Shrinkage Structure of Partial Least Squares

OLE C. LINGJÆRDE and NILS CHRISTOPHERSEN

University of Oslo

ABSTRACT. Partial least squares regression (PLS) is one method to estimate parameters in a linear model when predictor variables are nearly collinear. One way to characterize PLS is in terms of the scaling (shrinkage or expansion) along each eigenvector of the predictor correlation matrix. This characterization is useful in providing a link between PLS and other shrinkage estimators, such as principal components regression (PCR) and ridge regression (RR), thus facilitating a direct comparison of PLS with these methods. This paper gives a detailed analysis of the shrinkage structure of PLS, and several new results are presented regarding the nature and extent of shrinkage.

Key words: Krylov subspace, least squares, partial least squares, principal components, rank deficient, shrinkage estimators, subspace distance

1. Introduction

In the presence of nearly collinear data, least squares regression coefficients can have large variances, and alternative estimators are often considered. One such alternative is PLS (also known as PLS1) (Wold, 1975), which has been used extensively in chemometrics for more than two decades, both for prediction and for identification of latent structure models (Wold, 1993). Similar techniques are currently receiving much attention in other areas as methods for solving ill-posed inverse problems (Hansen, 1998).

The relationship between PLS and other regression methods has been the focus of much research (see e.g. Næs & Martens, 1985; Helland, 1988; Stone & Brooks, 1990). One way to compare estimators is to consider their shrinkage properties relative to the ordinary least squares (LS) estimator (Frank & Friedman, 1993). It is known that PLS shrinks relative to LS (De Jong, 1995; Goutis, 1996), although PLS may expand the LS solution in some directions (Frank & Friedman, 1993). However, the precise shrinkage properties of PLS depend in a very complicated manner on the given problem and has so far eluded a full theoretical analysis.

The purpose of this paper is to convey some new theoretical results concerning the shrinkage structure of PLS, and to discuss practical consequences of these new insights. The point of departure is the relation between PLS and constrained least squares regression over Krylov subspaces, first described by Manne (1987) and Helland (1988). Over the past few decades, new results in Krylov subspace optimization have appeared, with important implications for the understanding of the PLS estimator (Saad, 1992; Hansen, 1998). In this paper, these new results will be put in the context of PLS regression and will be further extended.

2. Preliminaries

Suppose we have data consisting of scalar response values y_i and corresponding vectors of predictors $x_i = (x_{i1}, \ldots, x_{ip})$, where $i = 1, 2, \ldots, n$. Here, we assume $n \geqslant p$; with some minor modifications, the results below can be established for the case $n < p$ as well. Throughout the paper, we let X denote the $n \times p$ predictor matrix with rows $x_1, \ldots, x_n$. The vector of responses $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and the columns of X are usually centred and

sometimes scaled to unit variance prior to the analysis, although no such transformations are required in this paper. We consider the linear regression model

$$y = Xb + \epsilon, \tag{1}$$

where b is an unknown $p \times 1$-parameter vector and ϵ is an $n \times 1$-vector of noise terms. Distributional conditions on the noise vector (e.g. that the components ϵ_i have identical means and are uncorrelated) are not needed for the discussions in this paper. Note that when X has full column rank, b is uniquely determined by the linear predictor $\eta = Xb$ as $b = X^{\dagger}\eta$, where $X^{\dagger}$ is the Moore–Penrose inverse of X.

We will make considerable use of the singular value decomposition (SVD) of X, given by $X = UDV^{\mathrm{T}}$, where $U^{\mathrm{T}}U = V^{\mathrm{T}}V = VV^{\mathrm{T}} = I_p$ (the $p \times p$ identity matrix) and where D is diagonal with the singular values $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_p$ on the diagonal. The columns $u_1 \ldots, u_p$ of U are denoted left singular vectors, and the columns $v_1, \ldots, v_p$ of V are denoted right singular vectors. In the sequel we also refer frequently to the eigenvalues $\lambda_i = \sigma_i^2$ of $X^{\mathrm{T}}X$. For later use, we note that $X^{\dagger} = VD^{-1}U^{\mathrm{T}}$. Although the theory in this paper could be carried over to treat matrices of arbitrary rank, we assume for clarity of exposition that X has full column rank, so that all p singular values are positive. In most practical cases where $n \geqslant p$, this condition will hold. However, note that the numerical rank of X, defined with respect to a given tolerance level δ and given as the number k satisfying $\sigma_k > \delta \geqslant \sigma_{k+1}$ (see Golub & van Loan, 1996), can be much smaller than p. For an introduction to the use of SVD in regression analysis, see Mandel (1982).

The LS estimator for the parameter vector in (1) is $\hat{b}_{\mathrm{LS}} = X^{\dagger}y = VD^{-1}U^{\mathrm{T}}y$. Expanding $\hat{b}_{\mathrm{LS}}$ in terms of the right singular vectors of X (i.e. the eigenvectors of $X^{\mathrm{T}}X$), we have

$$\hat{b}_{\mathrm{LS}} = VD^{-1}U^{\mathrm{T}}y = \sum_{i=1}^{p} \frac{u_i^{\mathrm{T}}y}{\sigma_i} v_i, \tag{2}$$

In the following, the inner products $f_i = u_i^{\mathrm{T}}y$ are denoted Fourier coefficients (of y). When y and the columns of X are centred, f_i can be expressed in terms of the sample covariance $s_i = p_i^{\mathrm{T}}y$ between y and the ith principal component $p_i = Xv_i = \sigma_i u_i$ of X as $f_i = s_i/\sigma_i$.

Several shrinkage estimators can be expressed in a similar manner to LS as a linear combination of the right singular vectors of X (see Huffel & Vandewalle, 1991). The PCR estimator (Massy, 1965) truncates the expansion in (2) after a certain term and is given by

$$\hat{b}_{\mathrm{PCR}} = \sum_{i=1}^{m} \frac{u_i^{\mathrm{T}}y}{\sigma_i} v_i, \tag{3}$$

where $m \leqslant p$ is the number of factors (or principal components) retained. The RR estimator (Hoerl & Kennard, 1970) can be written as

$$\hat{b}_{\mathrm{RR}} = \sum_{i=1}^{p} \frac{\sigma_i^2}{\sigma_i^2 + k} \frac{u_i^{\mathrm{T}}y}{\sigma_i} v_i, \tag{4}$$

for some $k \geqslant 0$. A related estimator is generalized ridge regression (GRR), which arises by replacing the scalar k in the ith term of (4) by k_i, $i = 1, 2, \ldots, p$ (see e.g. Hocking *et al.*, 1976). It will be shown in section 3 that PLS also has an expansion in terms of the right singular vectors of X. There are estimators that cannot be expressed in this manner, however. In latent root regression (Hawkins, 1973; Webster *et al.*, 1974) singular values and right singular vectors are extracted from the matrix $[X; y]$ rather than from X; as a consequence, the latent root regression estimator cannot in general be expressed as a linear combination of the vectors v_i used here (see Huffel & Vandewalle, 1991).

Algorithmic descriptions of PLS prevail in the literature (see, e.g. Wold, 1975; Martens & Næs, 1989; Frank & Friedman, 1993). Our point of departure, however, will be the algebraic description given by Helland (1988). For $m \geqslant 1$ define the Krylov space

$$\mathscr{K}_m = \text{span}\{X^T y, (X^T X)X^T y, \ldots, (X^T X)^{m-1} X^T y\}. \tag{5}$$

For some $M \leqslant p$ we have

$$\mathscr{K}_1 \subset \mathscr{K}_2 \subset \cdots \subset \mathscr{K}_M = \mathscr{K}_{M+1} = \cdots \tag{6}$$

where $\subset$ denotes proper set inclusion. The maximal Krylov space $\mathscr{K}_M$ is spanned by the vectors $P_\lambda X^T y$ where λ ranges over the eigenvalues of $X^T X$ and P_λ denotes the orthogonal projector onto the eigenspace of $X^T X$ corresponding to λ (Parlett, 1998). A simple calculation reveals that $P_\lambda X^T y = \sum_{i:\lambda_i=\lambda} \sigma_i (u_i^T y) v_i$; hence, $\mathscr{K}_1, \ldots, \mathscr{K}_M$ are subspaces of the space spanned by the vectors v_i for which $u_i^T y \neq 0$. In particular, M is the number of distinct eigenvalues λ_i of $X^T X$ for which the corresponding Fourier coefficient $u_i^T y$ is non-zero (see also Helland, 1988, 1990).

The estimate for b in (1) found by using PLS with $m \geqslant 1$ latent factors solves the constrained optimization problem

$$\begin{aligned} &\text{minimize} \|y - Xb\|_2 \\ &\text{such that } b \in \mathscr{K}_m. \end{aligned} \tag{7}$$

PLS regression with $m = M$ factors yields the LS estimate (2) (Helland, 1988). In the following we omit this special case and assume that $m < M$.

An explicit formula for the PLS estimate can be found as follows. Let R_m be a $p \times m$ matrix with columns that span the subspace $\mathscr{K}_m$ (we may, e.g. let the columns of R_m be the vectors in the right hand side of (5)). Any minimizer of (7) must have the form $\hat{b} = R_m \hat{z}$, where $\hat{z} \in \mathbb{R}^m$ is a minimizer of the unconstrained optimization problem $\min \|y - XR_m z\|_2$. A standard least squares argument then gives $\hat{b} = \hat{b}_{\text{PLS}}$, where

$$\hat{b}_{\text{PLS}} = R_m (R_m^T X^T X R_m)^{-1} R_m^T X^T y. \tag{8}$$

3. Filter factors

Many well-known estimators for the parameter vector b in (1) can be written as

$$\hat{b} = \sum_{i=1}^{p} \omega_i \frac{u_i^T y}{\sigma_i} v_i. \tag{9}$$

The weights ω_i will be referred to in this paper as filter factors. For the LS estimator we have $\omega_i = 1$ for all i. For PCR we have $\omega_i = 1$ for $i \leqslant m$ and $\omega_i = 0$ for $i > m$, while for RR we have $\omega_i = \sigma_i^2/(\sigma_i^2 + k)$, $i = 1, 2, \ldots, p$, and for GRR we have $\omega_i = \sigma_i^2/(\sigma_i^2 + k_i)$, $i = 1, 2, \ldots, p$. The PLS estimate (8) also has a representation of the form (9), since $\hat{b}_{\text{PLS}}$ is in the subspace spanned by the vectors v_i for which $u_i^T y \neq 0$ (see section 2).

Filter factors for PLS can be found as follows. Let Λ denote the $p \times p$ diagonal matrix with the components of $U^T y$ on the diagonal, and define $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T$. Writing (9) as $\hat{b}_{\text{PLS}} = VD^{-1}\Lambda\boldsymbol{\omega}$ and using (8), we have

$$\Lambda\boldsymbol{\omega} = DV^T \hat{b}_{\text{PLS}} = U^T (UDV^T \hat{b}_{\text{PLS}}) = U^T \mathscr{P} y \tag{10}$$

where $\mathscr{P} = XR_m (R_m^T X^T X R_m)^{-1} R_m^T X^T$. Note that $\mathscr{P}$ is an orthogonal projector onto the subspace $\mathscr{R}(XR_m)$. Unfortunately, these filter factors do not lend themselves to easy interpretations. Observe that $\boldsymbol{\omega}$ depends non-linearly on y and also appears to be non-

trivially related to the number of factors m. The following theorem provides a more useful characterization of $\boldsymbol{\omega}$, and will be central in the discussions below. Essentially the same result can be found in the numerical linear algebra literature (see e.g. Hansen, 1998), and we skip the proof.

Theorem 1
Assume that $\dim(\mathcal{K}_m) = m$. *The filter factors for PLS with m factors, as defined by the expansion* (9) *are given by*

$$\omega_i^{(m)} = 1 - \prod_{j=1}^{m}\left(1 - \frac{\lambda_i}{\theta_j^{(m)}}\right), \quad i = 1, 2, \ldots, p \tag{11}$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ *are the eigenvalues of* $X^{\mathrm{T}}X$, *and* $\theta_1^{(m)} \geqslant \theta_2^{(m)} \geqslant \cdots \geqslant \theta_m^{(m)}$ *are the eigenvalues of* $V_m^{\mathrm{T}}X^{\mathrm{T}}XV_m$, *where* V_m *is any* $p \times m$ *matrix with columns that form an orthonormal basis for* $\mathcal{K}_m$ *(Note: when no confusion can arise, we sometimes omit the superscripts and write* ω_i *and* θ_j*).*

The above theorem shows that the filter factors for PLS are completely determined by the eigenvalues of the matrices $X^{\mathrm{T}}X$ and $V_m^{\mathrm{T}}X^{\mathrm{T}}XV_m$, where V_m is any matrix with columns that form an orthonormal basis for the Krylov space $\mathcal{K}_m$ (the particular choice of orthonormal basis does not influence the eigenvalues).

4. Ritz values and their properties

The eigenvalues $\theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_m$ in theorem 1 are commonly referred to in numerical analysis at Ritz values and play a central role in numerical methods for large eigenvalue problems (see, e.g. Saad, 1992). To see how the Ritz values enter into these problems, consider the problem of finding a good approximation, from within the subspace $\mathcal{K}_m$, to an eigenvector-eigenvalue pair (v, λ) for $X^{\mathrm{T}}X$ (i.e. $X^{\mathrm{T}}Xv = \lambda v$). An orthogonal projection method approximates the eigenpair by $(\tilde{v}, \theta)$, where $\tilde{v} \in \mathcal{K}_m$ $(\tilde{v} \neq 0)$, $\theta \in \mathbb{R}$, and the orthogonality condition

$$X^{\mathrm{T}}X\tilde{v} - \theta\tilde{v} \perp \mathcal{K}_m \tag{12}$$

holds. The solutions for θ in (12) are exactly the Ritz values θ_i, $i = 1, \ldots, m$ (Saad, 1992). The best known method for computation of the Ritz values (and corresponding eigenvectors) is the celebrated Lanczos algorithm (Lanczos, 1950). A more general approach that works for other subspaces than $\mathcal{K}_m$ as well, is the Rayleigh–Ritz procedure (see e.g. Golub & van Loan, 1996).

We would expect the above eigenpair approximation to become gradually better as the dimension of the Krylov subspace $\mathcal{K}_m$ increases. Interestingly, approximations tend to be excellent even for small m, particularly for those eigenvalues of largest magnitude.

For future reference, we compile some known facts about the Ritz values, rephrasing them for our purposes. Proofs of these properties can be found elsewhere (see Saad, 1992; Fischer, 1996; Parlett, 1998). As before, we assume $\dim(\mathcal{K}_m) = m < M$. The acute angle between a vector u and a subspace $\mathcal{K}$ will be denoted $\varphi(u, \mathcal{K})$ and satisfies

$$\cos\varphi(u, \mathcal{K}) = \frac{\|\mathcal{P}_{\mathcal{K}}u\|_2}{\|u\|_2}$$

where $\mathcal{P}_{\mathcal{K}}$ is the orthogonal projector onto the subspace $\mathcal{K}$ (note that there is a one-to-one correspondence between orthogonal projectors and the subspaces they project onto). As

before, let $v_1, \ldots, v_p$ denote normalized eigenvectors of X^TX corresponding to the eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_p$.

Theorem 2
The Ritz values $\theta_1^{(m)} \geqslant \cdots \geqslant \theta_m^{(m)}$ satisfy the properties:

(a) $\lambda_1 > \theta_1^{(m)} > \theta_2^{(m)} > \cdots > \theta_m^{(m)} > \lambda_p$.
(b) $\lambda_{p-m+i} \leqslant \theta_i^{(m)} < \lambda_i$, $i = 1, 2, \ldots, m$.
(c) Each of the open intervals $(\theta_{i+1}^{(m)}, \theta_i^{(m)})$ contains one or more eigenvalues λ_j.
(d) The Ritz values $\{\theta_i^{(m)}\}_{i=1}^m$ and $\{\theta_i^{(m+1)}\}_{i=1}^{m+1}$ separate each other

$$\theta_1^{(m+1)} > \theta_1^{(m)} > \theta_2^{(m+1)} > \theta_2^{(m)} \cdots > \theta_m^{(m)} > \theta_{m+1}^{(m+1)}.$$

Thus for fixed k, the kth largest Ritz value $\theta_k^{(m)}$ increases with m and the kth smallest Ritz value $\theta_{m-k}^{(m)}$ decreases with m.
(e) The following error bound holds for $i \leqslant m \leqslant p-2$

$$0 \leqslant \lambda_i - \theta_i^{(m)} \leqslant \rho_i^{(m)}(\lambda_1 - \lambda_p)\tan^2 \varphi(X^Ty, v_i)$$

where $\rho_i^{(m)} = (\kappa_i^{(m)}/T_{m-i}(1+2\gamma_i))^2$, $\gamma_i = (\lambda_i - \lambda_{i+1})/(\lambda_{i+1} - \lambda_p)$, $\kappa_1^{(m)} = 1$,

$$\kappa_i^{(m)} = \prod_{j=1}^{i-1} \frac{\theta_j^{(m)} - \lambda_p}{\theta_j^{(m)} - \lambda_i} \quad \text{for } i > 1,$$

and T_k, $k \geqslant 0$ denote the Chebyshev polynomials of the first kind, which can be defined by the recurrence relation $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$ with initial functions $T_0(x) = 1$, $T_1(x) = x$. (These polynomials satisfy $T_k(x) \in [-1, 1]$ for $x \in [-1, 1]$ and grow very rapidly outside this interval. In fact, for $|x| \geqslant 1$ we have $T_k(x) = \frac{1}{2}[(x + \sqrt{x^2-1})^k + (x + \sqrt{x^2-1})^{-k}]$).

An example of the convergence properties of the Ritz values $\theta_i^{(m)}$ towards the eigenvalues λ_i is shown in Fig. 1. Note the rapid convergence of the first few Ritz values; a three- or four-dimensional Krylov subspace is sufficient in this case to approximate the largest eigenvalue almost exactly.

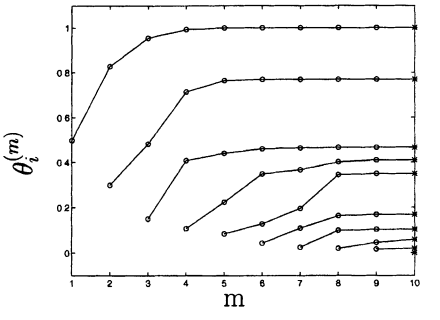

Fig. 1. Convergence of Ritz values towards the eigenvalues for a symmetric 10×10 matrix X^TX. The ith curve from above depicts $\theta_i^{(m)}$ for $m = i, \ldots, 10$ ($i = 1, 2, \ldots, 9$). Eigenvalues are shown as asterisks at the right edge of the plot.

5. Shrinkage properties of PLS

Few facts are known about the precise shrinking structure in PLS. Filter factors depend in a complex way on the singular value spectrum of the regression matrix, and the problem is further complicated by the fact that filter factors depend on the response. In contrast, filter factors for PCR, RR and GRR bear simple relations to the singular values and are independent of y. A complete analysis of the filter properties would thus have to take into account the detailed structure of the given problem. Some fundamental properties of the PLS shrinkage can nevertheless be established, based on the expression for the filter factors found in theorem 1. This section lays out some results in this direction. The properties of the Ritz values (as given in theorem 2) are fundamental to this analysis.

5.1. Two examples

Before we proceed with the theoretical analysis, it might be instructive to consider a few examples, to get a feel of what to expect from the theory. Figures 2 and 3 show results from two PLS estimation problems which are based on poorly conditioned regression

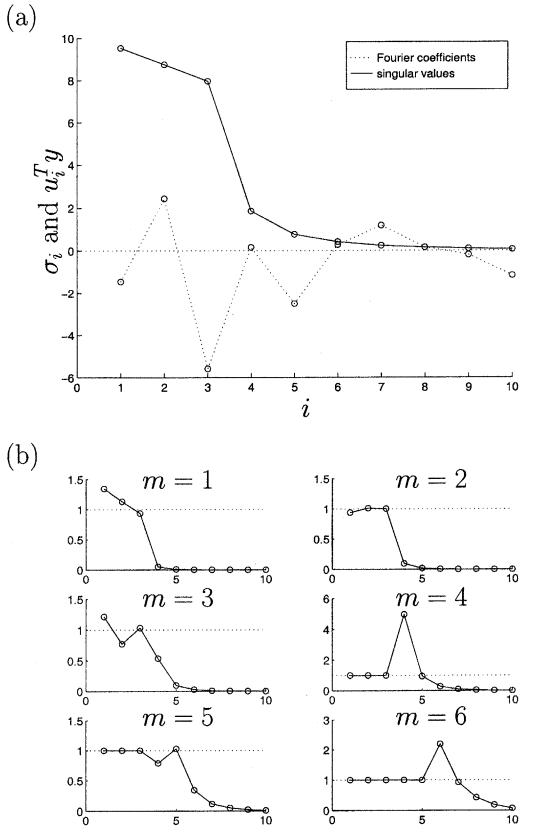

Fig. 2. (a) Singular values $\sigma_i = \sqrt{\lambda_i}$ and Fourier coefficients $u_i^{\mathrm{T}} y$ are shown for $i = 1, 2, \ldots, 10$, for a data set (X, y) where X is 50×10 and y is 50×1. Notice the gap in the singular value spectrum. (b) PLS filter factors for the data in (a) and for $m = 1, 2, \ldots, 6$ factors. In each graph, m is fixed and $\omega_i^{(m)}$ is plotted as a function of i.

matrices $X(50 \times 10)$ and corresponding response vectors $y(50 \times 1)$ for which some (but not all) of the Fourier coefficients are large. Figures 2a and 3a depict the singular values σ_i and the Fourier coefficients $u_i^{\mathrm{T}} y$ for the two problems, and Figs 2b and 3b show the PLS filter factors for the two problems when $m = 1, 2, \ldots, 6$ factors are used.

Notice the irregular behaviour of the filter factors; in particular, no simple relationship exists between filter factors and singular values. Furthermore, filter factors larger than one, and even negative filter factors, may occur. In contrast, filter factors ω_i for the estimators PCR, RR and GRR are non-increasing functions of i (i.e. $\omega_1 \geqslant \omega_2 \geqslant \cdots \geqslant \omega_p$) and always lie in [0, 1]. In all the cases shown in Figs. 2 and 3, the filter factors $\omega_i^{(m)}$ approach zero as i becomes large, but this effect always sets in after the mth filter factor, regardless of the value of m. In the following, it will be demonstrated that considerable insight into the shrinkage pattern of PLS can be obtained from the results in section 3 and 4.

5.2. *Fundamental properties of the filter factors*

Consider first the problem of determining under which conditions PLS performs an expansion or contraction in some direction. This question is easy to answer for the directions corresponding to the largest and smallest eigenvalues.

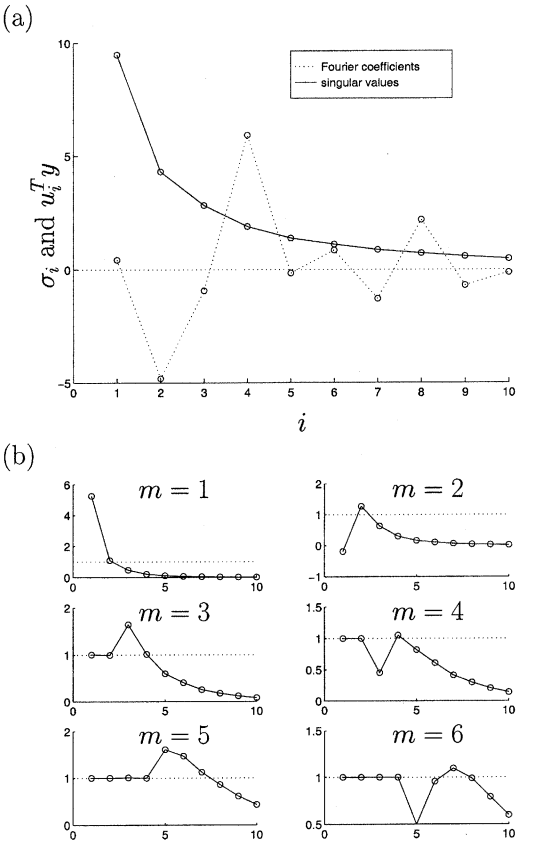

Fig. 3. (a) Singular values $\sigma_i = \sqrt{\lambda_i}$ and Fourier coefficients $u_i^{\mathrm{T}} y$ are shown for $i = 1, 2, \ldots, 10$, for a data set (X, y) where X is 50×10 and y is 50×1. (b) PLS filter factors for the data in (a) and for $m = 1, 2, \ldots, 6$ factors. In each graph, m is fixed and $\omega_i^{(m)}$ is plotted as a function of i.

Theorem 3

We have $\omega_p^{(m)} \leqslant 1$ *for all* m, *and*

$$\omega_1^{(m)} \geqslant 1 \text{ for } m = 1, 3, 5, \ldots$$

$$\omega_1^{(m)} \leqslant 1 \text{ for } m = 2, 4, 6, \ldots$$

Proof. According to theorem 1, the inequality $\omega_p^{(m)} \leqslant 1$ holds if and only if $\prod_1^m(\theta_j - \lambda_p) \geqslant 0$, and the latter inequality must hold, since $\theta_j \geqslant \lambda_p$ for all j. On the other hand, $\omega_1^{(m)} \leqslant 1$ is equivalent to $\prod_1^m(\theta_j - \lambda_1) \geqslant 0$. Since $\theta_j \leqslant \lambda_1$ for all j, all the m terms in the product are non-positive. Hence, the inequality is satisfied if m is even. Likewise, $\omega_1^{(m)} \geqslant 1$ is equivalent to $\prod_1^m(\theta_j - \lambda_1) \leqslant 0$, and the inequality is satisfied for odd integers m.

For the remaining filter factors $\omega_i^{(m)}$, $i = 2, \ldots, m-1$, one may have either $\omega_i^{(m)} \leqslant 1$ or $\omega_i^{(m)} \geqslant 1$, depending on the location of the Ritz values. One immediate observation from (11) is that if $\lambda_i \leqslant \theta_m^{(m)}$, then $0 \leqslant \omega_i^{(m)} \leqslant 1$. More insight can be gained by considering the mth degree polynomial

$$p_m(\lambda) = 1 - \prod_{j=1}^{m}\left(1 - \frac{\lambda}{\theta_j^{(m)}}\right)$$

According to theorem 1, we have $\omega_i^{(m)} = p_m(\lambda_i)$, $i = 1, \ldots, p$, and we also have $p_m(\theta_j^{(m)}) = 1$, $j = 1, \ldots, m$. The polynomial $p_m(\lambda)$ is shown in Fig. 4 for an example problem and for several values of m.

From theorem 2, $\theta_1^{(m)}, \ldots, \theta_m^{(m)}$ are distinct points, thus the polynomial $1 - p_m(\lambda)$ changes sign exactly m times, at the locations $\theta_j^{(m)}$. It is readily seen that for $\lambda \geqslant \theta_1^{(m)}$ we have $p(\lambda) \geqslant 1$ when m is an odd integer, and $p(\lambda) \leqslant 1$ when m is an even integer. A schematic illustration of the polynomial p_m is provided in Fig. 5 for ease of reference.

The next result is essentially a consequence of the observations above.

Theorem 4

For $m < M$ *there is a partitioning of the set of integers* $1, 2, \ldots, p$ *in* $m+1$ *non-empty disjoint sets* $I_1, \ldots, I_{m+1}$, *where each element in* I_j *is smaller than each element in* I_k *when* $j < k$, *and where*

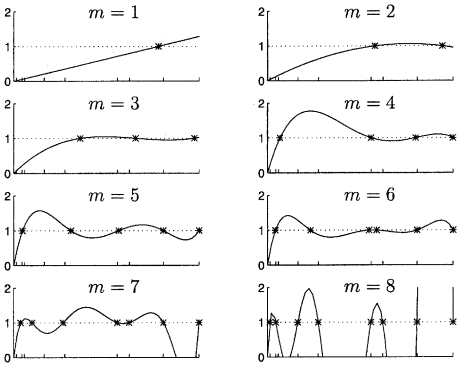

Fig. 4. The polynomial $p_m(\lambda)$, $m = 1, 2, \ldots, 8$, for a data set (X, y) where X is 10×10 and y is 10×1. The points $(\theta_j^{(m)}, 1)$ are indicated by asterisks (*), and eigenvalues λ_i are shown as tick marks on the abscissas.

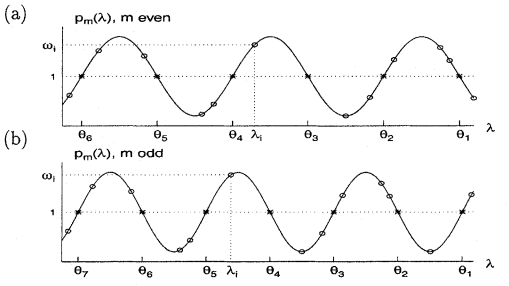

Fig. 5. Schematic illustration of the polynomial $p_m(\lambda)$ for (a) m an even integer and (b) m an odd integer.

$$\omega_i^{(m)} \leqslant 1 \quad \textit{for } i \in I_1 \cup I_3 \cup \cdots \cup I_{m+1}$$
$$\omega_i^{(m)} \geqslant 1 \quad \textit{for } i \in I_2 \cup I_4 \cup \cdots \cup I_m$$

when m is even, and

$$\omega_i^{(m)} \leqslant 1 \quad \textit{for } i \in I_2 \cup I_4 \cup \cdots \cup I_{m+1}$$
$$\omega_i^{(m)} \geqslant 1 \quad \textit{for } i \in I_1 \cup I_3 \cup \cdots \cup I_m$$

when m is odd.

Proof. Define $\theta_0 = \infty$ and $\theta_{m+1} = -\infty$. By theorem 2(a) and 2(c), each half-open interval $\mathscr{I}_i = [\theta_i, \theta_{i-1})$, $i = 1, \ldots, m+1$, contains at least one of the eigenvalues of $X^{\mathrm{T}}X$. Let I_i denote the set of indices of the eigenvalues in $\mathscr{I}_i$, $i = 1, \ldots, m+1$. Clearly, $I_1, \ldots, I_{m+1}$ are non-empty and disjoint. Suppose m is even. Since $\omega_i^{(m)} = p_m(\lambda_i)$, the first two inequalities in the theorem follow from the fact (see the top panel of Fig. 5) that $p_m(\lambda) \leqslant 1$ when $\lambda \in \mathscr{I}_i$, $i = 1, 3, \ldots, m+1$, and $p_m(\lambda) \geqslant 1$ when $\lambda \in \mathscr{I}_i$, $i = 2, 4, \ldots, m$. The case where m is odd follows in the same manner.

Corollary 1
The filter factors $\omega_1^{(m)}, \ldots, \omega_p^{(m)}$ in PLS regression with $m < M$ factors have the following properties:

(a) At least $\lfloor (m+1)/2 \rfloor$ filter factors satisfy $\omega_i^{(m)} \geqslant 1$.
(b) At least $\lfloor m/2 \rfloor + 1$ filter factors satisfy $\omega_i^{(m)} \leqslant 1$.
(c) There exists an $i \geqslant m$ such that $\omega_i^{(m)} \geqslant 1$.

Proof. The corollary follows easily from theorem 4.

These results merely represent some of several properties of the filter factors that may be deduced from theorem 4. For example, we easily find that, for some $i \leqslant p - m + 1$, we have $\omega_i^{(m)} \geqslant 1$. Note that theorem 4 shows that the oscillatory behaviour of the filter factor curves in Figs. 2b and 3b represents a general feature of PLS.

5.3. Non-shrinkage directions

As noted earlier, the PLS estimator with the maximal number of factors $m = M$ (see section 2) is identical to the LS estimator, and in this case all filter factors are equal to

one. In this case, no shrinkage is performed. Suppose on the other hand that $m < M$. Understanding the relation between directions of maximal shrinkage and properties of the data such as singular values and Fourier coefficients of the response, lies at the heart of understanding PLS. A simpler, but related problem, is to relate non-shrinkage directions (i.e. directions where filter factors are close to one) to such properties of the data. From (11) this question appears to be a complicated one, involving all the Ritz values. However, we have the following result.

Theorem 5
For $1 \leq i \leq m \leq p - 2$ *we have*

$$|\omega_i^{(m)} - 1| \leq \rho_i^{(m)} \frac{(\lambda_1 - \lambda_p)^m}{\lambda_p^m} \tan^2 \varphi(X^T y, v_i) \tag{13}$$

where $\rho_i^{(m)}$ *is defined as in theorem* 2(e).

Proof. Theorem 1 gives $|\omega_i^{(m)} - 1| = \alpha|\theta_i^{(m)} - \lambda_i|$ where $\alpha = (\theta_i^{(m)})^{-1} \prod_{j \neq i} |\theta_j^{(m)} - \lambda_i| / \theta_j^{(m)}$. From theorem 2(a) all Ritz values satisfy $\lambda_p \leq \theta_j^{(m)} \leq \lambda_1$, and so $\alpha \leq (\lambda_1 - \lambda_p)^{m-1} / \lambda_p^m$. Combining this upper bound with the upper bound on $|\theta_i^{(m)} - \lambda_i|$ given in theorem 2(e) proves the theorem.

Theorem 5 indicates that a small angle $\varphi(X^T y, v_i)$ between $X^T y$ and the ith eigenvector v_i of $X^T X$ should result in $\omega_i^{(m)}$ being close to unity. A straightforward calculation reveals that

$$\tan^2 \varphi(X^T y, v_i) = \sum_{\substack{j=1 \\ j \neq i}}^{p} \frac{\lambda_j (u_j^T y)^2}{\lambda_i (u_i^T y)^2}.$$

Thus, we would expect the ith filter factor to be close to one whenever both the ith eigenvalue λ_i and the corresponding Fourier coefficient $u_i^T y$ are large in magnitude. Returning to the first example in section 5.1, depicted in Fig. 2, we find that the third filter factor is the first one (as m increases) to settle at a value close to one, followed by the first, second, and fifth, and then number 4 and 7. Compare this with Fig. 6a, which shows the absolute value of each Fourier coefficient multiplied by the associated singular value for the same problem. Indeed, there seems to be a close correspondence between the magnitude of these products and the order in which the filter factors stabilize at unity. Corresponding results are found for the other example by comparing Fig. 3 with Fig. 6b.

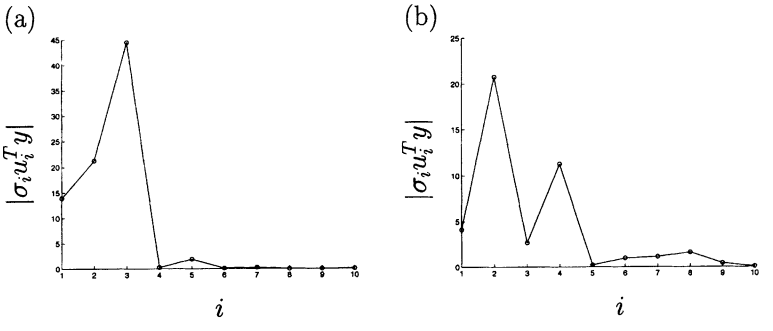

Fig. 6. (a) The absolute value of $\sigma_i u_i^T y$ as a function of i, for the same problem as depicted in Fig. 2. (b) The absolute value of $\sigma_i u_i^T y$ as a function of i, for the same problem as depicted in Fig. 3.

5.4. How many non-shrinkage directions can there be?

According to theorem 1, a filter factor $\omega_i^{(m)}$ becomes equal to one precisely when some Ritz value has become equal to the ith eigenvalue λ_i. Since there are only m Ritz values, no more than m different eigenvalues can match exactly with one of the Ritz values. Thus the number of filter factors identical to one is always less than or equal to the number of PLS factors m, unless some of the eigenvalues of $X^{\mathrm{T}}X$ coincide. However, except in rare cases, the filter factors do not become identical to one, and a more relevant question is whether we can have more than m filter factors close to one. The next theorem provides one answer to that.

Theorem 6
Let N be an integer such that $m \leqslant N \leqslant p$ and suppose $\delta^* = \min\{|\lambda_i - \lambda_k|: i \neq k \text{ and } i, k \leqslant N\}$ is positive. Given a positive $\delta \leqslant \delta^*/2\lambda_1$, suppose there is a set of m distinct indices $J \subseteq \{1, 2, \ldots, N\}$ such that $|\omega_i^{(m)} - 1| < \delta^m$ for $i \in J$. Then, for any index $l \in \{1, \ldots, p\} \backslash J$, we have

$$|\omega_l^{(m)} - 1| > \prod_{i \in J} \left|1 - \frac{\lambda_l}{\lambda_i}\right| - \delta \sum_{i \in J} \frac{\lambda_l}{\lambda_i} \prod_{k \in J \backslash \{i\}} \left|1 - \frac{\lambda_l}{\lambda_i}\right| + O(\delta^2). \tag{14}$$

Proof. See Appendix A.

Suppose m filter factors satisfy $\omega_i^{(m)} \approx 1$. Then, according to theorem 6, we have the approximate inequality

$$|\omega_l^{(m)} - 1| > \prod_{i \in J} \left|1 - \frac{\lambda_l}{\lambda_i}\right|.$$

Hence, for any l such that $\lambda_l \ll \lambda_i$ for all $i \in J$, the corresponding filter factor $\omega_l^{(m)}$ must deviate substantially from one. For example, assume that the first m filter factors $\omega_1^{(m)} \ldots, \omega_m^{(m)}$ are within the distance δ^m of unity. Then, for sufficiently small values of δ satisfying the condition in theorem 6, we have the approximate inequality

$$|\omega_{m+k}^{(m)} - 1| \geqslant \frac{(\lambda_1 - \lambda_{m+k})(\lambda_2 - \lambda_{m+k}) \cdots (\lambda_m - \lambda_{m+k})}{\lambda_1 \lambda_2 \cdots \lambda_m}$$

for any $k = 1, 2, \ldots, p - m$. Thus the larger the gap between λ_{m+k} and the m largest eigenvalues, the larger the distance is between the $(m + k)$th filter factor and one. In particular, if $\lambda_p \ll \lambda_m$, then (since $\omega_p^{(m)} \leqslant 1$) we must have $\omega_p^{(m)} \approx 0$.

5.5. Bounds on the amount of expansion

The next result provides an upper bound on the magnitude of a filter factor.

Theorem 7
Suppose $\omega_i^{(m)} \geqslant R + 1 \geqslant 2$. Then we have $\lambda_i \geqslant \lambda_p(1 + \sqrt[m]{R})$.

Proof. From the expression for the filter factors in (11) and the condition in the theorem, we have

$$\left| \prod_{j=1}^{m} \left(1 - \frac{\lambda_i}{\theta_j} \right) \right| \geq R. \tag{15}$$

For this to hold, some of the factors in (15) must satisfy $|1 - \lambda_i/\theta_j| \geq \sqrt[m]{R}$, which implies that $\lambda_i/\theta_j \geq 1 + \sqrt[m]{R}$. Thus,

$$\lambda_i \geq \theta_j(1 + \sqrt[m]{R}) \geq \lambda_p(1 + \sqrt[m]{R}).$$

where we have used that fact that $\theta_j \geq \lambda_p$ for all j.

To illustrate the above theorem, assume that for some i we have $\lambda_i < 2\lambda_p$. Then we must have $\omega_i^{(m)} < 2$ for all m. The bound in theorem 7 can be improved if we have additional information about the problem. For example, suppose the first k eigenvalues are distinct, and let the first k filter factors be equal to one, i.e. $\omega_1^{(m)} = \omega_2^{(m)} = \cdots = \omega_k^{(m)} = 1$. Then a slight modification of the proof of theorem 7 would give the following necessary condition for having $\omega_i^{(m)} \geq R + 1 \geq 2$, where $i > k$ and $\lambda_i < \lambda_k$:

$$\lambda_i \geq \lambda_p \left(1 + \sqrt[m-k]{\frac{\lambda_1\lambda_2 \cdots \lambda_k}{(\lambda_1 - \lambda_i)(\lambda_2 - \lambda_i) \cdots (\lambda_k - \lambda_i)} \cdot R} \right).$$

Bounds on the filter factors could also be derived from bounds on $|\lambda_i - \theta_i|$, using e.g. the result in theorem 2(e); another possibility is to make use of the fact that PLS shrinks relative to LS (De Jong, 1995; Goutis, 1996).

6. Final remarks

One of the more striking differences between PLS and other shrinkage methods such as PCR and RR is that filter factors for PLS oscillate between values below and above one (see section 5.2) whereas filter factors for these other methods are related in magnitude to the singular values. Nevertheless, there are certain similarities in the way PLS behave and the way PCR and RR behave.

The first filter factors for PLS quickly converge to one, whereas the last Ritz values in general are poor approximations to the corresponding eigenvalues. It can be shown that, under certain regularity conditions (essentially saying that, on average, the Fourier coefficients $u_i^T y$ decay faster to zero than the singular values σ_i), the Ritz values tend to approximate the eigenvalues in their natural order (Hansen, 1998). Under such conditions, the PLS filter factors resemble those of ridge regression for a suitably chosen ridge parameter.

If there is a clear gap between the mth and the smallest eigenvalue, the last filter factor will be close to zero. From the expression for the filter factors it is seen that filter factors are almost identical in directions corresponding to nearby eigenvalues. Consequently, in this case all filter factors corresponding to eigenvalues close to the smallest one should be close to zero. Thus, PLS effectively filters away signals along the last eigenvectors when there is a clear gap in the eigenvalue spectrum. This helps to explain the similar behaviour of PLS an PCR in situations where the latter method works well.

Intermediate filter factors (close to the mth one, where m is the number of factors) seem to follow a less consistent pattern, and can deviate substantially from one in either directions. Empirically, however, it seems to be the case that the mth filter factor often is one of the largest filter factors, except when it has converged to one.

One apparent anomaly of PLS, not present in PCR or RR, is the fact that filter factors may become negative or greater than one. It is well known that for linear estimators, both the

Table 1. *The size of the ith filter factor ω_i, as determined by the size of the singular value σ_i and absolute value $|u_i^T y|$ of the corresponding Fourier coefficient. The table does not cover all cases; intermediate singular values and Fourier coefficients are not included*

	$\lvert u_i^T y\rvert$ small	$\lvert u_i^T y\rvert$ large
σ_i small	$\omega_i \leq 1$	$\omega_i \leq 1$
σ_i large	?	$\omega_i \approx 1$

variance and the bias will be reduced by replacing filter factors larger than unity with unity. However, as pointed out by Frank & Friedman (1993), the same argument need not hold for PLS which is a non-linear function of the responses. A very rough guide to the size of the filter factors is shown in Table 1 which is based on the above observations.

Observe that anti-shrinkage mainly is constrained to the case where the corresponding singular value is large and the corresponding Fourier coefficient is small. The contribution of such terms to the PLS estimate (9) will be small, even for filter factors larger than one. Thus anti-shrinkage in PLS may not be too harmful.

Some questions concerning the shrinkage structure of PLS cannot easily be answered using the approach described in this paper. For example, it appears difficult to establish within our framework the fact (shown by De Jong (1995) and Goutis (1996) using other methods) that PLS shrinks in a global sense (i.e. $\|\hat{b}_{\mathrm{PLS}}\|_2 \leq \|\hat{b}_{\mathrm{LS}}\|_2$). Nevertheless, more insight may be gained by exploring further the potential of the filter factor representation of PLS discussed in this paper.

Acknowledgements

We thank Professor Inge Helland for helpful conversations and for pointing out some errors in an early version of the manuscript. We also thank three anonymous referees for valuable comments.

References

De Jong, S. (1995). PLS shrinks. *J. Chemometrics* **9**, 323–326.

Fischer, B. (1996). *Polynomial based iteration methods for symmetric linear systems*, Wiley Teubner, Chichester.

Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–148.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix computations*, 3rd edn. Johns Hopkins University Press, Baltimore, MD.

Goutis, C. (1996). Partial least squares algorithm yields shrinkage estimators. *Ann. Statist.* **24**, 816–824.

Hansen, P. C. (1998). *Rank-deficient and discrete ill-posed problems*, Monographs on Mathematical Modeling and Computation, SIAM, Philadelphia.

Hawkins, D. M. (1973). On the investigation of alternative regressions by principal component analysis. *Appl. Statist.* **22**, 275–286.

Helland, I. S. (1988). On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.* **17**, 581–607.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.* **17**, 97–114.

Hocking, R. R., Speed, F. M. and Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics* **18**, 425–437.

Hoerl, W. E. & Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics* **12**, 55–67.

Huffel, S. V. & Vandewalle, J. (1991). *The total least squares problem: computational aspects and analysis*, Frontiers in Applied Mathematics. SIAM, Philadelphia.
Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl Bureau Standards* **45**, 255–282.
Mandel, J. (1982). Use of the singular value decomposition in regression analysis. *Amer. Statist.* **36**, 15–24.
Manne, R. (1987). Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics Int. Lab. Syst.* **2**, 187–197.
Martens, H. & Næs, T. (1989). *Multivariate calibration*, Wiley, New York.
Massy, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc.* **60**, 234–246.
Næs, T. & Martins, H. (1985). Comparison of prediction methods for multi-collinear data. *Comm. Statist. Simulation Comput.* **14**, 545–576.
Parlett, B. N. (1998). *The symmetric eigenvalue problem*, Classics in Applied Mathematics. SIAM, Philadelphia.
Saad, Y. (1992). *Numerical methods for large eigenvalue problems*, Halsted Press, Manchester.
Stone, M. & Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. Roy. Statist. Soc. Ser. B* **52**, 237–269.
Webster, J. T., Gunst, R. F. & Mason, R. L. (1974). Latent root regression analysis. *Technometrics* **16**, 513–522.
Wold, H. (1975). Soft modeling by latent variables; the nonlinear iterative partial least squares approach, in *Perspectives in probability and statistics. Papers in honour of M. S. Bartlett* (ed. J. Gani). Academic Press, New York.
Wold, S. (1993). PLS in chemical practice (discussion). *Technometrics* **35**, 136–139.

Received September 1998, in final form October 1999

Ole Christian Lingjærde, Division of Zoology, P.O. Box 1050 Blindern, N-0316 Oslo, Norway.

Appendix A

Proof of theorem 6

Proof. Let $i \in J$. Imposing the conditions in the theorem on the filter factors given in (11), there exist an element $j = j(i)$ in 1, ..., m such that

$$\left|1 - \frac{\lambda_i}{\theta_j}\right| < \delta. \tag{16}$$

Let $k \neq i$, $k \leq N$ be given. Then

$$\left|1 - \frac{\lambda_k}{\theta_j}\right| = \left|\frac{\theta_j - \lambda_i + \lambda_i - \lambda_k}{\theta_j}\right| \geq \left|\frac{\lambda_i - \lambda_k}{\theta_j}\right| - \left|\frac{\theta_j - \lambda_i}{\theta_j}\right|.$$

The first term is strictly bounded from below by δ^*/θ_j, and the second term is bounded from above by δ. Using the fact (see theorem 2(a)) that $\theta_j \leq \lambda_1$ for all j, we then have

$$\left|1 - \frac{\lambda_k}{\theta_j}\right| > \frac{\delta^*}{\theta_j} - \delta \geq \frac{2\lambda_1\delta}{\theta_j} - \delta \geq \delta$$

Thus no θ_j can satisfy more than one of the inequalities in (16). Accordingly, the set of indices $j(i)$ (for $i \in J$) must all be different and hence must take on every value in 1, 2, ..., m. Now suppose $l \in \{1, 2, \ldots, p\} \backslash J$, and consider

$$|\omega_l^{(m)} - 1| = \prod_{j=1}^{m}\left|1 - \frac{\lambda_l}{\theta_j}\right| = \prod_{i \in J}\left|1 - \frac{\lambda_l}{\theta_{j(i)}}\right|. \tag{17}$$

We seek a lower bound for the expression above. First, observe that

$$\left|\frac{\lambda_l}{\theta_{j(i)}}-\frac{\lambda_l}{\lambda_i}\right|=\left|\frac{\lambda_i\lambda_l-\lambda_l\theta_{j(i)}}{\theta_{j(i)}\lambda_i}\right|<\frac{\lambda_l}{\lambda_i}\delta.$$

Thus,

$$\left|1-\frac{\lambda_l}{\theta_{j(i)}}\right|=\left|1-\frac{\lambda_l}{\lambda_i}+\frac{\lambda_l}{\lambda_i}-\frac{\lambda_l}{\theta_{j(i)}}\right|\geqslant\left|1-\frac{\lambda_l}{\lambda_i}\right|-\left|\frac{\lambda_l}{\lambda_i}-\frac{\lambda_l}{\theta_{j(i)}}\right|>\left|1-\frac{\lambda_l}{\lambda_i}\right|-\delta\frac{\lambda_l}{\lambda_i}$$

and (17) then gives

$$|\omega_l^{(m)}-1|>\prod_{i\in J}\left|\left|1-\frac{\lambda_l}{\lambda_i}\right|-\delta\frac{\lambda_l}{\lambda_i}\right|\geqslant\prod_{i\in J}\left(\left|1-\frac{\lambda_l}{\lambda_i}\right|-\delta\frac{\lambda_l}{\lambda_i}\right).$$

Expanding the right hand side above gives the theorem.