Working Paper 97-62
Statistics and Econometrics Series 23
September 1997

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

ON ROBUST PARTIAL LEAST SQUARES (PLS) METHODS.

Juan A. G. Torrubias and Rosario Romera*

Abstract

PLS regression methods have been used in applied fields for two decades. Techniques based on iteratively reweighted regression have appeared in the specialized literature with the contaminated data case. We propose a new robust PLS technique based on the Stahel-Donoho estimator. Computational results showing the better robustness and efficiency of the new method are included.

Keywords:
Partial least squares, robust regression methods, robust covariance matrices, Stahel-Donoho estimator.

*G. Torrubias, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid. C/ Madrid, 126 28903 Madrid. Spain. Ph: 34-1-624.98.90, Fax: 34-1-624.98.49, e-mail: torrubia@est-econ.uc3m.es; Romera, Departamento de Estadística y Econometría. Universidad Carlos III de Madrid. e-mail: mrromera@est-econ.uc3m.es

On Robust Partial Least Squares (PLS) Methods

Juan A. G. Torrubias[*] M. Rosario Romera[†]

September - 1997

Abstract

PLS regression methods have been used in applied fields for two decades. Techniques based on iteratively reweighted regression have appeared in the specialized literature with the contaminated data case. We propose a new robust PLS technique based on the Stahel-Donoho estimator. Computational results showing the better robustness and efficiency of the new method are included.

Key Words: Partial Least Squares; Robust Regression Methods; Robust Covariance Matrices; Stahel-Donoho Estimator.

[*] Universidad Carlos III de Madrid, e-mail: torrubia@est-econ.uc3m.es

[†] Universidad Carlos III de Madrid, e-mail: mrromera@est-econ.uc3m.es

1 Introduction

The interest in Partial Least Squares (PLS) methods from the statistical point of view is very recent (Stone and Brooks (1990), Frank and Friedman (1993), Breiman and Friedman (1997)). The traditional regression methods used with highly collinear data, such as Principal Components Regression (PCR) and Ridge Regression (RR), have been compared with PLS in term of prediction ability (Helland and Almøy (1994)). They conclude that in a lot of cases PLS is the better option.

Outliers in high dimensions are difficult to detect, but they generally affect the estimation. Robust techniques for ordinary least squares (OLS) regression have been widely developed, but the estimators are still affected by multicollinearity. The PLS algorithm solves the problem of collinearity, but it is still affected by the outliers. For example, the new Quantitative Structure-Activity Relationship (QSAR) procedures (in the Chemometrics field) have prediction problems when there are outliers in the data. Some robust solutions have been lately proposed, but the reflection in the statistic literature is still scarce.

In this section, we present the classical version of the PLS1[1] algorithm. Section 2 describes the robust methods that exist in the literature. Section 3 presents the methods that we propose. Section 4 contains a simulation study and graphics with the most significant results. Finally, section 5 collects some conclusions and possible extensions.

PLS1 algorithm

Let $\mathbf{X}$ be a matrix of dimension $n \times m$. We want to use this matrix to predict future values of the vector $\mathbf{y}$, which has dimension $n \times 1$. We suppose that the data are centred. PLS1 is a technique for dimensionality reduction, which is focused on maximizing the predictive power. The method calculates a new matrix $\mathbf{T}$ of dimension $n \times A$ (where $A \leq m$), whose columns are linear combinations of the columns of $\mathbf{X}$. The algorithm is the following:

1. Supose that A is a fixed and known value. Give starting values for the residual matrices.

$$\begin{aligned} a &= 1 \\ \mathbf{E}_0 &= \mathbf{X} \\ \mathbf{f}_0 &= \mathbf{y} \end{aligned} \tag{1}$$

[1] PLS regression methods with univariate response are known as PLS1. If the response is multivariate, then are known as PLS2. In this article, we often use PLS to refer to PLS1.

2. Calculate the loading[2] vector $\mathbf{w}_a$ and then calculate the corresponding scores vector $\mathbf{t}_a$.

$$\begin{array}{lcll} \mathbf{w}_a & \propto & \mathbf{E}'_{a-1}\mathbf{f}_{a-1} & \|\mathbf{w}_a\| = 1 \\ \mathbf{t}_a & = & \mathbf{E}_{a-1}\mathbf{w}_a & \end{array} \tag{2}$$

3. Calculate $m+1$ simple regressions between the columns of the residual matrices and the new scores vector $\mathbf{t}_a$. Since $\mathbf{t}_a$ is a linear combination of the columns of $\mathbf{X}$, it also has mean equal to zero. Thus the regressions have zero intercepts.

$$\begin{array}{lcll} \mathbf{f}_{a-1} & = & q_a\mathbf{t}_a + \mathbf{e}_{f_{a-1}} & \\ \mathbf{E}_{a-1,j} & = & p_{a,j}\mathbf{t}_a + \mathbf{e}_{E_{a-1,j}} & j = 1, 2, \ldots, m \end{array} \tag{3}$$

4. Update the residual matrices using the residuals estimated in the previous step as the new column vectors.

$$\begin{array}{lcl} \mathbf{f}_a & = & \hat{\mathbf{e}}_{f_{a-1}} \\ \mathbf{E}_a & = & [\hat{\mathbf{e}}_{E_{a-1,1}}, \hat{\mathbf{e}}_{E_{a-1,2}}, \ldots, \hat{\mathbf{e}}_{E_{a-1,m}}] \end{array} \tag{4}$$

5. If a is less than A, then increase its value by 1 and return to step 2. If a is equal than A, then the algorithm stops and the estimated vector $\mathbf{y}$ is

$$\hat{\mathbf{y}}_A = \hat{q}_1\mathbf{t}_1 + \hat{q}_2\mathbf{t}_2 + \ldots + \hat{q}_A\mathbf{t}_A \tag{5}$$

The above algorithm requires that A is known but in practice this value is unknown and must be estimated, for example by *leave-one-out* cross-validation methods. Let y_i be the ith element of the vector $\mathbf{y}$, and let $\hat{y}_{(i),A}$ be the estimate of $\mathbf{y}$ which comes from PLS with A components when we have eliminated the ith observation. Then, the value of Predictive Residual Error Sum of Squares (PRESS) is

$$PRESS(A) = \sum_{i=1}^{n} \frac{(y_i - (\hat{y}_{(i),A})_i)^2}{n} \tag{6}$$

and we choose the number of components A that minimizes this value

$$A = arg\min_k\{PRESS(k)\}$$

[2] In the literature, the vectors $\mathbf{w}_a$ are called *weight vectors* because they give a weight to each variable $\mathbf{x}_j$ to form the scores vectors $\mathbf{t}_a$. In this article, the vectors $\mathbf{w}_a$ are called *loading vectors* in order to avoid the confusion with the *weight vectors* of the robust methods

2 Robust-PLS based on Reweighting techniques

We start this section by emphasizing that all the robust PLS methods that have been proposed before in the literature use *iterative reweighting* techniques. Therefore, we first summarize iteratively reweighted regression and then robust PLS methods.

Iteratively Reweighted Least Squares (IRLS)

Suppose that we want to calculate the multiple regression of $\mathbf{y}$ on $\mathbf{X}$. The IRLS method is the following:

1. Calculate the initial value to the regression coefficient $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
2. Calculate the residuals of the regression $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$
3. Calculate weights[3] for each observation according to the corresponding residuals. The observations which have a small residuals have high weights (close to 1), and the observations which have a high residuals have small weights (close to 0). Construct a diagonal matrix $\boldsymbol{\Phi}$ from the weight vector.
4. Apply the weights and calculate a new regression coefficient
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{y}$$
5. Choose some convergence criterion and repeat the steps from 2 to 4 until the criterion has been satisfied.

2.1 Internal Iterative Reweighting: IRLS into PLS

Wakeling and Macfie (1992) were the first to write about the robustification of PLS. They work with the PLS2 algoritm, which they present in the form of many simple regressions. Next, they replace these regressions with weighted regressions. To avoid confusion, we do not include here the formulation of PLS2 since we are only working in this article with the PLS1 algorithm.

Following the previous methodology, Griep, Wakeling, Vankeerberghen, and Massart (1995) carry out comparison among three different methods of robust regression and they study their incorporation into PLS1 algorithm. To understand better the concepts, it can be seen from table 1 that PLS1 can be presented so that it uses four different regression steps. In principle, it is

[3] Cummins and Andrews (1995) present a checking of the weight functions used in IRLS.

PLS1 Algorithm		
Classical Version	Version with Regressions	
$\mathbf{w} = \mathbf{X}'\mathbf{y}$	$\mathbf{w} = \mathbf{X}'\mathbf{y}/\mathbf{y}'\mathbf{y}$	Regression
$\mathbf{w} = \mathbf{w}/\|\mathbf{w}\|$	$\mathbf{w} = \mathbf{w}/\|\mathbf{w}\|$	Normalization
$\mathbf{t} = \mathbf{X}\mathbf{w}$	$\mathbf{t} = (\mathbf{X}')'\mathbf{w}/\mathbf{w}'\mathbf{w}$	Regression
$\mathbf{p} = \mathbf{X}'\mathbf{t}/\mathbf{t}'\mathbf{t}$	$\mathbf{p} = \mathbf{X}'\mathbf{t}/\mathbf{t}'\mathbf{t}$	Regression
$b = \mathbf{y}'\mathbf{t}/\mathbf{t}'\mathbf{t}$	$b = \mathbf{y}'\mathbf{t}/\mathbf{t}'\mathbf{t}$	Regression
$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}'$	$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}'$	Residual Matrix $\mathbf{X}$
$\mathbf{y} = \mathbf{y} - b\mathbf{t}$	$\mathbf{y} = \mathbf{y} - b\mathbf{t}$	Residual Vector $\mathbf{y}$

Table 1: Two equivalent forms to present the steps of PLS1 algorithm.

possible to replace all the steps by one of the robust procedures which would make the procedure completely robust. On the other hand, replacement of one or two selected steps instead of all steps together could still have a good performance in terms of handling outliers. Such a procedure is called semirobust.

In their study, they replace the first step with three different methods of robust regression: Least Median of Squares(LMS), Siegel's Repeated Median (RM) and Iteratively Reweighted Least Squares (IRLS). On the basis of their results, the best option is to use IRLS.

In fact the first step of PLS is not just one regression but it is formed from the simple regressions of each variable $\mathbf{X}_j$ on $\mathbf{y}$. Thus the application of IRLS in this first step consists of the application of IRLS to each simple regression. Therefore, a criticism of the work of Griep, Wakeling, Vankeerberghen, and Massart (1995) is that they look for outliers in the projections of the data onto the planes $\{\mathbf{X}_1, \mathbf{y}\}, \{\mathbf{X}_2, \mathbf{y}\}, \ldots, \{\mathbf{X}_p, \mathbf{y}\}$, but they forget the multivariate nature of the data. In fact, there may exist outliers in high dimensions that can not be detected when we project onto these planes.

2.2 External Iterative Reweighting: IRPLS

Cummins and Andrews (1995) propose to generalize iteratively reweighted regression to the PLS1 algorithm. The idea is the same one as in IRLS but in this case using the residuals of the PLS regression. The steps are:

1. Perform an ordinary PLS regression analysis.

2. Pass the regression residuals from step1 into the weight function.

3. Perform a weighted PLS regression analysis using the weights just obtained.

4. Pass the residuals[4] from step 3 into the weight function.

5. If the convergence criterion is met then stop; else go to step 3.

A convergence criterion could compare the cross-validated R^2

$$R^2_{cv} = 1 - \frac{PRESS}{SS_{Total}} \tag{7}$$

from the current iteration with its value from the previous iteration and the stopping rule might be to stop if the value has changed by less than 5%. They found that convergence occurs very quickly.

In step 3, the weighted PLS is first done with cross-validation to choose the optimal number of components, then a final run is performed using that number of components. The criterion for choosing the optimal number of components is that of maximizing the cross-validated R^2, which is equivalent to minimizing the *PRESS* value.

The problem that this method could have is that the residuals for each sample would depend strongly on the number of components calculated in PLS. In this case, different criteria for choosing the number of components could cause different weights for each sample. It would be interesting to study if in this case there are convergence problems.

3 New Robust-PLS

Let's assume that the data come from a joint distribution with mean vector equal to 0 and population covariance matrix $\Sigma_{[y,X]}$ consisting of the following elements

$$\Sigma\Sigma_{[y,X]} = \begin{pmatrix} \sigma^2_y & \delta'_{y,X} \\ \delta_{y,X} & \Sigma_X \end{pmatrix} \tag{8}$$

Independently of the original data distribution, the *population* loading vectors are given by the following equation (Helland 1990):

$$\begin{array}{lll} \mathbf{w}_1 & \propto & \delta_{y,X} \\ \mathbf{w}_{a+1} & \propto & \delta_{y,X} - \Sigma_X \mathbf{W}_a(\mathbf{W}'_a\Sigma_X\mathbf{W}_a)^{-1}\mathbf{W}'_a\delta_{y,X} \end{array} \tag{9}$$

where $\mathbf{W}_a$ is the matrix whose column vectors are the $\mathbf{w}_i$ calculated previously

$$\mathbf{W}_a = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_a]$$

[4] Rather than passing the ordinary residuals into the weight function, they found that better results were obtained by passing the predicted residuals into the weight function.

Using the expressions given by (9) it is not necessary to calculate the components $\mathbf{t}_a$. In each step, $\mathbf{w}_{a+1}$ only depends on the value of the a previous vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_a$, on Σ_X and on $\delta_{y,X}$. Moreover, w_1 only depends on $\delta_{y,X}$. Then we can conclude that the calculating of $\mathbf{w}_{a+1}$ is completely fixed by the value of Σ_X and $\delta_{y,X}$.

With the same notation, Helland (1990) gives the equation for the *population* regression vector calculated by PLS

$$\beta_{a,PLS} = \mathbf{W}_a(\mathbf{W}'_a \Sigma_X \mathbf{W}_a)^{-1} \mathbf{W}'_a \delta_{y,X} \tag{10}$$

Just like $\mathbf{w}_a$, this vector $\beta_{a,PLS}$ only depends of the values of Σ_X and $\delta_{y,X}$.

The two previous equations (9) and (10) can be considered as an alternative definition of the PLS algorithm.

The first two methods proposed in the literature (Wakeling and Macfie (1992), Griep, Wakeling, Vankeerberghen, and Massart (1995)), which have been described in the section 2.1, present a robust calculation of the vector $\delta_{y,X}$ but they leave alone the matrix Σ_X. In the simulations, these methods of robustification seem to work well but one possible reason for this good behaviour is that in the simulation studies only the vector $\mathbf{y}$ is contaminated. Therefore, the matrix Σ_X is not affected by the outliers.

In this article, we propose three robustifications of the PLS algorithm based on robustification procedures, statisticaly tested, for the covariance matrix (Maronna and Yohai 1995), for example the Stahel-Donoho estimator (SDE) or the minimum volume ellipsoid estimator (MVEE). The first two could be considered partial robustification or semirobust, and the third is a global robustification. First we present the three methods and then we explain briefly the method of robustification of covariance matrices.

3.1 Partial Robustification

We follow the idea of the methods proposed by Wakeling and Macfie (1992) and Griep, Wakeling, Vankeerberghen, and Massart (1995). We are going to calculate the loading vector $\mathbf{w}_a$ in a robust form and to keep the remainding steps of the algorithm without any modification. The calculation of $\mathbf{w}_a$ can be seen as the normalization of the covariance vector between the matrix $\mathbf{E}_{a-1}$ and the vector $\mathbf{y}$. Next we present two possibilities.

PLSR

We can suppose that the data $[\mathbf{y}, \mathbf{X}]$ have a sample covariance matrix $\mathbf{V}$ with dimension $m+1$

$$\mathbf{V} = \begin{pmatrix} Var(\mathbf{y}) & Cov(\mathbf{y}, \mathbf{X}) \\ Cov(\mathbf{y}, \mathbf{X})' & Var(\mathbf{X}) \end{pmatrix} = \begin{pmatrix} \alpha & \delta' \\ \delta & \Sigma \end{pmatrix} \tag{11}$$

In this case, we calculate the matrix V in robust form and the resulting matrix is

$$\mathbf{V}_r = \begin{pmatrix} \alpha_r & \delta_r' \\ \delta_r & \Sigma_r \end{pmatrix} \tag{12}$$

Next we take $\mathbf{w}_1$ as the normalization of the vector δ_r

In the following iterations, we calculate the loading vector $\mathbf{w}_a$ in the same way, replacing $\mathbf{X}$ by $\mathbf{E}_{a-1}$.

PLSR2

In the previous method, we used the information about all the variables to estimate the covariance vector in a robust form. Now, following Griep, Wakeling, Vankeerberghen, and Massart (1995), we are going to use only the information about two variables in each step. The covariance vector δ, which is defined by equation (11), consists of the individual covariances between $\mathbf{y}$ and each column $\mathbf{x}_j$ of $\mathbf{X}$. In this case, we estimate in a robust form m matrices each with dimension two

$$\begin{array}{lll} \mathbf{U}_r^1 & = \begin{pmatrix} Var(\mathbf{y}) & Cov(\mathbf{y}, \mathbf{x}_1) \\ Cov(\mathbf{y}, \mathbf{x}_1) & Var(\mathbf{x}_1) \end{pmatrix} & = \begin{pmatrix} a_1 & b_1 \\ b_1 & c_1 \end{pmatrix} \\ \vdots & \qquad\qquad\vdots & \qquad\vdots \\ \mathbf{U}_r^m & = \begin{pmatrix} Var(\mathbf{y}) & Cov(\mathbf{y}, \mathbf{x}_m) \\ Cov(\mathbf{y}, \mathbf{x}_m) & Var(\mathbf{x}_m) \end{pmatrix} & = \begin{pmatrix} a_m & b_m \\ b_m & c_m \end{pmatrix} \end{array} \tag{13}$$

and we construct the new vector δ_r using the elements that have been calculated individually $\delta_r = (b_1, b_2, \ldots, b_m)'$. Finally, we normalize the vector δ_r and the result is the robust loading vector $\mathbf{w}_1$.

In the following iterations, we calculate the loading vector $\mathbf{w}_a$ in the same way, replacing $\mathbf{X}$ by $\mathbf{E}_{a-1}$.

3.2 Global Robustification: PLSMR

We have seen before that in the calculation of $\mathbf{w}_1, \ldots, \mathbf{w}_a$ using the iterative procedure (9) only the values of Σ_X and $\delta_{y,X}$ are necessary. If we replace these population values by the sample values, they can be influenced by the presence of outliers. A *global* robust version can come from using again the robust covariance matrix

$$\mathbf{V}_r = \begin{pmatrix} \alpha_r & \delta_r' \\ \delta_r & \Sigma_r \end{pmatrix}$$

As in PLSR, we take $\mathbf{w}_1$ as the normalization of the vector δ_r, but now the succeeding values of $\mathbf{w}_a$ can be calculated in a robust form using

$$\mathbf{w}_{a+1} \propto \delta_r - \Sigma_r \mathbf{W}_a (\mathbf{W}_a' \Sigma_r \mathbf{W}_a)^{-1} \mathbf{W}_a' \delta_r \tag{14}$$

3.3 Robustification of covariance matrices

The two robustification methods for covariance matrices most widely used are the Stahel-Donoho Estimator (SDE) and the Minimum Volume Ellipsoid Estimator (MVEE). Maronna and Yohai (1995) carry out a broad simulation study where they compare the behaviour, among others, of both estimators. They conclude that SDE has the best bias and variability properties and therefore, it will be the choosen method in this work.

Stahel-Donoho Estimator (SDE)

Let $\mathbf{X}$ be a matrix with n samples (columns) and p variables (rows). Denoting by $\mathbf{x}_i$ the ith column of $\mathbf{X}$, the location and scale SDE $(\mathbf{t}_{SDE}(X), \mathbf{V}_{SDE}(X))$ are defined as

$$\mathbf{t}_{SDE}(X) = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i}{\sum_{i=1}^{n} w_i} \tag{15}$$

and

$$\mathbf{V}_{SDE}(X) = \frac{\sum_{i=1}^{n} [w_i(\mathbf{x}_i - \mathbf{t}_{SDE}(X))(\mathbf{x}_i - \mathbf{t}_{SDE}(X))']}{\sum_{i=1}^{n} w_i} \tag{16}$$

with $w_i = \mathrm{w}(\mathrm{r}(\mathbf{x}_i, \mathbf{X}))$, where

$$\mathrm{r}(\mathbf{x}_i, \mathbf{X}) = \sup_{\|\mathrm{d}\|=1} \left\{ \frac{|\mathbf{d}'\mathbf{x}_i - \mu(\mathbf{d}'\mathbf{X})|}{\sigma(\mathbf{d}'\mathbf{X})} \right\} \tag{17}$$

Maronna and Yohai (1995) make some recommendations that have been used in this article:

- The weight function is *Huber*'s function.

$$\mathrm{w}(r) = \begin{cases} 1 & si \quad r \leq c \\ (\frac{c}{r})^2 & si \quad r > c \end{cases} \tag{18}$$

- A good choice for the constant c in *Huber*'s function is the following

$$c = \sqrt{\chi_p^2(.95)} \tag{19}$$

- The maximum breakdown point is attained when μ is the median and σ is the average of the k_1th and the k_2th smallest absolute deviations about μ, with

$$\begin{aligned} k_1 &= p - 1 + [(n+1)/2] \\ k_2 &= p - 1 + [(n+2)/2] \end{aligned}$$

Alternative subsampling scheme

The SDE is not usually calculated in its exact form because the computational cost is very high. In general, approximate methods with subsampling procedures are used. In this article, we have used an alternative subsampling procedure proposed by Juan and Prieto (1995) which has the following form

1. Construct N subsamples of size $p+2$.

2. Remove from each subsample the observation having the largest Mahalanobis distance.

3. Compute the direction orthogonal to each of the $p+1$ subsets of p observations that can be formed from the final subsample of size $p+1$.

4. Compute $\mathrm{r}(\mathbf{x}_i, \mathbf{X})$, replacing $||\mathbf{d}|| = 1$ with the set of directions obtained in step 3.

4 Comparative results

4.1 Measuring performance

Let $\mathbf{X}^c$ and $\mathbf{y}^c$ be the contaminated data. Using simulation, we want to compare the PLS algorithm and the robust methods, and see how they are affected by the contamination. With this intention, we need to define some criteria of comparison that depend on the *true values*.

Before defining the criteria of comparison, we must emphasize that there are at least two possibilities for defining the *true values* of PLS: the first of them consists of choosing the values resulting from applying PLS to the non-contaminated data; the second consists of choosing the population[5] values of PLS. In this work we use simulation and we will see the results of comparison for both criteria.

In classical regression the comparison between the true parameter vector $\boldsymbol{\beta}$ and the estimated $\hat{\boldsymbol{\beta}}$ can give an idea about the robustness of the estimation. In PLS, the values that define the estimation are the loading vectors $\mathbf{w}_1, \ldots, \mathbf{w}_A$. Wakeling and Macfie (1992) carry out the comparison between the real and the estimated loading vectors. We are going to compare these vectors too, but moreover we are going to compare the *true* regression vector $\boldsymbol{\beta}_{PLS}$ and the estimated vectors using the other methods. The proposed comparison measures are the following:

- Every vector $\mathbf{w}_i$ has norm equal to 1. Therefore, as one measure of comparison, it seems reasonable to choose the angle between them. Using the two comparison criteria, we have the following discrepancy measures:

$$\begin{aligned} angw_{1,i,j} &= ang(\mathbf{w}_{i,Pob}; \hat{\mathbf{w}}_{i,[y^c,X^c],j}) & (20) \\ angw_{2,i,j} &= ang(\hat{\mathbf{w}}_{i,[y,X],PLS}; \hat{\mathbf{w}}_{i,[y^c,X^c],j}) & (21) \end{aligned}$$

  where $ang(\mathbf{v}_1; \mathbf{v}_2)$ is the function that calculates the angle in degrees between the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$; $i = 1, 2, \ldots, A$ indicates the number of components calculated in PLS; and $j = 1, 2, \ldots, 6$ indicates the method used: PLS, PLSR, PLSR2, PLSMR, PLSIR, IRPLS.

  In this notation, $\mathbf{w}_{i,Pob}$ indicates the value of $\mathbf{w}_i$ when we use the population covariance matrix given by (8), $\hat{\mathbf{w}}_{i,[y,X],PLS}$ indicates the estimate of $\mathbf{w}_i$ when we apply PLS to the non contaminated data, and $\hat{\mathbf{w}}_{i,[y^c,X^c],j}$ indicates the estimate of $\mathbf{w}_i$ when we apply the jth method to the contaminated data.

- To compare the vectors $\boldsymbol{\beta}$ we can follow the method and the notation used with the vectors $\mathbf{w}_i$ to define the following discrepancy measures:

$$\begin{aligned} ang\beta_{1,j} &= ang(\boldsymbol{\beta}_{Pob}; \hat{\boldsymbol{\beta}}_{[y^c,X^c],j}) & (22) \\ ang\beta_{2,j} &= ang(\hat{\boldsymbol{\beta}}_{[y,X],PLS}; \hat{\boldsymbol{\beta}}_{[y^c,X^c],j}) & (23) \end{aligned}$$

[5] In real data, these population values are unknown but in simulated data we can know the theoretical covariance matrix which generates the data.

Additionally, the vectors $\boldsymbol{\beta}$ have different norms and therefore, a comparison measure among the norms is interesting too. Three possible comparison measures are: the norm of the difference of the vectors; the quotient between the norm of the difference of the vectors and the norm of the *true* vector $\boldsymbol{\beta}$; and finally, the quotient between the norm of the estimated vector and the *true* vector. We have choosen the last measure because it reflects better the difference in the behaviour of the different methods. The measures are the following

$$
\begin{aligned}
norm\beta_{1,j} &= \frac{||\hat{\boldsymbol{\beta}}_{[y^c,X^c],j}||}{||\boldsymbol{\beta}_{Pob}||} && (24)\\
norm\beta_{2,j} &= \frac{||\hat{\boldsymbol{\beta}}_{[y^c,X^c],j}||}{||\hat{\boldsymbol{\beta}}_{[y,X],PLS}||} && (25)
\end{aligned}
$$

The ideal behaviour of a robust method would consist of having values of *angw* and *ang*β close to 0, and the value of *norm*β close to 1. In this work, we analyze all the comparison measures for all the presented methods.

4.2 Design of the study

There are two possible options, to use real data or use simulated data, and we have choosen the latter. The reason is to avoid possible initial contaminations in the data, and make sure that the only contaminations in the data are those that we insert.

To simulate the data, we use a covariance matrix $\mathbf{V}$ that we take as population covariance matrix. A way to choose this matrix, in such a way that it respects the structure of the data used in PLS, is to calculate it as the sample covariance matrix of some data used previously in the literature. In this work, we have choosen the data of Næs (1985, table 3) that consist of 9 explanatory variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_9$ and one explained variable $\mathbf{y}$. To avoid problems we have used only the 38 first observations, since the others have been classified as abnormal by the author himself. The covariance matrix of the data is presented in the table 2.

From this covariance matrix $\mathbf{V}$ we can calculate the population values of the vectors $\mathbf{w}_a$ and $\boldsymbol{\beta}$ using the equations (9) and (10). These values are fixed for all the simulations.

The simulation procedure consists of generating N data matrices $[\mathbf{y}, \mathbf{X}]^{(i)}$, with dimension $n \times 10$, from a multivariate normal distribution with mean vector equal to zero and covariance matrix equal to $\mathbf{V}$. For each matrix $[\mathbf{y}, \mathbf{X}]^{(i)}$, we can calculate the vectors $\mathbf{w}_a$ and the vector $\boldsymbol{\beta}$ applying PLS and we can take these values as the *true* values.

	$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$
$\mathbf{y}$	**13.7467**	0.3458	0.3081	0.2692	0.2744	0.2580	0.2684	0.2836	0.2836	0.1716
$\mathbf{x}_1$	0.3458	**0.0147**	0.0135	0.0120	0.0120	0.0109	0.0107	0.0113	0.0114	0.0068
$\mathbf{x}_2$	0.3081	0.0135	**0.0124**	0.0110	0.0111	0.0100	0.0098	0.0103	0.0104	0.0062
$\mathbf{x}_3$	0.2692	0.0120	0.0110	**0.0099**	0.0099	0.0090	0.0088	0.0093	0.0093	0.0056
$\mathbf{x}_4$	0.2744	0.0120	0.0111	0.0099	**0.0100**	0.0091	0.0088	0.0093	0.0094	0.0056
$\mathbf{x}_5$	0.2580	0.0109	0.0100	0.0090	0.0091	**0.0083**	0.0081	0.0085	0.0086	0.0052
$\mathbf{x}_6$	0.2684	0.0107	0.0098	0.0088	0.0088	0.0081	**0.0080**	0.0084	0.0085	0.0051
$\mathbf{x}_7$	0.2836	0.0113	0.0103	0.0093	0.0093	0.0085	0.0084	**0.0089**	0.0089	0.0054
$\mathbf{x}_8$	0.2836	0.0114	0.0104	0.0093	0.0094	0.0086	0.0085	0.0089	**0.0090**	0.0054
$\mathbf{x}_9$	0.1716	0.0068	0.0062	0.0056	0.0056	0.0052	0.0051	0.0054	0.0054	**0.0033**

Table 2: Sample covariance matrix of the data of Næs (1985) using only the first 38 observations. The elements on the diagonal (in bold) are the variances of each variable

Let ϵ be the percentage of contamination in the data which we consider fixed through the N simulations. We choose ϵ % of the elements of the matrix $[\mathbf{y}, \mathbf{X}]^{(i)}$ randomly, and add to them an element that is generated from a normal distribution with mean equal to zero and variance M times that of the corresponding variable. The new matrix of contaminated data is denoted by $[\mathbf{y}^c, \mathbf{x}^c]^{(i)}$. For each $i = 1, \ldots, N$, we apply PLS and the other robust methods to the contaminated data and then compute the values of the comparison measures.

The values that have been chosen for the different parameters in this study are the following: the number of simulations, for each fixed value of ϵ and M, is $N = 100$; the number of observations in the data matrix is $n = 50$; the percentages of contamination used are $\epsilon = 0.5, 1, 3, 5, 10, 20, 30$; and the sizes of the contaminations are $M = 3, 5, 10, 20$.

In order to fix the number of components calculated by PLS in each simulation, we have carried out a previous simulation. We have generated 100 data matrices with $n = 50$ observations from a multivariate normal with mean vector equal to zero and covariance matrix equal to $\mathbf{V}$. For each of them, we have calculated the $PRESS$ value when the number of components A varies from 1 to 9. The results are presented in the figure 1. It must be emphasized that, though the minimum value of $PRESS$ is not always attained for $A = 3$, the decrease is very small from 3 onwards. This conclussion agrees with that of Hardy, MacLaurin, Haswell, de Jong, and Vandeginste (1996, page 126) from the real data.

4.3 Results

We have chosen a specific case with $\epsilon = 5$ % and $M = 3$ to present the results for graphic form and compare in the six described methods, including PLS, the values of the performance measures proposed previously. In all the

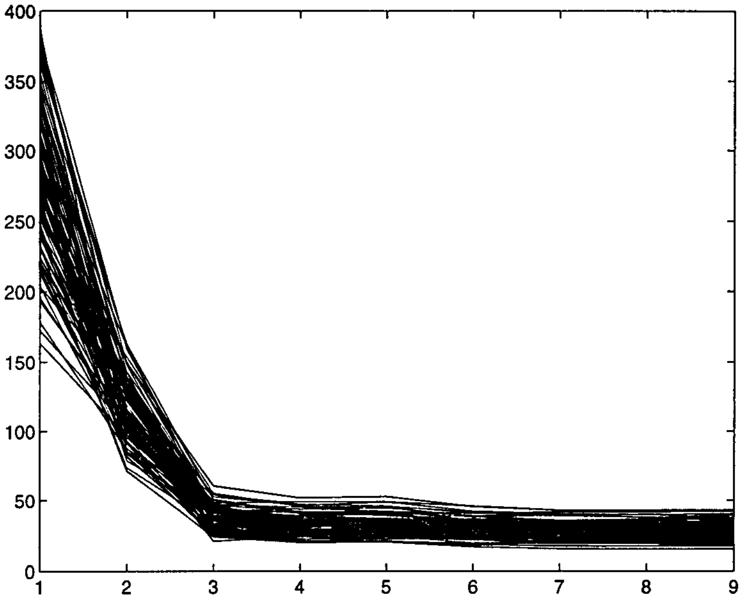

Figure 1: *PRESS* values calculate with 100 simulations of a multivariate normal distribution with mean vector equal to zero and covariance matrix Vgiven by the table 2. The number of generated observations in each simualtion has been fixed at $n = 50$

graphics presented in the figures 2, 3 and 4, the result on the left corresponds to PLS with the contaminated data. Next we comment on these graphics in more detail.

In figure 2 are presented the distributions that correspond to the angles defined by equation (21), for $j = 1, 2$. The results corresponding to $\mathbf{w}_1$ appear on the left. We can emphasize some important points. In the first place, the methods that obtain the better results (the smaller angles) are precisely those that use SDE jointly in all the variables, PLSR and PLSMR. In the second place, the remaining robust methods do not achieve significant improvements over PLS. The results corresponding to $\mathbf{w}_2$ appear on the right. In this case, all the robust methods except IRPLS achieve significant decreases in the angles compared with PLS. Again, PLSR and PLSMR are best.

In figure 3 are presented the distributions that correspond to the angles defined by the equation (21), for $j = 3$, and equation (23). The results corresponding to $\mathbf{w}_3$ appear on the left. Most outstanding are the small angles obtained by PLSMR in all the simulations. The other methods have similar behaviour to PLS, with a very high percentage of angles larger that 75 degrees. This could indicate that the power of robustification decreases as the number of components increases. The results corresponding to $\boldsymbol{\beta}$ appear on

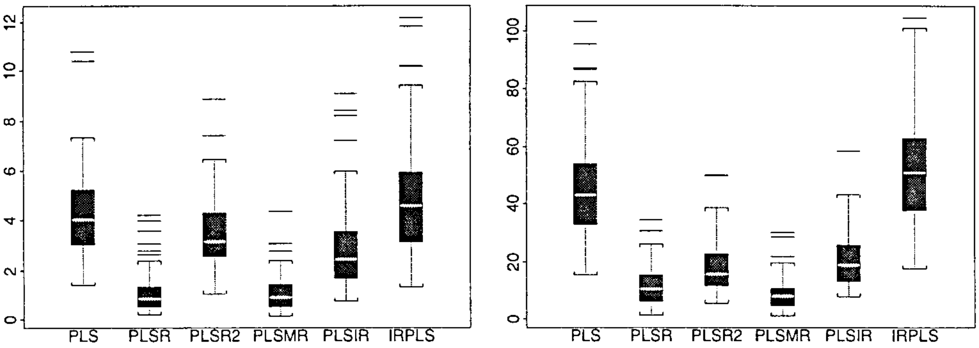

Figure 2: Boxplots of the angles, measured in degrees, between the vectors $\mathbf{w}_1$ (left graphic) and $\mathbf{w}_2$ (right graphic) using the contaminated data and the corresponding vectors calculated with the non-contaminated data. The values of the parameters are the following: $N = 100$ simulations; $n = 50$ observations; the percentage of contamination is $\epsilon = 5$ %; and the size of contamination is $M = 3$

the right. Again, most outstanding is the clear difference between PLSMR and the rest. The methods PLSR, PLSR2 and PLSIR have been defined as *semirobust*, so it is quite logical that its angles are a little smaller than the angles obtained by PLS, but without becoming as good as the global robust procedure. The case of IRPLS is more surprising since initially this method appears to be a global robust procedure. The lack of convergence which has been detected in some simulations is a possible reason for the bad results obtained by this method. Specifically, we fixed the maximum number of iterations at 500 and this number has been attained in 7 of the 100 simulations. We must emphasize at this point that we have implemented the IRPLS method keeping fixed the number of components, whereas Cummins and Andrews (1995) choose the number of components that minimizes R^2_{CV}, and this number is variable.

In figure 4 are presented the distributions of the ratios of the norms of the vectors $\boldsymbol{\beta}$ corresponding to the equations (24) and (25), left and right respectively. The PLSMR method seems to behave well in both cases with values around 1, though there is a larger variability in the comparison with the population vector $\boldsymbol{\beta}$. The other methods, including PLS, have a similar behaviour when they are applied to the contaminated data: the norms of the estimated vectors $\boldsymbol{\beta}$ are a lot smaller than the norms of the *true* vectors $\boldsymbol{\beta}$. The effect is to shrink even more the shrunk vector $\boldsymbol{\beta}$ that is obtained by PLS.

In summary, these previous graphis indicate the better behaviour of

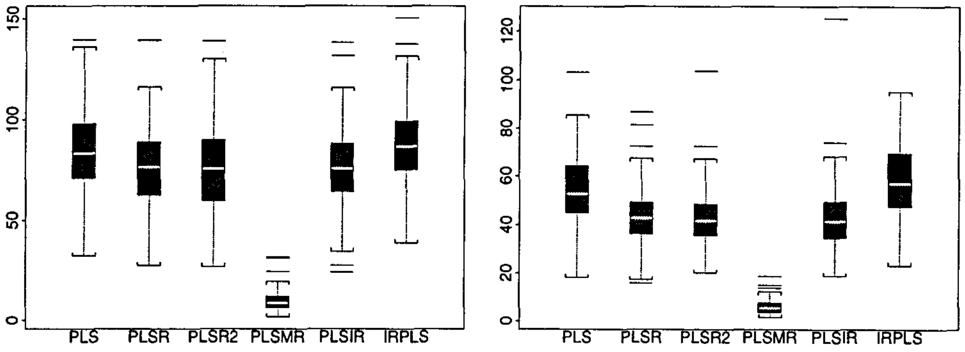

Figure 3: Boxplots of the angles, measured in degrees, between the vectors $\mathbf{w}_3$ (left graphic) and $\boldsymbol{\beta}$ (right graphic) using the contaminated data and the corresponding vectors calculated with the non-contaminated data. The values of the parameters are the following: $N = 100$ simulations; $n = 50$ observations; the percentage of contamination is $\epsilon = 5$ %; and the size of contamination is $M = 3$

$M \setminus \epsilon$	0.5	1	3	5	10	20	30
3	3.36(1.93)	3.53(1.71)	4.64(2.48)	5.25(2.52)	9.80(5.64)	35.43(16.00)	69.40(20.09)
5	2.99(1.34)	3.74(2.13)	4.95(2.83)	4.98(2.60)	9.96(4.72)	38.16(14.62)	70.31(19.88)
10	3.34(1.79)	3.12(1.60)	4.50(2.78)	5.69(3.01)	9.03(4.93)	43.89(21.79)	78.10(20.78)
20	3.13(1.93)	3.54(1.87)	4.51(2.31)	5.71(2.90)	9.90(5.17)	46.05(22.28)	76.36(21.71)

Table 3: Mean and standard deviation (in brackets) of the angles between $\hat{\boldsymbol{\beta}}_{[X,y],PLS}$ and $\hat{\boldsymbol{\beta}}_{[X^c,y^c],PLSMR}$. For each value of the parameters ϵ and M, the number of simulations is $N = 100$ with $n = 50$ observations.

PLSMR for all the comparison measures when we use the values $\epsilon = 5$ % and $M = 3$. The rest of the simulations with different values for ϵ and M have a behaviour very similar to this particular case. Therefore, we are going to concentrate on this method and study its behaviour when we change the values of ϵ and M.

Table 3 presents the mean and the standard deviation of the angles calculated by equation (23) for the different simulations (i.e. different values of ϵ and M). We note that the size of the outliers (M) does not seem to affect the size of the angles, therefore the angles only depend on the percentage ϵ of contamination. The behaviour of PLSMR can be considered very good, which means small angles, at least up to values of $\epsilon = 10$ %. A more comprehensive study would be necessary to determine exactly up to what value of ϵ the method gives acceptable results.

Finally, we have carried out a simulation with non-contaminated data and again we use the same 6 methods. In figure 5 are presented the results

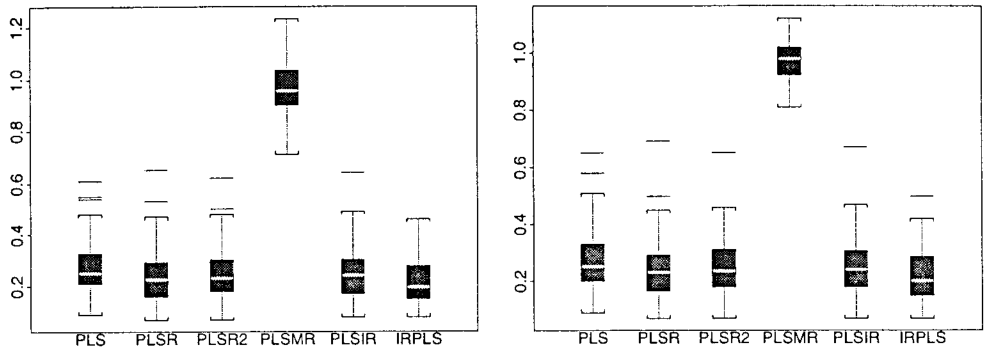

Figure 4: Boxplots of the discrepancy measures $norm\beta_{1,j}$ (left graphic) and $norm\beta_{2,j}$ (right graphic). The values of the parameters are the following: $N = 100$ simulations; $n = 50$ observations; the percentage of contamination is $\epsilon = 5$ %; and the size of contamination is $M = 3$

of this simulation for the measures given by the equations (22) and (23), left and right respectively. The angles for all the robust and semirobust methods are similar, with the exception of PLSR2 which has the worst behaviour.

5 Conclusions and extensions

The semirobust methods, those defined in the literature and the two proposed by us in this article, obtain small improvements over the clasical PLS, but the angles remain very high.

The global robust method PLSMR has a good behaviour with all the comparison measures that have been proposed in this article. In this specific simultation study, PLSMR has a large advantage over the semirobust methods. It would be interesting to repeat this simulation study with other population covariance matrices **V**, with other numbers of variables and observations, etc. It would also be interesting to apply PLSMR to real data and observe its behaviour in this case.

We have used the SDE to calculate the robust covariance matrix, but could be replaced by any other estimator with better properties that might appear in the future.

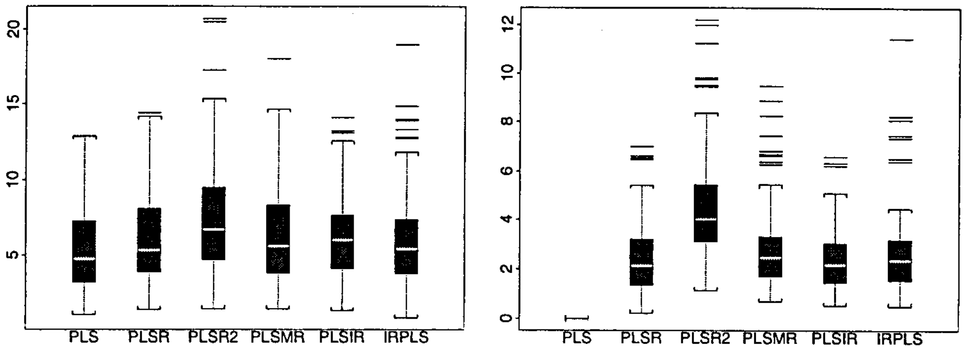

Figure 5: Boxplots of the discrepancy measures $ang\beta_{1,j}$ (left graphic) and $ang\beta_{2,j}$ (right graphic). The values have been calculated with $N = 100$ simulations of non contaminated data. The number of observations is $N = 50$.

Acknowledgment

This research has been supported by DGICYT, Spain through grant PB93-0232. We would like to thank Francisco J. Prieto and Jesus Juan for their comments and programs on the SDE method. We also would like to thank comments suggested by Tom Fearn (University College London) on earlier drafts of this document.

References

Breiman L. and Friedman J.H. (1997) "Predicting Multivariate Responses in Multiple Linear Regression" *Journal of the Royal Statistical Society. Series B* **59**(1), 3–54.

Cummins D.J. and Andrews C.W. (1995) "Iteratively Reweighted Partial Least Squares: a Performance Analysis by Monte Carlo Simulation" *Journal of Chemometrics* **9**, 489-507.

Frank I. and Friedman J. (1993) "A Statistical View of Some Chemometrics Regression Tools" *Technometrics* **35**, 109–148 With discussion.

Griep M., Wakeling I., Vankeerberghen P., and Massart D. (1995) "Comparison of semirobust and robust partial least squares procedures" *Chemometrics and Intelligent Laboratory Systems* **29**, 37–50.

Hardy A.J., MacLaurin P., Haswell S.J., de Jong S., and Vandeginste B.G. (1996) "Double-case diagnostics for outliers identification"

Chemometrics and Intelligent Laboratory Systems **34**, 117–129.

Helland I. (1990) "Partial Least Square Regression and Statistical Models" *Scandinavian Journal of Statistics* **17**, 97–114.

Helland I. and Almøy T. (1994) "Comparison of Prediction Methods When Only a Few Components Are Relevant" *Journal of the American Statistical Association* **89**, 583–591.

Juan J. and Prieto F.J. (1995) "A Subsampling Method for the Computation of Multivariate Estimators with High Breakdown Point" *Computational and Graphical Statistics* **4**(4), 319–334.

Maronna R.A. and Yohai V.J. (1995) "The Behavior of the Stahel-Donoho Robust Multivariate Estimator" *Journal of the American Statistical Association* **90**(429), 330–341.

Næs T. (1985) "Multivariate Calibration when the Error Covariance Matrix is Structured" *Technometrics* **27**(3), 301–311.

Stone M. and Brooks R. (1990) "Continuum Regression: Cross-validated Sequencial Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression (with discussion)" *Journal of the Royal Statistical Society, Series B* **52**, 237–269.

Wakeling I. and Macfie H. (1992) "A Robust PLS Procedure" *Journal of Chemometrics* **6**, 189–198.