RR-00-04

Another Interpretation of PLS and a New Dimensionality Reduction Method

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

April 2000

Another Interpretation of PLS and a New Dimensionality Reduction Method

G.M. Merola, Universitat Politècnica de Catalunya

and

B. Abraham University of Waterloo

Abstract

In this paper we give a novel interpretation of the well known dimensionality reduction method, partial least squares (PLS). Then we propose an alternative method, called PC-SLS, in which the predictive subspace is obtained from weighted principal components of the observed regressors. We compare this method with PLS through sets of simulated data.

Key Words: Partial Least Squares, Prediction, Dimensional Reduction, Principal Components of Simple Least Squares.

1 Introduction

In predicting responses using a multivariate linear regression model involving many (possibly correlated) explanatory variables and large number of observations sometimes a set of fewer linear combinations of the observed explanatory variables are used. These linear combinations are called latent variables (LVs) and the methods that generate the LVs take the generic name of dimensionality reduction methods (DRMs). The use of DRMs for prediction has been proven successful in fields like Chemometrics (e.g. Gelaldi and Kowalski (1986)), monitoring of chemical reactors (e.g. Kourti and MacGregor (1996)) and Quantitative Structure-Activity Relationships (e.g. Schmidli (1995)). These are contexts in which there are a large number of explanatory variables that could be highly correlated.

The LVs 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) have been proven to yield good predictions in some applications. However, optimal properties for the predictions obtained with these DRMs are not available. In particular, PLS was introduced as an algorithm (Wold (1982)) which, in the spirit of "soft modelling" , derives the LVs without optimizing an objective function connected to the predicted responses.

In this paper we give a new interpretation of PLS showing how the LVs are derived using the simple regressions of the response variables onto each observed explanatory variable. Our interpretation unifies the univariate and multivariate PLS algorithms. On the basis of this new interpretation, we derive a new method based on the optimization of an objective function, that can be used alternative to PLS. We call this method Principal Components of Simple Least Squares (PCSLS); it can be applied both for univariate and for multivariate regression. PCSLS amounts to deriving weighted principal components of the explanatory variables using the coefficients of determination of the simple regressions as weights.

In the next section, after introducing some notation, we briefly discuss Principal Component Analysis (PCA). Then we introduce PLS also giving it a new interpretation. We present PCSLS in the third section. In the fourth section we compare PLS and PCSLS through simulations showing how the predictions obtained with PCSLS are comparable with those of PLS. Finally in Section 5 we give some concluding remarks.

2 Dimensionally Reduced Prediction Models

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

DRMs consist of determining d orthogonal LVs $\mathbf{t}_k = \mathbf{X}\mathbf{a}_k$, $k = 1, \dots, d$, where the $\mathbf{a}_k$ are p-vectors containing the coefficients of the LVs. 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 LVs are then used as regressors for the responses and the predicted values are obtained by OLS as

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

$\hat{\mathbf{Y}}_{[d]}$ denotes the rank d predictions, $\mathbf{T}_{(d)} = \mathbf{X}\mathbf{A}_{(d)}$ 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 LVs. 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 rank d estimate of the regression coefficients, obtained with d LVs.

The approach of estimating the coefficients $\mathbf{a}_k$ by least squares, that is minimizing the Euclidean norm $||\mathbf{Y} - \mathbf{X}\mathbf{B}_{[d]}||$, often is not effective. In this approach, called Reduced Rank Regression (Izenman (1975)), the hypothesis made in ordinary regression that the observed regressors are the true explanatory variables is carried over to the LVs $\mathbf{t}_k$. In fact, for univariate response the only solution $\mathbf{a}_1$ is the vector of OLS regression coefficients and for multivariate responses the LVs are the principal components of the OLS solutions. These solutions simply minimize the additional error to OLS due to the rank constraints. Methods that are claimed to yield better predictions do not take the LS approach to estimating the coefficients. Next we illustrate these methods.

2.1 Principal Component Regression

PCR consists of regressing the responses onto the first d principal components. In PCA the LVs form a sequence of orthogonal axis of the space spanned by the $\mathbf{X}$ variables that sequentially minimize 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 optimization problem:

$$\min_{\mathbf{t}'_k\mathbf{t}_j=\delta_{kj}} ||\mathbf{X} - \mathbf{t}_k\mathbf{t}'_k\mathbf{X}|| \tag{2.2}$$

where δ_{kj} is equal to 1 if $k = j$ and to 0 otherwise. The solutions $\mathbf{a}_k$, $k = 1, \ldots, d$ are given by the eigenvectors of the matrix $\mathbf{X}'\mathbf{X}$ corresponding to the first d eigenvalues taken in non-increasing order. It is well known that PCA is very sensitive to the variance of the variables $\mathbf{X}$. Firstly PCA will minimize the variance of the residuals of the variables with larger variance. Hence 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 standardize them to unit length (autoscale) prior to PCA. That is PCA is performed on the scaled matrix $\tilde{\mathbf{X}}$. 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 solutions of weighted PCA are the eigenvectors 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 one of the procedures of "path modelling". It was derived from a modification of NIPALS, an algorithm for computing simultaneously the principal components of two matrices (Gelaldi and Kowalski (1986)). PLS was presented 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, Reilly and Penlidis (1992) 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; in the comment to a paper by Stone and Brooks (1990) Helland says: "I have always had difficulties in understanding the rationale for that method [PLS] ..". Nonetheless PLS has been extensively used in many fields such as, for example, chemometrics (e.g. Gelaldi and Kowalski (1986)) and in statistical process control (e.g. Kourti and MacGregor (1996)).

PLS can be applied to univariate and to multivariate regression but the multivariate version is not considered a straightforward generalization 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 2.1. At step (2) of the algorithm 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. The number d of components used for the prediction of $\mathbf{y}$ is generally different from the number of components that exhaust $\mathbf{X}$ and it is chosen independently, usually by cross-validation. It was shown by Hoskuldsson (1988) that the coefficients $\mathbf{a}_k$ can be computed directly from the data matrices as $\mathbf{a}_k = \mathbf{F}'_{(k-1)}\mathbf{y}/||\mathbf{F}'_{(k-1)}\mathbf{y}||$.

Algorithm 2.1 Simplified univariate PLS algorithm.

```
0 ] set: $\mathbf{F}_0 = \tilde{\mathbf{X}}$, $\mathbf{r}_0 = \mathbf{y}$ and $k = 1$
1 ] iterate until $\mathbf{t}_k$ converges
        $\mathbf{a}_k = \frac{\mathbf{F}'_{(k-1)}\mathbf{r}_{(k-1)}}{||\mathbf{F}'_{(k-1)}\mathbf{r}_{(k-1)}||}$
        $\mathbf{t}_k = \mathbf{F}_{(k-1)}\mathbf{a}_k$
        $b_k = \frac{\mathbf{y}'\mathbf{t}_k}{\sqrt{\mathbf{y}'\mathbf{t}_k}}$
        $\mathbf{r}_k = \mathbf{y}b_k$
2 ] $\mathbf{F}_k = \mathbf{F}_{(k-1)} - \mathbf{t}_k(\mathbf{t}'_k\mathbf{t}_k)^{-1}\mathbf{t}'_k\mathbf{F}_{(k-1)}$
3 ] if $||\mathbf{F}_k|| > \epsilon$:  $k \leftarrow k + 1$, goto 1
4 ]$d \leftarrow k$; $\hat{\mathbf{y}}_{[d]} = \sum_{i=1}^{d} \mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1}\mathbf{t}'_i\mathbf{y}$
```

Garthwaite (1994) shows that the PLS LVs are expressible as the weighted averages of the simple regressions of the response on each explanatory variable as

$$\mathbf{t}_1 \propto \sum_{j=1}^{p} \mathbf{x}_j\mathbf{x}'_j\mathbf{y} = \sum_{j=1}^{p} \hat{\mathbf{y}}(\mathbf{x}_j)(\mathbf{x}'_j\mathbf{x}_j) \tag{2.3}$$

where $\hat{\mathbf{y}}(x_j) = \mathbf{x}_j(\mathbf{x}_j'\mathbf{x}_j)^{-1}\mathbf{x}_j'\mathbf{y}$. However, noting that the $\mathbf{X}$ matrix is always autoscaled, the first PLS LV amounts to the simple average of $(\hat{\mathbf{y}}(\mathbf{x}_1), \dots, \hat{\mathbf{y}}(\mathbf{x}_p))$. If we let

$$\hat{\mathbf{Y}}_u = (\hat{\mathbf{y}}(x_1), \dots, \hat{\mathbf{y}}(x_p)) = \tilde{\mathbf{X}}\overset{\Delta}{\mathbf{B}}, \tag{2.4}$$

where $\overset{\Delta}{\mathbf{B}}$ is a diagonal matrix with diagonal elements equal to $\{\tilde{\mathbf{x}}_j'\mathbf{y}\}$, the first PLS LV can be expressed as

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

where 1_p is the p-vector of ones. A well known property of the simple average is that it minimizes the sum of squared distances of a set of values from a point. That is, the first PLS LV minimizes the quantity $\sum_{j=1}^{p}(\hat{\mathbf{y}}(x_j) - \mathbf{t}_1)'(\hat{\mathbf{y}}(x_j) - \mathbf{t}_1)$.

In the subsequent iterations the deflated matrix $\mathbf{F}_k$ is not autoscaled, hence the successive LVs are weighted averages of the simple regressions with weights proportional to the variances of the residuals (the proportionality constant is irrelevant for the prediction of $\mathbf{y}$). That is:

$$\mathbf{t}_k = \mathbf{F}_{(k-1)}\mathbf{F}'_{(k-1)}\mathbf{y} = \sum_{j=1}^{p} \hat{\mathbf{y}}(\mathbf{f}_j)(\mathbf{f}_j'\mathbf{f}_j), \quad k > 1$$

where $\mathbf{f}_j$ is the j-th column of $\mathbf{F}_{(k-1)}$. It is difficult to find a justification for which these LVs should be good predictors of the response. The use of weighted averages for the LVs 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 PCA. Autoscaling the residuals $\mathbf{F}_k$ at each iteration would render these LVs homogeneous with the first one. We will refer to this modified PLS procedure as PLSSF.

Multivariate partial least squares

Algorithm 2.2 outlines a simplified multivariate PLS algorithm. Hoskuldsson (1988) shows that the solutions $\mathbf{a}_k$ can be computed directly from the matrix $\mathbf{F}'_{(k-1)}\mathbf{Y}\mathbf{Y}'\mathbf{F}_{(k-1)}$ 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:

$$\mathbf{F}'_{(k-1)}\mathbf{Y}\mathbf{Y}'\mathbf{F}_{(k-1)}\mathbf{a}_k = \mathbf{a}_k\phi_k.$$

Also the multivariate PLS algorithm can be interpreted using the simple regression of each response on each $\mathbf{x}$ variable.

Algorithm 2.2 Simplified multivariate PLS algorithm.

```
0 ] set $\mathbf{F}_0 = \tilde{\mathbf{X}}$, $\mathbf{r}_0 = 1_n$, and $k = 1$
1 ] iterate until $\mathbf{t}_k$ converges
      $\mathbf{a}_k = \frac{\mathbf{F}'_{(k-1)}\mathbf{r}_{(k-1)}}{||\mathbf{F}'_{(k-1)}\mathbf{r}_{(k-1)}||}$
      $\mathbf{t}_k = \mathbf{F}_{(k-1)}\mathbf{a}_k$
      $\mathbf{b}_k = \frac{\mathbf{Y}'\mathbf{t}_k}{||\mathbf{Y}'\mathbf{t}_k||}$
      $\mathbf{r}_k = \mathbf{Y}\mathbf{b}_k$
2 ] $\mathbf{F}_k = \mathbf{F}_{(k-1)} - \mathbf{t}_k(\mathbf{t}'_k\mathbf{t}_k)^{-1}\mathbf{t}'_k\mathbf{F}_{(k-1)}$
3 ] if $||\mathbf{F}_k|| > \epsilon$: $k \leftarrow k+1$, goto 1
4 ] $d \leftarrow k$; $\hat{\mathbf{Y}}_{[d]} = \sum_{i=1}^{d} \mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1}\mathbf{t}'_i\mathbf{Y}$
```

Let

$$\hat{\mathbf{Y}}(j) = (\hat{\mathbf{y}}_1(\mathbf{x}_j), \dots, \hat{\mathbf{y}}_q(\mathbf{x}_j)) = \tilde{\mathbf{x}}_j\tilde{\mathbf{x}}'_j\mathbf{Y} = \tilde{\mathbf{x}}_j\mathbf{b}'(j) \tag{2.5}$$

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 ”weights” $\mathbf{b}'(j) = \tilde{\mathbf{x}}'_j\mathbf{Y} = \{\tilde{\mathbf{x}}'_j\mathbf{y}_i\}$. PLS determines a unit-norm vector of q coefficients, $\mathbf{c} = (c_1, \dots, c_q)'$ 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

$$\max_{\mathbf{c}'\mathbf{c}=1} \sum_{j=1}^{p} \mathbf{v}'_j\mathbf{v}_j = \max_{\mathbf{c}'\mathbf{c}=1} \sum_{j=1}^{p} \mathbf{c}'\mathbf{Y}'\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}'_j\mathbf{Y}\mathbf{c}. \tag{2.6}$$

Since $||\hat{\mathbf{y}}_i(\mathbf{x}_j)||$ is a measure of predictability, the c_i's are coefficients that maximize the overall prediction of each variable. The solution of (2.6) for $\mathbf{c}$ is found by writing:

$$\max_{\mathbf{c}'\mathbf{c}=1} \sum_{j=1}^{p} \mathbf{v}_j'\mathbf{v}_j = \max_{\mathbf{c}'\mathbf{c}=1} \mathbf{c}'\mathbf{Y}'\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{Y}\mathbf{c} \tag{2.7}$$

which is an eigen-problem with solution:

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

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

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

where $\tilde{\mathbf{a}} = (\tilde{a}_1, \ldots, \tilde{a}_p)' = \tilde{\mathbf{X}}'\mathbf{Y}\mathbf{c}$. If we premultiply (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 $\mathbf{a}_1$ 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 $\mathbf{v}_j = \tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j'\mathbf{Y}\mathbf{c}$ by the norm of the corresponding $\mathbf{x}$ variable, $d_j^{\frac{1}{2}} = ||\mathbf{x}_j||$, then these are given by:

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

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

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

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

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

This is an eigen-problem with solution:

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

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

$$\mathbf{t} = \sum_{j=1}^{p} d_j^{\frac{1}{2}} \mathbf{v}_j = \sum_{j=1}^{p} \mathbf{x}_j\mathbf{x}_j'\mathbf{Y}\mathbf{c} = \mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{c} = \mathbf{X}\mathbf{a} \tag{2.16}$$

where $\mathbf{a} = \mathbf{X}'\mathbf{Y}\mathbf{c}$. Substituting this expression of $\mathbf{a}$ into the solution (2.15) gives the PLS2 solution:

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

Hence the first multivariate PLS LV on the unscaled $\mathbf{x}$ variables is obtained weighting twice the $\mathbf{x}$ axis. 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 $\hat{\mathbf{Y}}(\mathbf{X})$.

Also for the multivariate algorithm, the autoscaling of the $\mathbf{x}$ variables gives the plain average as first LV, which is optimal with respect to minimizing 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 LVs from the simple regressions in an optimal way. For the univariate case consider the matrix $\hat{\mathbf{Y}}_u = \tilde{\mathbf{X}}\overset{\Delta}{\mathbf{B}}$ defined in (2.4). We require that the LVs are a set of orthogonal linear combinations of the $\mathbf{X}$ variables with minimal sum of squared orthogonal distances from the matrix $\hat{\mathbf{Y}}_u$. That is, we consider the problem

$$\min_{\mathbf{t}_k=\tilde{\mathbf{X}}\tilde{\mathbf{a}}_k} ||\hat{\mathbf{Y}}_u - \mathbf{t}_k(\mathbf{t}_k'\mathbf{t}_k)^{-1}\mathbf{t}_k'\hat{\mathbf{Y}}_u||^2, \;\; \mathbf{t}_k'\mathbf{t}_l = 0, \; l \neq k. \tag{3.1}$$

Observing that we can write $\mathbf{t}_k = \hat{\mathbf{Y}}_u\mathbf{c}_k$, where $\mathbf{c}_k = \overset{\Delta}{\mathbf{B}}^{-1}\tilde{\mathbf{a}}_k$, we have that (3.1) is a weighted principal component problem, whose solutions are the eigenvectors

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

Hence we define the LVs of PCSLS as the weighted principal components of $\mathbf{X}$ with weights proportional to the regression coefficients $(\frac{\mathbf{x}_j'\mathbf{y}}{\mathbf{x}_j'\mathbf{x}_j})$. Figure 3.1 shows the geometry of the construction of the first LV in PLS and in PCSLS for two explanatory variables.

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

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

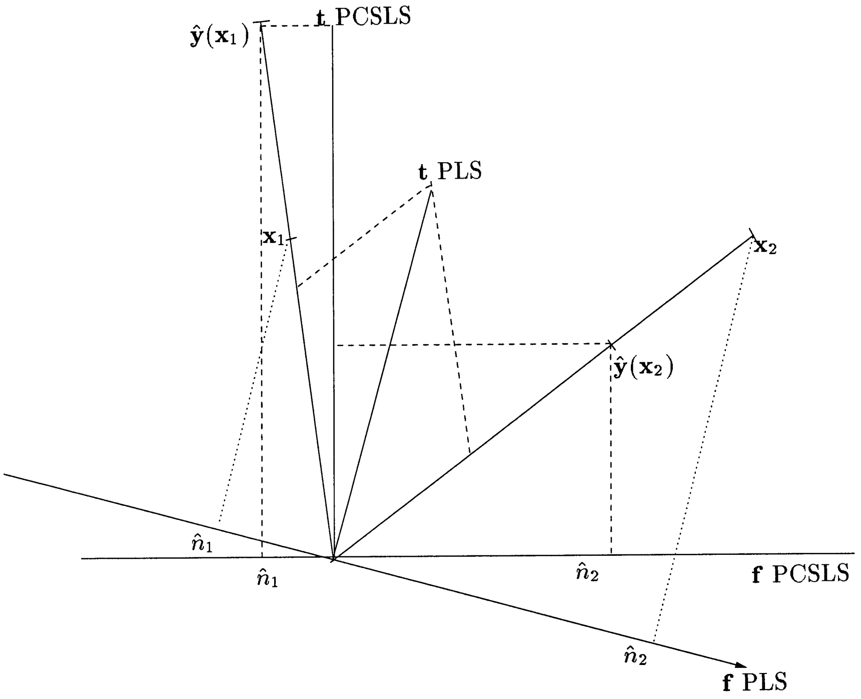

Figure 3.1: Construction of the first latent variable in PLS and in PCSLS. $\mathbf{x}_j = \mathbf{t}p_j + \mathbf{f}n_j$ where $\mathbf{t}$ and $\mathbf{f}$ are two unitary orthogonal variables. The symbols $\hat{n}_j$ denote the length of the residuals of the $\mathbf{x}$ variables.

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 $\hat{\mathbf{Y}}(j)$, $j = 1, \dots, p$, defined in (2.5), as

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

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

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

The solutions of this problem are the eigenvectors:

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

Note that

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

Hence the matrix $\mathbf{W} = \mathbf{BB}'$ is diagonal and the LVs are the weighted principal components of $\mathbf{X}$ and the coefficients are given by

$$\mathbf{W}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\tilde{\mathbf{a}} = \tilde{\mathbf{a}}\phi.$$

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

4 Simulation Study

We compare the predictive efficiency of PLS with that of PCSLS and PLSSF (with the autoscaling of the residuals $\mathbf{F}$) on simulated data-sets. Each data-set was generated according to the following model:

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

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 and test the predictions. Sets of 5000 repetitions were performed for different values of the parameters d, $\mathbf{N}$, $\mathbf{M}$, $\mathbf{P}$ and $\mathbf{Q}$. The signal-to-noise-ratios (SNR), are either constant or randomly generated, and are given by the ratio $\frac{\sum_{k=1}^{d} p_{k,j}^2}{n_j^2}$ and $\frac{\sum_{k=1}^{d} q_{k,j}^2}{m_j^2}$. We compare the methods by their predictive efficiency measured by the average prediction error sum of squares $(PRESS(met)_j)$, defined as

$$PRESS(met)_j = \frac{1}{5000}\sum_{k=1}^{5000}\frac{1}{10}\sum_{i=1}^{10}(y_{ki} - \hat{y}_{ki[j](met)})^2 \tag{4.2}$$

where $\hat{y}_{ki[j]}$ is the prediction of the i-th observation in the test sample using rank j predictions in the k-th run. *met* refers to which method is used.

4.1 Univariate Prediction

For the univariate case we also considered different summaries related to the predictive efficiency of PLS and PCSLS; these are:

$$\begin{aligned}
\text{ratio}_j &= \frac{1}{5000}\sum_{k=1}^{5000}\frac{\frac{1}{10}\sum_{i=1}^{10}(y_{ki}-\hat{y}_{ki[j]}(PCSLS))^2}{\frac{1}{10}\sum_{i=1}^{10}(y_{ki}-\hat{y}_{ki[j]}(PLS))^2}\\
\text{max.abs.err.}_{.j}(met) &= \frac{1}{5000}\sum_{k=1}^{5000}\max_{i=1,\dots,10}\left\{|y_{ki}-\hat{y}_{ki[j](met)}|\right\}\\
\text{pos.min.}_{.j}(met) &= \underset{k=,\dots,5000}{\text{num. of times}}\left\{\arg\min_{l=1,\dots,10}\frac{1}{10}\sum_{i=1}^{10}(y_{ki}-\hat{y}_{ki[l]}(met))^2=j\right\}
\end{aligned}$$

$$\mathbf{P}=\begin{pmatrix}
1.0 & 1.0 & 1.0 & 0.4 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0\\
0.0 & 0.4 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0\\
0.0 & 0.0 & 0.0 & 0.0 & 4.0 & 0.0 & 0.0 & 0.0 & 0.0 & 2.0\\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 4.0 & 0.0 & 0.0 & 2.0 & 2.0\\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 4.0 & 4.0 & 2.0 & 0.0\\
1.0 & 1.0 & 1.0 & 0.4 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0\\
0.0 & 0.4 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0\\
0.0 & 0.0 & 0.0 & 0.0 & 4.0 & 0.0 & 0.0 & 0.0 & 0.0 & 2.0\\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 4.0 & 0.0 & 0.0 & 2.0 & 2.0\\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 4.0 & 4.0 & 2.0 & 0.0
\end{pmatrix}$$

$\mathbf{x}$-variables with full rank (p=d=10). The other parameters were the following:

$$\mathbf{q}' = (\ 3.0\quad 3.0\quad 1.0\quad 1.0\quad 1.0\quad 3.0\quad 3.0\quad 1.0\quad 1.0\quad 1.0\ ) \tag{4.3}$$

$$SNR_X = (\ 1.0\quad 1.0\quad 1.0\quad 1.0\quad 10.0\quad 10.0\quad 10.0\quad 10.0\quad 10.0\quad 10.0\ ) \tag{4.4}$$

and $SNR_y = 3$. The influence of the noises on the 6 $\mathbf{x}$ variables with SNRs equal to 10 is negligible, on the other hand the other 4 regressors contain equal amounts of error and explanatory term. 2980 times (59.6%) PCSLS gave lower $PRESS$ than PLS (regardless of the rank with which it was achieved), with an average ratio $minPRESS(PCSLS)/minPRESS(PLS)$

of 0.9875. However, comparing the average *PRESS* for fixed rank prediction shown in Figure 4.1, that of PCSLS is higher than the other two methods. The average *PRESS* of PLSSF is very close to that of PLS but it is lower when 3 or more components are used. Other summaries of the simulations comparing PLS and PCSLS are shown in Table 1. PLS reaches its minimum *PRESS* with one component 44% of the times, PCSLS yields a better average *PRESS* for more than 3 components and overall. The average number of components for which the minimum *PRESS* is achieved is about 3 for PLS and about 5 for PCSLS. Note how the values of the *PRESS* are better than those of OLS (rank 10) for both methods.

Table 1 Comparison between PLS and PCSLS with one response and 10 full rank explanatory variables. The meaning of the columns is explained above.

Summaries for univariate response and 10 full rank x's					
rank	ratio	max abs. err.		pos.	min.
j	PCS/PLS	PLS	PCSLS	PCSLS	PLS
1	1.202	0.99	1.07	633	2200
2	1.129	1.01	1.03	667	933
3	0.990	1.05	1.00	717	333
4	0.952	1.06	1.00	483	300
5	0.938	1.07	1.00	417	283
6	0.916	1.09	1.00	533	333
7	0.949	1.09	1.04	350	267
8	0.964	1.09	1.05	300	117
9	0.983	1.10	1.08	333	84
10	1.000	1.10	1.10	567	150

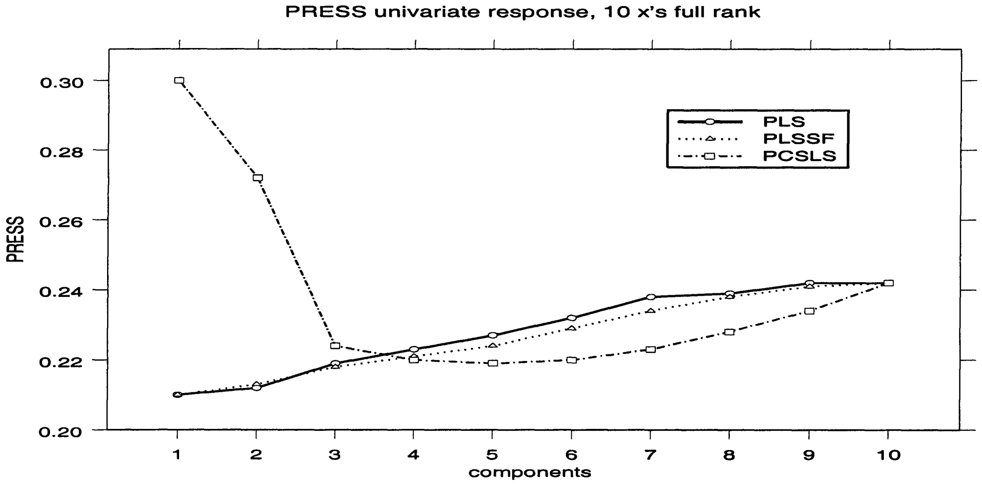

Figure 4.1: Average total *PRESS*. 10 **x** variables with underlying dimension of 10.

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 ($d = 5$). Hence the matrix $\mathbf{P}$ is constituted by the first 5 rows of the one above and the vector $\mathbf{q}$ consists of the first 5 elements of the one above. In these runs 2880 times (57.6%) PCSLS yielded a minimum $PRESS$ lower than that of PLS, with an average ratio $minPRESS(PCSLS)/minPRESS(PLS)$ equal to 0.9756. Other summary statistics for these runs are shown in Table 2. Also in this case the two methods have similar behaviours, PLS reaches its lowest values of $PRESS$ (higher than

Table 2 Comparison between PLS and PCSLS with one response and 10 explanatory variables with underlying rank of 5. The meaning of the columns is explained above.

Summaries for univariate response and 10 x's of rank 5					
rank	ratio	max abs. err.		pos.	min.
j	PCS/PLS	PLS	PCSLS	PCSLS	PLS
1	1.147	1.12	1.19	776	2303
2	1.098	1.12	1.16	842	1120
3	0.967	1.14	1.11	609	280
4	0.939	1.15	1.10	620	270
5	0.925	1.15	1.09	598	282
6	0.904	1.16	1.09	842	373
7	0.953	1.16	1.13	399	290
8	0.967	1.16	1.14	255	62
9	0.977	1.16	1.14	25	9
10	1.000	1.16	1.16	34	11

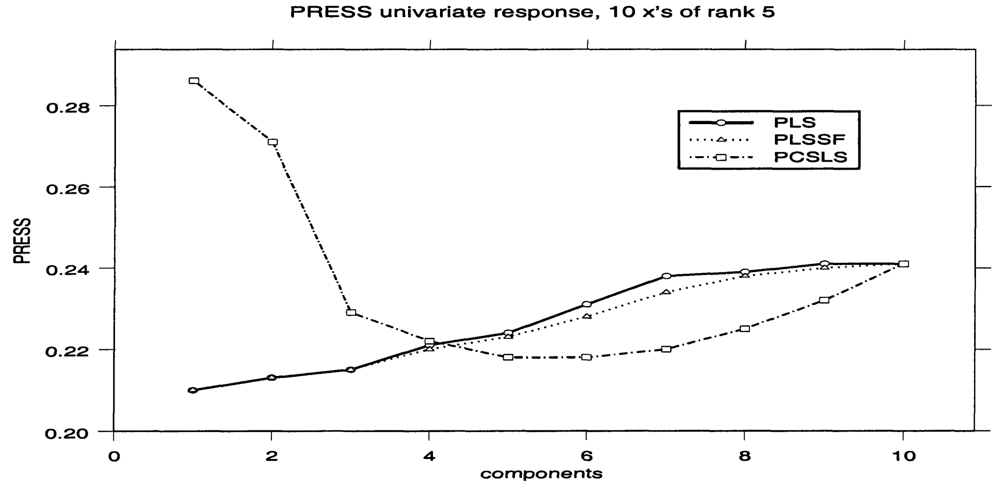

Figure 4.2: Average total $PRESS$. 10 $\mathbf{x}$ variables with underlying dimension of 5.

those of PCSLS on average) consistently with a lower number of components. Even if in this case the real rank of the predictive variables is 5, only 26.4% of the times (less than before) PLS has its minimum *PRESS* with one component. Also in this case both methods yield better predictions than OLS with less than full rank estimates. Figure 4.2 shows the average *PRESS* for all methods, including PLSSF, which behaves closely to PLS. The minimum average *PRESS* also in this case is that of PLS and PLSSF with one component.

4.2 Multivariate Prediction

For multivariate predictions we generated the elements of the matrices $\mathbf{P}$ and $\mathbf{Q}$ as independent uniform variables in the interval $[-1, 1]$ at each repetition. This avoids the problem of the choice of the model, adding generality to the results. We considered 25 explanatory variables and 10 responses running simulations for ranks 1, 5 and 10. We repeated the simulations using fixed SNRs and also generating these randomly at each repetition. 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.

Constant SNRs

For real rank equal to 1 the plot of the average *PRESS* is shown in Figure 4.3. All methods give very close minimae using 1 LV but PCSLS gives lower *PRESS* for all ranks of the predictions. For real rank of 5 the average *PRESS* is shown in Figure 4.4. Also in this case all methods give very close minimae for a number of LVs equal to the real rank of the data. PCSLS gives lower *PRESS* than the other methods using a number of LVs higher than 5. Figure 4.5 shows the *PRESS* for real rank equal to 10. In this case the *PRESS* for PLS and PLSSF are very similar and comparable with those of PCSLS but for the first two methods *PRESS* reaches its lower values with a smaller number of LVs.

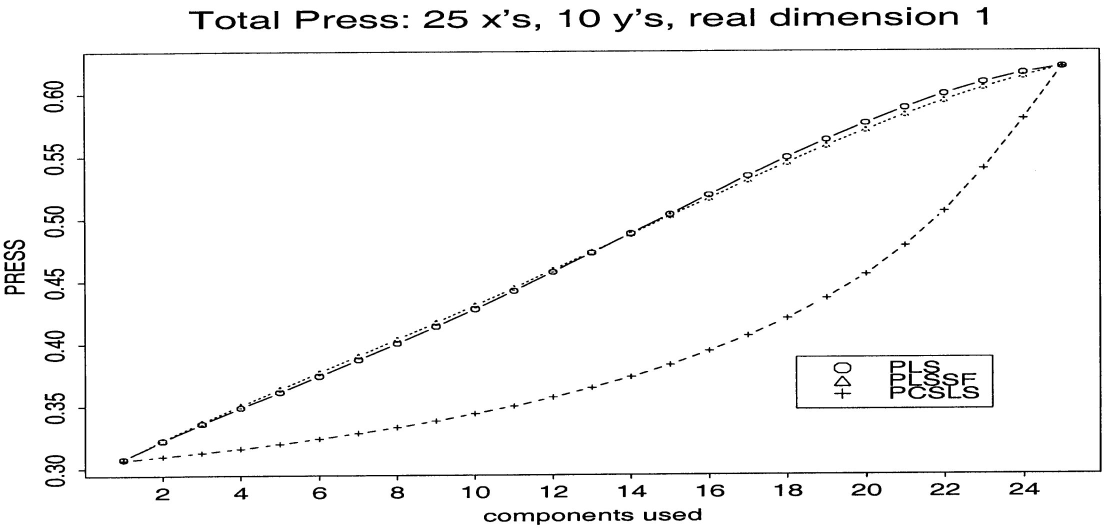

Figure 4.3: Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 1 and constant SNRs.

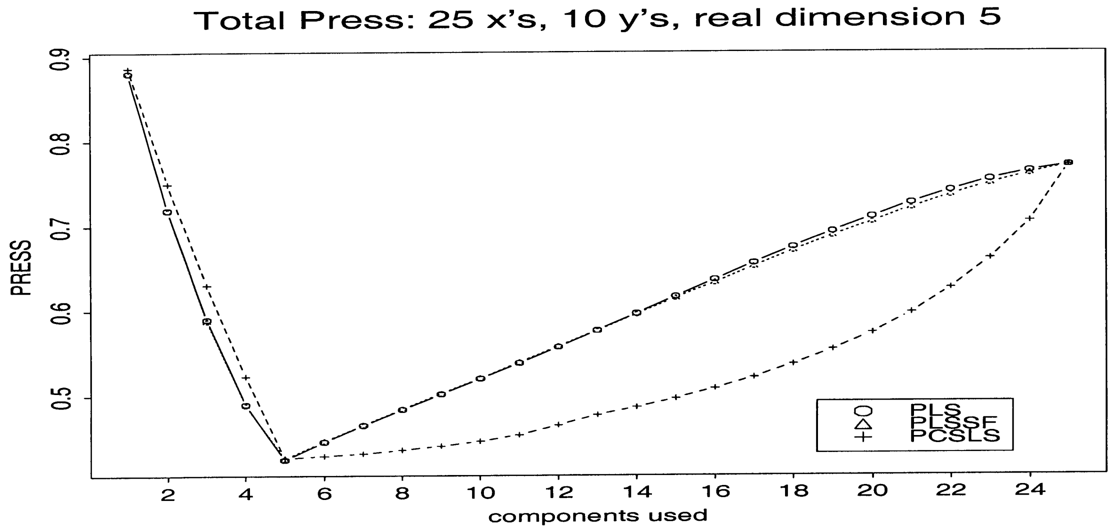

Figure 4.4: Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 5 and constant SNRs.

Note that also in this case PCSLS yields lower *PRESS* than the other methods for a number of LVs higher than that which gives the minimum (equal to 11).

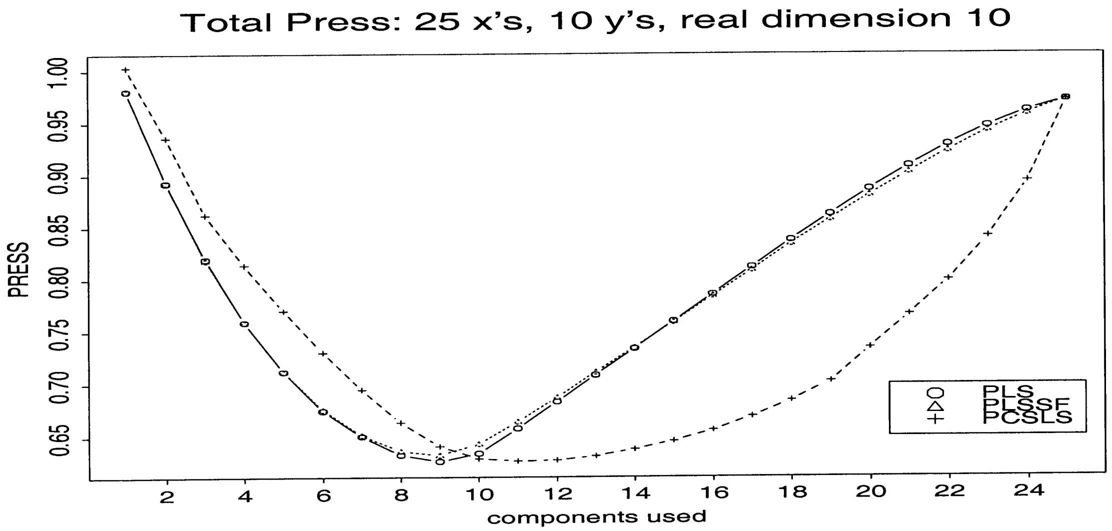

Figure 4.5: Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 10 and constant SNRs.

Random SNRs

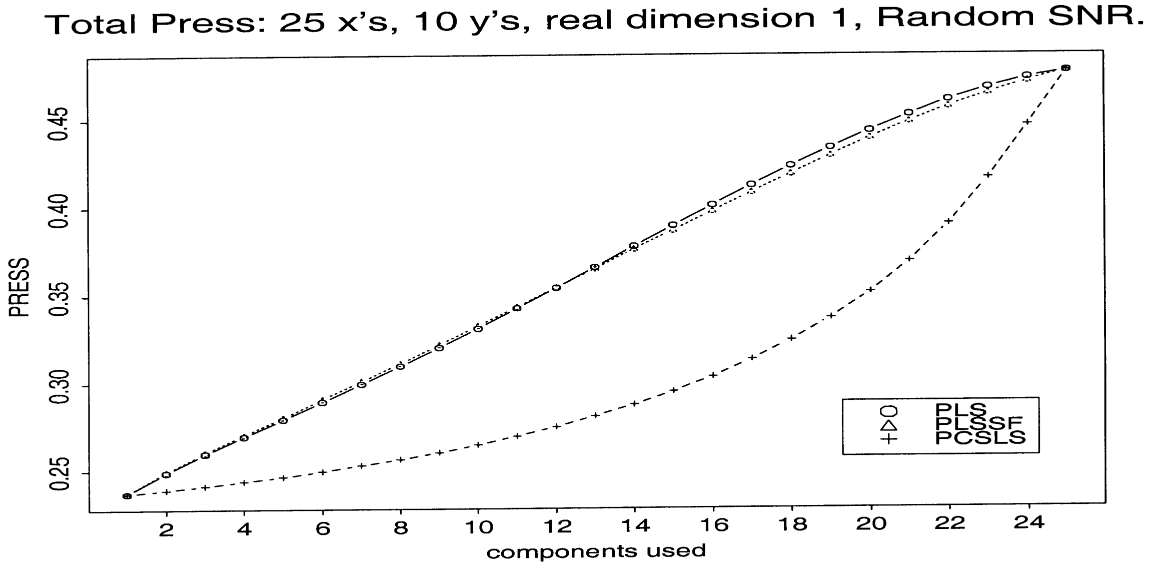

Figure 4.6: Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 1 and random SNRs.

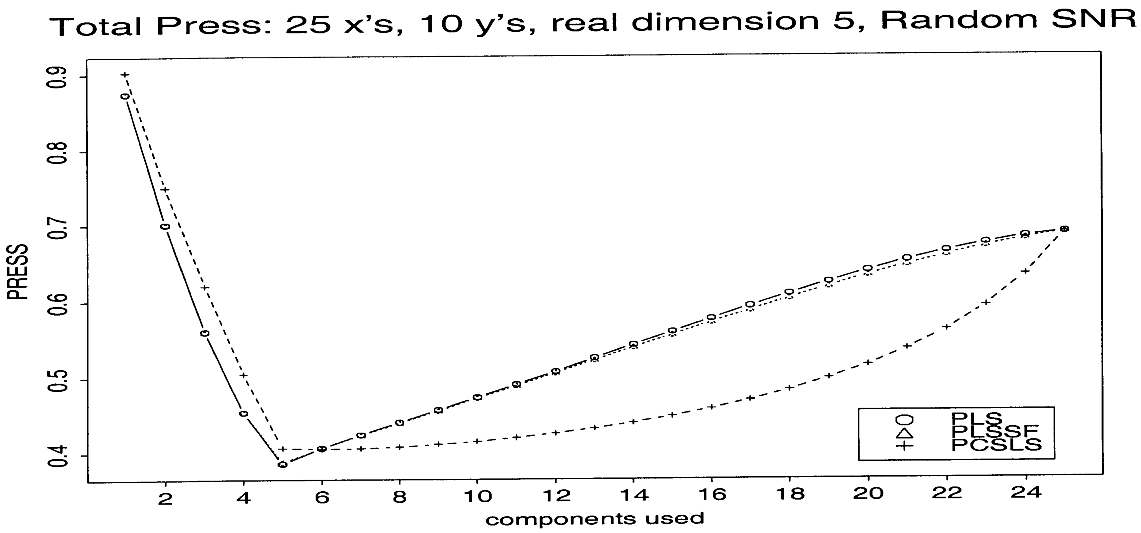

Figure 4.7: Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 5, random SNRs.

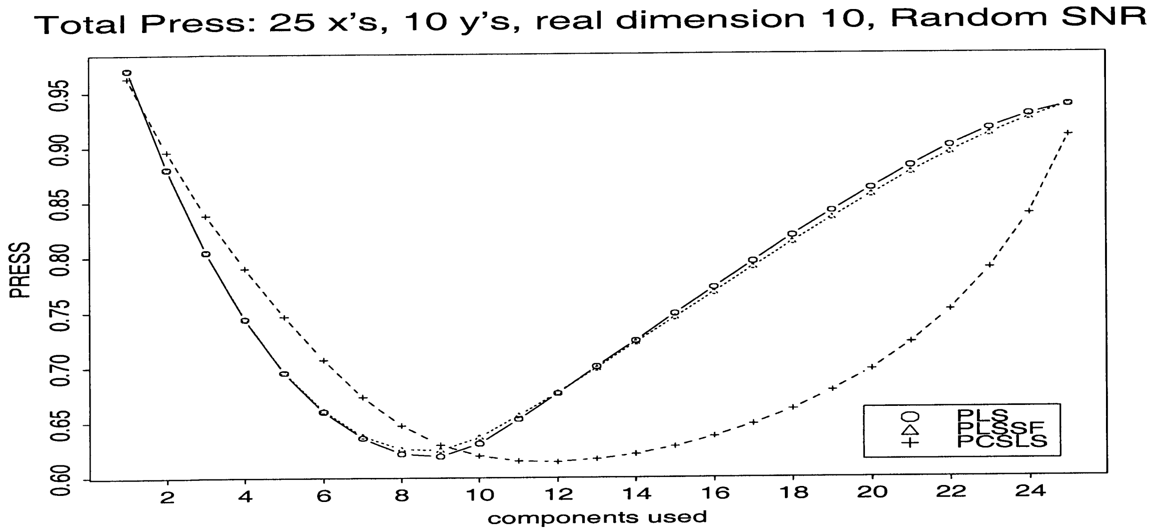

Figure 4.8: Average *PRESS* for different methods with 25 explanatory variables, 10 responses. Real rank equal to 10, random SNRs.

The pattern of the *PRESS* with random SNRs is similar to that obtained with fixed SNRs. This was expected since the SNRs were generated in small intervals, centered around the fixed values used before. Differently from before PCSLS gives higher minimum *PRESS* for real rank of 5 and lower for real rank of 10 than the other methods. The plots are shown in Figures 4.6, 4.7 and 4.8.

5 Conclusions

Based on the simulated results we can conclude that the results of 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 LVs than PCSLS. However, PCSLS consistently gave lower minimae of $PRESS$ and showed a better behaviour for higher number of LVs. The higher methodological simplicity of PCSLS can ease the interpretation of the results. The computation of the PCSLS solutions is much less computer intensive than that of PLS which can be an important feature when dealing with large data-sets. Furthermore, in PCSLS each response can be arbitrarily weighted (in the sense of re-weighting the matrix that generates the solutions) adding flexibility to this method. Of course, the distribution of the estimates of these methods depends on the distribution of the eigenvectors of random matrices and remains an open problem.

Acknowledgments

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).

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.

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *J. Royal Stat. Soc. B*, 52(2):237–269.

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.