{x}|| = \sqrt{\mathbf{x}'\mathbf{x}}$, see, e.g., [13], 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

$$\hat{\mathbf{b}}^{(S)}_{PLSE} = \mathbf{W}_S(\mathbf{V}'_S\mathbf{W}_S)^{-1}\mathbf{T}'_S\mathbf{z} \tag{1.2}$$

(see, e.g., [1]). 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 ([9]) 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

$$\mathscr{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z}) = \mathrm{Sp}(\mathbf{K}_S), \tag{1.3}$$

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

$$\mathbf{K}_s = [\mathbf{G}'\mathbf{z}, (\mathbf{G}'\mathbf{G})\mathbf{G}'\mathbf{z}, \cdots, (\mathbf{G}'\mathbf{G})^{S-1}\mathbf{G}'\mathbf{z}] \tag{1.4}$$

is called the Krylov matrix of order S. Because $\mathrm{Sp}(\mathbf{W}_S) = \mathscr{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$ (see [5], proposition 3.1. and [9]) $\mathbf{b}$ can be re-parameterized as $\mathbf{b} = \mathbf{W}_S\mathbf{a}$ for some $\mathbf{a}$. Then Equation (1.1) can be rewritten as

$$\mathbf{z} = \mathbf{G}\mathbf{W}_S\mathbf{a} + \mathbf{e}. \tag{1.5}$$

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

$$\hat{\mathbf{a}} = (\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{G}'\mathbf{z}, \tag{1.6}$$

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

$$\hat{\mathbf{b}}_{CLSE}^{(S)} = \mathbf{W}_S\hat{\mathbf{a}} = \mathbf{W}_S(\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{G}'\mathbf{z}. \tag{1.7}$$

To show that (1.7) is indeed equivalent to (1.2), we need several well-known results in the PLS literature [3, 4, 5, 9]. First of all, $\mathbf{W}_S$ is column-wise orthogonal, that is,

$$\mathbf{W}_S'\mathbf{W}_S = \mathbf{I}_S. \tag{1.8}$$

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

$$\mathbf{T}_S'\mathbf{T}_S = \mathbf{I}_S, \tag{1.9}$$

and

$$\mathbf{T}_S\mathbf{L}_S = \mathbf{G}\mathbf{W}_S, \tag{1.10}$$

where $\mathbf{L}_S$ is an upper bidiagonal matrix. Relations (1.8) through (1.10) imply that

$$\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S = \mathbf{L}_S'\mathbf{L}_S = \mathbf{H}_S, \tag{1.11}$$

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

$$\mathbf{V}_S' = \mathbf{T}_S'\mathbf{G}, \tag{1.12}$$

so that

$$\mathbf{L}_S = \mathbf{T}_S'\mathbf{G}\mathbf{W}_S = \mathbf{V}_S'\mathbf{W}_S. \tag{1.13}$$

Now it is straightforward to show that

$$\begin{aligned}
\hat{\mathbf{b}}_{CLSE}^{(S)} &= \mathbf{W}_S(\mathbf{W}_S'\mathbf{G}'\mathbf{G}\mathbf{W}_S)^{-1}\mathbf{W}_S'\mathbf{G}'\mathbf{z} \\
&= \mathbf{W}_S\mathbf{H}_S^{-1}\mathbf{L}_S'\mathbf{T}_S'\mathbf{z} \\
&= \mathbf{W}_S(\mathbf{L}_S'\mathbf{L}_S)^{-1}\mathbf{L}_S'\mathbf{T}_S'\mathbf{z} \\
&= \mathbf{W}_S\mathbf{L}_S^{-1}\mathbf{T}_S'\mathbf{z} \\
&= \mathbf{W}_S(\mathbf{V}_S'\mathbf{W}_S)^{-1}\mathbf{T}_S'\mathbf{z} \\
&= \hat{\mathbf{b}}_{PLSE}^{(S)},
\end{aligned} \tag{1.14}$$

and this establishes the equivalence between Equations (1.7) and (1.2).

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

1.3 Relations to the Lanczos Bidiagonalization Method

It has been pointed out [5] 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 $i = 2, \cdots, S$ (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 γ_{i-1} and δ_i ($i = 2, \cdots, S$) 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 $i = 1, \cdots, S$. It has been shown ([5] 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

$$\mathbf{w}_1 = \mathbf{u}_1 \quad \text{and} \quad \mathbf{t}_1 = \mathbf{q}_1. \tag{1.15}$$

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

$$\begin{aligned}
\mathbf{w}_2 &\propto \mathbf{G}'\mathbf{Q}_{Gw_1}\mathbf{z} \quad \text{(from Step 2.4 of the PLS1 algorithm)} \\
&= \mathbf{G}'\mathbf{z} - \mathbf{G}'\mathbf{G}\mathbf{w}_1(\mathbf{w}_1'\mathbf{G}'\mathbf{G}\mathbf{w}_1)^{-1}\mathbf{w}_1'\mathbf{G}'\mathbf{z} \\
&= \alpha_1(\mathbf{w}_1 - \mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1^2) \qquad (1.16) \\
&\propto -\mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1 + \delta_1\mathbf{w}_1, \qquad (1.17)
\end{aligned}$$

where $\propto$ means "proportional." To obtain the last expression, we multiplied Equation (1.16) by δ_1/α_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

$$\begin{aligned}
\mathbf{t}_2 &\propto \mathbf{Q}_{Gw_1}\mathbf{G}\mathbf{G}'\mathbf{Q}_{Gw_1}\mathbf{z} \quad \text{(from Step 2.2 of the PLS1 algorithm)} \\
&= \alpha_1(\mathbf{G}\mathbf{w}_1 - \mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{w}_1/\delta_1^2 - \mathbf{G}\mathbf{w}_1 + \frac{\beta_1^2}{\delta_1^4}\mathbf{G}\mathbf{w}_1) \qquad (1.18) \\
&\propto -\mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{w}_1 + \frac{\beta_1^2}{\delta_1^2}\mathbf{G}\mathbf{w}_1. \qquad (1.19)
\end{aligned}$$

To obtain Equation (1.19), we multiplied (1.18) by δ_1^2/α_1 (> 0). On the other hand, we have

$$\begin{aligned}
\mathbf{q}_2 &\propto \frac{1}{\delta_1\gamma_1}(\mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{u}_1 - \delta_1^2\mathbf{G}\mathbf{u}_1 - \gamma_1^2\mathbf{G}\mathbf{u}_1) \quad \text{(from Step 2(b) of the Lanczos algorithm)} \\
&\propto \mathbf{G}\mathbf{G}'\mathbf{G}\mathbf{u}_1 - (\delta_1^2 + \gamma_1^2)\mathbf{G}\mathbf{u}_1. \qquad (1.20)
\end{aligned}$$

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

$$\gamma^2 + \delta^2 = \beta_1^2/\delta_1^2. \qquad (1.21)$$

From Step 2(a) of the Lanczos algorithm,

$$\begin{aligned}
\gamma^2 &= (\mathbf{G}'\mathbf{G}\mathbf{u}_1/\delta_1 - \delta_1\mathbf{u}_1)'(\mathbf{G}'\mathbf{G}\mathbf{u}_1/\delta_1 - \delta_1\mathbf{u}_1) \\
&= \beta^2/\delta^2 - \delta^2, \qquad (1.22)
\end{aligned}$$

and so indeed (1.21) 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 bidiagonaliation method is given by

$$\hat{\mathbf{b}}_{LBD}^{(S)} = \mathbf{U}_s(\mathbf{L}_S^*)^{-1}\mathbf{Q}_S'\mathbf{z}, \qquad (1.23)$$

where

$$\mathbf{L}_S^* = \mathbf{Q}_S' \mathbf{G} \mathbf{U}_S, \tag{1.24}$$

which is upper bidiagonal, as is $\mathbf{L}_S$ (defined in Equation 1.13). 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

$$\begin{aligned}
\mathbf{W}_S \mathbf{L}_S^{-1} \mathbf{T}_S' &= \sum_{i=1}^{s} (\ell_{i,i} \mathbf{w}_i \mathbf{t}_i' + \ell_{i,i+1} \mathbf{w}_i \mathbf{t}_{i+1}') \\
&= \sum_{i=1}^{s} (\ell_{i,i}^* \mathbf{u}_i \mathbf{q}_i' + \ell_{i,i+1}^* \mathbf{u}_i \mathbf{q}_{i+1}') \\
&= \mathbf{U}_s (\mathbf{L}_s^*)^{-1} \mathbf{Q}_s',
\end{aligned} \tag{1.25}$$

where $\ell_{i,j}$ and $\ell_{i,j}^*$ are the ij-th element of (respectively) $\mathbf{L}_S$ and $\mathbf{L}_S^*$. Note that $\ell_{i,i} = \ell_{i,i}^*$, $\mathbf{w}_i \mathbf{t}_i' = \mathbf{u}_i \mathbf{q}_i'$, $\ell_{i,i+1} = -\ell_{i,i+1}^*$, and $\mathbf{w}_i \mathbf{t}_{i+1}' = -\mathbf{u}_i \mathbf{q}_{i+1}'$.

It is widely known (see, e.g., [11]) that the matrix of orthogonal basis vectors generated by the Arnoldi orthogonalization of $\mathbf{K}_S$ [2] 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 $\mathbf{u}_{i+1}$ $(i = 1, \cdots, S-1)$ by successively orthogonalizing $\mathbf{G}'\mathbf{G}\mathbf{u}_i$ $(i = 1, \cdots, S-1)$ 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

$$\mathbf{U}_S' \mathbf{G}' \mathbf{G} \mathbf{U}_S = \mathbf{L}_S^{*'} \mathbf{L}_S^* = \mathbf{H}_S^*, \tag{1.26}$$

where $\mathbf{H}_S^*$ is tridiagonal as is $\mathbf{H}_S$ defined in Equation (1.11). 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 $\mathbf{G}'\mathbf{G}$.

1.4 Relations to the Conjugate Gradient Method

It has been pointed out [9] that the conjugate gradient (CG) algorithm [7] for solving a system of linear simultaneous equations $\mathbf{G}'\mathbf{G}\mathbf{b} = \mathbf{G}'\mathbf{y}$ gives solutions identical to $\hat{\mathbf{b}}_{PLSE}^{(s)}$ $[s = 1, \cdots, \text{rank}(\mathbf{G})]$, 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 $i = 0, \cdots, s-1$, 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 $\mathrm{Sp}(\mathbf{G}'\mathbf{G}\mathbf{d}_i)$ along $\mathrm{Sp}(\mathbf{d}_i)$ (Its transpose, on the other hand, is the projector onto the space orthogonal $\mathrm{Sp}(\mathbf{d}_i)$ along $\mathrm{Sp}(\mathbf{G}'\mathbf{G}\mathbf{d}_i)$).
(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 $\mathbf{R}_j = [\mathbf{r}_0, \cdots, \mathbf{r}_{j-1}]$ and $\mathbf{D}_j = [\mathbf{d}_0, \cdots, \mathbf{d}_{j-1}]$ for $j \leq S$. We first show that

$$\mathrm{Sp}(\mathbf{R}_j) = \mathrm{Sp}(\mathbf{D}_j) = \mathscr{K}_j(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z}) \tag{1.27}$$

by induction, where, as before, $\mathrm{Sp}(\mathbf{A})$ 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 $\mathrm{Sp}(\mathbf{R}_1) = \mathrm{Sp}(\mathbf{D}_1) = \mathscr{K}_1(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. From Step 2(c) of the CG algorithm, we have

$$\mathbf{r}_1 = \mathbf{Q}'_{d_i/G'G}\mathbf{r}_0 = \mathbf{r}_0 - \mathbf{G}'\mathbf{G}\mathbf{d}_0 c_0 \tag{1.28}$$

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

$$\mathbf{d}_1 = \mathbf{Q}_{d_0/G'G}\mathbf{r}_1 = \mathbf{r}_1 - \mathbf{d}_0 c_0^* \tag{1.29}$$

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

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

$$\hat{\mathbf{b}}_{CG}^{(S)} = \mathbf{D}_S(\mathbf{D}_S'\mathbf{G}\mathbf{G}\mathbf{D}_S)^{-1}\mathbf{D}_S'\mathbf{G}\mathbf{z} \tag{1.30}$$

is identical to $\hat{\mathbf{b}}_{CLSE}^{(S)}$ as defined in Equation (1.7), which in turn is equal to $\hat{\mathbf{b}}_{PLSE}^{(S)}$ defined in Equation (1.2) [9] by virtue of Equation 1.14. It remains to show that $\hat{\mathbf{b}}_{CG}^{(S)}$ defined in (1.30) 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), Equation 1.30 can be rewritten as

$$\hat{\mathbf{b}}_{CG}^{(S)} = \sum_{i=0}^{S-1} \mathbf{d}_i(\mathbf{d}_i'\mathbf{G}'\mathbf{G}\mathbf{d}_i)^{-1}\mathbf{d}_i'\mathbf{G}'\mathbf{z}. \tag{1.31}$$

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

$$\mathbf{b}_1 = \mathbf{d}_0(\mathbf{d}_0'\mathbf{G}'\mathbf{G}\mathbf{d}_0)^{-1}\mathbf{d}_0'\mathbf{r}_0 = \mathbf{d}_0(\mathbf{d}_0'\mathbf{G}'\mathbf{G}\mathbf{d}_0)^{-1}\mathbf{d}_0'\mathbf{G}\mathbf{z} = \hat{\mathbf{b}}_{CG}^{(1)}, \tag{1.32}$$

and

$$\begin{aligned}\mathbf{b}_3 &= \hat{\mathbf{b}}_{CG}^{(1)} + \mathbf{d}_1(\mathbf{d}_1'\mathbf{G}'\mathbf{G}\mathbf{d}_1)^{-1}\mathbf{d}_1'\mathbf{r}_1, \\ &= \hat{\mathbf{b}}_{CG}^{(1)} + \mathbf{d}_1(\mathbf{d}_1'\mathbf{G}'\mathbf{G}\mathbf{d}_1)^{-1}\mathbf{d}_1'\mathbf{G}'\mathbf{z} = \hat{\mathbf{b}}_{CG}^{(2)}, \end{aligned} \tag{1.33}$$

since $\mathbf{d}_1'\mathbf{r}_1 = \mathbf{d}_1'\mathbf{Q}_{d_0/G'G}'\mathbf{r}_0 = \mathbf{d}_1'\mathbf{r}_0 = \mathbf{d}_1'\mathbf{G}\mathbf{z}$. (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

$$\begin{aligned}\mathbf{b}_3 &= \hat{\mathbf{b}}_{CG}^{(2)} + \mathbf{d}_2(\mathbf{d}_2'\mathbf{G}'\mathbf{G}\mathbf{d}_2)^{-1}\mathbf{d}_2'\mathbf{r}_2, \\ &= \hat{\mathbf{b}}_{CG}^{(2)} + \mathbf{d}_2(\mathbf{d}_2'\mathbf{G}'\mathbf{G}\mathbf{d}_2)^{-1}\mathbf{d}_2'\mathbf{G}'\mathbf{z} = \hat{\mathbf{b}}_{CG}^{(3)}, \end{aligned} \tag{1.34}$$

since $\mathbf{d}_2'\mathbf{r}_2 = \mathbf{d}_2'\mathbf{Q}_{d_1/G'G}'\mathbf{r}_1 = \mathbf{d}_2'\mathbf{r}_1 = \mathbf{d}_2'\mathbf{Q}_{d_0/G'G}'\mathbf{r}_0 = \mathbf{d}_2'\mathbf{r}_0 = \mathbf{d}_2'\mathbf{G}\mathbf{z}$. This extends to S larger than 3. This proves the claim made above that (1.30) 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 [6]. The proofs given in [6] 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

$$\mathbf{d}_{j-1}'\mathbf{r}_j = 0. \tag{1.35}$$

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

$$\mathbf{d}_{j-1}'\mathbf{r}_j = \mathbf{d}_{j-1}'\mathbf{Q}_{d_{j-1}/G'G}'\mathbf{r}_{j-1} = \mathbf{d}_{j-1}'(\mathbf{I} - \mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}_{j-1}'\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}_{j-1}')\mathbf{r}_{j-1} = 0, \tag{1.36}$$

as claimed above. We next show that

$$\mathbf{d}_{j-2}'\mathbf{r}_j = 0, \tag{1.37}$$

based on (1.35). From Step 2(c) of the algorithm, we have

$$\begin{aligned}\mathbf{d}_{j-2}'\mathbf{r}_j &= \mathbf{d}_{j-2}'\mathbf{Q}_{d_{j-1}/G'G}'\mathbf{r}_{j-1} \\ &= \mathbf{d}_{j-2}'(\mathbf{I} - \mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}_{j-1}'\mathbf{G}\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}_{j-1}')\mathbf{r}_{j-1} \\ &= \mathbf{d}_{j-2}'\mathbf{r}_{j-1} = 0, \end{aligned} \tag{1.38}$$

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 (1.38) holds due to (1.35). By repeating essentially the same process, we can prove that $\mathbf{d}_{j-k}'\mathbf{r}_j = 0$ for $k = 3, \cdots, j$, which implies

$$\mathbf{D}_j'\mathbf{r}_j = \mathbf{0}, \tag{1.39}$$

and

$$\mathbf{R}_j'\mathbf{r}_j = \mathbf{0}, \tag{1.40}$$

since $\mathrm{Sp}(\mathbf{D}_j) = \mathrm{Sp}(\mathbf{R}_j) = \mathscr{K}_j(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. 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

$$\mathbf{d}'_{j-1}\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0. \tag{1.41}$$

To do so, we first need to show that

$$\mathbf{d}'_j\mathbf{r}_j = ||\mathbf{r}_j||^2, \tag{1.42}$$

and also that

$$\mathbf{d}'_j\mathbf{r}_{j-1} = ||\mathbf{r}_j||^2. \tag{1.43}$$

For Equation 1.42, we note that

$$\begin{aligned}\mathbf{d}'_j\mathbf{r}_j &= \mathbf{r}'_j\mathbf{Q}'_{d_{j-1}/G'G}\mathbf{r}_j \quad \text{(by Step 2(e))} \\ &= ||\mathbf{r}_j||^2 - \mathbf{r}'_j\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1}(\mathbf{d}'_{j-1}\mathbf{G}'\mathbf{G}\mathbf{d}_{j-1})^{-1}\mathbf{d}'_{j-1})\mathbf{r}_j = ||\mathbf{r}_j||^2,\end{aligned} \tag{1.44}$$

due to Equation 1.35. For Equation 1.43, we have

$$\begin{aligned}\mathbf{d}'_j\mathbf{r}_{j-1} &= \mathbf{r}'_j\mathbf{r}_{j-1} + b_{j-1}\mathbf{d}'_{j-1}\mathbf{r}_{j-1} \quad \text{(by Step 2(e))} \\ &= 0 + (||\mathbf{r}_j||^2/||\mathbf{r}_{j-1}||^2)||\mathbf{r}_{j-1}||^2 = ||\mathbf{r}_j||^2.\end{aligned} \tag{1.45}$$

To show that (1.41) holds is now straightforward. We note that

$$\mathbf{r}'_j\mathbf{d}_j = \mathbf{r}'_{j-1}\mathbf{d}_j - a_{j-1}\mathbf{d}'_{j-1}\mathbf{G}'\mathbf{G}\mathbf{d}_j \tag{1.46}$$

by Step 2(c), and that $\mathbf{r}'_j\mathbf{d}_j = \mathbf{r}'_{j-1}\mathbf{d}_j = ||\mathbf{r}_j||^2$ by Equations 1.42 and 1.43. 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

$$\mathbf{d}'_{j-2}\mathbf{G}'\mathbf{G}\mathbf{d}_j = 0. \tag{1.47}$$

We first note that

$$\begin{aligned}\mathbf{r}'_{j-2}\mathbf{d}_j &= \mathbf{r}'_{j-2}\mathbf{r}_j + b_{j-1}\mathbf{r}'_{j-2}\mathbf{d}_{j-1} \quad \text{(by Step 2(e))} \\ &= 0 + (||\mathbf{r}_j||^2/||\mathbf{r}_{j-1}||^2)||\mathbf{r}_{j-1}||^2 \quad \text{(by (1.43))} \\ &= ||\mathbf{r}_j||^2.\end{aligned} \tag{1.48}$$

We also have

$$\mathbf{r}'_{j-1}\mathbf{d}_j = \mathbf{r}'_{j-2}\mathbf{d}_j - a_{j-2}\mathbf{d}'_{j-2}\mathbf{G}'\mathbf{G}\mathbf{d}_j \tag{1.49}$$

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 (1.47). 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 $k = 3, \cdots, j$. 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 (1.35). This implies that $\mathbf{R}'_1\mathbf{r}_1 = 0$ (since $\mathrm{Sp}(\mathbf{D}_1) = \mathrm{Sp}(R_1)$), 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 [12], 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 $\mathrm{Sp}(\mathbf{W}_S) = \mathrm{Sp}(\mathbf{R}_S) = \mathscr{K}_S(\mathbf{G}'\mathbf{G}, \mathbf{G}'\mathbf{z})$. 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 $\mathbf{R}_S$.

1.5 Concluding Remarks

The PLS1 algorithm was initially invented as a heuristic technique to solve LS problems [14]. 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

1. H. Abdi, "Partial least squares regression," in *Encyclopedia of Measurement and Statistics*, N.J. Salkind, ed, pp. 740–54, Thousand Oaks (CA): Sage, 2007.
2. W.E., Arnoldi, "The principle of minimized iterations in the solution of the matrix eigenvalue problem," *Quarterly of Applied Mathematics* **9**, pp. 17–29, 1951.
3. R. Bro and L. Eldén, "PLS works," *Journal of Chemometrics* **23**, pp. 69–71, 2009.
4. S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Journal of Chemometrics* **18**, pp. 251–263, 1993.
5. L. Eldén, "Partial least-squares vs Lanczos bidiagonalization–I: Analysis of a projection method for multiple regression," *Computational Statistics and Data Analysis* **46**, pp. 11–31, 2004.

6. G.H. Golub and C.F. van Loan, *Matrix Computations (Second Edition)*, Baltimore: The Johns Hopkins University Press, 1989.
7. M. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems"' *Journal of Research of the National Bureau of Standards* **49**, pp. 409–436, 1951.
8. J.B. Lohmöller, *Latent Variables Path-Modeling with Partial Least Squares*, Heidelberg: Physica-Verlag, 1989.
9. A. Phatak, and F. de Hoog, "Exploiting the connection between PLS, Lanczos methods and conjugate gradients: Alternative proofs of some properties of PLS," *Journal of Chemometrics* **16**, pp. 361–367, 2002.
10. R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *SLSFS 2005, LNCS 3940*, C. Saunders et al., eds., pp. 34–51, Berlin: Springer, 2006.
11. Y. Saad, *Iterative Methods for Sparse Linear Systems, (Second edition)*, Philadelphia: Society of Industrial and Applied Mathematics, 2003.
12. Y. Saad and M.H. Schultz, "A generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM Journal of Scientific Computing* **7**, pp. 856–869, 1986.
13. Y. Takane, *Constrained Principal Component Analysis and Related Techniques*, Boca Raton (FL): CRC Press, 2014.
14. H. Wold, "Estimation of principal components and related models by iterative least squares" in *Multivariate Analysis*, P.R. Krishnaiah, ed., pp. 391–420, New York: Academic Press, 1966.
15. H. Wold, 1982. "Soft modeling: The basic design and some extensions," in *Systems under indirect observations, Part 2*, K.G. Jöreskog and H. Wold, eds., pp. 1–54, Amsterdam: North-Holland, 1982.