

PLSR, Lanczos, and Conjugate Gradients

ALOKE PHATAK AND FRANK DE HOOG[†]
CSIRO Mathematical & Information Sciences
Private Bag No 5, Wembley, WA 6913
[†]GPO Box 664, Canberra, ACT 2601

Report No. CMIS 01/122
July 2001

Abstract

The connection between partial least squares regression (PLSR) and Lanczos methods for approximating the extremal eigenvalues of a symmetric matrix has long been known. Less well known, however, is that PLSR is in fact identical to a common implementation of the conjugate gradient algorithm for solving the normal equations. In this report, we outline the connections between, on the one hand, PLSR, and on the other, Lanczos methods and conjugate gradients. In addition to shedding more light on PLSR, these connections allow us to provide alternative and somewhat simpler proofs of two of its well-known properties: first, that it yields a shrinkage estimator, and second, that it 'fits better' than principal component regression.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 2 | Notation, Conventions, and Preliminaries | 1 |
| 3 | Partial Least Squares Regression | 2 |
| 3.1 | PLSR Estimator | 2 |
| 3.2 | Polynomial Representation of $\hat{\beta}_{PLS}^m$ | 3 |
| 3.3 | Shrinkage and 'Closeness' of PLSR | 3 |
| 4 | Connections to Lanczos Methods | 4 |
| 5 | Conjugate Gradient Method | 6 |
| 6 | Alternative Proofs | 8 |
| 6.1 | Shrinkage of $\hat{\beta}_{PLS}^m$ | 8 |
| 6.2 | PLSR Fits Closer than PCR | 10 |
| A | Polynomial Representation of $\hat{\beta}_{PLS}^m$ | 12 |
| A.1 | Proof I | 13 |
| A.2 | Proof II | 15 |

1 Introduction

Some early papers on univariate partial least squares regression (PLSR)^{12,17,18} cite its connection to the Lanczos method of approximating the extremal eigenvalues of a symmetric matrix.^{7,20} Moreover, Wold *et al.*²⁴ mention in passing that PLSR is identical to the method of conjugate gradients for solving the normal equations, although, as Manne¹⁷ pointed out, they did not discuss the ‘classic’ conjugate gradient (CG) algorithm of Hestenes and Stiefel¹¹ but an algorithm for bidiagonalizing a matrix \mathbf{X} .

It turns out, however, that a common implementation of the CG algorithm for solving the normal equations *does in fact* yield iterates that are identical to the PLSR estimator of corresponding dimensionality. In this report, we summarize the connections between PLSR, CG, and Lanczos methods and exploit them to provide alternative and simpler proofs of two well-known properties of PLSR: first, that it yields a shrinkage estimator,^{4,8} and second, that it ‘fits better’ than principal component regression (PCR).²

Section 2 introduces some notation while Section 3 outlines some well-known properties of the PLSR estimator. We will also show that estimators of increasing dimensionality can be written in terms of certain characteristic polynomials, a representation that arises naturally out of the fact that Krylov subspaces play a large role in PLSR. Section 4 then summarizes its connections to the Lanczos method of approximating the extremal eigenvalues of a symmetric matrix. The material in this section is not new, but is perhaps not as well known as it should be and is scattered throughout the chemometric literature. In Section 5, we outline the conjugate gradient algorithm of Hestenes and Stiefel¹¹ and show that it yields results that are identical to the sequence of PLSR estimates. Finally, Section 6 presents alternative proofs of the properties mentioned above, and in doing so we will make use of some identities and results presented in earlier parts of the report.

2 Notation, Conventions, and Preliminaries

Scalars are written in lower-case italics (a , β), column vectors in lower-case boldface type (\mathbf{a} , $\boldsymbol{\beta}$), and matrices in upper-case boldface letters (\mathbf{A} , $\boldsymbol{\Gamma}$). When we need to refer to the columns of a matrix \mathbf{A} , they will be written using the corresponding lower-case letter, e.g., $\mathbf{a}_1, \mathbf{a}_2, \dots$. Transposition is denoted by the superscript $'$ (\mathbf{a}' , \mathbf{A}'). Some special matrices and vectors that we require include the identity matrix \mathbf{I} which may have a subscript denoting its order; a vector of ones ($\mathbf{1}$); and the Moore-Penrose inverse¹⁶ of a matrix \mathbf{A} , which we write as \mathbf{A}^- . The Euclidean norm of a vector \mathbf{a} with respect to a symmetric, positive definite matrix \mathbf{A} is defined as $\|\mathbf{a}\|_{\mathbf{A}} \equiv (\mathbf{a}'\mathbf{A}\mathbf{a})^{1/2}$; when $\mathbf{A} = \mathbf{I}$ no subscript is used. When \mathbf{A} is symmetric, positive semidefinite, $\|\mathbf{a}\|_{\mathbf{A}}$ is defined only when $\mathbf{a} \in \text{range}(\mathbf{A})$ and is referred to as a seminorm.

One way of looking at PLSR is that it is a means of estimating the coefficients in the conventional linear regression model

$$\mathbf{y} = \mathbf{1}_n\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1)$$

Here, the $n \times 1$ vector \mathbf{y} consists of observations on a response variable; the vector $\mathbf{1}_n$ consists of ones; \mathbf{X} is an $n \times \tilde{p}$ matrix containing the values of \tilde{p} explanatory variables whose columns have been centered about their means, so that $\mathbf{1}_n'\mathbf{X} = \mathbf{0}$; β_0 is an unknown constant; $\boldsymbol{\beta}$ represents a $\tilde{p} \times 1$ vector of coefficients that is of primary interest; and $\boldsymbol{\epsilon}$ is an n -vector of

errors that are independently and identically distributed with zero expectation and constant variance. In the discussion that follows, it will be useful to define $p \equiv \min(n - 1, \tilde{p})$.

When $n > \tilde{p}$, the ordinary least squares estimator of β is $\hat{\beta}_{OLS} = \mathbf{A}^{-1}\mathbf{b}$, where $\mathbf{A} = \mathbf{X}'\mathbf{X}$ and $\mathbf{b} = \mathbf{X}'\mathbf{y}$. In most situations in which PLSR is used, however, $n < \tilde{p}$ so $\hat{\beta}_{OLS}$ does not exist. Nevertheless, we can still calculate the minimum-length least squares estimator, $\hat{\beta}_{MLLS} = \mathbf{A}^-\mathbf{b}$. Now let the spectral decomposition of \mathbf{A} be $\mathbf{A} = \Gamma\Lambda\Gamma' = \sum_{i=1}^p \lambda_i \gamma_i \gamma_i'$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ contains the p non-zero eigenvalues of \mathbf{A} ordered so that $\lambda_1 > \lambda_2 > \dots > \lambda_p$ and the γ_i are unit-norm eigenvectors of \mathbf{A} . Then, an explicit expression for the Moore-Penrose inverse of \mathbf{A} is $\mathbf{A}^- = \sum_{i=1}^p \lambda_i^{-1} \gamma_i \gamma_i'$. As a consequence, both $\hat{\beta}_{OLS}$ and $\hat{\beta}_{MLLS}$ can be written as

$$\hat{\beta}_{OLS} \text{ or } \hat{\beta}_{MLLS} = \sum_{i=1}^p \lambda_i^{-1} \gamma_i \gamma_i' \mathbf{b} \quad (2)$$

Of course, when $n - 1 > \tilde{p}$, $\mathbf{A}^- = \mathbf{A}^{-1}$, so hereafter, we will use $\hat{\beta}_{OLS}$ to refer to $\hat{\beta}_{MLLS}$. Where the distinction is required, we will refer to one or the other.

With spectroscopic data, for example, many of the λ_i in eq. (2) will be very small, and hence any estimator that incorporates them will have very large variance. One alternative is principal component regression, which deletes from eq. (2) the eigenvectors associated with very small eigenvalues. If say we retain the eigenvectors associated with the largest $m < p$ eigenvalues, the ‘ m -dimensional’ PCR estimator is

$$\hat{\beta}_{PCR}^m = \sum_{i=1}^m \lambda_i^{-1} \gamma_i \gamma_i' \mathbf{b} \quad (3)$$

Here the eigenvectors, or principal components $\mathbf{X}\gamma_i$, have been taken in their ‘natural’ – decreasing variance – order. An alternative is to use only that subset of components M that is in some way most relevant for prediction.¹³ In either case, using all components yields $\hat{\beta}_{OLS}$ or $\hat{\beta}_{MLLS}$.

3 Partial Least Squares Regression

3.1 PLSR Estimator

The PLS estimate of β based on m PLS dimensions or components is given by

$$\hat{\beta}_{PLS}^m = \mathbf{K}_m (\mathbf{K}_m' \mathbf{A} \mathbf{K}_m)^{-1} \mathbf{K}_m' \mathbf{b} \quad (4)$$

The columns of the $\tilde{p} \times m$ ($m \leq p$) matrix \mathbf{K}_m are the first m vectors of the Krylov sequence $\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \dots$. Thus $\mathbf{K}_m = (\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \dots, \mathbf{A}^{m-1}\mathbf{b})$; moreover, let us define the Krylov subspace $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$ to be $\mathcal{K}_m(\mathbf{A}, \mathbf{b}) = \text{span}(\mathbf{K}_m)$.

Note that in eq. (4) any matrix \mathbf{V}_m can be used in place of \mathbf{K}_m as long as its columns form a basis for $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$. Indeed eq. (4) is not practical for calculating $\hat{\beta}_{PLS}^m$, and all algorithms for PLSR generate an alternative basis for $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$.

A particularly convenient basis, $\mathbf{W}_m = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m)$, arises from the NIPALS algorithm¹⁰ and is formed by an implicit Gram-Schmidt orthonormalization of the columns of \mathbf{K}_m . As a consequence, the $m \times m$ matrix $\mathbf{T}_m = \mathbf{W}_m' \mathbf{A} \mathbf{W}_m$ is tridiagonal. In the NIPALS algorithm, the \mathbf{w}_j are calculated together with PLS dimensions or components, \tilde{t}_j , $j = 1, 2, \dots, m$, which are n -vectors that are analogous to principal components. They

form an orthogonal basis for $\mathbf{X}\mathbf{K}_m$. The first dimension is calculated as $\tilde{\mathbf{t}}_1 = \mathbf{E}_0\mathbf{w}_1$, where $\mathbf{E}_0 = \mathbf{X}$ and $\mathbf{w}_1 = \mathbf{b}/\|\mathbf{b}\|$. Subsequent dimensions are calculated as $\tilde{\mathbf{t}}_j = \mathbf{E}_{j-1}\mathbf{w}_j$, where $\mathbf{E}_{j-1} = \mathcal{H}_{\tilde{\mathbf{t}}_{j-1}^\perp}\mathbf{E}_{j-2}$, $\mathbf{w}_j = \mathbf{E}'_{j-1}\mathbf{y}/\|\mathbf{E}'_{j-1}\mathbf{y}\|$, and $\mathcal{H}_{\tilde{\mathbf{t}}_{j-1}^\perp}$ denotes the orthogonal projector onto $\tilde{\mathbf{t}}_{j-1}^\perp$. Note that with the exception of $\tilde{\mathbf{t}}_1$, the dimensions are not obtained as linear combinations of \mathbf{X} but of deflated matrices \mathbf{E}_j . It is possible, however, to express them in terms of \mathbf{X} . If we define $\mathbf{P}_m = \mathbf{X}'\tilde{\mathbf{T}}_m(\tilde{\mathbf{T}}_m'\tilde{\mathbf{T}}_m)^{-1}$, then $\tilde{\mathbf{D}}_m = \mathbf{W}_m(\mathbf{P}'_m\mathbf{W}_m)^{-1}$ yields weight vectors such that $\mathbf{X}\tilde{\mathbf{D}}_m = \tilde{\mathbf{T}}_m$. Because the $\tilde{\mathbf{t}}_j$ are orthogonal, the weights $\tilde{\mathbf{d}}_j$ are \mathbf{A} -orthogonal. Furthermore, $\|\tilde{\mathbf{d}}_j\|$ is determined by the way in which deflation is carried out and by the constraint $\|\mathbf{w}_j\| = 1$. Consequently, although $\|\tilde{\mathbf{d}}_1\| = 1$, the norms of subsequent $\tilde{\mathbf{d}}_j$ will not be equal to unity. However, it is straightforward to show that $\tilde{\mathbf{d}}_j'\mathbf{w}_j = 1$. By contrast with NIPALS, the SIMPLS algorithm of de Jong³ calculates the $\tilde{\mathbf{d}}_j$ directly under an arbitrary constraint on $\|\tilde{\mathbf{d}}_j\|$.

3.2 Polynomial Representation of $\hat{\boldsymbol{\beta}}_{PLS}^m$

Any vector \mathbf{s} in $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$ has a convenient representation in terms of a polynomial in \mathbf{A} . Thus, we can write

$$\begin{aligned}\mathbf{s} &= \mathbf{K}_m\boldsymbol{\pi}_m \\ &= \sum_{i=0}^{m-1} (\mathbf{A}^i\mathbf{b})\pi_i = \sum_{i=0}^{m-1} (\pi_i\mathbf{A}^i)\mathbf{b} \\ &= \pi(\mathbf{A})\mathbf{b}\end{aligned}$$

where $\pi(\xi) \equiv \sum \pi_i\xi^i$ is a polynomial of degree $< m$. A natural question to ask, therefore, is what is the polynomial representation of $\hat{\boldsymbol{\beta}}_{PLS}^m$? The answer, as Appendix A shows, is related to the characteristic polynomial of the sequence of tridiagonal matrices $\mathbf{T}_j = \mathbf{W}_j'\mathbf{A}\mathbf{W}_j$.

If we denote the characteristic polynomial of the tridiagonal matrix \mathbf{T}_m by $\chi_m(\xi)$, then $\chi_m(\xi) \equiv \det(\xi\mathbf{I} - \mathbf{T}_m) = \xi^m + \chi_{m,m-1}\xi^{m-1} + \dots + \chi_{m,1}\xi + \chi_{m,0}$. Furthermore, $\chi_{m,0} = (-1)^m \det(\mathbf{T}_m)$. Then, defining $\tilde{\chi}_m(\xi) \equiv \chi_m(\xi)/\chi_{m,0}$, we can write

$$\hat{\boldsymbol{\beta}}_{PLS}^m = \{\mathbf{I} - \tilde{\chi}_m(\mathbf{A})\}\mathbf{A}^-\mathbf{b} \quad (= \pi(\mathbf{A})\mathbf{b}) \quad (5)$$

$$= \{\mathbf{I} - \tilde{\chi}_m(\mathbf{A})\}\hat{\boldsymbol{\beta}}_{OLS} \quad (6)$$

where $\pi(\mathbf{A}) = \{\mathbf{I} - \tilde{\chi}_m(\mathbf{A})\}\mathbf{A}^-$ is a polynomial of order $m - 1$. Note that $\chi_m(\xi)$ is a monic polynomial (leading coefficient of one) but that $\tilde{\chi}_m(\xi)$ has been normalized so that $\tilde{\chi}(0) = 1$. It turns out that $\chi_m(\xi)$ and $\tilde{\chi}_m(\xi)$ have some useful minimizing characteristics that we will outline in Sections 4 and 5 and then exploit in Section 6.

3.3 Shrinkage and ‘Closeness’ of PLSR

By using the spectral decomposition of \mathbf{A} , we can rewrite the ordinary least squares estimator as $\hat{\boldsymbol{\beta}}_{OLS} = \sum_{i=1}^p \boldsymbol{\alpha}_i$, where $\boldsymbol{\alpha}_i = \gamma_i(\gamma_i'\mathbf{b})/\lambda_i$ is the component of $\hat{\boldsymbol{\beta}}_{OLS}$ along the i th eigenvector of \mathbf{A} . Using this result, we can rewrite the expression for $\hat{\boldsymbol{\beta}}_{PLS}^m$ as

$$\hat{\boldsymbol{\beta}}_{PLS}^m = \sum_{i=1}^p \{1 - \tilde{\chi}_m(\lambda_i)\}\boldsymbol{\alpha}_i \quad (7)$$

The scalar $f_m(\lambda_i) = 1 - \tilde{\chi}_m(\lambda_i)$ is the i th *shrinkage factor*,^{1,6} and it represents the extent to which PLSR shrinks (or expands) in the direction of the i th eigenvector of \mathbf{A} . As Frank and Friedman,⁶ and more recently, Butler and Denham,¹ show, PLSR shrinks in some directions and expands in others. Yet, like the ridge and principal component estimators, $\hat{\boldsymbol{\beta}}_{PLS}^m$ is a shrinkage estimator in that $\|\hat{\boldsymbol{\beta}}_{PLS}^m\| \leq \|\hat{\boldsymbol{\beta}}_{OLS}\|$. This result was proved geometrically by Goutis⁸ and algebraically by de Jong.⁴ In addition, the latter showed $\hat{\boldsymbol{\beta}}_{PLS}^m$ to be a strictly non-decreasing function of m . In Section 6, we will provide two additional proofs, one based on the correspondence of PLSR with conjugate gradients, the other on the properties of quadratic forms. We should also note that Butler and Denham¹ were able to deduce some of the properties of the polynomial $f_m(\xi)$, *without* knowing its relationship to $\chi_m(\xi)$. Some of their theorems are, however, easier to prove when we combine the expression for $f_m(\xi)$ with the fact that the zeros of $\chi_j(\xi)$, $j = 1, 2, \dots, p$ are strictly interlacing (see Section 4). Their work, in part, motivated our search for the relationship between PLSR, Lanczos, and CG.

The other property for which we will provide an alternative proof can be written as follows.² Let the vector of fitted values from OLS, PCR and PLSR be $\hat{\mathbf{y}}_{OLS} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, $\hat{\mathbf{y}}_{PCR}^m = \mathbf{X}\hat{\boldsymbol{\beta}}_{PCR}^m$ and $\hat{\mathbf{y}}_{PLS}^m = \mathbf{X}\hat{\boldsymbol{\beta}}_{PLS}^m$, respectively. Then, if $\mathcal{R}^2(\mathbf{y}, \hat{\mathbf{y}})$ denotes the squared correlation between the vectors \mathbf{y} and $\hat{\mathbf{y}}$,

$$\mathcal{R}^2(\hat{\mathbf{y}}_{OLS}, \hat{\mathbf{y}}_{PLS}^m) \geq \mathcal{R}^2(\hat{\mathbf{y}}_{OLS}, \hat{\mathbf{y}}_{PCR}^m)$$

when the principal components are taken in their ‘natural’ order, from largest to smallest variance. Alternatively, we can show that $\|\hat{\mathbf{y}}_{OLS} - \hat{\mathbf{y}}_{PLS}^m\| \leq \|\hat{\mathbf{y}}_{OLS} - \hat{\mathbf{y}}_{PCR}^m\|$ because both $\hat{\mathbf{y}}_{PCR}^m$ and $\hat{\mathbf{y}}_{PLS}^m$ are orthogonal projections onto an m -dimensional subspace. In other words, the vector of fitted values from PLSR is ‘closer’ to $\hat{\mathbf{y}}_{OLS}$ (and hence to \mathbf{y}) than its PCR counterpart.

4 Connections to Lanczos Methods

The appearance of \mathbf{K}_m and \mathbf{W}_m in expressions for the PLS estimator indicates the close connection of PLSR to Lanczos methods for estimating the extremal eigenvalues of a symmetric matrix. In the PLS literature, this connection has long been known.^{12,17–19,21}

The Lanczos method^{7,15,20} is a technique for approximating the extremal eigenvalues of large, sparse, symmetric matrices, \mathbf{A} . Given some starting vector \mathbf{b} , the procedure constructs a sequence of tridiagonal matrices $\mathbf{T}_j = \mathbf{W}_j^t \mathbf{A} \mathbf{W}_j$, $j = 1, 2, \dots, p$, where the columns of \mathbf{W}_j form the distinguished orthonormal basis of $\mathcal{K}_j(\mathbf{A}, \mathbf{b})$. They are sometimes known as *Lanczos vectors*, and they arise from a Gram-Schmidt orthonormalization of the first j vectors of the Krylov matrix \mathbf{K}_j . The usefulness of Lanczos methods lies in the fact that for large sparse \mathbf{A} , the extremal eigenvalues of \mathbf{T}_j converge to the extremal eigenvalues of \mathbf{A} well before $j = p$.

Below, we list some of the properties (in exact arithmetic) of the matrices and vectors generated during Lanczos and PLSR algorithms and in so doing make more explicit their connection. Except where explicit attribution has been made, most of these results may be found in Chapter 7 of Parlett.²⁰ Many of them are used in Appendix A to derive eq. (6).

1. \mathbf{K}_m can be written as

$$\mathbf{K}_m = \mathbf{W}_m \mathbf{R}_m^{-1} \tag{8}$$

where \mathbf{R}_m^{-1} is an $(m \times m)$ upper-triangular matrix with positive diagonal entries; hence, eq.(8) represents the ‘skinny’ (Golub and Van Loan,⁷ p. 217) QR factorization of \mathbf{K}_m . Moreover, \mathbf{R}_m^{-1} is the upper-triangular factor in the Cholesky decomposition of $\mathbf{K}_m^t \mathbf{K}_m$, and it can be written explicitly as

$$\mathbf{R}_m^{-1} = \|\mathbf{b}\|[\mathbf{i}_1, \mathbf{T}_m \mathbf{i}_1, \mathbf{T}_m^2 \mathbf{i}_1, \dots, \mathbf{T}_m^{m-1} \mathbf{i}_1] \quad (9)$$

where the notation \mathbf{i}_1 is used to denote the first column of the identity matrix of appropriate order.

2. \mathbf{T}_m is a symmetric tridiagonal matrix, and if \mathbf{K}_m is of full column rank, it is an *unreduced* tridiagonal matrix, that is, none of the supra- or sub-diagonal elements is zero. We write \mathbf{T}_m as

$$\mathbf{T}_m = \begin{bmatrix} \alpha_1 & \beta_1 & & \cdots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & \beta_{m-1} \\ 0 & \cdots & & \beta_{m-1} & \alpha_m \end{bmatrix}$$

Without loss of generality, we will assume that $\beta_i > 0$.

3. Let $\theta_i^{(m)}$ and $\mathbf{s}_i^{(m)}$, $i = 1, 2, \dots, m$ be the eigenvalues and unit norm eigenvectors, respectively, of \mathbf{T}_m . The $\theta_i^{(m)}$ are known as *Ritz values* and the vectors calculated as $\boldsymbol{\varphi}_i^{(m)} = \mathbf{W}_m \mathbf{s}_i^{(m)}$ as *Ritz vectors*. Then the Ritz pairs

$$(\theta_i^{(m)}, \boldsymbol{\varphi}_i^{(m)}), \quad i = 1, 2, \dots, m$$

constitute the best set of approximations to the eigenpairs (λ_i, γ_i) of \mathbf{A} that can be derived from $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$ (Parlett,²⁰ §11.4). (Occasionally, we will drop the superscript m where it is not required.)

4. For $m < p$, the Ritz vectors $\boldsymbol{\varphi}_i^{(m)}$ can be interpreted as the eigenvectors of \mathbf{A} restricted to $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$. So,

$$\mathbf{W}_m \mathbf{W}_m^t \mathbf{A} \boldsymbol{\varphi}_i^{(m)} = \theta_i^{(m)} \boldsymbol{\varphi}_i^{(m)}$$

5. In the sequence of matrices \mathbf{T}_j , $j = 1, 2, \dots, p$, \mathbf{T}_j is the leading principal $(j \times j)$ submatrix of \mathbf{T}_{j+1} . Hence, by the Cauchy interlace theorem, the eigenvalues $\theta_i^{(j)}$ of \mathbf{T}_j interlace those of \mathbf{T}_{j+1} . Moreover, the eigenvalues of \mathbf{T}_p are identical to those of \mathbf{A} .

6. By using the Ritz values and vectors, we can rewrite the expressions for $\hat{\boldsymbol{\beta}}_{PLS}^m$ into a form that is identical to (3), the expression for $\hat{\boldsymbol{\beta}}_{PCR}^m$. It is straightforward to show^{5,21} that

$$\hat{\boldsymbol{\beta}}_{PLS}^m = \sum_{i=1}^m (\theta_i^{(m)})^{-1} \boldsymbol{\varphi}_i^{(m)} \boldsymbol{\varphi}_i^{(m)t} \mathbf{b} \quad (10)$$

which is identical to the form given in Naes and Martens,¹⁹ who derived it without making an explicit connection to Lanczos.

7. The characteristic polynomials of the sequence of \mathbf{T}_j are sometimes known as *Lanczos polynomials*, and they have some useful properties:

- (a) **Orthogonality** (Parlett,²⁰ Theorem 7.8.1): Let $\mathbf{T}_p = \mathbf{S}\mathbf{\Theta}\mathbf{S}'$ represent the spectral decomposition of \mathbf{T}_p . Furthermore, let ϕ and ψ represent polynomials of degree less than p and define the inner product $\langle \phi, \psi \rangle$ to be

$$\langle \phi, \psi \rangle = \sum_{i=1}^p \omega_i \phi(\xi_i) \psi(\xi_i)$$

Then, if $\omega_i = s_{1i}^2$ where s_{1i} is the first element of the i th unit-norm eigenvector of \mathbf{T}_p ,

$$\langle \chi_j, \chi_k \rangle = \sum_{i=1}^p s_{1i}^2 \chi_j(\theta_i) \chi_k(\theta_i) = 0, \quad j \neq k \quad (11)$$

Using 3.–5. above, it is straightforward to show that this result can be written in matrix notation as $\mathbf{b}'\chi_j(\mathbf{A})\chi_k(\mathbf{A})\mathbf{b} = \mathbf{b}'\tilde{\chi}_j(\mathbf{A})\tilde{\chi}_k(\mathbf{A})\mathbf{b} = 0$ for $j \neq k$.

- (b) **Minimizing Property I** (Parlett,²⁰ Theorem 7.3.1): Let Π_j denote the set of monic polynomials of degree j . Then,

$$\|\chi_j(\mathbf{A})\mathbf{b}\| = \min\|\psi(\mathbf{A})\mathbf{b}\| \text{ over all } \psi \text{ in } \Pi_j \quad (12)$$

The quantity $\|\chi_m(\mathbf{A})\mathbf{b}\|$ is also the distance of $\mathbf{A}^m\mathbf{b}$ from $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$. Hence, as Greenbaum and Trefethen⁹ point out, the symmetric Lanczos procedure can be seen as one which finds the best approximation of $\mathbf{A}^m\mathbf{b}$ in $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$, and it does so by constructing the tridiagonal matrix \mathbf{T}_m of which χ_m is the characteristic polynomial. The polynomial $\tilde{\chi}_j(\xi)$ also has a minimizing property, but we defer discussing it to the next Section.

5 Conjugate Gradient Method

The method of conjugate gradients (CG) is a means of solving a system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ arising out of the minimization of the quadratic function $\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}'\mathbf{A}\mathbf{x} - \mathbf{b}'\mathbf{x}$, where \mathbf{A} is positive semi-definite. The connection between CG and Lanczos was known long ago, as far back as the original papers on the topics themselves.^{11,15} Minimization of $\phi(\mathbf{x})$ takes place over a sequence of subspaces – Krylov subspaces \mathcal{K}_j as it turns out – of increasing dimensionality. The CG method has a number of desirable properties, including the fact that given an arbitrary initial estimate \mathbf{x}_0 , the sequence of iterates \mathbf{x}_j , $j = 1, 2, \dots, p$, that it generates converges (in exact arithmetic) to the solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ in p steps. It does so by successively minimizing $\phi(\mathbf{x})$ along p directions \mathbf{d}_j , $j = 0, 1, \dots, p-1$, that are *conjugate* (orthogonal) with respect to \mathbf{A} .

Our interest in CG stems from the fact that when $\mathbf{A}\mathbf{x} = \mathbf{b}$ represents the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$, the conjugate gradient iterates \mathbf{x}_j are identical to the PLS estimates $\hat{\boldsymbol{\beta}}_{PLS}^j$, $j = 1, 2, \dots, p$, when the CG algorithm starts with initial estimate $\mathbf{x}_0 = \mathbf{0}$. It should also be noted that when \mathbf{A} is positive semi-definite, CG converges to $\mathbf{A}^{-1}\mathbf{b}$ only when $\mathbf{b} \in \text{span}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_p)$, the eigenvectors of \mathbf{A} associated with non-zero eigenvalues. This, of course, is true for the normal equations.

The canonical algorithm of Hestenes and Stiefel¹¹ can be written as follows:

1. Initialization: $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{d}_0 = \mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0 = \mathbf{b}$
2. $a_i = \frac{\mathbf{d}_i' \mathbf{r}_i}{\mathbf{d}_i' \mathbf{A} \mathbf{d}_i}$
3. $\mathbf{x}_{i+1} = \mathbf{x}_i + a_i \mathbf{d}_i$
4. $\mathbf{r}_{i+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{i+1} (= \mathbf{r}_i - a_i \mathbf{A} \mathbf{d}_i)$
5. $b_i = -\frac{\mathbf{r}_{i+1}' \mathbf{A} \mathbf{d}_i}{\mathbf{d}_i' \mathbf{A} \mathbf{d}_i}$
6. $\mathbf{d}_{i+1} = \mathbf{r}_{i+1} + b_i \mathbf{d}_i$

Although more efficient implementations use alternative expressions in Steps 2, 4, and 5, the sequence shown above shows more clearly the mechanics of the CG algorithm. The terminology of CG differs somewhat from its statistical counterpart: for example, the \mathbf{r}_i are known as *residuals*, while *errors* are calculated as $\mathbf{e}_i = \mathbf{x}_i - \mathbf{x} = \mathbf{x}_i - \mathbf{x}_p$.

In the CG method, \mathbf{x}_{i+1} is obtained as the sum of the current estimate \mathbf{x}_i and a vector $a_i \mathbf{d}_i$ along the i th search direction. How far along the latter do we proceed to obtain the next estimate? It is straightforward to show that the value of a_i in Step 2 can be obtained by minimizing the function $\phi(\mathbf{x}_{i+1}) = \phi(\mathbf{x}_i + a \mathbf{d}_i)$ with respect to a . Hence, CG is a line search method that chooses the value of a_i that minimizes ϕ along the line $\mathbf{x}_i + a \mathbf{d}_i$.

The initial search direction \mathbf{d}_0 is identical to the initial residual, but how should subsequent directions be chosen? For a number of reasons⁷ it is important to insist that the direction vectors be *A-conjugate*, that is,

$$\mathbf{d}_i' \mathbf{A} \mathbf{d}_j = 0, \quad i \neq j \quad (13)$$

Hence, \mathbf{d}_{i+1} must lie in the space orthogonal to $\text{span}(\mathbf{A} \mathbf{D}_i)$, where \mathbf{D}_i denotes the matrix whose columns are the first i direction vectors. Furthermore, since the objective is to reduce the size of the residual, one way of doing so is to generate \mathbf{d}_{i+1} by projecting \mathbf{r}_{i+1} onto $\text{span}(\mathbf{A} \mathbf{D}_i)^\perp$. Indeed, it turns out to be sufficient to project onto $\text{span}(\mathbf{A} \mathbf{d}_i)^\perp$. By substituting the expression for b_i (Step 5) into the expression in Step 6, we obtain

$$\mathbf{d}_{i+1} = (\mathbf{I} - \mathbf{d}_i (\mathbf{d}_i' \mathbf{A} \mathbf{d}_i)^{-1} \mathbf{d}_i' \mathbf{A}) \mathbf{r}_{i+1} \quad (14)$$

which shows \mathbf{d}_{i+1} to be the projection of \mathbf{r}_{i+1} onto $\text{span}(\mathbf{A} \mathbf{d}_i)^\perp$ along \mathbf{d}_i ; hence, it is an *oblique* projection. Projection along \mathbf{d}_i ensures that \mathbf{d}_{i+1} can be written in terms of \mathbf{d}_i only; \mathbf{d}_j , $j < i$ are not required and storage requirements are reduced as a consequence.

There are several important properties of, and relationships among, the vectors generated during the CG algorithm that have a direct bearing on PLSR. We will neither prove nor summarize them all, only those that are required for our purposes. Exhaustive summaries can be found in the original paper of Hestenes and Stiefel¹¹ and elsewhere.^{7,23}

1. For $j < p$,

$$\begin{aligned} \text{span}(\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{j-1}) &= \text{span}(\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_{j-1}) \\ &= \text{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{j-1}\mathbf{b}) \equiv \mathcal{K}_j(\mathbf{A}, \mathbf{b}) \end{aligned} \quad (15)$$

Hence, the residuals and the conjugate directions span the same Krylov subspace $\mathcal{K}_j(\mathbf{A}, \mathbf{b})$.

2. By induction, it is straightforward to show that

$$\mathbf{x}_i = \sum_{j=0}^{i-1} \mathbf{d}'_j (\mathbf{d}'_j \mathbf{A} \mathbf{d}_j)^{-1} \mathbf{d}'_j \mathbf{b} \quad (16)$$

Comparing this with eq. (4) and using the results in 1. above, we can see that $\hat{\boldsymbol{\beta}}_{PLS}^m$ and \mathbf{x}_m are identical. As a consequence, we can also write

$$\mathbf{x}_j = \{\mathbf{I} - \tilde{\chi}_j(\mathbf{A})\} \mathbf{x} \quad (17)$$

3. The residuals $\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_{p-1}$ are orthogonal so that $\mathbf{r}'_j \mathbf{r}_k = 0, j \neq k$. This result is often proved by induction; alternatively, we can make use of the orthogonality of the polynomials $\tilde{\chi}_j$. The i th residual can be written as $\mathbf{r}_i = \mathbf{b} - \mathbf{A} \{\mathbf{I} - \tilde{\chi}_i(\mathbf{A})\} \mathbf{x} = \tilde{\chi}_i(\mathbf{A}) \mathbf{b}$, and hence the orthogonality of the residuals follows from the orthogonality of the polynomials as defined in Section 4, 7(a).

4. As a consequence of 1. and 3. the residuals are the Lanczos vectors except for a scale factor. Thus, $\mathbf{w}_i = \pm \mathbf{r}_{i-1} / \|\mathbf{r}_{i-1}\|$. Furthermore, because the direction vectors also span $\mathcal{K}_j(\mathbf{A}, \mathbf{b})$ and are \mathbf{A} -conjugate, the \mathbf{d}_i are essentially the weight vectors $\tilde{\mathbf{d}}_{i+1}$ (up to a scale factor) generated directly by the SIMPLS (cf Section 3.1) algorithm for PLSR. Because $\mathbf{d}'_i \mathbf{r}_i = \|\mathbf{r}_i\|^2$ (Hestenes and Stiefel,¹¹ Theorem 5.1 or see Proof I, §6.1), $\mathbf{d}_i / \|\mathbf{r}_i\|$ yields weight vectors that have identical scaling to those obtained by the NIPALS algorithm.

5. **Minimizing Property II** (van der Sluis and van der Vorst,²³ Property 2.8): An important property of CG estimates \mathbf{x}_j is that they minimize $\|\mathbf{x} - \mathbf{x}_j\|_{\mathbf{A}}$, the error norm with respect to \mathbf{A} , over $\mathcal{K}_j(\mathbf{A}, \mathbf{b})$. Using eq. (17) the errors can be rewritten as $\mathbf{e}_j = \mathbf{x} - \mathbf{x}_j = \mathbf{A}^{-1} \tilde{\chi}_j(\mathbf{A}) \mathbf{b}$. Hence, an equivalent statement can be made in terms of the polynomial $\tilde{\chi}(\xi)$. Let $\tilde{\Pi}_j$ denote the set of polynomials of degree j whose constant term is one. Then,

$$\|\tilde{\chi}_j(\mathbf{A}) \mathbf{b}\|_{\mathbf{A}^{-1}} = \min \|\psi(\mathbf{A}) \mathbf{b}\|_{\mathbf{A}^{-1}} \text{ over all } \psi \text{ in } \tilde{\Pi}_j \quad (18)$$

The quantity $\|\tilde{\chi}_j(\mathbf{A}) \mathbf{b}\|_{\mathbf{A}^{-1}}$, or alternatively, $\|\tilde{\chi}_j(\mathbf{A}) \mathbf{x}\|_{\mathbf{A}}$, has a nice interpretation: it is the distance of $\mathbf{A}^{-1/2} \mathbf{b}$ from the subspace $\mathbf{A}^{1/2} \mathcal{K}_j(\mathbf{A}, \mathbf{b})$. Hence, both CG and PLSR can be seen as finding the best approximation of $\mathbf{A}^{-1/2} \mathbf{b}$ in $\mathbf{A}^{1/2} \mathcal{K}_j(\mathbf{A}, \mathbf{b})$.

6 Alternative Proofs

6.1 Shrinkage of $\hat{\boldsymbol{\beta}}_{PLS}^m$

We provide two proofs of the shrinkage of $\hat{\boldsymbol{\beta}}_{PLS}^m$. The first relies explicitly on the connection to conjugate gradients, the second on the properties of certain quadratic forms.

Proof I The original paper of Hestenes and Stiefel¹¹ does not consider the norms of the iterates, though it contains all of the results necessary to prove that $\|\mathbf{x}_j\|$ is strictly nondecreasing. Indeed, Steihaug²² and Kaasschieter¹⁴ use the results in Hestenes and Stiefel to provide similar proofs albeit in somewhat different contexts. We sketch out below the essential elements of their proofs.

The starting point is Step 3 in the CG algorithm. The squared norm of \mathbf{x}_{j+1} can be written as

$$\|\mathbf{x}_{j+1}\|^2 = \|\mathbf{x}_j\|^2 + 2a_j^2\|\mathbf{d}_j\|^2 + a_j\mathbf{d}'_j\mathbf{x}_j \quad (19)$$

and hence to prove that $\|\mathbf{x}_{j+1}\|^2 \geq \|\mathbf{x}_j\|^2$ we need only show that $a_j\mathbf{d}'_j\mathbf{x}_j \geq 0$. Indeed, it is sufficient to show that $\mathbf{d}'_j\mathbf{x}_j \geq 0$ because it turns out that $a_j > 0$. To see this, we first multiply both sides of the expression in Step 6 by \mathbf{r}'_{j+1} to get $\mathbf{r}'_{j+1}\mathbf{d}_{j+1} = \mathbf{r}'_{j+1}\mathbf{r}_{j+1} + b_j\mathbf{r}'_{j+1}\mathbf{d}_j$. Combining eq. (15) and the fact that the residuals are orthogonal, \mathbf{r}_{j+1} must be orthogonal to \mathbf{d}_j , from which it follows that $a_j = \mathbf{d}'_j\mathbf{r}_j/\|\mathbf{d}_j\|_{\mathbf{A}}^2 = \|\mathbf{r}_j\|^2/\|\mathbf{d}_j\|_{\mathbf{A}}^2 > 0$.

By repeated application of Step 3, we can show that $\mathbf{x}_j = \sum_{i=0}^j a_i\mathbf{d}_i$ so that $\mathbf{d}'_j\mathbf{x}_j = \sum_{i=0}^j a_i\mathbf{d}'_j\mathbf{d}_i$ and hence we now need to show that $\mathbf{d}'_j\mathbf{d}_i > 0$, $i \neq j$. A transparent way of doing so is to first express the direction vectors in terms of the residuals. By repeated substitution of Step 6, $\mathbf{d}_j = \mathbf{r}_j + \sum_{i=0}^{j-1} \left(\prod_{k=i}^{j-1} b_k \right) \mathbf{r}_i$. Thus, to complete the proof, we require $b_j > 0$.

Using the expression in parentheses in Step 4, $\mathbf{A}\mathbf{d}_i$ can be written as $a_i^{-1}(\mathbf{r}_i - \mathbf{r}_{i+1})$. Substituting this into the numerator of the expression for b_j yields $b_j = -\mathbf{r}'_{j+1}(\mathbf{r}_j - \mathbf{r}_{j+1})/a_j\|\mathbf{d}_j\|_{\mathbf{A}}^2 = \|\mathbf{r}_{j+1}\|^2/a_j\|\mathbf{d}_j\|_{\mathbf{A}}^2 > 0$, which completes the result. Using some of the intermediate results above, a further simplification leads to $b_j = \|\mathbf{r}_{j+1}\|^2/\|\mathbf{r}_j\|^2$.

Proof II We pointed out in Section 5, CG estimates minimize $\|\mathbf{x}_m - \mathbf{x}\|_{\mathbf{A}}$ over $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$. By defining $\tilde{\mathbf{x}} = \mathbf{A}^{1/2}\mathbf{x}$ and $\tilde{\mathbf{x}}_m = \mathbf{A}^{1/2}\mathbf{x}_m$ we can restate the problem that CG and PLSR solve as

$$\min_{\tilde{\mathbf{x}}_m \in \tilde{\mathcal{K}}_m(\mathbf{A}, \mathbf{b})} \{ (\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_m)'(\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_m) \}$$

where $\tilde{\mathcal{K}}_m(\mathbf{A}, \mathbf{b}) = \mathbf{A}^{1/2}\mathcal{K}_m(\mathbf{A}, \mathbf{b})$ and $\tilde{\mathcal{K}}_1 \subset \tilde{\mathcal{K}}_2 \subset \dots \subset \tilde{\mathcal{K}}_m$. Consequently, $\tilde{\mathbf{x}}_m$ is the orthogonal projection of $\tilde{\mathbf{x}}$ onto $\tilde{\mathcal{K}}_m(\mathbf{A}, \mathbf{b})$; hence, for $m \leq p$, the following properties are straightforward to establish:

1. $\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_1\| \geq \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_2\| \geq \dots \geq \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_m\|$
2. $\tilde{\mathbf{x}}'_1\tilde{\mathbf{x}}_1 \leq \tilde{\mathbf{x}}'_2\tilde{\mathbf{x}}_2 \leq \dots \leq \tilde{\mathbf{x}}'_m\tilde{\mathbf{x}}_m$
3. $\tilde{\mathbf{x}}'_j\tilde{\mathbf{x}} = \|\tilde{\mathbf{x}}_j\|^2$ for $j \leq p$
4. $\tilde{\mathbf{x}}'_j\tilde{\mathbf{x}}_k = \|\tilde{\mathbf{x}}_j\|^2$ for $j \leq k \leq p$

Recall from eq. (14) that the residuals $\mathbf{r}_0, \dots, \mathbf{r}_{m-1}$ span $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$. Hence, \mathbf{x}_m can be written as

$$\mathbf{x}_m = \sum_{j=0}^{m-1} \mathbf{r}_j(\mathbf{r}'_j\mathbf{r}_j)^{-1}\mathbf{r}'_j\mathbf{x}_m$$

from which it follows that

$$\|\mathbf{x}_m\|^2 = \sum_{j=0}^{m-1} (\mathbf{x}'_m\mathbf{r}_j)^2/(\mathbf{r}'_j\mathbf{r}_j) \quad (20)$$

Using the fact that $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{x}_j = \mathbf{A}(\mathbf{x} - \mathbf{x}_j)$, we can insert this into eq. (20) so that we have

$$\begin{aligned}
\|\mathbf{x}_m\|^2 &= \sum_{j=0}^{m-1} \{ \mathbf{x}'_m \mathbf{A}(\mathbf{x} - \mathbf{x}_j) \}^2 / (\mathbf{r}'_j \mathbf{r}_j) \\
&= \sum_{j=0}^{m-1} (\mathbf{x}'_m \mathbf{A}\mathbf{x} - \mathbf{x}'_m \mathbf{A}\mathbf{x}_j)^2 / (\mathbf{r}'_j \mathbf{r}_j) \\
&= \sum_{j=0}^{m-1} \{ (\tilde{\mathbf{x}}'_m \tilde{\mathbf{x}} - \tilde{\mathbf{x}}'_m \tilde{\mathbf{x}}_j) / \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_j\|_{\mathbf{A}} \}^2 \\
&= \sum_{j=0}^{m-1} \left\{ (\tilde{\mathbf{x}}'_m \tilde{\mathbf{x}}_m - \tilde{\mathbf{x}}'_j \tilde{\mathbf{x}}_j) / \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_j\|_{\mathbf{A}} \right\}^2 \tag{21}
\end{aligned}$$

The last simplification occurs as a result of properties 3 and 4 above. Using property 2, we can write that $\tilde{\mathbf{x}}'_{m+1} \tilde{\mathbf{x}}_{m+1} \geq \tilde{\mathbf{x}}'_m \tilde{\mathbf{x}}_m$. Then, on using this inequality in eq. (21),

$$\begin{aligned}
\|\mathbf{x}_m\|^2 &\leq \sum_{j=0}^{m-1} \left\{ (\tilde{\mathbf{x}}'_{m+1} \tilde{\mathbf{x}}_{m+1} - \tilde{\mathbf{x}}'_j \tilde{\mathbf{x}}_j) / \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_j\|_{\mathbf{A}} \right\}^2 \\
&\leq \sum_{j=0}^m \left\{ (\tilde{\mathbf{x}}'_{m+1} \tilde{\mathbf{x}}_{m+1} - \tilde{\mathbf{x}}'_j \tilde{\mathbf{x}}_j) / \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_j\|_{\mathbf{A}} \right\}^2 \tag{22}
\end{aligned}$$

By comparing it with eq. (21), we can see that the right-hand side of the inequality in eq. (22) is equal to $\|\mathbf{x}_{m+1}\|^2$ and hence that $\|\mathbf{x}_{m+1}\|^2 \geq \|\mathbf{x}_m\|^2$.

6.2 PLSR Fits Closer than PCR

The proof that PLSR fits closer than PCR (when the principal components are taken in their natural order) relies on the minimizing property of the polynomial $\tilde{\chi}_j(\xi)$ (Minimizing Property II), and it is somewhat more compact than the proof of de Jong.²

We need to show that $\|\hat{\mathbf{y}}_{OLS} - \hat{\mathbf{y}}_{PLS}^m\| \leq \|\hat{\mathbf{y}}_{OLS} - \hat{\mathbf{y}}_{PCR}^m\|$, which, in the notation introduced in Proof II above, amounts to showing $\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_m\| \leq \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_m^*\|$, where $\mathbf{x}_m^* = \hat{\boldsymbol{\beta}}_{PCR}^m$. Now after some manipulation of eq. (3), \mathbf{x}_m^* can be written as $\mathbf{x}_m^* = \Gamma_m \Gamma'_m \mathbf{x}$ and hence $\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_m^*\| = \|(\mathbf{I} - \Gamma_m \Gamma'_m) \tilde{\mathbf{x}}\|$. Furthermore, $\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_m\| = \|\tilde{\chi}_m(\mathbf{A}) \tilde{\mathbf{x}}\|$, which is the quantity minimized by CG or PLSR over $\tilde{\Pi}_m$, the set of polynomials of degree m whose constant term is one. The proof to follow consists of two steps: first, we show that $\|(\mathbf{I} - \Gamma_m \Gamma'_m) \mathbf{u}\| \geq \|\tilde{\zeta}_m(\mathbf{A}) \mathbf{u}\|$, where $\tilde{\zeta}_m \in \tilde{\Pi}_m$ and \mathbf{u} is an arbitrary vector. Then, because of the minimizing property of $\tilde{\chi}_m$, we conclude that $\|\tilde{\chi}_m(\mathbf{A}) \tilde{\mathbf{x}}\| \leq \|\tilde{\zeta}_m(\mathbf{A}) \tilde{\mathbf{x}}\| \leq \|(\mathbf{I} - \Gamma_m \Gamma'_m) \tilde{\mathbf{x}}\|$ and hence that PLSR fits closer than PCR.

The expression $(\mathbf{I} - \Gamma_m \Gamma'_m) \mathbf{u}$ represents the orthogonal projection of \mathbf{u} onto the space orthogonal to the first m eigenvectors $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_m$. Hence,

$$\|(\mathbf{I} - \Gamma_m \Gamma'_m) \mathbf{u}\|^2 = \sum_{i=m+1}^p (\mathbf{u}' \boldsymbol{\gamma}_i)^2 \tag{23}$$

Consider now the normalized characteristic polynomial of the diagonal matrix $\boldsymbol{\Lambda}_m (= \Gamma'_m \mathbf{A} \Gamma_m)$, which consists of the m largest eigenvalues of \mathbf{A} arranged in decreasing order:

$$\tilde{\zeta}_m(\lambda) \equiv \left((-1)^m \prod_{i=1}^m \lambda_i \right)^{-1} \det(\lambda \mathbf{I} - \boldsymbol{\Lambda}_m) = (-1)^m \frac{(\lambda - \lambda_1)}{\lambda_1} \frac{(\lambda - \lambda_2)}{\lambda_2} \dots \frac{(\lambda - \lambda_m)}{\lambda_m} \tag{24}$$

Clearly, $\tilde{\zeta}_m(\lambda) \in \tilde{\Pi}_m$. Using the spectral decomposition of \mathbf{A} , we can write $\tilde{\zeta}_m(\mathbf{A})\mathbf{u}$ as $\Gamma\tilde{\zeta}_m(\mathbf{A})\Gamma'\mathbf{u}$, from which it follows that

$$\|\tilde{\zeta}_m(\mathbf{A})\mathbf{u}\|^2 = \sum_{i=m+1}^p \tilde{\zeta}_m^2(\lambda_i)(\mathbf{u}'\boldsymbol{\gamma}_i)^2 \quad (25)$$

The summation is from $m+1$ to p because, as eq. (24) shows, $\tilde{\zeta}_m(\lambda_i) = 0$ for $i = 1, 2, \dots, m$. The difference between the expressions in eqs. (23) and (25) is the factor $\tilde{\zeta}_m^2(\lambda_i)$. Because of the ordering of the eigenvalues of \mathbf{A} , that is, $\lambda_1 > \lambda_2 > \dots > \lambda_p$, it is easy to see that $\tilde{\zeta}_m^2(\lambda_i) < 1$ for $i = m+1, \dots, p$ and consequently that $\|\tilde{\zeta}_m(\mathbf{A})\mathbf{u}\|^2 \leq \|(\mathbf{I} - \Gamma_m\Gamma'_m)\mathbf{u}\|^2$. Finally, after substituting $\tilde{\mathbf{x}}$ for \mathbf{u} , we have that

$$\|\tilde{\chi}_m(\mathbf{A})\tilde{\mathbf{x}}\| \leq \|\tilde{\zeta}_m(\mathbf{A})\tilde{\mathbf{x}}\| \leq \|(\mathbf{I} - \Gamma_m\Gamma'_m)\tilde{\mathbf{x}}\| \quad (26)$$

and hence that PLSR fits closer than PCR.

Acknowledgments

We would like to thank Sijmen de Jong of Unilever, Vlaardingen, for suggesting a much simpler proof of eq. (6) and for his close reading of the manuscript and many suggestions for improving it. We also received helpful pointers to the numerical analysis literature from several members of the NA-Digest (http://www.netlib.org/na-net/na_home.html).

References

- [1] N. BUTLER AND M. DENHAM, *The peculiar shrinkage properties of partial least squares regression*, Tech. Rep. 99/7, Department of Applied Statistics, University of Reading, July 1999.
- [2] S. DE JONG, *PLS fits closer than PCR*, *J. Chemomet.*, 7 (1993), pp. 551–557.
- [3] ———, *SIMPLS: An alternative approach to partial least squares regression*, *Chem. Intell. Lab. Systems*, 18 (1993), pp. 251–263.
- [4] ———, *PLS shrinks*, *J. Chemomet.*, 9 (1995), pp. 323–326.
- [5] S. DE JONG AND A. PHATAK, *Partial least squares regression*, in *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, S. Van Huffel, ed., SIAM, Philadelphia, 1997, pp. 25–36.
- [6] I. FRANK AND J. FRIEDMAN, *A statistical view of some chemometrics regression tools*, *Technometrics*, 35 (1993), pp. 109–135.
- [7] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, 1989.
- [8] C. GOUTIS, *Partial least squares algorithm yields shrinkage estimators*, *Ann. Stat.*, 24 (1996), pp. 816–824.

- [9] A. GREENBAUM AND L. N. TREFETHEN, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.
- [10] I. HELLAND, *On the structure of partial least squares regression*, Commun. Statist. - Simula., 17 (1988), pp. 581–607.
- [11] M. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. NBS, 49 (1952), pp. 409–436.
- [12] A. HÖSKULDSSON, *PLS regression methods*, J. Chemomet., 2 (1988), pp. 211–228.
- [13] I. JOLLIFFE, *Principal Component Analysis*, Springer Verlag, New York, 1986.
- [14] E. KAASSCHIETER, *A practical termination criterion for the conjugate gradient method*, BIT, 28 (1988), pp. 308–322.
- [15] C. LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. NBS, 45 (1950), pp. 225–280.
- [16] J. MAGNUS AND H. NEUDECKER, *Matrix Differential Calculus With Applications in Statistics and Econometrics*, Wiley, Chichester, 1988.
- [17] R. MANNE, *Analysis of two partial-least-squares algorithms for multivariate calibration*, Chem. Intell. Lab. Systems, 2 (1987), pp. 187–197.
- [18] R. MARBACH AND H. HEISE, *Calibration modeling by partial least-squares and principal component regression and its optimization using an improved leverage correction for prediction testing*, Chem. Intell. Lab. Systems, 9 (1990), pp. 45–63.
- [19] T. NAES AND H. MARTENS, *Comparison of prediction methods for multicollinear data*, Commun. Statist. – Simula. Computa., 14 (1985), pp. 545–576.
- [20] B. PARLETT, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998.
- [21] A. PHATAK, *Evaluation of Some Multivariate Methods and their Applications in Chemical Engineering*, PhD thesis, University of Waterloo, 1993.
- [22] T. STEIHAUG, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
- [23] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.
- [24] S. WOLD, A. RUHE, H. WOLD, AND W. DUNN, *The collinearity problem in linear regression. The PLS approach to generalized inverses*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 735–743.

A Polynomial Representation of $\hat{\beta}_{PLS}^m$

We provide two proofs of eq. (6). The first one is direct algebraic simplification of eq. (4), and is somewhat lengthy but is interesting for the use it makes of a number of interesting results on tridiagonal matrices. The second one is much more compact and was suggested to us by Sijmen de Jong.

A.1 Proof I

Using eq. (8), the QR decomposition of $\mathbf{K}_m (= \mathbf{W}_m \mathbf{R}_m^{-1})$, another way of writing eq. (refeq:bpls) is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{PLS}^m &= \mathbf{K}_m \mathbf{R}_m \mathbf{T}_m^{-1} \mathbf{W}_m' \mathbf{b} \\ &= \mathbf{K}_m \left(\|\mathbf{b}\| \mathbf{R}_m \mathbf{T}_m^{-1} \mathbf{i}_1 \right)\end{aligned}\quad (27)$$

where the simplification occurs because $\mathbf{W}_m' \mathbf{b} = \|\mathbf{b}\| \mathbf{W}_m' \mathbf{w}_1 = \|\mathbf{b}\| \mathbf{i}_1$. Hence the coefficients in the polynomial representation (see Section 3.2) $\hat{\boldsymbol{\beta}}_{PLS}^m = \boldsymbol{\pi}(\mathbf{A}) \mathbf{b}$ are given by the elements of the vector $\boldsymbol{\pi}_m = \|\mathbf{b}\| \mathbf{R}_m \mathbf{T}_m^{-1} \mathbf{i}_1$. In simplifying this expression below, we make heavy use of the results on tridiagonal matrices and Lanczos methods in Chapters 7, 12, and 13 of Parlett,²⁰ so we will not explicitly cite each result that we employ. Moreover, most of the notation and ideas used here are first introduced in Section 4 of the report, so it should be read first before attempting to follow the derivation.

The matrix \mathbf{R}_m in eq. (27) is also upper triangular, and it can be written as

$$\mathbf{R}_m = \mathbf{U}_m \boldsymbol{\Delta}_m \quad (28)$$

where \mathbf{U}_m is upper triangular with ones on the diagonal and $\boldsymbol{\Delta}_m$ is a diagonal matrix with elements given by $1/\text{diag}(\mathbf{R}_m^{-1})$. To begin with then, what are the diagonal elements of \mathbf{R}_m^{-1} ? If we recall from eq. (9) the expression for \mathbf{R}_m^{-1} , it is easy to show that the last non-zero element of $\mathbf{T}_m^j \mathbf{i}_1$ is the product of $\beta_1 \beta_2 \cdots \beta_j$ and is in row $(j+1)$. So, the diagonal elements of \mathbf{R}_m^{-1} are given by $\|\mathbf{b}\| (1, \beta_1, \beta_1 \beta_2, \dots, \beta_1 \beta_2 \cdots \beta_{m-1})$ and therefore,

$$\boldsymbol{\Delta}_m = \|\mathbf{b}\|^{-1} \text{diag}(1, \beta_1, \beta_1 \beta_2, \dots, \beta_1 \beta_2 \cdots \beta_{m-1})^{-1} \quad (29)$$

Furthermore, the j th column of \mathbf{U}_m contains the coefficients of the $(j-1)$ th *Lanczos polynomial* $\chi_{j-1}(\xi)$, which is defined as $\chi_j(\xi) \equiv \det(\xi \mathbf{I} - \mathbf{T}_j)$, $j = 1, 2, \dots, m$, where \mathbf{T}_j denotes the leading principal $(j \times j)$ submatrix of \mathbf{T}_m . For convenience, we let $\chi_0(\xi) = 1$. Now if we substitute eq. (28) into the expression for $\boldsymbol{\pi}_m$, we have

$$\boldsymbol{\pi}_m = \mathbf{U}_m \tilde{\boldsymbol{\Delta}}_m \mathbf{T}_m^{-1} \mathbf{i}_1$$

where $\tilde{\boldsymbol{\Delta}}_m = \|\mathbf{b}\| \boldsymbol{\Delta}_m$.

To continue simplifying the expression for $\boldsymbol{\pi}_m$, we need an explicit expression for the eigenvectors of \mathbf{T}_m^{-1} (hence of \mathbf{T}_m). Again, the answer involves Lanczos polynomials. It turns out that if $\theta_i^{(m)}$ is an eigenvector of \mathbf{T}_m , the associated eigenvector is

$$[1, \chi_1(\theta_i^{(m)})/\beta_1, \dots, \chi_{m-1}(\theta_i^{(m)})/(\beta_1 \beta_2 \cdots \beta_{m-1})]'$$

The proof relies on the well-known (see, for example, Parlett [20, §7.8] three-term recurrence for $\{\chi_j(\xi)\}$. Armed with the above result, it is straightforward to show that the *rows* of $\mathbf{V}_\theta' \mathbf{U}_m \tilde{\boldsymbol{\Delta}}_m$ are eigenvectors of \mathbf{T}_m^{-1} or \mathbf{T}_m , where the matrix \mathbf{V}_θ is a Vandermonde matrix of the form

$$\mathbf{V}_\theta = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \theta_1 & \theta_2 & \cdots & \theta_m \\ \vdots & \vdots & & \vdots \\ \theta_1^{m-1} & \theta_2^{m-1} & \cdots & \theta_m^{m-1} \end{bmatrix}$$

Here, we have dropped the superscript (m) that denotes that the θ_i are eigenvalues of \mathbf{T}_m to avoid clutter. Hence, we can now write

$$\begin{aligned}\boldsymbol{\pi}_m &= \mathbf{V}'_{\theta}{}^{-1} \mathbf{V}'_{\theta} \mathbf{U}_m \tilde{\boldsymbol{\Delta}}_m \mathbf{T}_m^{-1} \mathbf{i}_1 \\ &= \mathbf{V}'_{\theta}{}^{-1} \boldsymbol{\Theta}^{-1} \mathbf{V}'_{\theta} \mathbf{U}_m \tilde{\boldsymbol{\Delta}}_m \mathbf{i}_1 \\ &= \mathbf{V}'_{\theta}{}^{-1} \boldsymbol{\Theta}^{-1} \mathbf{V}'_{\theta} \mathbf{i}_1\end{aligned}$$

The matrix $\boldsymbol{\Theta}$ is diagonal and contains the m eigenvalues of \mathbf{T}_m . The last expression arises out of the fact that post-multiplication of a matrix by \mathbf{i}_1 picks out its first column and that the first column of $\mathbf{U}_m \tilde{\boldsymbol{\Delta}}_m$ is also \mathbf{i}_1 .

The characteristic polynomial of \mathbf{T}_m is $\chi_m(\xi)$ and hence when it is evaluated at the eigenvalues θ_i of \mathbf{T}_m , $\chi_m(\theta_i) = 0$. Its coefficients are denoted by $\chi_{m,j}$, $j = 0, 1, \dots, m$, and because of the definition of $\chi_m(\xi)$, $\chi_{m,m} = 1$. Then it is straightforward to verify that the matrix

$$\mathcal{C}_m = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -\chi_{m,0}/\chi_{m,m} \\ 1 & 0 & 0 & \cdots & 0 & -\chi_{m,1}/\chi_{m,m} \\ 0 & 1 & 0 & \cdots & 0 & -\chi_{m,2}/\chi_{m,m} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\chi_{m,m-1}/\chi_{m,m} \end{bmatrix}$$

has the same characteristic polynomial as \mathbf{T}_m and that it satisfies

$$\mathcal{C}'_m \mathbf{V}_{\theta} = \mathbf{V}_{\theta} \boldsymbol{\Theta} \quad (30)$$

The matrix \mathcal{C}_m is known as the *companion matrix* of the characteristic polynomial of \mathbf{T}_m . Rearranging eq. (30) yields $\mathcal{C}_m^{-1} \mathbf{V}'_{\theta}{}^{-1} = \mathbf{V}'_{\theta}{}^{-1} \boldsymbol{\Theta}^{-1}$ and inserting this into the previous expression for $\boldsymbol{\pi}_m$ yields

$$\begin{aligned}\boldsymbol{\pi}_m &= \mathcal{C}_m^{-1} \mathbf{V}'_{\theta}{}^{-1} \mathbf{V}'_{\theta} \mathbf{i}_1 \\ &= \mathcal{C}_m^{-1} \mathbf{i}_1\end{aligned}$$

So, after all that, we see that the elements of $\boldsymbol{\pi}_m$ are given by the elements in the first column of \mathcal{C}_m^{-1} . The inverse of \mathcal{C}_m has an equally simple form, and it is easy to verify that it is given by

$$\mathcal{C}_m^{-1} = \begin{bmatrix} -\chi_{m,1}/\chi_{m,0} & & & & \\ -\chi_{m,2}/\chi_{m,0} & \mathbf{I}_{m-1} & & & \\ \vdots & & & & \\ -\chi_{m,m}/\chi_{m,0} & 0 & \cdots & 0 & \end{bmatrix}$$

Thus,

$$\boldsymbol{\pi}_m = \begin{bmatrix} -\chi_{m,1}/\chi_{m,0} \\ -\chi_{m,2}/\chi_{m,0} \\ \vdots \\ -\chi_{m,m}/\chi_{m,0} \end{bmatrix} \quad (31)$$

and finally

$$\hat{\boldsymbol{\beta}}_{PLS}^m = \sum_{i=0}^{m-1} \left(\frac{-\chi_{m,i+1}}{\chi_{m,0}} \mathbf{A}^i \right) \mathbf{b}$$

$$\begin{aligned}
&= - \left(\frac{\chi_{m,1}}{\chi_{m,0}} \mathbf{I} + \frac{\chi_{m,2}}{\chi_{m,0}} \mathbf{A} + \dots + \frac{\chi_{m,m}}{\chi_{m,0}} \mathbf{A}^{m-1} \right) \mathbf{b} \\
&= \left\{ \mathbf{I} - \left(\frac{\chi_{m,0}}{\chi_{m,0}} \mathbf{I} + \frac{\chi_{m,1}}{\chi_{m,0}} \mathbf{A} + \dots + \frac{\chi_{m,m}}{\chi_{m,0}} \mathbf{A}^m \right) \right\} \mathbf{A}^{-1} \mathbf{b} \\
&= \{ \mathbf{I} - \tilde{\chi}_m(\mathbf{A}) \} \hat{\boldsymbol{\beta}}_{OLS}
\end{aligned} \tag{32}$$

where $\tilde{\chi}_m(\xi) = \chi_m(\xi)/\chi_{m,0}$.

A.2 Proof II

The expression for $\hat{\boldsymbol{\beta}}_{PLS}^m$ in eq. (4) can be written in terms of the Lanczos vectors as

$$\hat{\boldsymbol{\beta}}_{PLS}^m = \mathbf{W}_m \mathbf{T}_m^{-1} \mathbf{W}_m' \mathbf{b} \tag{33}$$

Now from the Cayley-Hamilton theorem, $\chi_m(\mathbf{T}_m) = \mathbf{0}$, and we can use this result to write

$$\begin{aligned}
\mathbf{T}_m^{-1} &= -\frac{1}{\chi_{m,0}} \left(\chi_{m,m} \mathbf{T}_m^{m-1} + \chi_{m,m-1} \mathbf{T}_m^{m-2} + \dots + \chi_{m,1} \mathbf{I} \right) \\
&= \pi(\mathbf{T}_m)
\end{aligned}$$

where, as in Section 3.2, $\pi(\mathbf{T}_m)$ is a polynomial of degree $m - 1$. Substituting this result into eq. (33) and writing \mathbf{T}_m as $\mathbf{W}_m' \mathbf{A} \mathbf{W}_m$ yields

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{PLS}^m &= \mathbf{W}_m \pi(\mathbf{T}_m) \mathbf{W}_m' \mathbf{b} \\
&= \pi(\mathbf{W}_m \mathbf{W}_m' \mathbf{A} \mathbf{W}_m \mathbf{W}_m') \mathbf{b}
\end{aligned} \tag{34}$$

Note that although $\mathbf{W}_m \mathbf{W}_m' \neq \mathbf{I}$, $\mathbf{W}_m \mathbf{W}_m' \mathbf{b} = \mathbf{I} \mathbf{b}$, because $\mathbf{W}_m \mathbf{W}_m'$ is an orthogonal projector onto $\mathcal{K}_m(\mathbf{A}, \mathbf{b}) = \text{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{m-1}\mathbf{b})$.

Consider now the highest order term in eq. (34):

$$(\mathbf{W}_m \mathbf{W}_m' \mathbf{A} \mathbf{W}_m \mathbf{W}_m')^{m-1} \mathbf{b} = \mathbf{W}_m \mathbf{W}_m' \mathbf{A} \mathbf{W}_m \mathbf{W}_m' \mathbf{A} \dots \mathbf{W}_m \mathbf{W}_m' \mathbf{A} \mathbf{W}_m \mathbf{W}_m' \mathbf{b}$$

By starting at the right hand side and then successively projecting onto $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$, it is easy to see that $(\mathbf{W}_m \mathbf{W}_m' \mathbf{A} \mathbf{W}_m \mathbf{W}_m')^{m-1} \mathbf{b} = \mathbf{A}^{m-1} \mathbf{b}$. This result extends, of course, to the lower order terms, and hence $\pi(\mathbf{W}_m \mathbf{W}_m' \mathbf{A} \mathbf{W}_m \mathbf{W}_m') \mathbf{b} = \pi(\mathbf{A}) \mathbf{b}$. Substituting this into eq. (34) yields $\hat{\boldsymbol{\beta}}_{PLS}^m = \pi(\mathbf{A}) \mathbf{b} = \{ \mathbf{I} - \tilde{\chi}_m(\mathbf{A}) \} \hat{\boldsymbol{\beta}}_{OLS}$.