J. R. Statist. Soc. B (2000)
62, *Part* 3, *pp.* 585–593

The peculiar shrinkage properties of partial least squares regression

Neil A. Butler and Michael C. Denham

University of Reading, UK

[Received July 1999. Revised January 2000]

Summary. Partial least squares regression has been widely adopted within some areas as a useful alternative to ordinary least squares regression in the manner of other shrinkage methods such as principal components regression and ridge regression. In this paper we examine the nature of this shrinkage and demonstrate that partial least squares regression exhibits some undesirable properties.

Keywords: Ordinary least squares; Partial least squares; Principal components; Regularized regression; Ridge regression

1. Introduction

Partial least squares (PLS) regression has become an established alternative to ordinary least squares (OLS) regression for relating observed responses to a set of explanatory variables, particularly within the field of chemometrics where it is probably the most frequently used regression method of all. It is most often used in situations where the explanatory variables are highly collinear and where they outnumber the observations. For example, in the context of chemometrics, the explanatory variables are often absorbance values from a 'digitization' of an underlying infra-red spectrum, whereas the responses may be the amount of a chemical of interest in a sample.

The method has its roots in work by H. Wold and co-workers in the 1960s (Wold, 1966, 1973, 1982) but is most often presented as a latent factor regression method in its own right (Martens and Jensen, 1983; Helland, 1988; Stone and Brooks, 1990). Initially there appears to have been little interest in the method by the wider statistical community, but more recently statisticians have attempted to shed some light on the method and its properties (Helland, 1988; Stone and Brooks, 1990; Frank and Friedman, 1993). In particular, it has been established that PLS regression is a shrinkage estimator (Denham, 1991; Goutis, 1996).

Frank and Friedman (1993) attempted to gain insight into the shrinkage behaviour of PLS by examining its behaviour in various situations. They obtained an expression for the shrinkage of PLS which depends only on the OLS coefficients and the eigenvalues of the '$X'X$' matrix. This can be used to calculate the shrinkage behaviour of PLS in any given situation. Frank and Friedman considered four numerical examples corresponding to different combinations of OLS coefficients and eigenvalues and drew general conclusions about the nature of PLS regression from them. They concluded that, for k-factor PLS regression, the OLS coefficients associated with the eigenvalues closest to the kth-largest

Address for correspondence: Neil A. Butler, Department of Applied Statistics, University of Reading, PO Box 240, Earley Gate, Reading, RG6 6FN, UK.
E-mail: N.A.Butler@reading.ac.uk

 1369–7412/00/62585

eigenvalue are expanded although no formal proof was given. They also appeared to conclude that PLS regression with k factors always expands the OLS coefficient associated with the kth-largest eigenvalue.

In this paper we derive an alternative representation of the shrinkage of PLS regression relative to OLS. Using this representation we can show that the actual amounts of shrinkage applied by k-factor PLS are 'connected' by an underlying polynomial of degree k. We find that the conclusions drawn by Frank and Friedman are only true in special circumstances. By a careful consideration of this polynomial we show that k-factor PLS divides the eigenvalues (sorted in ascending order) into $k+1$ consecutive non-empty disjoint sets for which all the associated coefficients within each set are either shrunk or expanded. Furthermore, the coefficients associated with the set of smallest eigenvalues are always shrunk, with subsequent sets of coefficients being alternately expanded and shrunk. By a further examination of this shrinkage polynomial we provide additional insight into the behaviour of k-factor PLS regression when applied to data with $k+1$ non-zero eigenvalues and when the coefficients associated with the smallest $m-k$ of m non-zero eigenvalues are all shrunk. We illustrate our findings by using two artificial examples and one based on real data.

2. Related methods of regression

Consider linear regression of a univariate response y on a set of p explanatory variables $x_1, x_2, \ldots, x_p$. For a sample of size n we write the linear regression model as $Y = X\beta + \epsilon$ where $Y = (y_1, y_2, \ldots, y_n)$ is the response vector, X is an $n \times p$ design matrix, β is a vector of p unknown parameters and ϵ is a vector of errors whose elements are independent and identically distributed $N(0, \sigma^2)$. To avoid the need for an intercept term we assume that Y and X are centred, i.e. $\mathbf{1}_n' Y = 0$ and $\mathbf{1}_n' X = \mathbf{0}_p'$, and so X has rank $m \leqslant \min(n-1, p)$.

There are various methods of estimating the parameter vector β. These include OLS, ridge regression (RR), principal components regression (PCR) and PLS regression. When $n < p$, OLS is not uniquely defined but is usually made unique by choosing the minimum length least squares estimator. For all these methods, the estimator $\hat{\beta}$ may be expressed in a general form by using the spectral decomposition

$$X'X = \sum_{j=1}^{m} \lambda_j u_j u_j'$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m$ are the positive eigenvalues of $X'X$ and u_j are the corresponding eigenvectors. Then

$$\hat{\beta} = \sum_{j=1}^{m} f(\lambda_j)\hat{\alpha}_j u_j, \qquad \hat{\alpha}_j = \frac{1}{\lambda_j} z_j \tag{1}$$

where the $f(\lambda_j)$ are shrinkage (or scale) factors (Frank and Friedman, 1993), the $\hat{\alpha}_j$ are the coefficients of the OLS estimator in the principal directions u_j and the $z_j = u_j' X' Y$ are the canonical covariances. Note that the coefficients $\hat{\alpha}_j$ are mutually orthogonal and independent.

Setting $f(\lambda_j)$ to values other than 1 introduces bias into the estimation process, but if $f(\lambda_j) < 1$ it will also lead to a reduction in the variance of $\hat{\beta}$. Shrinkage is beneficial because

for some values of $f(\lambda_j)$ the reduction in variance more than offsets the increase in bias so that the mean-squared error (MSE) of $\hat{\beta}$ is reduced. In contrast, setting $f(\lambda_j) > 1$ is guaranteed to increase the MSE since it simultaneously increases the bias and variance of $\hat{\beta}$.

The function $f(\lambda)$ determining the shrinkage factors is called the shrinkage function. Under OLS, $f(\lambda) = 1$ by definition, whereas, under RR, $f(\lambda) = \lambda/(\lambda + k)$ for some constant $k > 0$. Similarly, under PCR, $f(\lambda_j) = 1$ if the eigenvector u_j is included in the model and $f(\lambda_j) = 0$ otherwise. Both RR and PCR are regarded as shrinkage methods as by shrinking some of the coefficients $\hat{\alpha}_j$ they shrink $\hat{\beta}$.

PLS regression is already known to be a shrinkage estimator in the sense that $|\hat{\beta}_{\mathrm{PLS}}| \leqslant |\hat{\beta}_{\mathrm{OLS}}|$ (Goutis, 1996). However, the shrinkage properties of PLS are shown in Section 3 to be complicated as some of the coefficients $\hat{\alpha}_j$ are actually expanded, i.e. $f(\lambda_j) > 1$.

3. Partial least squares regression

PLS regression is usually described as a two-stage approach. The first stage determines the $k \leqslant m$ factors t_i of length n to be included in the regression. The ith factor $t_i = Xc_i$ is chosen to maximize $t_i'Y$ subject to the constraints that $|c_i| = 1$ and that t_i is orthogonal to the space spanned by the basis $\{t_1, t_2, \ldots, t_{i-1}\}$. The second stage of PLS regresses Y on the factors t_i $(i = 1, 2, \ldots, k)$. This involves minimizing the sum of squares

$$\left(Y - \sum_{i=1}^{k} b_i t_i\right)' \left(Y - \sum_{i=1}^{k} b_i t_i\right) = (Y - X\beta_{\mathrm{PLS}})'(Y - X\beta_{\mathrm{PLS}}) \tag{2}$$

with respect to b_i where $\beta_{\mathrm{PLS}} = \Sigma_{i=1}^{k} b_i c_i$ is the PLS parameter vector. Predictions $\hat{y}_{x'}$ of future responses can then be made by using

$$\hat{y}_{x'} = \bar{y} + \sum_{j=1}^{p} \hat{\beta}_{\mathrm{PLS},j}(x_j' - \bar{x}_j)$$

where $\bar{y}$ and $\bar{x}_j$ $(j = 1, 2, \ldots, p)$ are the sample means used in centring Y and X. The prediction depends on the number of factors k included in the regression, which is often determined by using cross-validation (see, for example, Brown (1993)).

From Helland (1988) we have

$$\mathrm{span}\{c_1, c_2, \ldots, c_k\} = \mathrm{span}\{X'Y, (X'X)X'Y, \ldots, (X'X)^{k-1}X'Y\} \tag{3}$$

so an alternative form for the PLS regression estimate of β is

$$\hat{\beta}_{\mathrm{PLS}} = \sum_{i=1}^{k} \hat{\gamma}_i (X'X)^{i-1} X'Y \tag{4}$$

where the parameters $\hat{\gamma}_i$ minimize equation (2). Using the spectral decomposition of $X'X$, it follows (Frank and Friedman, 1993) that $\hat{\beta}_{\mathrm{PLS}}$ is given by equations (1) with shrinkage function

$$f(\lambda) = \sum_{i=1}^{k} \hat{\gamma}_i \lambda^i. \tag{5}$$

The estimator $\hat{\beta}_{\mathrm{PLS}}$ may also be obtained by minimizing

$$(\hat{\beta}_{\mathrm{OLS}} - \beta_{\mathrm{PLS}})' X'X (\hat{\beta}_{\mathrm{OLS}} - \beta_{\mathrm{PLS}}) \tag{6}$$

rather than equation (2) since

$$(Y - X\beta_{\mathrm{PLS}})'(Y - X\beta_{\mathrm{PLS}}) = (Y - X\hat{\beta}_{\mathrm{OLS}})'(Y - X\hat{\beta}_{\mathrm{OLS}}) + (\hat{\beta}_{\mathrm{OLS}} - \beta_{\mathrm{PLS}})'X'X(\hat{\beta}_{\mathrm{OLS}} - \beta_{\mathrm{PLS}}).$$

Again, using the spectral decomposition of $X'X$, expression (6) is equivalent to

$$\sum_{j=1}^{m} \frac{z_j^2}{\lambda_j}\left(1 - \sum_{i=1}^{k} \gamma_i \lambda_j^i\right)^2. \tag{7}$$

Differentiating this with respect to γ_q, we have for $q = 1, 2, \ldots, k$

$$\sum_{j=1}^{m} z_j^2 \lambda_j^{q-1}\left(1 - \sum_{i=1}^{k} \hat{\gamma}_i \lambda_j^i\right) = 0. \tag{8}$$

An immediate consequence of equations (5) and (8) required below is that

$$\sum_{j=1}^{m} z_j^2 \, p(\lambda_j)\{1 - f(\lambda_j)\} = 0 \tag{9}$$

for any polynomial $p(\lambda)$ of order less than k.

The theorems below provide insight into the shrinkage properties of PLS regression. They make the reasonable assumptions for empirical data that $z_j \neq 0$ and that the eigenvalues λ_j are distinct. We assume $k < m$ since when $k = m$ it is already known that $\hat{\beta}_{\mathrm{PLS}} = \hat{\beta}_{\mathrm{OLS}}$ (Helland, 1988).

Let $S = \{\lambda_j: 0 \leqslant f(\lambda_j) < 1, j = 1, \ldots, m\}$ be the set of eigenvalues λ_j where the coefficients $\hat{\alpha}_j$ are shrunk, and let $E = \{\lambda_j: f(\lambda_j) > 1, j = 1, \ldots, m\}$ be the set where the coefficients are expanded. Also, we say that R is a set of consecutive eigenvalues if $R = \{\lambda_j: a \leqslant j \leqslant b\}$ for some integers $a \leqslant b$, and $R_1 < R_2$ ('R_1 is less than R_2') if the largest element in R_1 is less than the smallest element in R_2.

Theorem 1. Under PLS regression with $k < m$ factors,

$$S = \bigcup_{i \text{ odd}} R_i, \qquad E = \bigcup_{i \text{ even}} R_i, \qquad i = 1, 2, \ldots, k+1,$$

where R_i are non-empty disjoint sets of consecutive eigenvalues and $R_1 < R_2 < \ldots < R_{k+1}$.

Proof. From equation (5), the function $g(\lambda) = 1 - f(\lambda)$ has k roots. It will first be shown that the roots of $g(\lambda)$ are distinct and that they are all on the open interval (λ_m, λ_1). Suppose that this is not true and that there are only $k' < k$ distinct roots of $g(\lambda)$ on (λ_m, λ_1). Call these roots $a_1, a_2, \ldots, a_{k'}$. Now consider the polynomial

$$h(\lambda) = g(\lambda) \prod_{i=1}^{k'} (a_i - \lambda)^{b_i}$$

where $b_i = 1$ if a_i is a root of $g(\lambda)$ of odd order and $b_i = 0$ if a_i is a root of even order. Observe that $h(\lambda)$ is clearly either greater than or equal to 0 or less than or equal to 0 on the whole of the closed interval $[\lambda_m, \lambda_1]$. Also, $h(\lambda) = 0$ at only $k' < m$ points on the interval, and therefore $\Sigma z_j^2 h(\lambda_j) \neq 0$. However, equation (9) implies that the sum equals 0 and so there is a contradiction. Therefore, $g(\lambda)$ must have k distinct roots on (λ_m, λ_1). Call the roots of $g(\lambda)$ in increasing order $a_1, a_2, \ldots, a_k$ and note that they are all positive since $\lambda_1 > 0$. Also define $a_0 = 0$ and $a_{k+1} = \infty$, and consider the interval $(0, \infty)$. Now, as $g(0) = 1$, it follows that $g(\lambda)$ is positive on the intervals (a_{i-1}, a_i) when i is odd, and $g(\lambda)$ is negative on the intervals

(a_{i-1}, a_i) when i is even. Equivalently, $f(\lambda) < 1$ and $f(\lambda) > 1$ on the intervals (a_{i-1}, a_i) when i is odd and i is even respectively.

Now define $R_i = \{\lambda_j: \lambda_j \in (a_{i-1}, a_i),\ j = 1, \ldots, m\}$ for $i = 1, 2, \ldots, k+1$. The theorem then follows except to show that the sets R_i are non-empty. The two extreme sets R_1 and R_{k+1} are non-empty as they contain λ_m and λ_1 respectively. Now suppose that the set R_i for some $2 \leqslant i \leqslant k$ is empty, and consider the polynomial

$$h_i(\lambda) = g(\lambda) \prod_{r=1}^{k} (a_r - \lambda)^{b_r}$$

where $b_r = 0$ if $r = i$ or $r = i-1$, and $b_r = 1$ otherwise. Then, using exactly the same procedure as for $h(\lambda)$ above, it can be shown that there is a contradiction, and therefore all the sets R_i $(i = 1, 2, \ldots, k+1)$ are non-empty. □

Theorem 1 demonstrates the oscillatory behaviour of the shrinkage factors between S and E. This is in marked contrast with other shrinkage methods such as RR and PCR where the set E is empty. The corollaries below follow immediately from theorem 1 and highlight some of the unusual shrinkage properties of PLS regression.

Corollary 1. $\lambda_m \in S$.

Corollary 2. $\lambda_1 \in S$ if k is even, and $\lambda_1 \in E$ if k is odd.

Corollary 3. If $k = 1$, $S = \{\lambda_j: j \geqslant b\}$ and $E = \{\lambda_j: j \leqslant a\}$ where $1 \leqslant a < b \leqslant m$.

The following two theorems provide further insight into the shrinkage factors.

Theorem 2. If $k = m - 1$, then $S = \{\lambda_j: m - j \text{ is even}\}$, $E = \{\lambda_j: m - j \text{ is odd}\}$ and

$$f(\lambda_j) = 1 - C\left\{z_j^2 \prod_{i=1, i\neq j}^{k+1} (\lambda_i - \lambda_j)\right\}^{-1}, \qquad j = 1, 2, \ldots, m, \tag{10}$$

for some positive constant C.

Proof. The first part of theorem 2 follows from theorem 1 and so it is sufficient to show that the shrinkage factors are of the correct form. To do this, choose any two elements $j_1, j_2 \in \{1, 2, \ldots, m\}$ and consider the polynomial of order $k - 1$

$$h(\lambda) = \prod_{i \in M} (\lambda_i - \lambda)$$

where $M = \{1, 2, \ldots, m\} \setminus \{j_1, j_2\}$. Then

$$\sum_{j=1}^{m} z_j^2\, h(\lambda_j)\{1 - f(\lambda_j)\} = z_{j_1}^2\, h(\lambda_{j_1})\{1 - f(\lambda_{j_1})\} + z_{j_2}^2\, h(\lambda_{j_2})\{1 - f(\lambda_{j_2})\}$$

and by equation (9) this equals 0. As a consequence, for all $j \in \{1, 2, \ldots, m\}$

$$z_j^2\{1 - f(\lambda_j)\} \prod_{i=1, i\neq j}^{k+1} (\lambda_i - \lambda_j) = C$$

for some constant C. Theorem 2 then follows on rearranging terms. It is possible to find an explicit expression for C in terms of λ_j and z_j but this is not included here as it is not of direct interest. □

Note that the term in theorem 2

$$\prod_{i=1,\, i\neq j}^{k+1} (\lambda_i - \lambda_j) \tag{11}$$

is positive for the smallest eigenvalue λ_m and then alternates between being positive and negative for subsequent eigenvalues. Consequently, the effects on the coefficients $\hat{\alpha}_j$ alternate between shrinkage and expansion, as anticipated by theorem 1.

Theorem 3. If the set R_1 contains $m - k$ elements, then the remaining sets R_2, R_3, . . ., R_{k+1} contain one element each, and

$$f(\lambda_j) = 1 - C_j \left\{ z_j^2 \prod_{i=1,\, i\neq j}^{k+1} (\lambda_i - \lambda_j) \right\}^{-1}, \qquad j = 1, 2, \ldots, k, \tag{12}$$

where $C_j > 0$ is a decreasing function of j and $C_j/(1 - \lambda_{k+1}/\lambda_j)$ is an increasing function of j.

Proof. If R_1 contains $m - k$ elements, the remaining sets have one element each due to theorem 1. Now consider the polynomial of order $k - 1$

$$h(\lambda) = \prod_{i=1,\, i\neq j}^{k} (\lambda_i - \lambda), \qquad j \in \{1, 2, \ldots, k\}.$$

We have

$$\sum_{j'=1}^{m} z_{j'}^2\, h(\lambda_{j'})\{1 - f(\lambda_{j'})\} = z_j^2\, h(\lambda_j)\{1 - f(\lambda_j)\} + \sum_{r=k+1}^{m} z_r^2\, h(\lambda_r)\{1 - f(\lambda_r)\}$$

which equals 0 by equation (9). On rearranging the right-hand side of the above equation and using equation (12),

$$C_j = (\lambda_j - \lambda_{k+1}) \sum_{r=k+1}^{m} \frac{q(\lambda_r)}{\lambda_j - \lambda_r} \tag{13}$$

where

$$q(\lambda_r) = z_r^2 \{1 - f(\lambda_r)\} \prod_{i=1}^{k} (\lambda_i - \lambda_r).$$

Equation (13) can then be used to show that C_j has the properties stated in theorem 3. □

Theorem 3 deals with a common scenario as is highlighted by examples 2 and 3 in Section 4. The shrinkage factors (12) of the k largest eigenvalues are the same as in theorem 2 except that the constant C has been replaced by C_j. Nevertheless, the shrinkage factors share a similar dependence on the terms (11) and z_j^2. In practice, the magnitudes of both terms are generally smaller at the lower eigenvalues, and consequently the lower eigenvalues tend to exhibit more expansion and shrinkage. This behaviour is seen very clearly in the examples considered by Frank and Friedman (1993) (see example 2 in Section 4).

4. Examples

To illustrate the shrinkage behaviour of PLS regression and the theorems given above, we give two artificial examples and one based on real data.

In the first example we have deliberately chosen equispaced eigenvalues ranging from 1 to 11 and have set $z_j^2 = \lambda_j$ for all j. Although this may not reflect the sort of data that PLS regression is typically applied to, these values show the shrinkage behaviour particularly clearly.

Fig. 1 gives shrinkage factors for one-, two-, four- and seven-factor PLS regressions. This clearly illustrates the expansionist behaviour of PLS regression for some coefficients even when relatively large numbers of factors are included. For PLS regression with one factor, the coefficients associated with the four largest eigenvalues are expanded whereas the remainder are shrunk. For two-factor PLS, only the coefficients associated with the medium-sized eigenvalues are expanded. Note that for two-, four- and seven-factor PLS regressions the greatest expansion does not occur for the eigenvalues near to the kth-largest eigenvalue.

In their study, Frank and Friedman (1993) considered four separate situations. Here we consider their third situation for which $\lambda_j = 1/j^2$ and $z_j = 1/j^2$ for $j = 1, \ldots, 10$. They regarded this as corresponding to moderate multicollinearity in the X-variables and having covariances which are favourable to the use of PLS regression. Fig. 2 reproduces their results in the case of two and four factors, but with the shrinkage plotted against the corresponding eigenvalue (on the log-scale) rather than against j. In the case of two-factor PLS regression we find that the coefficient associated with the second-largest eigenvalue is expanded whereas the remaining coefficients are shrunk. In the case of four-factor regression the coefficient associated with the fourth-largest eigenvalue is expanded and to a much smaller extent so is the coefficient for the second-largest eigenvalue. In both cases we observe substantial shrinkage for coefficients associated with small eigenvalues. Frank and Friedman appeared to conclude from this that k-factor PLS regression will expand coefficients at and nearby the

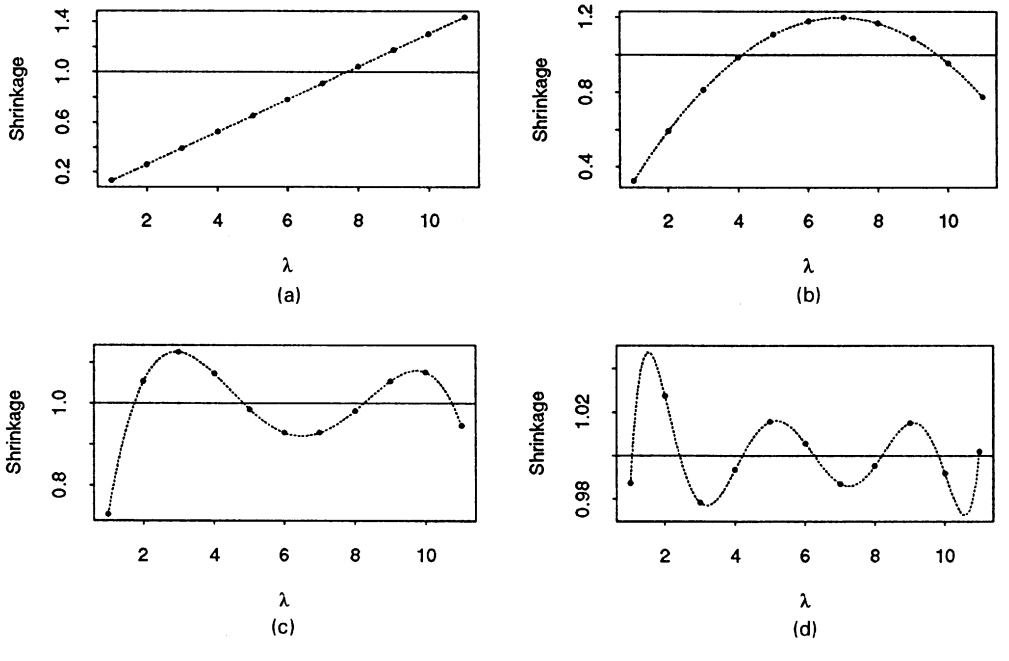

Fig. 1. Shrinkage functions (·····) for example 1 for (a) one-factor PLS regression, (b) two-factor PLS regression, (c) four-factor PLS regression and (d) seven-factor PLS regression: points correspond to the shrinkage factors for the eigenvalues of the $X'X$ matrix

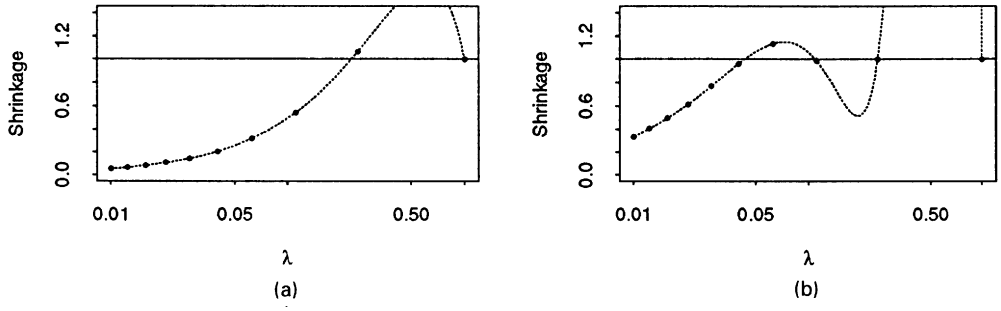

Fig. 2. Shrinkage functions (·····) for example 2 for (a) two-factor PLS regression and (b) four-factor PLS regression: points correspond to the shrinkage factors for the eigenvalues of the $X'X$ matrix

kth-largest eigenvalue. However, this is not necessarily the case as was illustrated in our first example.

For our final example we use data from a calibration experiment to determine the chemical composition of liquid detergent by using mid-infra-red spectroscopy (Brown, 1990). The original data consisted of 12 detergent samples made from mixtures of four chemicals together with water. For each detergent sample a mid-infra-red spectrum of absorbances at 1168 frequencies from 3100 to 750 cm^{-1} was recorded. Here we consider regression of the first chemical component on the spectral measurements. After centring, the eigenvalues Λ of $X'X$ and the canonical covariances z are

$$\Lambda = \{8.0059,\ 6.0324,\ 1.4529,\ 0.8665,\ 0.0201,\ 0.0122,\ 0.0053,\ 0.0033,\ 0.0020,\ 0.0017,\ 0.0014\},$$

$$z = \{22.2858,\ -8.0025,\ -10.5844,\ 3.6489,\ -0.0070,\ -0.0341,\ 0.0004,\ -0.0018,\ -0.0039,\ 0.0005,\ 0.0039\}$$

Table 1 gives the canonical shrinkage for PLS regression assuming various numbers of factors. Here, we observe that the coefficient associated with the kth eigenvalue is expanded

Table 1. Shrinkage factors $f(\lambda_j)$ for PLS regression applied to the detergent data†

j	*Results for the following numbers of factors:*							
	1	*2*	*3*	*4*	*5*	*7*	*9*	*10*
1	**1.2106**	0.9390	**1.0017**	1.0000	**1.0000**	**1.0000**	**1.0000**	1.0000
2	0.9122	**1.6663**	0.9733	**1.0000**	1.0000	1.0000	1.0000	**1.0000**
3	0.2197	0.9371	**1.0404**	1.0000	**1.0000**	**1.0000**	**1.0000**	1.0000
4	0.1310	0.5998	0.7232	**1.0003**	1.0000	1.0000	1.0000	**1.0000**
5	0.0030	0.0153	0.0206	0.0424	**1.6111**	**1.0046**	**1.0000**	1.0000
6	0.0018	0.0093	0.0125	0.0258	0.9944	0.9993	1.0000	**1.0000**
7	0.0008	0.0040	0.0055	0.0113	0.4383	**1.8250**	**1.3489**	0.9999
8	0.0005	0.0025	0.0034	0.0070	0.2741	**1.5271**	0.9316	**1.0001**
9	0.0003	0.0015	0.0021	0.0043	0.1666	**1.1003**	**1.0200**	0.9999
10	0.0003	0.0013	0.0017	0.0036	0.1417	0.9716	**1.0196**	**1.0082**
11	0.0002	0.0011	0.0014	0.0030	0.1167	0.8306	0.9892	0.9999

†Figures have been calculated exactly and then rounded to four decimal places. Numbers in bold indicate shrinkage factors which exceed 1.

when using k-factor PLS and that the shrinkage factors associated with the largest eigenvalues approach 1 as the number of factors in the PLS regression is increased. For this example there is a tendency for the magnitude of the canonical covariances z_j to be large when the corresponding eigenvalues are large. As a consequence, the shrinkage factor for these terms tends to be relatively close to 1, as anticipated by theorem 3.

5. Discussion

This paper has demonstrated the peculiar shrinkage properties of PLS regression. Despite this the method has often been shown to work reasonably well in some fields of application. One reason why this may be so is that these situations correspond to having large canonical covariances for large eigenvalues of the $X'X$ matrix, in which case expanded coefficients are not expanded by much but those coefficients associated with small eigenvalues and small canonical covariances are shrunk substantially. However, as we have demonstrated, the PLS regression method can break down quite badly when the canonical covariances are of a more similar magnitude (see example 1).

In conclusion, the shrinkage properties derived in this paper suggest that PLS regression should be used with caution and should not be adopted as an automatic solution to multicollinearity in regression problems. At the very least users should examine the canonical coefficients of PLS relative to OLS to ensure that none are expanded substantially.

References

Brown, P. J. (1990) Partial least squares in perspective. *Anal. Proc. R. Soc. Chem.*, **27**, 303–306.

———(1993) *Measurement, Regression, and Calibration*. Oxford: Clarendon.

Denham, M. C. (1991) Calibration in infrared spectroscopy. *PhD Thesis*. University of Liverpool, Liverpool.

Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics tools. *Technometrics*, **35**, 109–135.

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

Helland, I. (1988) On the structure of partial least squares regression. *Communs Statist. Simuln Computn*, **17**, 581–607.

Martens, H. and Jensen, S. A. (1983) Partial least squares regression: a new two-stage NIR calibration method. In *Progress in Cereal Chemistry and Technology: Proc. 7th Wrld Cereal and Bread Congr. Prague, June 1982* (eds J. Holas and J. Kratochvil), pp. 607–647. Amsterdam: Elsevier.

Stone, M. and Brooks, R. J. (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. R. Statist. Soc.* B, **52**, 237–269.

Wold, H. (1966) Nonlinear estimation by iterative least squares procedures. In *Research Papers in Statistics: Festschrift for J. Neyman* (ed. F. N. David), pp. 411–444. New York: Wiley.

———(1973) Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivariate Analysis III* (ed. P. Krishnaiah), pp. 383–407. New York: Academic Press.

———(1982) Soft modeling: the basic design and some extensions. In *Systems under Indirect Observation: Causality–Structure–Prediction* (eds K. G. Jöreskog and H. Wold), vol. II, ch. 1, pp. 1–54. Amsterdam: North-Holland.