Selecting both latent and explanatory variables in the PLS1 regression model

Aziz Lazraq\textsuperscript{a}, Robert Cléroux\textsuperscript{b,\,*}, Jean-Pierre Gauchi\textsuperscript{c}

\textsuperscript{a}Ecole Nationale de l’Industrie Minérale, Rabat, Morocco
\textsuperscript{b}Department of Mathematics and Statistics, University of Montreal, Montreal, Canada
\textsuperscript{c}National Institute for Agricultural Research, Biometrics Unit, Jouy-en-Josas, France

Accepted 7 January 2003

Abstract

In this paper, two inferential procedures for selecting the significant predictors in the PLS1 regression model are introduced. The significant PLS components are first obtained and the two predictor selection methods, called PLS–Forward and PLS–Bootstrap, are applied to the PLS model obtained. They are also compared empirically to two other methods that exist in the literature with respect to the quality of fit of the model and to their predictive ability. Although none of the four methods is uniformly best, it is seen that PLS–Forward and PLS–Bootstrap perform well and can be very useful in practical situations in identifying the important explanatory variables.

Keywords: PLS regression; Explanatory variable selection; PLS component selection; Inferential procedures

1. Introduction

In the PLS regression model of $Y$ on $X$, it is often the case that the number of explanatory variables or predictors $X_{(j)}$, $j=1,2,\ldots,q$ is big, even bigger than the sample size. The problem of variable selection is thus of interest. The objective in some practical situations is to have a good model with as few $X_{(j)}$s as possible. However, a model that fits the data well may not have a good predictive ability. This happens, for instance, when there is an overfitting effect, or when the nature of the random process has changed between model building and prediction. On the other hand, a model that does not fit the data well will not predict well either. It is thus necessary to eliminate the useless predictors and to identify the most important ones to be included in the model.

Many authors have worked on the problem of explanatory variable selection in PLS regression (see, for instance, the following recent papers and the references therein: Forina et al. [1], Höskuldsson [2], Gauchi and Chagnon [3], Lindgren et al. [7], and Sarabia et al. [10]; more general references on variable selection in regression are in Miller [8] and Thompson [13]). The method introduced by Forina et al. [1] is based on the cyclic repetition of the PLS
algorithm, each time multiplying the predictors by their importance computed in the previous cycle. Höskuldsson [2] makes use of intervals of variables instead of the variables themselves. Lindgren et al. [7] reweight, for each dimension, the elements in the PLS weight vector, and Saribia et al. [10] propose a variable selection procedure also by introducing weights to the loadings in the PLS model. Finally, Gauchi and Chagnon [3] compared empirically 20 methods of variable selection on five data sets in the case of univariate PLS regression (also called the PLS1 regression model). They concluded that the Backward $Q^2_{cum}$ (BQ) method (see Section 7) is the one to be recommended in practical situations.

The selection procedures are of two types: (i) dimension-wise selection (or elimination), where one works with one dimension at a time or where one selects (or eliminates) variables on one PLS component at a time; and (ii) model-wise selection, where the PLS model is first obtained and then a variable selection procedure is applied.

As we have mentioned before, in the PLS model, the number $q$ of predictors is usually big and is often bigger than the sample size $n$. In Lazraq and Cléroux [6], an inferential algorithm has been obtained for selecting the significant PLS latent variables. In this paper, we develop two different model-wise inferential procedures to select the significant predictors $X(j)$ among $X(1), X(2), \ldots, X(q)$. This will be done in the PLS univariate regression model PLS1. Then these procedures will be compared empirically with the two following methods: the BQ method introduced by Gauchi and Chagnon [3], found to be the best procedure among the 20 that they have compared; and the Variable Importance in the Projection (VIP) method, which is implemented in the SIMCA-P computer package and is widely used (see, for example, SIMCA-P [14] or Tenenhaus [12], p. 139). The choice of the VIP method here is explained by the fact that it has not been considered by Gauchi and Chagnon [3] in their empirical comparisons. Of course, comparing the procedures introduced in this paper with the ones mentioned in the references above should also be done. But this would be the object of another paper. The comparisons will be made on four of the data sets used in Gauchi and Chagnon [3] with respect to quality of fit of the model as measured by the multiple correlation coefficient and the adjusted multiple correlation coefficient, and also with respect to the predictive ability of the model as measured by the cross-validation prediction sum of squares.

The paper is organized as follows. Section 2 recalls a redundancy index on which the inferential algorithms are based. Section 3 presents a forward variable selection procedure in multivariate regression. This procedure is used in the sequel. Section 4 deals with latent variable selection in PLS1 models and Section 5 with predictor selection. A first method, called PLS–Forward, is presented. Section 6 presents a second method called PLS–Bootstrap. In Section 7, we recall the BQ method and in Section 8 the VIP method. The four methods are compared empirically in Section 9 and a conclusion follows in Section 10.

2. A redundancy index

In this section, we recall the redundancy index on which are based the new selection algorithms introduced in this paper.

Let $\{ \chi_i \}$ be a $(p+q) \times 1$ random vector where $Y: p \times 1$ is the vector to be predicted and $X: q \times 1$ is the vector of predictors, with covariance matrix $\Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_{XX} \end{bmatrix}$ where $\Sigma_{YY}: p \times q$, $\Sigma_{XX}: q \times q$, and $\Sigma_{XY}=\Sigma_{XY}^T: p \times q$.

Let $\{ \chi_1, \chi_2, \ldots, \chi_n \}$ be a random sample of size $n$. The covariance matrix $\Sigma$ is estimated as usual by the matrix $\hat{S} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) (Y_i - \bar{Y})^T$; $\hat{S}_{XX} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}) (X_i - \bar{X})^T$; $\hat{S}_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y}) (Y_i - \bar{Y})^T$; $\bar{Y} = (1/n) \sum_{i=1}^{n} Y_i$ and $\bar{X} = (1/n) \sum_{i=1}^{n} X_i$.

The redundancy index introduced by Stewart and Love [11], and later generalized by Gleason [4], is given by:

$$\text{RI}(Y, X) = \frac{\text{tr}(S_{YY}^{-1} S_{XY} S_{XY}^{-1} S_{YY})}{\text{tr}(S_{YY})}$$

and it can also be written as:

$$\text{RI}(Y, X) = \frac{\sum_{i=1}^{p} s^2_Y Y(i)^{2} R^2_Y(i):X}{\sum_{i=1}^{p} s^2_Y Y(i)^{2}}$$

(2)
where $R_{(i,X)}^2$ is the sample multiple correlation coefficient squared between the $i$th component $Y(i)$ of $Y$ and the vector $X$ and where $s_{(i)}^2$ is the sample variance of $Y(i)$. It is the fraction of the variance of $Y$, which is accounted for by the multivariate linear regression of $Y$ on $X$. When $p=1$, $RI(Y,X)$ reduces to the multiple correlation coefficient squared between the variable $Y$ and the vector $X$. The reader interested in other properties of $RI(Y,X)$ is referred to Lazraq and Cléroux [5].

3. Forward variable selection in multivariate regression

We mentioned earlier that we work in the PLS1 regression framework. However, in Section 5, in order to find the significant variables $X(j)$ in the prediction of a set of PLS latent variables by $X$, we need a variable selection procedure in a multivariate regression model. The following algorithm generalizes the forward variable selection in multiple regression. It is adapted from an inferential algorithm developed in Lazraq and Cléroux [5]. The inferential part of the algorithm is not necessary here for reasons to be discussed later.

The problem is thus to predict linearly a vector $Y$: $p \times 1$ from a vector of predictors $X'=(X(1),X(2),\ldots,X(q))$. Let $G$ be a subset of the set of indices $\{1,2,\ldots,q\}$ and let $X_G$ be the corresponding subvector of $X$ indexed by $G$. The selection algorithm is the as follows:

- **Step 1:** Initialize $Q \leftarrow \{1,2,\ldots,q\}$, $G \leftarrow \emptyset$, $C \leftarrow 0$.
- **Step 2:** Compute max$_{1 \leq j \leq q}RI(Y,X(j))=RI(Y,X(h))=RI$.
- **Step 3:** If $RI - C > \epsilon$.
  - 3.1: $G \leftarrow G \cup \{h\}$, $C \leftarrow RI$.
  - 3.2: Compute max$_{j \in C}RI(Y,X_{G \cup \{j\}}) = RI(Y,X_{G \cup \{j\}}) = RI$.
- **Step 4:** Go to Step 3. Else STOP.

Here $C^G_G$ is the complement of $G$ with respect to $Q$ and $\epsilon$ is a predetermined tolerance threshold. It is clear that the choice of $\epsilon$ has a direct influence on the output $X_G$. Our experience with this algorithm suggests using a value of $\epsilon$ in the interval $[0.01,0.05]$. One could also make different choices for $\epsilon$, obtain several sets $X_G$, and make a decision afterwards. The choice of $C$ will be discussed in Section 5, where this algorithm is used in the PLS context.

In the sequel, this algorithm will be denoted $X_G \leftarrow \text{Forward (Y,X)}$.

4. Latent variable selection in the PLS1 model

Here we let $p=1$ and the problem of interest is the selection of the significant PLS latent variables in the prediction of the random variable $Y$ from $X'=X(1),X(2),\ldots,X(q))$.

We assume that $\sigma_{YY} = \sigma_{XY} = 1 \times q$ and $\Sigma_{XX} = q \times q$.

We recall the PLS1 method with the notation used in Lazraq and Cléroux [6]:

(i) Let $Y(0)=Y,X(0)=X$, and $k=1$.
(ii) At step $k$, we look for a latent variable $t_k = x_k^T X(k-1)$ such that $\text{cov}(Y,X_k^T X(k-1))$ is maximized under the constraint $\|z_k\| = 1$.
(iii) Using Lagrange multipliers, the solution is:

$$z_k = \frac{\sigma_{XY}(k-1)}{(\sigma_{YY}(k-1)\sigma_{XX}(k-1))^{1/2}}$$

(iv) We make a linear regression of $X(k-1)$ on $t_k$ and of $Y(k-1)$ on $t_k$ to get $X_{(k-1)} = \mathbb{P}(k) + X(k)$ and $Y_{(k-1)} = \mathbb{R}(k) + Y(k)$ where $\mathbb{P}(k) = \Sigma_{XX}(k-1) z_k / \text{var}(t_k)$, $\mathbb{R}(k) = \Sigma_{YY}(k-1) z_k / \text{var}(t_k)$ and $\text{var}(t_k) = z_k^T \Sigma_{XX}(k-1) z_k$ (here we write $\Sigma_{XX}(k-1)$ for $\Sigma_{X(k-1)}$).
(v) It is known that if $(Y(k-1),X(k-1))$ is multivariate normal with covariance matrix $\Sigma(k-1)$, then $(Y(k),X(k))$ is also multivariate normal with covariance matrix:

$$\sum_{(k)} = \Sigma_{(k-1)} - \left( \frac{\sigma_{YY}(k-1) z_k}{\sum_{XX}(k-1) z_k} \right) \text{var}^{-1}(t_k) \times \left( \frac{\sigma_{YY}(k-1) z_k}{\sum_{XX}(k-1) z_k} \right)$$

(vi) We test the significance of $t_k$ for predicting $Y$. If $t_k$ is significant, we GO TO (ii) above with $k+1$ in
place of \( k \). The test is made explicit in the next algorithm.

4.1. Inferential PLS1 algorithm

Let \( \tilde{Y} = (Y_1, Y_2, \ldots, Y_q) \) be a random sample of \( Y \) and let:

\[
\tilde{X} = (X_q) = \begin{pmatrix} X_1^t \\ X_2^t \\ \vdots \\ X_n^t \end{pmatrix}
\]

be the \( n \times q \) matrix formed from the random sample \( X_1, X_2, \ldots, X_n \) of \( X \). In the algorithm, the following recurrence relation, proven in Lazraq and Cleroux [6], is used: if \( k \leq q \), then \( \sigma_{XY(k-1)} = \text{cov}(Y_{(k-1)}, X_{(k-1)}) = \text{cov}(Y, X_{(k-1)}) \).

Step 1: Read the data \( \tilde{Y} \) and \( \tilde{X} \). Transform \( \tilde{Y} \) and \( \tilde{X} \) so that each column has a mean of 0.

Step 2: Initialize \( \tilde{X}_{(0)} \leftarrow X, k \leftarrow 1, \text{SL} \leftarrow 0 \).

Step 3: If \( k \leq q \), compute \( s_{XY(k-1)}^{(k-1)} = (1/n - 1) \tilde{Y}^t\tilde{X}_{(k-1)}, s_{XY(k-1)}^{(k-1)} = s_{XY(k-1)^t}, \) and \( S_{XY(k-1)} = (1/n - 1)\tilde{X}_{(k-1)}^t\tilde{X}_{(k-1)} \).

Step 4: Compute \( a_k = s_{XY(k-1)}^{(k-1)}/(s_{XY(k-1)}^{(k-1)})^{1/2} \).

Step 5: Compute \( t_{(k)} = \tilde{X}_a \) (here \( t_{(k)} \) is a vector) and \( r^2 = \text{RI}(\tilde{Y}, t_{(k)}) = r^2(\tilde{Y}, t_{(k)}) \), with the simple correlation coefficient squared.

Step 6: If \( \rho \) denotes the corresponding population correlation coefficient, test and \( H_0: \rho^2 = 0 \) against \( H_1: \rho^2 > 0 \) using an \( F \) test. If \( H_0 \) is rejected, compute \( P_k = s_{XY(k-1)}^{(k-1)}a_k/\text{var}(t_{(k)}), \tilde{X}_{(k)} = \tilde{X}_{(k-1)} - t_{(k)}p_k \) and \( \text{SL} = \text{SL} + r^2 \). Then let \( k \leftarrow k + 1 \) and GO TO Step 3. If \( H_0 \) is accepted, STOP.

Step 7: For each significant PLS latent variable, print \( r^2 \), \( \text{SL} \), and the \( p \) value of the test.

Note that Steps 5 and 6 are based on the fact, proven in Lazraq and Cleroux [6] in a more general setup, that the \( t_{(k)} \)'s are independent.

5. Selecting the predictors \( X_{(j)} \): the PLS–Forward algorithm

The preceding algorithm allows selecting the significant PLS latent variables in the prediction of \( Y \): \( 1 \times 1 \) by \( X \): \( q \times 1 \). Suppose that the significant latent variables are \( t_{(1)}, t_{(2)}, \ldots, t_{(k)} \) and let \( \tilde{T} = (t_{(1)}, t_{(2)}, \ldots, t_{(k)}) \): \( n \times k \). This algorithm will be denoted by \( \tilde{T} \leftarrow \text{PLS}(\tilde{Y}, \tilde{X}) \).

The next algorithm, called PLS–Forward, will now allow the selection of a set of predictors among \( X_{(1)}, X_{(2)}, \ldots, X_{(q)} \), after the latent variables have been selected:

Step 1: Read the data \( (Y_z, X_z), z = 1, 2, \ldots, n \).

Step 2: Apply \( \tilde{T} \leftarrow \text{PLS}(\tilde{Y}, \tilde{X}) \).

Step 3: Apply \( X_G \leftarrow \text{Forward}(\tilde{T}, X) \).

Step 4: Output: \( X_G \) and \( \text{RI}(Y, X_G) \).

The set \( X_G \) obtained is the best subset of predictors (in the sense of Section 3) from \( \{X_{(1)}, X_{(2)}, \ldots, X_{(q)}\} \) to explain \( \tilde{T} \) linearly, which in turn explains \( Y \). The overall quality of \( X_G \) in predicting \( Y \) is measured by \( R^2 = \text{RI}(Y, X_G) \), the multiple correlation coefficient squared.

Since \( T=(t_1, t_2, \ldots, t_p) \) is, by construction, a perfect linear function of \( X \), it is certain that from a certain set \( G \) of cardinality \( g \), in general much smaller than \( q \), we will have \( \text{RI}(T, X_G) = 1 \). There is thus no need to perform tests of significance in the Forward (\( \tilde{T}, X \)) algorithm. On the other hand, tests of significance are necessary in the PLS(\( Y, X \)) algorithm.

We thus see that the value of \( C \) in the Forward(\( \tilde{T}, X \)) algorithm can be chosen near 1, say 0.95 or 0.99. It is the target proportion of the total value of \( \text{RI}(T, X_G) \) algorithm.

6. Selecting the predictors \( X_{(j)} \): the PLS–Bootstrap method

In this section, we also assume that \( (\tilde{Y}/\tilde{X})^{(1+q) \times 1} \) has the multivariate normal distribution, and the following results, which have been proven in Lazraq and Cleroux [6], will be used:

(i) \( T \) is given by:

\[
T = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{pmatrix} = \begin{pmatrix} \tilde{a}_1 \\ \tilde{a}_2 \\ \vdots \\ \tilde{a}_k \end{pmatrix} X = \tilde{A}X,
\]
where the $\hat{a}_j$'s are given below. Moreover, $T$ is multivariate normal if $X$ is multivariate normal.

(ii) \[
Y = \sum_{j=1}^{k} R_j t_j + Y_{(k)} = \sum_{j=1}^{k} R_j \hat{a}_j X + Y_{(k)}
\]
\[
= \left( \sum_{j=1}^{k} R_j \hat{a}_j^T \right) X + Y_{(k)} = BX + Y_{(k)}
\]

where $B$ is the $1 \times q$ vector of PLS regression coefficients, that is,
\[
B = (R_1, R_2, \ldots, R_k) \cdot \begin{pmatrix}
\hat{a}_1^T \\
\hat{a}_2^T \\
\vdots \\
\hat{a}_k^T
\end{pmatrix}
\]

with
\[
R_j = \frac{\sigma_{YX(j-1)Z_j}}{\text{var}(t_j)} = \frac{\sigma_{YX(j-1)}^2 Z_j}{\text{var}(t_j)}.
\]

Now, in practice, the parameters $z_j$, $\text{var}(t_j)$, $\sigma_{YX}^{(j-1)}$, $\hat{a}_j$, and $R_j$ are unknown and will have to be replaced by convergent estimators. They are all functions of the covariance matrix $\Sigma$ and their natural estimators are the same functions of the sample covariance matrix $S$. Let $\hat{z}_j$, $\hat{\text{var}}(t_j)$, $\hat{\sigma}_{YX}^{(j-1)}$, $\hat{a}_j$, and $\hat{R}_j$ be their estimators. Since $S \overset{p}{\rightarrow} \Sigma$ as $n \to \infty$, these estimators are convergent estimators and, by Slutsky’s theorem, $\hat{R}_j$ has the same asymptotic distribution as $Z_j = \hat{\sigma}_{YX}^{(j-1)} Z_j / \text{var}(t_j)$.

We will now find the asymptotic distribution of $\hat{B}$, which is needed in the algorithm.

In Muirhead [9] (p. 113), we can find the following result: let $S$ be the sample covariance matrix obtained from a sample of size $n$ of the population $Z$: $m \times 1$ with covariance matrix $\Sigma$. Then:
\[
\sqrt{n}(\text{vec}(S) - \text{vec}(\Sigma)) \xrightarrow{p} N(0, A)
\]
where $A = (I_m^2 + K)(\Sigma \otimes \Sigma)$, where $K = \Sigma_{j-1}^{\alpha} \Sigma_{j-1}^{\beta}$, \(H_{ij} \otimes H_{ij}'\) and where $H_{ij}$ is the $m \times m$ matrix having element $(i,j)$ equal to 1 and all the others equal to 0. Let $U$ be the $(1 + qk) \times 1$ vector defined by:
\[
U = \begin{pmatrix}
Y \\
X^{(1)} \\
\vdots \\
X_k
\end{pmatrix} = \begin{pmatrix}
Y \\
A_1 X \\
\vdots \\
A_k X
\end{pmatrix}
\]

the second equality being obtained in Lazraq and Cléroux [6] (p. 529). We can write $U = \hat{g}_U \left( \frac{Y}{X} \right)$ where:
\[
G_k = \begin{pmatrix}
1 & 0 \\
A_1 & A_2 \\
& \ddots & \ddots
\end{pmatrix}
\]

and $U$ has the multivariate normal distribution with covariance matrix $\Sigma_U = G_k \Sigma G_k'$. The sample covariance matrix $S_U$, which estimates $\Sigma_U$, is such that $\sqrt{n} (\text{vec}(S_U) - \text{vec}(\Sigma_U)) \xrightarrow{d} N(0, A_U)$ by Eq. (3). But since $S_U$ contains as submatrices the matrix $s_{YX}^{(j-1)}$ for $j = 1, 2, \ldots, k$, it follows that $\sqrt{n} \hat{B}$ has the same asymptotic distribution as:
\[
\sum_{j=1}^{k} \sqrt{n} Z_j \hat{a}_j = \sum_{j=1}^{n} \sqrt{n} s_{YX}^{(j-1)} \hat{a}_j / \text{var}(t_j)
\]

which is a linear function of jointly normally distributed random variables. Thus, $\sqrt{n}(\hat{B} - B) \xrightarrow{d} N(0, \Sigma_B)$, where $\Sigma_B$ is a complicated function of $A_U$, $z_j$, $a_j$, and $\text{var}(t_j)$ for $j = 1, 2, \ldots, q$. This function will not be made explicit here. We will use instead a bootstrap approach. More precisely, instead of finding confidence intervals and making tests of hypotheses for the components of $B = (b_1, b_2, \ldots, b_q)$ (the PLS1 regression coefficients), using an asymptotic normal distribution with unknown parameters that have to be estimated, we choose a bootstrap procedure that also takes into account the asymptotic normality of each $\hat{b}_j$, $j = 1, 2, \ldots, q$. But before writing down the algorithm, a few remarks are necessary:

(i) since the inferential PLS1 algorithm of Section 4 outputs $\hat{T}$ and since $\hat{B}$ is computed from $\hat{T}$, we will assume that the algorithm directly outputs $\hat{B}$ and we write $\hat{B} \leftarrow \text{PLS} (Y, X)$.

(ii) The following notation will be used:

(ii.1) $\ell$: number of bootstrap iterations or samples.

(ii.2) $W^\ell : \left( \frac{Y^\ell}{X^\ell} \right) = \left( \frac{Y_{\ell k}}{X_{\ell k}} \right), \ell = 1, 2, \ldots, n$, the $\ell$th bootstrap sample of size $n$ with replacement, $\ell = 1, 2, \ldots, \ell$, with $Y^\ell : n \times 1$, $X^\ell : n \times q$, and $W^\ell : n \times (1 + qk)$. 

(ii.3) For each bootstrap iteration $\ell$, we write the result of the inferential PLS1 algorithm as $B_0^* = \text{PLS}(Y_0^*, X_0^*)$ with $B_0^* = (b_{10}^*, b_{20}^*, \ldots, b_{q0}^*)$.

$1 \times q$.

### 6.1. The PLS–Bootstrap algorithm

Step 1: Repeat for $\ell = 1, 2, \ldots, L$.

1.1: Take a sample of size $n$: $u_0 - \left(\begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right)$ with replacement.

1.2: Obtain $B_0^* = \text{PLS}(Y_0^*, X_0^*)$.

Step 2: Repeat for $k = 1, 2, \ldots, q$.

2.1: $E_k \leftarrow (b_{1k}^*, b_{2k}^*, \ldots, b_{qk}^*)$: $1 \times L$, where $E_k$ is a bootstrap sample of size $L$ of $b_k$, the regression coefficient of $Y$ on $X_k$, and the $k$th component of the vector $X$ of predictors.

2.2: Compute $\hat{b}_k^* = (1/L) \sum_{\ell=1}^{L} b_{\ell k}^*$.

2.3: Compute $s_k^* = (1/L - 1) \sum_{\ell=1}^{L} (b_{\ell k}^* - \hat{b}_k^*)^2$.

2.4: Obtain a confidence interval $I_k^*$ for $b_k$.

2.5: If $0 \in I_k^*$, delete variable $X_k$.

Step 3: Output the remaining variables $X_{G_k}$.

The confidence interval of Step 2.4 is obtained by using the symmetry property of the asymptotic normal distribution and is given by $b_k^* \pm c s_k^*$ where $c$ depends on the confidence level. Thus, to retain less variables in $X_{G_k}$, it is sufficient to increase the confidence level and consequently the value of $c$. In the examples below, $c$ varies from 1.4 to 5.2.

The two proceeding algorithms, namely the PLS–Forward algorithm and the PLS–Bootstrap algorithm, which both select the PLS significant latent variables, together with the significant predictors $X_{G_k}$, will be compared empirically to the BQ method (Gauchi and Chagnon [3]) and the VIP method (SIMCA-P [14]). These methods are recalled in Sections 7 and 8 of the paper.

### 7. The BQ method

The BQ method (Backward $Q_{\text{cum}}^2$) has been found by Gauchi and Chagnon [3] to be the best variable selection procedure among the 20 methods they compared empirically. It is based on the prediction criteria $Q_{\text{cum}}^2$, which should be made, by cross-validation, as close to one as possible. In the PLS1 regression model, it is defined by:

$$Q_{\text{cum}}^2 = 1 - \frac{\mathbf{I}_{\ell} - \sum_{j=1}^{k} \text{PRESS}_j}{\text{RSS}_{j-1}}$$

where $k$ is the number of PLS latent variables in the model; PRESS is the prediction error sum of squares defined by $\text{PRESS} = \sum_{x=1}^{n_1} (Y_x - \hat{Y}_{(-x)})^2$ where $n_1$ is the size of the learning set and $\hat{Y}_{(-x)}$ is the prediction of $Y_x$ when $Y_x$ is removed from the data before constructing the model; PRESS$_j$ is the PRESS when $j$ PLS latent variables are in the model; RSS is the residual sum of squares defined by $\text{RSS} = \sum_{x=1}^{n_1} (Y_x - \hat{Y}_x)^2$ where $\hat{Y}_x$ is the prediction of $Y_x$ when $Y_x$ is included in the data for constructing the model; RSS$_{j-1}$ is the RSS when $j - 1$ PLS latent variables are in the model.

The selection procedure is as follows:

(i) First obtain a PLS model containing all $q$ variables $X_{(j)}$ and retain $k$ (say) PLS latent components by cross-validation. Compute the value of $Q_{\text{cum}}^2$.

(ii) Among the $q$ variables $X_{(j)}$, exclude the one having the smallest (in absolute value) PLS regression coefficient. Then obtain a new PLS model with the remaining $q - 1$ variables. Compute the corresponding value of $Q_{\text{cum}}^2$.

(iii) Keep on excluding in this way one variable $X_{(j)}$ at the time. The identification of the excluded variable, together with the value of $Q_{\text{cum}}^2$ at each step, is stored.

(iv) Draw the graph of $Q_{\text{cum}}^2$ as a function of the number of variables included in the model and choose the model corresponding to the maximum value of $Q_{\text{cum}}^2$. List the variables in that model.

The authors state that there often exists only one maximum value of $Q_{\text{cum}}^2$.

### 8. The VIP method

The VIP method (Variable Importance in the Projection) is implemented in the SIMCA-P computer
package [14]. It has not been considered by Gauchi and Chagnon [3] in their empirical study of several variable selection methods. The algorithm is as follows:

1. Make a PLS regression of $Y$ on $X$ and retain $k$ (say) PLS latent variables $T=(t_1,t_2,\ldots,t_k)'$ by cross-validation where, for any $\ell=1,2,\ldots,k$, $t_\ell=x_\ell X_\ell(X_\ell'X_\ell)^{-1}$ and $x_\ell=(x_1, x_2, \ldots, x_q)'$.

2. For $j=1,2,\ldots,q$: compute $\text{VIP}(j) = \left(\frac{\sum_{i=1}^{n} h_{ii} r_{ij}^2}{\sum_{i=1}^{n} r_{ij}^2}\right)^{1/2}$.

3. Write the $q$ VIP values in decreasing order. To retain $q_o<q$ variables $X(j)$ in the model, delete the $q-q_o$ variables $X(j)$ corresponding to the $q-q_o$ smallest values of VIP.

Since $\|x_\ell\| = 1$ for all $\ell$, it is easy to see that $\sum_{j=1}^{q} \text{VIP}^2(j) = q$. A variant of this method consists of first retaining only the variables $X(j)$ with a large VIP value ($>1$) and then going back to Step 1. In the computations of Section 9, we used the three steps above.

9. Empirical comparisons

In this section, the four methods of variable selection will be compared on four data sets used in Gauchi and Chagnon [3]. The descriptions of these data sets are found in that reference and will not be reproduced here. They are the LATEX data set with $n=262$, $p=1$, $q=117$; the OXY data set with $n=25$, $p=1$, $q=95$; the ADPN data set with $n=71$, $p=1$, $q=100$; and the SPIRA data set with $n=145$, $p=1$, $q=96$. The computations have been made on the standardized data.

The variable selection methods PLS–Forward and PLS–Bootstrap yield best subsets $X_G$ for fitting purposes, whereas the BQ and VIP methods are designed for prediction purposes. The four methods are compared with respect to both objectives.

In order to assess the overall quality of fit of the final model obtained by each method (i.e., the model from the set $X_G$ of variables retained by each method), we compute the value of $R^2(Y, X_G)$, the multiple correlation coefficient squared between $Y$, and the retained variables in each case together with the adjusted $R^2(Y, X_G)$. We also compute the ordinary least squares (OLS) prediction sum of squares:

$$\text{PRESS}_\text{OLS} = \sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$$

where $y_i$ is the $i$th data point; $\hat{y}_{-i}$ is the OLS estimator of $y_i$ obtained without using the $i$th observation; $e_i = y_i - \hat{y}_i$; $h_{ii}$ is the $i$th diagonal element of the $n \times n$ matrix $X_G^T(X_G^T X_G)^{-1} X_G^T$; where $X_G^T$ is the $n \times \text{card}(X_G)$ matrix formed by the $n$ observations on each variable in $X_G$. The leaving-one-out cross-validation value $R^2(Y, X_G)$ is also computed. Table 1 gives the cardinality of $X_G$ together with these values for the four variable selection methods for each data set. The following comments can be made from Table 1:

(i) For the LATEX data set, all four methods yield models that fit well the data and have a good predictive value as far as OLS is concerned.

Table 1
<table>
<thead>
<tr>
<th>Data set</th>
<th>Variable selection method</th>
<th>Card</th>
<th>$R^2(Y, X_G)$</th>
<th>Adjusted $R^2(Y, X_G)$</th>
<th>$\text{PRESS}_\text{OLS}$</th>
<th>$R^2(Y, X_G)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LATEX</td>
<td>Forward</td>
<td>24</td>
<td>0.82</td>
<td>0.80</td>
<td>56.77</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Bootstrap</td>
<td>22</td>
<td>0.79</td>
<td>0.77</td>
<td>66.02</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>BQ</td>
<td>24</td>
<td>0.83</td>
<td>0.81</td>
<td>55.75</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>VIP</td>
<td>24</td>
<td>0.81</td>
<td>0.79</td>
<td>64.90</td>
<td>0.75</td>
</tr>
<tr>
<td>OXY</td>
<td>Forward</td>
<td>12</td>
<td>0.91</td>
<td>0.84</td>
<td>12.71</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Bootstrap</td>
<td>12</td>
<td>0.93</td>
<td>0.86</td>
<td>7.75</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>BQ</td>
<td>12</td>
<td>0.85</td>
<td>0.70</td>
<td>15.38</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>VIP</td>
<td>13</td>
<td>0.78</td>
<td>0.69</td>
<td>33.82</td>
<td>0.52</td>
</tr>
<tr>
<td>ADPN</td>
<td>Forward</td>
<td>13</td>
<td>0.66</td>
<td>0.65</td>
<td>22.93</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Bootstrap</td>
<td>13</td>
<td>0.88</td>
<td>0.86</td>
<td>9.10</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>BQ</td>
<td>13</td>
<td>0.63</td>
<td>0.54</td>
<td>40.63</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>VIP</td>
<td>15</td>
<td>0.65</td>
<td>0.61</td>
<td>66.02</td>
<td>0.54</td>
</tr>
<tr>
<td>SPIRA</td>
<td>Forward</td>
<td>15</td>
<td>0.64</td>
<td>0.64</td>
<td>66.02</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Bootstrap</td>
<td>15</td>
<td>0.64</td>
<td>0.64</td>
<td>72.68</td>
<td>0.50</td>
</tr>
</tbody>
</table>
PLS–Bootstrap method retains less variables than the other methods.

(ii) For the OYY data set, PLS–Forward, PLS–Backward, and BQ have good fitting values but poor prediction values.

(iii) For the ADPN data set, BQ offers the best fit and prediction values and PLS–Forward stands in second place.

(iv) For the SPIRA data set, none of the four methods performs well.

We now compare the $X_G$ sets obtained by the four variable selection methods on the grounds of PLS prediction. In order to obtain this, we proceed as follows: starting with each $X_G$ set, we obtain the significant PLS components using the inferential procedure of Lazraq and Cléroux [6] and, from each PLS model, we compute the leaving-one-out cross-validation value of $R^2_{CV,j}$, where $J$ is the number of PLS components in the model. This is a measure of the global quality of prediction of the model. We also compute the value of $Q^2_{GJ}$ (see Eq. (5) below), which takes into account the different steps in the PLS model building:

$$Q^2_{Gj} = 1 - \prod_{j=1}^{J} (1 - Q^2_{Gj})$$

where $Q^2_{Gj} = 1 - (PRESS_j/RSS_{j-1})$; $PRESS_j = \sum_{i=1}^{n} (e_{ij}/(1 - h_{ij}))^2$ (same as Miller [8], p. 202); $e_{ij} = y_i - \hat{y}_i^{PLS}$; $\hat{y}_i^{PLS}$ is PLS prediction of $y$ with $j$ PLS components $t_1,t_2,\ldots,t_j$; $h_{ij}$ is the $i$th diagonal element of the $n \times j$ matrix $T(T^T)^{-1}T$ with $T=(t_1,t_2,\ldots,t_j)$; $n \times j$ and where $RSS_{j-1} = \sum_{i=1}^{n} (y_i - \hat{y}_i^{PLS_{j-1}})^2 = \sum_{i=1}^{n} e_{i,j-1,i}^2$.

The results are given in Table 2 and the following comments can be made:

(i) For the LATEX data set, the PLS models corresponding to PLS–Forward, PLS–Bootstrap, and BQ have good predictive values. PLS–Bootstrap and BQ require less PLS significant components. From Table 1, PLS–Bootstrap also retains less variables.

(ii) For the OXY data set, PLS–Forward, PLS–Bootstrap, and BQ yield models that fit well and predict well also. But the $X_G$ sets are quite different. The experimenter thus has some choice if he/she wishes to consider other experimental factors before selecting the variables. However, some variables are important since they are often selected. For instance, variables 53 and 96 are selected by all four methods and variables 51, 57, 74, 95, 97, and 98 are selected in three $X_G$ sets.

(iii) For the SPIRA data set, none of the methods performs well.

To go a little further in the analysis, Table 3 compares the $X_G$ sets for each method and each data set. It gives the indices of the variables selected. The following comments can be made:

(i) For the LATEX data set, it was seen that PLS–Forward, PLS–Bootstrap, and BQ yield models that fit well and predict well also. But the $X_G$ sets are quite different. The experimenter thus has some choice if he/she wishes to consider other experimental factors before selecting the variables. However, some variables are important since they are often selected. For instance, variables 53 and 96 are selected by all four methods and variables 51, 57, 74, 95, 97, and 98 are selected in three $X_G$ sets.

(ii) For the OXY data sets, it was seen that PLS–Forward, PLS–Bootstrap, and BQ yield models that both fit and predict well. Variables 53 and 90 are selected in the four $X_G$ sets and variables 57, 79, and 95 in three sets. The two sets $X_G$ (PLS–Bootstrap) and $X_G$ (VIP) contain nine common variables out of 12 or 13. But the remaining variables are sufficient to make VIP perform badly.
in terms of prediction. This shows that the PLS models are very sensitive to variable selection.

(iii) No more information is gained by looking at the \(X_G\) (ADPN) or \(X_G\) (SPIRA) sets of variables.

Let us summarize the information relative to reducing the size of a problem when selecting the PLS components and the explanatory variables:

(a) LATEX: \(n = 262\); \(q = 117\); \(\text{card}(X_G) = 22\) or 24; \(J = 5, 7, 8\). 

(b) OXY: \(n = 25\); \(q = 95\); \(\text{card}(X_G) = 12\) or 13; \(J = 1\) or 2.

(c) ADPN: \(n = 71\); \(q = 100\); \(\text{card}(X_G) = 11\) or 13; \(J = 2\).

(d) SPIRA: \(n = 145\); \(q = 96\); \(\text{card}(X_G) = 13\) or 15; \(J = 2\) or 4.

In every case, \(X_G\) contains less than 20\% of the original explanatory variables and the member of significant PLS component is small.

Let us end this section with a few comments relative to the computations:

(i) For the PLS–Bootstrap method, we took \(L = 100\), and for the four data sets, we took \(c = 4, 5.2, 2.35,\) and 1.4, respectively.

(ii) The PLS–Bootstrap method gives very stable results in the sense that if the algorithm is repeated many times, we obtain practically always the same set \(X_G\).

(iii) The PLS–Forward and PLS–Bootstrap algorithms have been coded in S-Plus. The data are given as input and the set \(X_G\) is obtained as the output of the algorithms. The S-Plus version used is S-Plus 3.4 release 1 (copyright 1988 and 1996, Mathsoft).

(iv) The execution times for these two algorithms are reasonable even for the big data sets considered here for the empirical comparisons.

(v) Graphics showing the histograms of the bootstrapped values of some representative regression coefficients have been obtained. Symmetry and normality of the distributions are reasonable hypotheses.

10. Conclusion

In this paper, we introduced two model-wise inferential procedures to select the significant predictor set \(X_G\) in PLS1 regression and compared them with two other methods that exist in the literature. The PLS–
Forward and PLS–Forward perform well, although no one is uniformly better between them and the BQ procedure. It was seen that two different $X_G$ sets may yield good fitting and predictive results, but slightly different $X_G$ sets may yield very different results. The PLS model is very sensitive to predictor selection. Therefore, it would be wise in practical situations to use more than one selection procedure and examine the different $X_G$ sets obtained in order to identify the important predictors (i.e., those that are selected in all or most sets $X_G$). Then, as suggested by Miller [8] (pp. 13–14), these important predictors can be forced into future subsets for prediction. Obtaining more than one set $X_G$ yielding a PLS model, which fits well the data and has good predictive ability, offers the experimenter the possibility of taking into account some other experimental factors before selecting his own $X_G$ set.

Acknowledgements

The authors express their gratitude to the anonymous referees for their helpful comments. This paper has been supported, in part, by the National Sciences and Engineering Research Council of Canada.

References