RR-00-06

Principal Components of Simple Least Squares: A New Weighting Scheme for Principal Component Regression

G.M. Merola
Universitat Politécnica de Catalunya
Bovas Abraham
University of Waterloo

June 2000

**This is a revised version of research report RR-00-04.*

Principal Components of Simple Least Squares: A New Weighting Scheme for Principal Component Regression

G.M. Merola and B. Abraham

University of Waterloo

Abstract

In this paper we give a novel interpretation of a well known dimensionality reduction method for prediction, Partial Least Squares (PLS). We propose an alternative method, called PCSLS, in which the predictive subspace is obtained from weighted principal components of the observed regressors. We compare this method with PLS through simulated and published data.

Key Words: *Partial Least Squares, Principal Component Regression, Prediction, Dimension Reduction, Principal Components of Simple Least Squares.*

1 Introduction

In the context of a linear regression model involving many (possibly correlated) explanatory variables sometimes better predictions can be obtained from a set of fewer linear combinations of the explanatory variables. These linear combinations are called latent variables (lv's) and the methods that generate the lv's take the generic name of dimensionality reduction methods (DRMs).

The use of DRMs for prediction has been proven successful in many fields, such as Chemometrics (e.g. Gelaldi and Kowalski (1986)), monitoring of chemical reactors (e.g. Kourti and MacGregor (1996)), Quantitative Structure Activity Relationships (e.g. Schmidli (1995)) and Sensory Analysis (e.g. Næs and Risvik (1996)).

The lv's can be obtained by means of several different methods. The regression of the responses on the first few principal components of the explanatory variables (PCR) and Partial Least Squares (PLS) (Wold (1982)) have been reported to yield accurate predictions in several applications. However, model based justifications and optimal properties for the predictions obtained with these DRMs are not available. Merola and Abraham (2001) discuss optimal estimation of DRMs for prediction and suggest a different class of methods.

In the next section, after introducing some notation, we briefly discuss Principal Component Regression (PCR). Subsequently we introduce PLS, also giving a new interpretation of it. Based on this interpretation, in the third section we introduce a new method called Principal Components of Simple Least Squares (PCSLS). In the fourth section we compare PLS and PCSLS through simulations and published data. Finally in Section 5 we give some concluding remarks.

2 Dimensionally Reduced Prediction Models

Let $\mathbf{Y}$ be an $(n \times q)$ matrix and $\mathbf{X}$ an $(n \times p)$ matrix whose columns consist of n independent observations on q responses and on p explanatory variables respectively. For univariate regression we denote with $\mathbf{y}$ the vector containing the observations on the response variable while for multivariate responses we denote with $\mathbf{y}_j$ the column of $\mathbf{Y}$ containing the observations on the j-th response. For simplicity but without lost of generality, we take all the variables to be centred to zero-mean.

DRMs consist of determining d orthogonal lv's $\mathbf{t}_k = \mathbf{X}\mathbf{a}_k,\ k = 1, \dots, d$, where the $\mathbf{a}_k$ are p-vectors containing the coefficients of the lv's. The $\mathbf{X}$ space is then partitioned into two parts: the latent space $\mathbf{T}_{(d)} = (\mathbf{t}_1, \dots, \mathbf{t}_d)$ and its orthogonal complement $(\mathbf{X} \perp \mathbf{T}_{(d)})$. The lv's are then used as regressors for the responses and the predicted values are obtained by Ordinary Least Squares (OLS) as

$$\hat{\mathbf{Y}}_{[d]} = \mathbf{T}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y} = \sum_{k=1}^{d} \mathbf{t}_k(\mathbf{t}'_k\mathbf{t}_k)^{-1}\mathbf{t}'_k\mathbf{Y} \qquad (2.1)$$

where $\hat{\mathbf{Y}}_{[d]}$ denotes the predictions obtained using d lv's, $\mathbf{T}_{(d)} = \mathbf{X}\mathbf{A}_{(d)}$ the $(n \times d)$ matrix whose columns are the latent variables and $\mathbf{A}_{(d)}$ the $(p \times d)$ matrix whose columns are the coefficients of the lv's. Substituting $\mathbf{T}_{(d)} = \mathbf{X}\mathbf{A}_{(d)}$ into (2.1) we obtain

$$\hat{\mathbf{Y}}_{[d]} = \mathbf{X}\mathbf{A}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y} = \mathbf{X}\mathbf{B}_{[d]}$$

where $\mathbf{B}_{[d]} = \mathbf{A}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y}$ is the estimate of the regression coefficients obtained with d lv's. Note that when the lv's are to be used for prediction their length is irrelevant, therefore we can, and will occasionally, refer to the generic direction in the $\mathbf{X}$ space rather than to the actual lv, without loss of generality.

The approach of estimating the coefficients $\mathbf{a}_k$ by least squares, that is minimising the Euclidean norm $||\mathbf{Y}-\mathbf{X}\mathbf{B}_{[d]}||$, often is not effective. In this approach, called Reduced Rank Regression (RRR) (Izenman (1975)), the solutions are obtained minimising the residual sum of squares for the ordinary regression model over the additional rank constraints. In fact, for univariate response the only solution $\mathbf{a}_1$ is the vector of OLS regression coefficients and for multivariate responses the lv's are the principal components of the OLS solutions (e.g. Merola (1998)).

Methods that are claimed to yield better predictions do not take the LS approach to estimating the coefficients. Next we illustrate these methods dropping the reference to d, as they are invariant to the number of lv's computed.

2.1 Principal Component Regression

PCR consists of regressing the responses onto the first d principal components. The principal components form a sequence of orthogonal axis of the space spanned by the $\mathbf{X}$ variables that sequentially minimise the norm of the residual orthogonal space. That is the $k-th$ principal component $\mathbf{t}_k = \mathbf{X}\mathbf{a}_k$ is the solution of the optimisation problem:

$$\min_{\mathbf{t}'_k\mathbf{t}_j=\delta_{kj}} ||\mathbf{X} - \mathbf{t}_k\mathbf{p}'_k|| \qquad (2.2)$$

where δ_{kj} is equal to 1 if $k = j$ and to 0 otherwise and the $\mathbf{p}_k$'s are vectors of unknown parameters. The solutions $\mathbf{a}_k$, $k = 1, \dots, d$ are given by the eigenvectors of the matrix $\mathbf{X}'\mathbf{X}$ corresponding to the first d eigenvalues taken in non-increasing order. This procedure is generally referred to as Principal Components Analysis (PCA).

The resulting latent variables, known as principal components, are the linear combinations of the x variables that have maximal norm.

It is well known that PCA is very sensitive to the variance of the x variables and that the first principal component will be "closer" to the variables with larger variance. This property may be undesirable, especially when the units of measure of the variables are not comparable. Furthermore, in a predictive context there is no a-priori reason for which the regressors with larger variance should be better predictors of the response than those with smaller variance.

In order to overcome the problem connected with the variance of the explanatory variables it is customary to standardise them to unit length (autoscale) prior to PCA. That is, PCA is performed on the scaled matrix $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{D}^{-1}$, where $\mathbf{D}$ is a diagonal matrix with diagonal elements $d_{jj} = \sqrt{\mathbf{x}_j'\mathbf{x}_j}$, $j = 1, \dots, p$. Also autoscaling presents some drawbacks and other scaling policies may be adopted. In general, PCA performed on the matrix $\mathbf{X}$ post-multiplied by a diagonal positive definite matrix of weights $\mathbf{W}$ is called *weighted* PCA. The lv's in weighted PCA are the eigen-vectors of the matrix $\mathbf{W}^2\mathbf{X}'\mathbf{X}$.

Obviously, the principal components are "independent" of the responses and PCR can be applied to univariate and multivariate regression.

2.2 Partial Least Squares

PLS was introduced by Wold (1982) as an algorithm for prediction without any "hard" modelling behind, hence without any explicit optimality property. The mathematical functioning of the algorithm was explained by Hoskuldsson (1988), Helland (1988) and de Jong (1993); Phatak et al. (1992) and Merola (1998) contributed to explaining its geometry. However, nobody seems to have succeeded in finding a convincing optimality property for the prediction of the responses or even a rationale for its use; nonetheless PLS is extensively used in many fields.

PLS can be applied to univariate and multivariate regression but the multivariate version is not considered a straightforward generalisation of the univariate one. We examine the univariate case first and then the multivariate case.

Univariate partial least squares

A simplified univariate PLS algorithm is outlined in Algorithm A.1 in the appendix. Hoskuldsson (1988) showed that at each step PLS determines the coefficient as $\mathbf{a}_k = \mathbf{F}_k'\mathbf{y}/||\mathbf{F}_k'\mathbf{y}||$, were $\mathbf{F}_k = (\mathbf{I}_n - \mathbf{t}_k(\mathbf{t}_k'\mathbf{t}_k)^{-1}\mathbf{t}_k')\mathbf{F}_{(k-1)}$ is the matrix of orthogonal residuals, called "deflated $\mathbf{X}$ matrix" $(\mathbf{F}_1 = \tilde{\mathbf{X}})$.

Garthwaite (1994) shows that the PLS lv's are proportional to the weighted averages of the simple regressions of the response on each explanatory variable. That is:

$$\mathbf{t}_k \propto \mathbf{F}_k\mathbf{F}_k'\mathbf{y} = \sum_{j=1}^{p} \mathbf{f}_{(k)j}\mathbf{f}_{(k)j}{}'\mathbf{y} = \sum_{j=1}^{p} \hat{\mathbf{y}}(\mathbf{f}_{(k)j})(\mathbf{f}_{(k)j}{}'\mathbf{f}_{(k)j}) \qquad (2.3)$$

where $\mathbf{f}_{(k)j}$ is the j-th column of $\mathbf{F}_k$ and $\hat{\mathbf{y}}(\mathbf{f}_{(k)j}) = \mathbf{f}_{(k)j}(\mathbf{f}_{(k)j}'\mathbf{f}_{(k)j})^{-1}\mathbf{f}_{(k)j}'\mathbf{y}$. However, noting that in PLS the x variables are always autoscaled, the first PLS lv amounts to the simple average of the projections of $\mathbf{y}$ onto the individual $\mathbf{x}_j$'s, $(\hat{\mathbf{y}}(\mathbf{x}_j), j = 1, \ldots, p$. Let

$$\hat{\mathbf{Y}}_u = (\hat{\mathbf{y}}(x_1), \ldots, \hat{\mathbf{y}}(x_p)) = \tilde{\mathbf{X}}\tilde{\mathbf{B}}_u = \mathbf{X}\mathbf{B}_u, \qquad (2.4)$$

where $\tilde{\mathbf{B}}_u$ is a diagonal matrix with diagonal elements equal to $\{\tilde{\mathbf{x}}_j'\mathbf{y}\}$ and $\mathbf{B}_u$ is diagonal with diagonal elements equal to $\{\frac{\mathbf{x}_j'\mathbf{y}}{\mathbf{x}_j'\mathbf{x}_j}\}$. Then the first PLS lv satisfies:

$$\mathbf{t}_1 \propto \sum_{j=1}^{p} \hat{\mathbf{y}}(\mathbf{x}_j) = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{y} = \hat{\mathbf{Y}}_u\mathbf{1}_p,$$

where $\mathbf{1}_p$ is the p-vector of ones. A well known property of the simple average is that it minimises the sum of squared distances of a set of values from a point. That is, the first PLS lv is along the direction that minimises the quantity

$$||\hat{\mathbf{Y}}_u - \mathbf{t}_1 \mathbf{1}_p'||^2 = \sum_{j=1}^{p} (\hat{\mathbf{y}}(x_j) - \mathbf{t}_1)'(\hat{\mathbf{y}}(x_j) - \mathbf{t}_1). \tag{2.5}$$

In the subsequent iterations the deflated matrix $\mathbf{F}_k$ is not autoscaled, hence the successive lv's are weighted averages of the simple regressions with weights proportional to the squared norms of the residuals, as given in equation (2.3).

The use of weighted averages for the lv's successive to the first one gives higher weight to the $\mathbf{x}_j$'s that have not been well "explained" by the previous components, like in PCR. In a predictive context one can see this weighting policy as the continuous analogue of the "tolerance threshold" in the step-wise algorithm for selecting regressors, for which the variables that are well explained by the others in the model are eliminated from the candidates to enter (e.g. Weisberg (1985)). Autoscaling the residuals $\mathbf{F}_k$ at each iteration would render these lv's homogeneous with the first one. We will refer to this modified PLS procedure as Partial Least Squared with Scaled F's (PLSSF).

Multivariate partial least squares

A simplified multivariate PLS algorithm is outlined in Algorithm A.2 in the appendix. Hoskuldsson (1988) shows that the solutions $\mathbf{a}_k$ can be computed directly from the matrix $\mathbf{F}_k'\mathbf{Y}\mathbf{Y}'\mathbf{F}_k$ as the eigen-vector corresponding to the largest eigen-value. Therefore, if we let ϕ_k be these eigen-values, at each iteration the coefficients $\mathbf{a}_k$ satisfy:

$$\mathbf{F}_k'\mathbf{Y}\mathbf{Y}'\mathbf{F}_k\mathbf{a}_k = \mathbf{a}_k\phi_k.$$

Also the multivariate PLS algorithm can be interpreted using the simple regressions of the responses on the individual $\mathbf{x}$ variables. Let

$$\hat{\mathbf{Y}}(j) = (\hat{\mathbf{y}}_1(\mathbf{x}_j), \hat{\mathbf{y}}_2(\mathbf{x}_j), \ldots, \hat{\mathbf{y}}_q(\mathbf{x}_j)) = \tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j'\mathbf{Y} = \tilde{\mathbf{x}}_j\tilde{\mathbf{b}}'(j) \tag{2.6}$$

be the $(n \times q)$ consensus matrices whose i-th column is the projection of $\mathbf{y}_i$ on $\mathbf{x}_j$. To each variable $\tilde{\mathbf{x}}_j$ corresponds a vector of q "weights" $\tilde{\mathbf{b}}'(j) = \tilde{\mathbf{x}}_j'\mathbf{Y} = \{\tilde{\mathbf{x}}_j'\mathbf{y}_i\}$. PLS determines a unit-norm vector of q coefficients, $\mathbf{c} = (c_1, \ldots, c_q)'$ so that the sum of the squared norms of the vectors $\mathbf{v}_j = \hat{\mathbf{Y}}(j)\mathbf{c}$ is maximal. That is, $\mathbf{c}$ is the solution to

$$\max_{\mathbf{c}'\mathbf{c}=1} \sum_{j=1}^{p} \mathbf{v}_j'\mathbf{v}_j = \max_{\mathbf{c}'\mathbf{c}=1} \sum_{j=1}^{p}\sum_{i=1}^{q} ||\hat{\mathbf{y}}_i(\mathbf{x}_j)||^2 c_i^2 = \max_{\mathbf{c}'\mathbf{c}=1} \mathbf{c}'\mathbf{Y}'\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{Y}\mathbf{c}. \tag{2.7}$$

Since $||\hat{\mathbf{y}}_i(\mathbf{x}_j)||$ is a measure of predictability, the c_i's are coefficients that maximise the overall prediction of each variable. The solution of (2.7) for $\mathbf{c}$ is the eigen-vector:

$$\mathbf{Y}'\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{Y}\mathbf{c} = \mathbf{c}\phi_1,\ \phi_1 > \phi_l,\ l > 1 \tag{2.8}$$

In terms of the autoscaled variables $\tilde{\mathbf{X}}$ the vectors $\mathbf{v}_j$ are

$$\mathbf{v}_j = \tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j'\mathbf{Y}\mathbf{c} = \tilde{\mathbf{x}}_j\tilde{a}_j \tag{2.9}$$

where $\tilde{a}_j = \tilde{\mathbf{x}}_j'\mathbf{Y}\mathbf{c}$. Let $\mathbf{t}$ be proportional to the sum of the $\mathbf{v}_j$ vectors, then we have

$$\mathbf{t} \propto \sum_{j=1}^{p} \mathbf{v}_j = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{Y}\mathbf{c} = \tilde{\mathbf{X}}\tilde{\mathbf{a}} \tag{2.10}$$

where $\tilde{\mathbf{a}} = (\tilde{a}_1, \ldots, \tilde{a}_p)' = \tilde{\mathbf{X}}'\mathbf{Y}\mathbf{c}$. If we pre-multiply (2.8) by $\tilde{\mathbf{X}}'\mathbf{Y}$ we have:

$$\tilde{\mathbf{X}}'\mathbf{Y}\mathbf{Y}'\tilde{\mathbf{X}}\mathbf{a} = \mathbf{a}\phi_1,\ \phi_1 > \phi_l,\ l > 1 \tag{2.11}$$

which is the PLS solution $\tilde{\mathbf{a}}_1$ when $\mathbf{F}_0 = \tilde{\mathbf{X}}$. The successive latent variables are obtained by applying the PLS algorithm on the unscaled orthogonal residuals $\mathbf{F}_k$. PLS on the unscaled explanatory variables is not equivalent to that on the scaled variables. It can be explained by a double weighting.

Consider multiplying (weighting) each vector $\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j'\mathbf{Yc}$ by the norm of the corresponding $\mathbf{x}$ variable, $d_j^{\frac{1}{2}} = ||\mathbf{x}_j||$, then the $\mathbf{v}_j$'s are given by:

$$\mathbf{v}_j = \hat{\mathbf{Y}}(\mathbf{x}_j)d_j^{\frac{1}{2}}\mathbf{c} = \mathbf{x}_j d_j^{-\frac{1}{2}}\mathbf{x}_j'\mathbf{Yc}. \tag{2.12}$$

The squared norm of $\mathbf{v}_j$ is

$$\mathbf{v}_j'\mathbf{v}_j = \mathbf{c}'\mathbf{Y}'\mathbf{x}_j\mathbf{x}_j'\mathbf{Yc}. \tag{2.13}$$

The vector of weights $\mathbf{c}$ is again determined as that maximising the sum of these squared norms, that is as the solution of

$$\max_{\mathbf{c}'\mathbf{c}=1}\sum_{j=1}^{p}\mathbf{v}_j'\mathbf{v}_j = \max_{\mathbf{c}'\mathbf{c}=1}\mathbf{c}'\mathbf{Y}'\mathbf{XX}'\mathbf{Yc}. \tag{2.14}$$

This is an eigen-problem with solution:

$$\mathbf{Y}'\mathbf{XX}'\mathbf{Yc} = \mathbf{c}\theta_1,\ \theta_1 > \theta_l,\ l > 1. \tag{2.15}$$

If we take the weighted average of the $\mathbf{v}_j$, taking the norms $d_j^{\frac{1}{2}}$ as weights again, we have:

$$\mathbf{t} \propto \sum_{j=1}^{p} d_j^{\frac{1}{2}}\mathbf{v}_j = \sum_{j=1}^{p}\mathbf{x}_j\mathbf{x}_j'\mathbf{Yc} = \mathbf{XX}'\mathbf{Yc} = \mathbf{Xa} \tag{2.16}$$

where $\mathbf{a} = \mathbf{X}'\mathbf{Yc}$. Substituting this expression of $\mathbf{a}$ into the solution (2.15) gives the PLS solution:

$$\mathbf{X}'\mathbf{YY}'\mathbf{Xa} = \mathbf{a}\theta_1,\ \theta_1 > \theta_l,\ l > 1. \tag{2.17}$$

Hence the first multivariate PLS lv on the unscaled $\mathbf{x}$ variables is obtained weighting the $\mathbf{x}$ axis twice. The first time the norms $d_j^{\frac{1}{2}}$ weight the sum of squares of the $\hat{\mathbf{y}}_i(\mathbf{x}_j)$ for determining the coefficients $\mathbf{c}$ and then they weight the vectors $\mathbf{v}_j$ to obtain the lv as a weighted average. This double weighting procedure agrees with the geometrical interpretation of the PLS algorithm given by Phatak et al. (1992), which consists of a double rotation of the LS solutions $\hat{\mathbf{Y}}(\mathbf{X})$. When the x variables have been autoscaled, the subsequent lv's are obtained as above but substituting the residuals $\mathbf{f}_j$ for the variables $\mathbf{x}_j$.

Also for the multivariate algorithm, the autoscaling of the $\mathbf{x}$ variables gives the plain average as the first lv, which is optimal with respect to minimising the distances of the projections $\hat{\mathbf{y}}(\mathbf{x}_j)$ from one point. The same considerations about the scaling of the residuals $\mathbf{F}_k$ and the optimality of the procedure made for the univariate algorithm can be extended to the multivariate algorithm.

It is interesting to note that this interpretation of PLS gives univariate PLS as a special case of the multivariate PLS algorithm. In fact, when there is only one response variable the vector of coefficients $\mathbf{c}$ reduces to a scalar, simply equal to 1. In this light it is possible to remove the duality between univariate and multivariate algorithms.

3 Principal Components of Simple Least Squares

In this section we present a method, that we name principal components of simple least squares (PCSLS), which determines the lv's from the simple regressions in an optimal way. We propose to use the least squares solutions of the simple regressions to weight the x variables.

For the univariate case we require that the lv's $\mathbf{t}_k = \tilde{\mathbf{X}}\tilde{\mathbf{a}}_k$ are a set of orthogonal linear combinations of the $\mathbf{X}$ variables with minimal sum of squared orthogonal distances from the matrix $\hat{\mathbf{Y}}_u$ defined in (2.4). That is, we consider optimising:

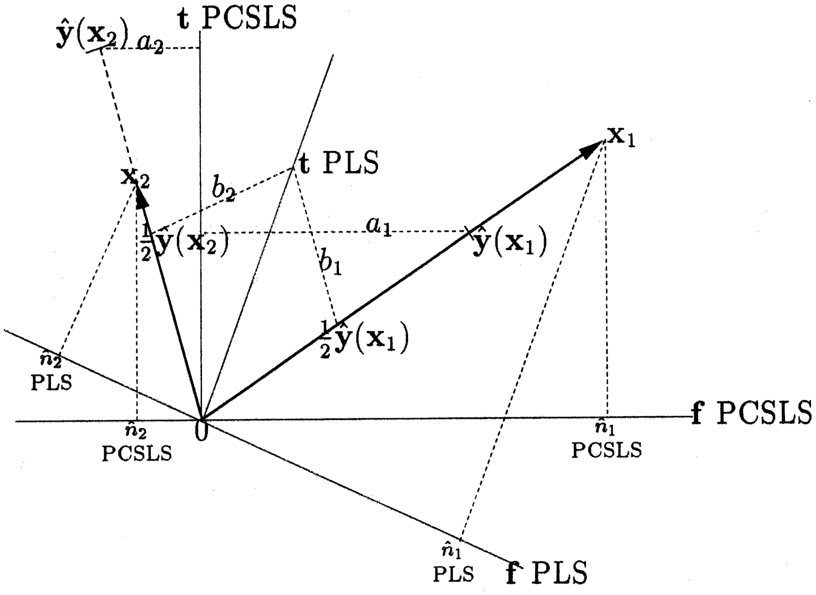
$$\min_{\mathbf{t}'_k\mathbf{t}_l=\delta_{kl}} ||\hat{\mathbf{Y}}_u - \mathbf{t}_k\mathbf{p}'_k||^2 \tag{3.1}$$

where $\mathbf{p}_k$ is a vector of p unknown coefficients. Note that this is more general than the objective function of PLS (2.5). Observing that we can write $\mathbf{t}_k = \hat{\mathbf{Y}}_u\mathbf{c}_k$, where $\mathbf{c}_k = \mathbf{B}_u{}^{-1}\mathbf{a}_k$, we have that (3.1) is a principal component problem, whose solutions are weighted principal components of $\mathbf{X}$, given by:

$$\mathbf{B}_u{}^2\mathbf{X}'\mathbf{X}\mathbf{a}_k = \mathbf{a}_k\phi_k,\ k = 1,\dots,d,\ \phi_k > \phi_l,\ l > k. \tag{3.2}$$

Hence the lv's of PCSLS are weighted principal components of $\mathbf{X}$ with weights proportional to $\sqrt{\hat{\mathbf{y}}(\mathbf{x}_j)'\hat{\mathbf{y}}(\mathbf{x}_j)} = \tilde{\mathbf{b}}_j$. Figure 3.1 shows the difference in the construction of the first lv in PLS and in PCSLS for two explanatory variables. The PLS lv is proportional to the vector sum of $b_1 = \frac{1}{2}\hat{\mathbf{y}}(\mathbf{x}_1)$ and $b_2 = \frac{1}{2}\hat{\mathbf{y}}(\mathbf{x}_2)$. In PCSLS the direction of the lv minimises the sum of the squared distances a_1 and a_2.

Figure 3.1 Construction of the first latent variable in PLS and in PCSLS. $\mathbf{x}_j = \mathbf{t}p_j + \mathbf{f}n_j$ where $\mathbf{t}$ and $\mathbf{f}$ are two unitary orthogonal variables. The symbols $\hat{n}_j$ denote the length of the residuals of the $\mathbf{x}$ variables.



Also for the multivariate regression problem we find the lv's that minimise the sum of the variances of the orthogonal residuals from the simple regressions of each response. That is we want to find the vectors $\mathbf{a}_k$ as solutions of

$$\min_{\mathbf{t}_k'\mathbf{t}_l=\delta_{il}} \sum_{i=1}^{q}\sum_{j=1}^{p}(\hat{\mathbf{y}}_i(\mathbf{x}_j) - \mathbf{t}_k p'_{k(i,j)})'(\hat{\mathbf{y}}_i(\mathbf{x}_j) - \mathbf{t}_k p'_{k(i,j)}). \tag{3.3}$$

Note that we substituted, without loss of generality, the orthogonality constraints with orthonormality ones. Let us denote the $(n \times pq)$ matrix $\hat{\mathbf{Y}}_m$, obtained setting next to each other the matrices $\hat{\mathbf{Y}}(j)$, $j = 1, \ldots, p$, defined in (2.6), as

$$\hat{\mathbf{Y}}_m = \mathbf{X} \begin{pmatrix} \mathbf{b}'(1) & \mathbf{0}'_q & \cdots & \mathbf{0}'_q \\ \mathbf{0}'_q & \mathbf{b}'(2) & \cdots & \mathbf{0}'_q \\ \vdots & \mathbf{0}'_q & \cdots & \vdots \\ \mathbf{0}'_q & \mathbf{0}'_q & \cdots & \mathbf{b}'(p) \end{pmatrix} = \mathbf{XB}.$$

Then we can express (3.3) as a principal components problem:

$$\min_{\mathbf{t}'_k \mathbf{t}_l = \delta_{kl}} ||\hat{\mathbf{Y}}_m - \mathbf{t}_k \mathbf{p}'_k||^2.$$

The solutions to this problem is given by:

$$\mathbf{BB}'\mathbf{X}'\mathbf{X}\mathbf{a}_k = \mathbf{a}_k \phi_k, \; \phi_k > \phi_l, \; l > k \tag{3.4}$$

But

$$\{\mathbf{BB}'\}_{ij} = \begin{cases} 0 & i \neq j \\ \mathbf{b}(j)'\mathbf{b}(j) = \sum_{l=1}^{q} ||\hat{\mathbf{y}}_l(j)||^2 & i = j, \end{cases}$$

therefore the matrix $\mathbf{W}^2 = \mathbf{BB}'$ is diagonal and the lv's solutions to (3.3) are weighted principal components of $\mathbf{X}$ and the coefficients are given by

$$\mathbf{W}^2\mathbf{X}'\mathbf{X}\mathbf{a}_k = \mathbf{a}_k \phi_k \; \phi_k > \phi_l, \; l > k.$$

The weights $w_{jj} = \sqrt{\sum_{l=1}^{q} ||\hat{\mathbf{y}}_l(j)||^2}$ for the multivariate PCSLS are the generalisation of those of the univariate case.

4 Simulation Study and Example

We compare the predictive accuracy of PLS with that of PCSLS and PLSSF on simulated and published data-sets.

4.1 Simulation study

Each data-set was generated according to the following model:

$$\begin{cases} x_{ij} = \sum\limits_{k=1}^{d} t_{ik} p_{kj} + f_{ij} n_j; \; i = 1, \dots, 60; \; j = 1, \dots, p \\ y_{ij} = \sum\limits_{k=1}^{d} t_{ik} q_{kj} + e_{ij} m_j; \; i = 1, \dots, 60; \; j = 1, \dots, q \end{cases} \tag{4.1}$$

where the variables t_{ik}, f_{ij} and e_{ij} are independent standard Normal variables. 50 observations are used to estimate the coefficients and the remaining 10 to compute the PRESS. Sets of 5000 repetitions were performed for different values of the parameters d, $\{n_j\}$, $\{m_i\}$, $\{p_{kj}\}$ and $\{q_{kj}\}$. The signal-to-noise-ratios (SNRs), given by the ratios $\frac{\sum_{k=1}^{d} p_{k,j}^2}{n_j^2}$ and $\frac{\sum_{k=1}^{d} q_{k,i}^2}{m_i^2}$ for the x_j's and y_i's respectively, are either constant or randomly generated. The predictive accuracy of each method is measured by the prediction error sum of squares $(PRESS(met)_j)$, defined as

$$PRESS(met)_j = \frac{1}{q} \sum_{k=1}^{q} \frac{1}{10} \sum_{i=1}^{10} (y_{ki} - \hat{y}_{ki[j](met)})^2 \tag{4.2}$$

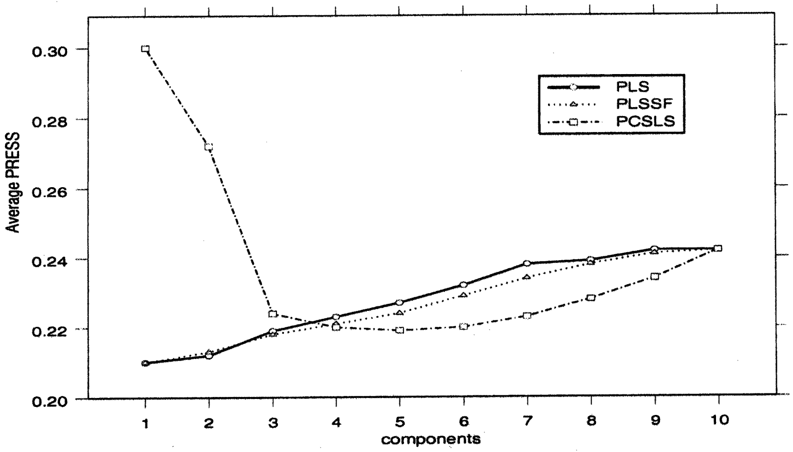
where $\hat{y}_{ki[j]}$ is the prediction of the i-th observation on the k-th response in the test sample using j lv's and met refers to which method is used. The methods are compared averaging the PRESS over the 5000 repetitions.

4.1.1 Univariate Prediction

A first set of 5000 repetitions were run with a set of 10 **x**-variables with full rank latent space (p=d=10). The first 4 x's had SNR equal to 1 and the last 6 had SNR of 10; the SNR for y was 3. The influence of the noises on the last 6

x variables is negligible, on the other hand the other 4 regressors contain equal amounts of error and explanatory term. Regardless of the number of lv's used, PCSLS gave global lower minimum $PRESS$ than PLS 2980 times out of 5000 (59.6%), with an average ratio $minPRESS(PCSLS)/minPRESS(PLS)$ of 0.9875. However, consistently PLS achieved its global minimum average PRESS with less lv's than PCSLS. The average number of components for which the global minimum $PRESS$ is achieved is about 3 for PLS and about 5 for PCSLS.

Figure 4.1 Average $PRESS$. 10 **x** variables with underlying dimension of 10.



Note how the average $PRESS$ is lower than that of OLS (10 lv's) for all methods.

In Figure 4.1 the average $PRESS$ for different number of components is compared. We note that average PRESS of PCSLS is lower than the other two methods when more than 3 lv's are used and that PLSSF behaves very similarly to PLS. Another set of 5000 repetitions were run with the same parameters as the previous ones but reducing the rank of the "true" explanatory variables to 5 ($d = 5$). In this case the behaviour of the methods resembles the previous case. PCSLS yielded a global minimum $PRESS$ lower than that of PLS 2880 times out of 5000 (57.6%), with an average ratio $minPRESS(PCSLS)/minPRESS(PLS)$ equal to 0.9756. PLS reaches its lowest value of $PRESS$ consistently with a lower number of components but then PCSLS yields better predictions for higher number of lv's used. Also in this case all methods yield better predictions than OLS with fewer lv's.

4.1.2 Multivariate Prediction

For multivariate predictions we generated the parameters $\{p_{kj}\}$ and $\{q_{kj}\}$ as independent uniform variables in the interval $[-1, 1]$ at each repetition. This avoids the problem of the choice of a fixed model, adding generality to the results. We considered 25 explanatory variables and 10 responses and $d = 1, 5, 10$. We run one set of simulations using fixed SNRs and another generating the SNRs randomly at each repetitions. The fixed SNRs were equal to 2 for the explanatory variables and to 4 for the responses. The random SNRs were generated as uniform variables in the intervals $[1, 3]$ for the explanatory variables and $[3, 5]$ for the responses. For each case 5000 repetitions were run. We only report on the random SNRs as the results were almost identical to the fixed SNR case. For real rank equal to 1 the plot of the average *PRESS* is shown in Figure 4.2.

Figure 4.2 Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 1 and random SNRs.

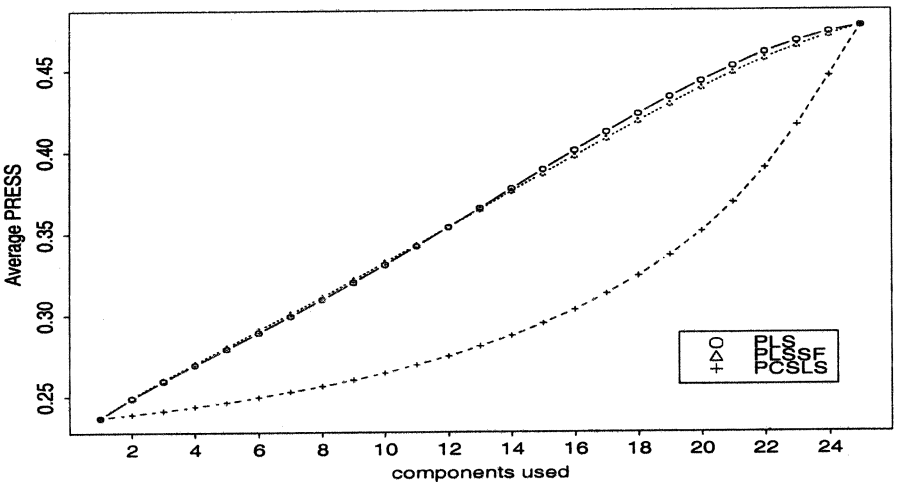
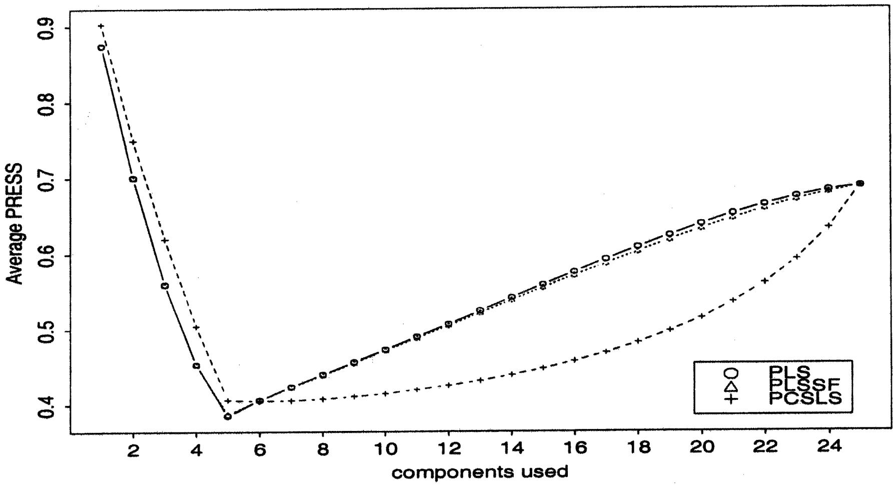
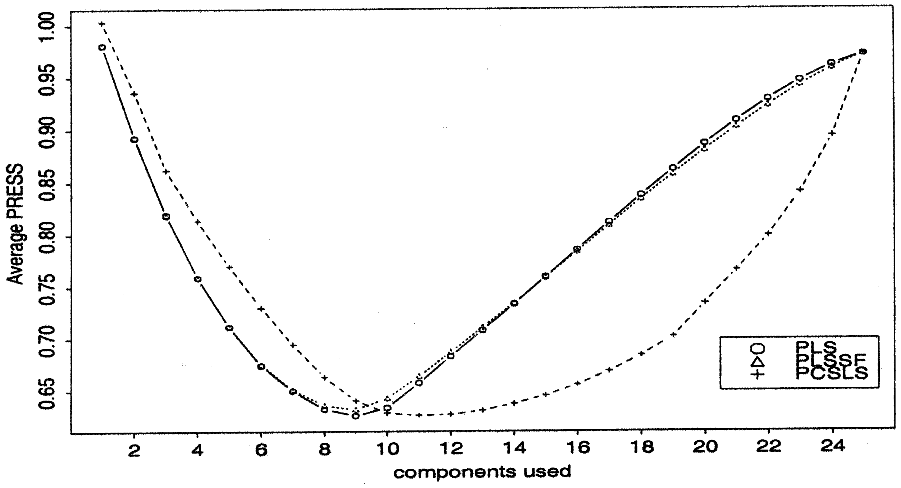


Figure 4.3 Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 5 and random SNRs.



In this case all methods give very close minima using 1 lv but PCSLS gives lower average *PRESS* for all number of lv's used.

Figure 4.4 Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 10 and random SNRs.

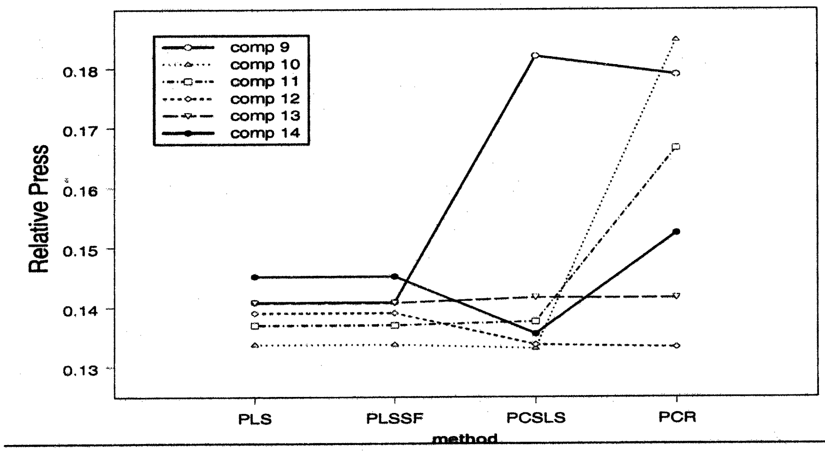


For real rank of 5 the average *PRESS* is shown in Figure 4.3. Also in this case all methods give very close minima when the number of lv's is equal to the real rank of the data. PCSLS gives lower *PRESS* than the other methods when more than 5 lv's are used. Figure 4.4 shows the average *PRESS* for real rank equal to 10. Also in this case PLS and PLSSF have similar behaviour. The average *PRESS* for the three methods is comparable but PCSLS gives lower average PRESS using 10 (real dimension) or more lv's and it achieves the overall minimum with 11 lv's. However, PLS and PLSSF give a lower average PRESS for lower number of lv's.

4.2 Example

In this section we compare some of the dimensionality reduction techniques we discussed on a set of data published in Skagerberg et al. (1992). The data consist of a simulation of a Low-Density Poly-Ethylene (LDPE) production process. The authors produced a set of 32 in-control observations and a set of 24 out-of-control observations. We only use the first 32 of these to test our methods. The explanatory variables consist of 2 input variables and 20 readings of temperatures inside the reactor. Clearly these last 20 variables are highly correlated. The measurements on 6 properties of the polymer were used as responses.

Figure 4.5 Chemical reactor example. Cross-validated PRESS for different number of components relative to the OLS.



The authors applied PLS to these data for implementing multivariate control charts. We compare the predictive efficiency by the leave-one-out cross-validated PRESS. For graphical reasons we will present the plots of the PRESS divided by the PRESS of the OLS solutions. Figure 4.5 shows the leave-one-out cross-validated PRESS for these data. The values of the PRESS for these data are consistent with the simulations. PLS and PLSSF give very close results; PCSLS gives a very close minimum and more accurate predictions for higher number of components used. We note that PCR gives the worst results.

5 Conclusions

Based on the simulated results and the example we can conclude that the PLS and PCSLS are comparable. The autoscaling of the deflated matrix in PLS does not seem to change the overall behaviour. PLS seems to achieve its best performance with a lower number of lv's than PCSLS. However, PCSLS consistently gave lower minima of $PRESS$ and showed a better behaviour for higher number of lv's. The higher methodological simplicity of PCSLS can ease the interpretation of the results. PCSLS is much less computer intensive than PLS, which can be an important feature when dealing with large data-sets. Furthermore, in PCSLS it is still possible to weight each x variable based on prior knowledge adding flexibility to this method.

Acknowledgements

B. Abraham was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

de Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemom. and Intell. Lab. Systems*, 18:251–263.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *JASA Th. and Met.*, 89(425):122–127.

Gelaldi, P. and Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17.

Helland, I. S. (1988). On the structure of partial least squares. *Comm. Stat.-sim*, 17(2):581–607.

Hoskuldsson, P. (1988). Pls regression methods. *J. of Chemometrics*, 2:211–228.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate bilinear model. *J. of Multivariate Analysis*, 5:248–264.

Kourti, T. and MacGregor, J. F. (1996). Multivariate spc methods for process and product monitoring. *J. Quality Eng.*, 28(4).

Merola, G. M. (1998). *Dimensionality reduction methods in multivariate prediction.* PhD thesis, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada.

Merola, G. M. and Abraham, B. (2001). Dimensionality reduction approach to multivariate prediction. To appear in:. *Canadian J. of Stat.*, 29(2). Downloadable at http://www.mat.ulaval.ca/cjs/pub/merola/merola.pdf.

Næs, T. and Risvik, E. (1996). *Multivariate analysis of data in sensory science.* Elsevier.

Phatak, A., Reilly, P. M., and Penlidis, A. (1992). The geometry of 2-block partial least squares regression. *Comm. in Statistics, Part A–Th. and Meth.*, 21:1517–1553.

Schmidli, H. (1995). *Reduced Rank Regression.* Contributions to Statistics. Physica-Verlag.

Skagerberg, B., MacGregor, J. F., and Kiparissides, C. (1992). Multivariate data analysis applied to low-density polythylene reactors. *Chemom. and Intel. Lab. Systems*, 14:341–356.

Weisberg, S. (1985). *Applied Linear Regression.* Wiley and Sons.

Wold, H. (1982). Soft modelling, the basic design and some extensions. In Joresorg, K. and Wold, H., editors, *Systems Under Indirect Observation*, volume II, pages 589–591. John Wiley and Sons.

Appendix

Algorithm A.1 Simplified univariate PLS algorithm.

```
0 ] set: $\mathbf{F}_1 = \tilde{\mathbf{X}}$, $\mathbf{r}_1 = \mathbf{1}_q$ and $k = 1$
1 ] iterate until $\mathbf{a}_k$ converges
        $\mathbf{a}_k = \frac{\mathbf{F}'_k \mathbf{r}_{(k)}}{||\mathbf{F}'_k \mathbf{r}_{(k)}||}$
        $\mathbf{t}_k = \mathbf{F}_k \mathbf{a}_k$
        $b_k = \frac{\mathbf{y}'\mathbf{t}_k}{\sqrt{\mathbf{y}'\mathbf{t}_k}}$
        $\mathbf{r}_k = \mathbf{y} b_k$
2 ] $\mathbf{F}_{(k+1)} = \mathbf{F}_k - \mathbf{t}_k(\mathbf{t}'_k \mathbf{t}_k)^{-1} \mathbf{t}'_k \mathbf{F}_k$
3 ] if $||\mathbf{F}_k|| > \epsilon$: $k \leftarrow k + 1$, goto 1
4 ] exit
```

PLS (Wold (1982)) was derived as one of the procedures of "path modelling" from a modification of NIPALS, an algorithm for computing simultaneously the principal components of two matrices (Gelaldi and Kowalski (1986)). A simplified univariate PLS algorithm is outlined in Algorithm A.1 and a simplified version of the multivariate algorithm is given in Algorithm A.2. At step (2) of the algorithms the matrix of explanatory variables is substituted with the matrix of orthogonal residuals $\mathbf{F}_k$, called "deflated $\mathbf{X}$ matrix". In this way each latent variable automatically satisfies the constraint of being orthogonal to the preceding ones. The process is iterated until the $\mathbf{X}$ matrix is exhausted by requiring that $||\mathbf{F}_k||$ is small enough.

Algorithm A.2 Simplified multivariate PLS algorithm.

```
0 ] set $\mathbf{F}_1 = \tilde{\mathbf{X}}$, $\mathbf{r}_1 = \mathbf{1}_n$, and $k = 1$
1 ] iterate until $\mathbf{a}_k$ converges
        $\mathbf{a}_k = \frac{\mathbf{F}'_k \mathbf{r}_{(k)}}{||\mathbf{F}'_k \mathbf{r}_{(k)}||}$
        $\mathbf{t}_k = \mathbf{F}_k \mathbf{a}_k$
        $\mathbf{b}_k = \frac{\mathbf{Y}' \mathbf{t}_k}{||\mathbf{Y}' \mathbf{t}_k||}$
        $\mathbf{r}_k = \mathbf{Y} \mathbf{b}_k$
2 ] $\mathbf{F}_{k+1} = \mathbf{F}_k - \mathbf{t}_k (\mathbf{t}'_k \mathbf{t}_k)^{-1} \mathbf{t}'_k \mathbf{F}_k$
3 ] if $||\mathbf{F}_k|| > \epsilon$:  $k \leftarrow k + 1$, goto 1
4 ] exit
```

The coefficients $\mathbf{a}_k$ can be also computed as $\mathbf{a}_k = \mathbf{F}'_k \mathbf{y} / ||\mathbf{F}'_k \mathbf{y}||$ for the univariate case and as the eigen-vector corresponding to the largest eigen-value of the matrix $\mathbf{F}'_k \mathbf{Y} \mathbf{Y}' \mathbf{F}_k$ for multivariate response. The advantage of the recursive algorithm is that it allows for missing observations.

The coefficients $\mathbf{a}_k$ determined by PLS for $k > 1$ cannot be used to compute the lv's from the x variables because they refer to the residuals $\mathbf{f}_{(k)}$. The actual coefficients for the x variables, $\mathbf{C}$ say, have to be computed from the a_k's as $\mathbf{C} = \mathbf{S}^{-1}\mathbf{A}$ where $\mathbf{QS}$ is the QR decomposition of the matrix $\mathbf{XA}$.

The algorithm terminates when k latent variables exhaust the $\mathbf{X}$ matrix. However, the optimal number of lv's used for prediction is not necessarily k but is instead chosen by other means, often by cross-validation.