UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 07-13
Statistic and Econometric Series 04
March 2007

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

A ROBUST PARTIAL LEAST SQUARES METHOD WITH APPLICATIONS*

Javier González[1], Daniel Peña [2] and Rosario Romera[3]

Abstract

Partial least squares regression (PLS) is a linear regression technique developed to relate many regressors to one or several response variables. Robust methods are introduced to reduce or remove the effect of outlying data points. In this paper we show that if the sample covariance matrix is properly robustified further robustification of the linear regression steps of the PLS algorithm becomes unnecessary. The robust estimate of the covariance matrix is computed by searching for outliers in univariate projections of the data on a combination of random directions (Stahel-Donoho) and specific directions obtained by maximizing and minimizing the kurtosis coefficient of the projected data, as proposed by Peña and Prieto (2006). It is shown that this procedure is fast to apply and provides better results than other procedures proposed in the literature. Its performance is illustrated by Monte Carlo and by an example, where the algorithm is able to show features of the data which were undetected by previous methods.

Keywords: Kurtosis, projections, robust covariance matrix, Stahel-Donoho estimator.

*

[1, 2, 3] Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe, Madrid, Spain.

This research has been supported by Spanish MEC grant SEJ2004-03303.

A Robust Partial Least Squares Regression Method with Applications

Javier González, Daniel Peña y Rosario Romera
Departamento de Estadística. Universidad Carlos III de Madrid.

March 21, 2007

Abstract

Partial least squares regression (PLS) is a linear regression technique developed to relate many regressors to one or several response variables. Robust methods are introduced to reduce or remove the effect of outlying data points. In this paper we show that if the sample covariance matrix is properly robustified further robustification of the linear regression steps of the PLS algorithm becomes unnecessary. The robust estimate of the covariance matrix is computed by searching for outliers in univariate projections of the data on a combination of random directions (Stahel-Donoho) and specific directions obtained by maximizing and minimizing the kurtosis coefficient of the projected data, as proposed by Peña and Prieto (2006). It is shown that this procedure is fast to apply and provides better results than other procedures proposed in the literature. Its performance is illustrated by Monte Carlo and by an example, where the algorithm is able to show features of the data which were undetected by previous methods.

Key words: Kurtosis, projections, robust covariance matrix, Stahel-Donoho estimator.

1 Introduction

Partial least squares (PLS) is a useful procedure for relating a set of responses to many explanatory variables. It can be seen as a general dimension reduction technique which takes into account the linear relationship between the responses and the regressors. When the responses are dummy variables, as in linear discriminant problems, PLS has been found to work very well for classification (see Ngueyen and Rocke, 2002 and Barker ad Rayens, 2003). However, it is well known that the popular algorithms for PLS regression (NIPALS and SIMPLS) are very sensitive to outliers in the data set. For univariate or multivariate response several robustified versions have already been proposed. Wakeling and Macfie (1992) worked with the PLS with multivariate response (which will be called PLS2) and their idea was to replace the set of regressions involved in the

standard PLS2 algorithm by M estimates based on weighted regressions. Griep et al. (1995) compared least median of squares (LMS), Siegel's repeated median (RM) and iterative reweighted least squares (IRLS) for PLS with univariate response (PLS1 algorithm), but these methods are not resistant to high leverage outliers. Procedures combining robust covariance matrices and robust regression methods have been proposed by Gil and Romera (1998) and Hubert and Branden (2003). For an extensive review of the commonly used multivariate regression methods that have appeared since 1996 in the field of Chemometrics see Moller at al. (2005). There is a special emphasis on the robust versions of PCA and PLS.

In this paper we show that if the sample covariance matrix is properly robustified the PLS algorithm will be robust and therefore, further robustification of the linear regression steps of the PLS algorithm is unnecessary. We present a procedure which applies the standard PLS algorithm to a robust covariance matrix. The covariance matrix is estimated by projecting the data in some directions, finding outliers on this directions, deleting them from the sample and using the clean data to compute the covariance matrix. The key ingredient in this approach is selecting the projecting directions. A popular way to generate directions is at random, as in the Stahel-Donoho robust multivariate estimator (SDE), and this was the procedure used in the robust PLS algorithm by Gil and Romera (1998). See Maronna and Yohai (1995) for the theoretical behavior of the SDE and practical recommendations for its implementation. However, as shown in Peña and Prieto (2001) the SDE fails with concentrated contamination if p is large, as it is usually the case in PLS. Peña and Prieto (2006) proposed to search for outliers in multivariate data by using a combination of random directions, obtained in a much more efficient way than in the usual SDE method, and specific directions, obtained by maximizing and minimizing the kurtosis coefficient of the projected data. The combination of both types of directions leads to a procedure with useful theoretical properties and good performance: it is affine equivariant, inherits the good theoretical properties of the SDE, inherits the good properties for finding high leverage concentrated outliers of the kurtosis procedure and it is fast to compute, so that it can be applied for large data sets. In this paper we adapt this algorithm for robust PLS estimations.

As Moller et al. (2005) pointed out a theoretical high breakdown point might be helpful for initially assessing the data quality and detecting outliers. Nevertheless, usually there is a price to pay for such high robustness: high computational complexity or low statistical efficiency and rate convergence. However, the proposed robust PLS method is fast and, as shown in the Monte Carlo results, is able to deal with a highly contaminated data.

The rest of the paper is organized as follows. Section 2 reviews briefly the PLS algorithm for a one-dimensional response variable and analyses the implication of the robustification of the covariance matrix for the regression steps. Section 3 presents the new procedure. Section 4 reports a Monte Carlo experiment where the performance of the new method is compared to other robust procedures. Section 5 illustrates the performance of the proposed method

in a well known set of data where we show that the present algorithm is able to find outliers that were undetected by previous methods.

2 Robust PLS Methods

2.1 The PLS algorithm

Suppose that we have a sample of size n of a $1+p$ dimensional vector $z = (y, x)^T$, which can be decomposed as a set of p regressors, x, and a response variable y. Let S_z be the sample covariance matrix of z, consisting of the elements

$$S_z = \begin{pmatrix} s_y^2 & S_{y,x}^T \\ S_{y,x} & S_x \end{pmatrix}, \tag{1}$$

where $S_{y,x}$ is the $p \times 1$ vector of covariances between y and the x variables. We are interested in estimating the linear regression $\widehat{y} = \widehat{\beta}^T x$, and we assume that the response can be linearly explained by a set of a factors $t = (t_1, ..., t_a)$, with $a \ll p$, which are linear functions of the x variables. More precisely, calling X to the $n \times p$ data matrix of the regressors, and x_i^T to its ith row, the following model holds

$$x_i = Pt(i) + \epsilon_i \tag{2}$$
$$y_i = q^T t(i) + \eta_i, \tag{3}$$

where P is the $p \times a$ matrix of the loadings of the vector $t(i) = (t_1(i), ..., t_a(i))^T$ and q is the a-dimensional vector of the y-loadings. The vectors ϵ_i and η_i have normal distribution and are uncorrelated. Then it can be shown (Helland, 1992) that the maximum likelihood estimation of the matrix of components, $t = (t_1, ..., t_a)$ is given by

$$t = XW_a \tag{4}$$

where $W_a = [w_1, w_2, ..., w_a]$ is the $p \times a$ matrix of coefficients and the vectors w_i verify $w_i^T w_i = 1$ and $w_i^T S_x w_j = 0$ for $i \neq j$, so that the $n \times 1$ variables t_i are orthogonal. It can be shown that these vectors are the solution of

$$w_{j+1} = \underset{w^T w=1 \text{ and } w^T S_x w_i=0 \text{ for } i=1,...,j}{arg\max}(cov^2(Xw, y))$$

and are found as the eigenvectors linked to the largest eigenvalues of the matrix

$$(I - P_x(j))S_{y,x}S_{y,x}^T$$

where $P_x(j)$ is the projection matrix on the space spanned by $S_x W_j$, given by $P_x(j) = (S_x W_j)\left[(S_x W_j)^T (S_x W_j)\right]^{-1} (S_x W_j)^T$. From these results it is easy to see that the vectors w_i can be computed recursively as

$$w_1 \propto S_{y,x} \tag{5}$$
$$w_{j+1} \propto S_{y,x} - S_x W_j (W_j^T S_x W_j)^{-1} W_j^T S_{y,x}. \tag{6}$$

The selected number of PLS components,a, is usually estimated by *leave-one-out* cross-validation methods. Note that by using the expressions given by (5) and (6), it is not necessary to calculate the PLS components t_j. In each step of the algorithm, w_{j+1} only depends on the value of the j previous vectors $w_1, w_2, ..., w_j$, on S_x and on $S_{y,x}$. Moreover, as w_1 only depends on $S_{y,x}$, the calculation of W is completely fixed by the values of S_x and $S_{y,x}$. Finally, as the regression coefficients in (3) are uncorrelated, due to the uncorrelation of the t variables, it is easy to see that the regression coefficients $\widehat{\beta}_a^{PLS}$ are given by

$$\widehat{\beta}_a^{PLS} = W_a(W_a^T S_x W_a)^{-1} W_a^T S_{y,x} \tag{7}$$

The application of this algorithm can be seen as a two step procedure: (1) the weights w_j, that define the new orthogonal regressor t_j, are computed with (5) and (6) by using the covariance matrix of the observations; (2) the regression coefficients q_j are computed from a simple regression between the response, y, and the regressor, t_j. Thus, several authors (see Gil and Romera (1998) and Hubert and Branden (2003)) have proposed a two steps robustification. First a robust covariance matrix is computed in order to obtain robust weights and, second, a robust regression estimate, usually an M-estimate, is computed to obtain the robust q weights. However, as is shown in (7), these two steps depend only on the covariance matrix of the observations and we may think that if this matrix is properly robustified the procedure will be robust. In the next section we discuss this intuition.

2.2 Multivariate versus regression outliers in PLS

Given the sample $z_i = (y_i, x_i)^T$ for $i = 1, ..., n$ of a $1+p$ dimensional multinormal vector with covariance matrix (1), let us analyze the result of detecting outliers with respect to the joint distribution, which we will call multivariate outliers, versus the detection of outliers with respect to the conditional distribution of $y|t$, which we will call regression outliers. Let us assume without loss of generality that all the variables have zero mean. Then, observation z_i will be considered as a multivariate outlier if it verifies

$$z_i^T S_z^{-1} z_i > c_1 \tag{8}$$

where c_1 is some percentile of the Chi-square distribution with $1 + p$ degrees of freedom. In this expression we are assuming that $n > 1 + p$, so that the covariance matrix is not singular. The case $1 + p > n$ will be considered later.

On the other hand, the regression outliers will be found as extreme observations in the estimated distributions of $y|t_j$, as the t_j are orthogonal. That is, an observation will be an outlier in the regression $\widehat{y}(t_j) = \widehat{q}_j t_j$ if it has high leverage and relatively large residual. We show that by deleting observations which are outliers with respect to the multivariate distribution: (1) we cannot have high leverage outliers in the PLS regressions, and (2) the squared standardized residuals $(y_i - \widehat{y}_i(t_j))^2/s_j^2$ from the regression of y on the variable t_j with residual variance s_j^2 are bounded.

It is easy to see from (1) that

$$S_z^{-1} = \begin{pmatrix} s_e^{-2} & -s_e^{-2}\widehat{\beta}^T \\ -\widehat{\beta}s_e^{-2} & S_x^{-1} + \widehat{\beta}s_e^{-2}\widehat{\beta}^T \end{pmatrix}$$

where $\widehat{\beta} = S_x^{-1}S_{y,x}$, $\widehat{y} = X\widehat{\beta}$ estimates $E(y|X)$, and $s_e^2 = (y-\widehat{y})^T(y-\widehat{y})/n$ estimates $\mathrm{var}(y|X)$. Then we can write (8) as

$$\frac{(y_i - \widehat{y}_i)^2}{s_e^2} + v_i > c_1 \tag{9}$$

where $v_i = x_i^T S_x^{-1} x_i$. We will show that this equation indicates that by dropping multivariate outliers with (8) the data in the regression have the property that both the size of the least squares (LS) residuals $|y_i - \widehat{y}_i|$ and the leverage of the x variables are bounded. Note that points with large values of v_i will be deleted as multivariate outliers, and thus the leverage of the regressor $t_j = Xw_j$ is bounded. The leverage of $t_j(i) = w_j' x_i$ will be given by

$$l_i = \frac{w_j' x_i x_i' w_j}{w_j'(X'X)w_j}$$

and it is straightforward to show that the maximum value of this leverage is achieved for $w_j = k(X'X)^{-1}x_i$, which leads to a leverage $l_i = \frac{1}{n}v_i$. As observations with v_i large are deleted by (8) the maximum leverage of the regressor is bounded.

Next, as the t_j regressor is included in the space generated by the columns of X

$$\sum (y_i - \widehat{y}_i(t_j))^2 = \sum (y_i - \widehat{y}_i)^2 + \sum (\widehat{y}_i - \widehat{y}_i(t_j))^2$$

and calling s_j^2 to the residual variance in the regression $\widehat{y}(t_j)$ we have that $s_e^2 \leq s_j^2$. Also, we have that the standardized residuals of the regression of the response on t_j can be expressed as

$$\frac{(y_i - \widehat{y}_i(t_j))^2}{s_j^2} = \frac{(y_i - \widehat{y}_i)^2}{s_j^2} + \frac{(x_i'(\widehat{\beta} - \widehat{q}_j w_j))^2}{s_j^2} + 2\frac{(y_i - \widehat{y}_i)}{s_j}\frac{x_i'(\widehat{\beta} - \widehat{q}_j w_j)}{s_j}$$

and it is easy to see that they are bounded. The first term is bounded by (9) and the result $s_e^2 \leq s_j^2$, and the second by the bound on the leverage of the x points coming from (9). Thus the standardized residuals of the regression of the response on t_j are bounded and cannot be very large. Note that this applies to the regression of the response on one or several t_j variables, because when we include all the relevant ones to explain the response the predictive value will be close to the LS fit and the residuals are obviously bounded by (9).

Finally, suppose the case $1 + p > n$. Then $S_z = \sum_{i=1}^{r} \lambda_i u_i u_i^T$, where $r = \min(n, 1+p)$, and instead of S_z^{-1}, which does not exits, we can use the generalized

inverse $S_z^- = \sum_{i=1}^r \lambda_i^{-1} u_i u_i^T$. The Mahalanobis distances are given by

$$\sum_{i=1}^{r} \lambda_i^{-1} (z_i^T u_i)^2$$

and the main argument of the analysis is the same, but now applied to the principal components of the data.

3 The proposed algorithm

We have shown in the previous section that if we delete multivariate outliers we will not have influential points or outliers in the regression between the response and the factors. The algorithm we propose is designed to obtain a robust covariance matrix free from multivariate outliers and works as follows. Without loss of generality we assume that the original data have zero mean and covariance matrix S_z. The points are transformed by using

$$\widetilde{z}_i = S_z^{-1/2} z_i, \quad i = 1, \ldots, n. \tag{10}$$

and when the covariance matrix is singular we will use the generalized inverse, as defined in the previous section. The algorithm has three steps. In the first, two specific directions are generated by maximizing and minimizing the kurtosis coefficient of the projections and a univariate search for outliers is done in these directions. In the second, random directions are generated by following the procedure presented in Peña and Prieto (2006) of stratified sampling and again outliers are identified. In the third, all the suspicious observations are tentatively deleted from the sample and the mean and covariance matrix of the remaining data is computed. Then by using the Mahalanobis distance all the suspicious observations are checked. The points considered as outliers are deleted from the sample and the three steps of the procedure are now again applied to the new cleaned sample until no more outliers are found. We explain next the details of these steps

Step I: Compute the directions which maximizes and minimizes the kurtosis coefficient of the projection and also the normalized univariate distances of the data in these two directions. The first direction is obtained as the solution of the problem

$$\begin{aligned} d_1 = \quad & \arg\max_d \quad \frac{1}{n} \sum_{i=1}^{n} (d' \widetilde{z}_i)^4 \\ & \text{s.t.} \qquad\quad d'd = 1. \end{aligned} \tag{11}$$

The same process is applied to the computation of the direction minimizing the kurtosis coefficient. Let d_2 be this second direction and let $p_i^{(j)} = d_j' \widetilde{z}_i$ be the projected values on these two directions, $j = 1, 2$. The normalized univariate distances $r_i^{(j)}$ for these projected values are computed as

$$r_i^{(j)} = \frac{1}{\beta_p} \frac{|p_i^{(j)} - \text{median}_i(p_i^{(j)})|}{\text{MAD}_i(p_i^{(j)})}, \qquad j = 1, 2, \tag{12}$$

β_p is a predefined reference value which depends on the dimension p and is obtained by Monte Carlo to ensure a type I error equal to 0.05. Note that if $n < 1+p$ we consider S_z^- instead of S_z and them $\dim(S_z^-) = \min(n, 1+p)$ replaces dimension p. This step will identify outliers forming clusters as observations with large $r_i^{(j)}$, as shown by Peña and Prieto (2001). If the size of the group of outliers is small (roughly smaller than 20%) the useful direction for identifying the outliers will be d_1, whereas if the size is large will be d_2.

Step II: Compute random directions from a stratified sampling procedure. Then, search for outliers in these directions. These random directions are generated by the procedure proposed by Peña and Prieto (2006) which is much more efficient to detect outliers than the standard one. Each direction is generated in two stages. In the first one two observations are chosen randomly from the sample, the direction defined by these two observations is computed, and the observations are then projected onto this direction. This is repeated for $l = 1, .., L$. For each l, the second stage builds a set of K stratified samples as follows. The projections are ordered and partitioned into K intervals of size n/K, where K is a prespecified number. From each of these k intervals, $1 \leq k \leq K$, a subsample of p observations is chosen without replacement, and the direction, $\widetilde{d}_j$, orthogonal to the hyperplane generated by these p observations is computed. This direction $\widetilde{d}_j$ is now used to search for outliers, as in Step I. The corresponding projections $\widetilde{p}_i^{(j)} = \widetilde{d}_j \widetilde{z}_i$ provide the normalized univariate distances $\widetilde{r}_i^{(j)}$,

$$\widetilde{r}_i^{(j)} = \frac{1}{\beta_p} \frac{|\widetilde{p}_i^{(j)} - \text{median}_i(\widetilde{p}_i^{(j)})|}{\text{MAD}_i(\widetilde{p}_i^{(j)})}, \qquad j = 1, ..., LK. \tag{13}$$

Step III: For each observation i its corresponding normalized outlyingness measure r_i is obtained as:

$$r_i = \max \left\{ r_i^{(1)}, r_i^{(2)}, \widetilde{r}_i^{(1)}, ..., \widetilde{r}_i^{(LK)} \right\}.$$

Those observations having values $r_i > 1$ are labeled as outliers and, if their number is smaller than $n - \lfloor (n+1+p)/2 \rfloor$, removed from the sample. Otherwise, only those $n - \lfloor (n+1+p)/2 \rfloor$ observations having the largest values of r_i are labeled as outliers.

Finally, let U denote the set of all observations not labeled as outliers. The algorithm computes the Mahalanobis distance of the original observations labeled as outliers with respect to the good observations as follows:

$$\begin{aligned} \tilde{m} &= \frac{1}{|U|} \sum_{i \in U} z_i, \\ \tilde{S}_z &= \frac{1}{|U|-1} \sum_{i \in U} (z_i - \tilde{m})(z_i - \tilde{m})', \\ \widetilde{v}_i &= (z_i - \tilde{m})^T \tilde{S}_z^{-1} (z_i - \tilde{m}), \qquad \forall i \notin U. \end{aligned}$$

Those observations $i \notin U$ such that $\widetilde{v}_i < \chi^2_{p-1,0.99}$ are considered not to be outliers, and are included in U. When $\tilde{S}_z^{-1}$ does not exit the Mahalanobis distances are computed by using the generalized inverse.

The three steps are repeated until no more outliers are found (or U becomes the set of all observations). Note that the final covariance matrix does not need to be scaled for consistency, as in Peña and Prieto (2006), because the PLS coefficients and the directions obtained are invariant to scale changes.

The preceding algorithm includes a certain number of parameters. The values assigned to them in the implementation have been chosen as recommended by Peña and Prieto (2006) to ensure adequate theoretical and efficiency properties. The parameter β_p is chosen to ensure a reasonable level of Type I errors, and depends on the sample space dimension p. Table 1, taken from Peña and Prieto (2006) shows the values used for several sample space dimensions. The values for other dimensions could be obtained by interpolating $\log \beta_p$ linearly in $\log p$.

Table 1: Cutoff values for univariate projections for steps I and II

Sample space dimension p	5	10	20
Cutoff value β_p	3.46	3.86	4.67

In step II, the number of intervals, K, is chosen as $K = 3$ or 5, depending on p and L is equal to $10p$.

The proposed robust PLS algorithm is based on the robust covariance matrix $\tilde{S}_z$. We take w_1 as the normalization of the vector $\widetilde{S}_{y,x}$, the first column of $\tilde{S}_z$ dropping the first element, as defined in (1), and the succeeding values of w_j are also calculated in a robust form by using

$$w_{j+1} \propto \widetilde{S}_{y,x} - \tilde{S}_x W_j (W_j^T \tilde{S}_x W_j)^{-1} W_j^T \widetilde{S}_{y,x}, \qquad 1 < j \leq a.$$

Note that the proposed algorithm includes some modifications of previous algorithms: (1) when $p > n$ and the covariance does not exist we use the generalized inverse to compute the robust Mahalanobis distance and to clean from outliers; (2) we do not require computing scale factors to make consistent the estimated robust covariance matrix as the proposed PLS procedure is invariant to scale changes.

4 Numerical Results

4.1 Monte Carlo experiments study

We have performed several Monte Carlo simulations to compare the performance of the proposed method to other robust algorithm. The initial model used to generate the data is:

$$\begin{aligned} t &\sim N_a(0_a, \Sigma_t) \\ x &= I_{p,a} t + N_p(0_p, 0.1 I_p), \qquad p > a \\ y &= q^T t + N(0,1) \end{aligned}$$

Table 2: Simulation study

n	p	a	Σ_t	ϵ
100	5	2	diag(4,2)	0.1 and 0.3

where $(I_{p,a})_{i,j} = 1$ for $i = j$ and $(I_{p,a})_{i,j} = 0$ otherwise, q is a p-dimensional vector of ones, I_p is the $p \times p$-dimensional identity matrix. We assume a as known, two contamination levels, ϵ, equal to 0.10 and 0.30, and four types of contamination:

1. Bad leverage contamination:

$$\begin{aligned} x_\epsilon &= I_{p,a} t_\epsilon + N_p(0_p, 0.1 I_p) \\ t_\epsilon &\sim N_a(10_a, \Sigma_t) \end{aligned}$$

2. Vertical outliers :

$$y_\epsilon = q' t + N(10, 0.1)$$

3. Orthogonal outliers:

$$x_\epsilon = I_{a,p} t_\epsilon + N_p((0_a, 10_{p-a}), 0.1 I_p)$$

4. Concentrated contamination:

$$\begin{aligned} x_\epsilon &= I_{a,p} t_\epsilon + N_p(10_p, 0.001 I_p) \\ t_\epsilon &\sim N_a(10_a, \Sigma_t) \end{aligned}$$

For each contamination level ϵ we have generated sets of $100(1-\epsilon)$ observations from the previous model, and we have added 100ϵ additional observations generated from one of these contamination models. Three robust procedures are compared in this study to the standard PLS algorithm. The first, PLS-KurSD, is the one proposed in this paper. The second, PLS-SD, is the one proposed by Gil and Romera (1998). The third, RSIMPLS, is the algorithm proposed by Hubert and Branden (2003). The comparison is made by using three classical regression measures. The first two are based on the angle between the true parameter vector β and the estimated vector $\hat{\beta}_{[y_\epsilon, X_\epsilon], a}$. After $m = 1000$ replications we evaluate the $Mean(angle)$ and the $MSE_a(\hat{\beta}) = \frac{1}{m} \sum_{l=1}^{m} \|\hat{\beta}_a^{(l)} - \beta\|^2 = Norm(\beta)$, where $\|\|$ denotes Euclidean norm and $\hat{\beta}_a^{(l)}$ is the estimated regression vector in the replication l. In addition, we compute a third measure $MSE_a = \sqrt{\frac{1}{n_t} \sum_{i=1}^{n_t} (y_i - \hat{y}_{i,a})^2}$ with $n_t = 50$ observations generated from the initial model and where $\hat{y}_{i,a}$ is the predicted value for the observation i.

Table 3: Estimation results in 1000 replications. The sample size is n=100, p=5 and the proportion of outliers is 10% .

Algorithm	PLS	PLS-SD	PLS-KurSD	RSIMPLS
No Contamination				
Mean(Angle)	0.06(0.03)	0.07(0.03)	0.07(0.03)	0.08(0.03)
Norm(β)	0.01(0.01)	0.01(0.01)	0.01(0.01)	0.01(0.01)
MSE_a	0.16(0.08)	0.17(0.09)	0.17(0.09)	0.17(0.09)
10% Bad leverage points				
Mean(Angle)	1.13(0.22)	0.11(0.06)	0.07(0.03)	0.08(0.03)
Norm(β)	1.23(0.15)	0.07(0.04)	0.01(0.01)	0.02(0.01)
MSE_a	2.07(0.23)	0.48(0.16)	0.18(0.10)	0.18(0.09)
10% Vertical outliers				
Mean(Angle)	1.14(0.21)	0.11(0.06)	0.07(0.03)	0.08(0.03)
Norm(β)	1.23(0.14)	0.07(0.05)	0.02(0.01)	0.02(0.01)
MSE_a	2.08(0.24)	0.47(0.17)	0.18(0.10)	0.18(0.10)
10% Orthogonal outliers				
Mean(Angle)	1.13(0.21)	0.11(0.06)	0.07(0.04)	0.08(0.03)
Norm(β)	1.22(0.15)	0.07(0.04)	0.02(0.01)	0.02(0.01)
$\text{MSE}(\sigma_e)$	2.06(0.22)	0.48(0.16)	0.18(0.10)	0.18(0.10)
10% Concentrated outliers				
Mean(Angle)	1.14(0.21)	0.11(0.06)	0.08(0.04)	0.08(0.04)
Norm(β)	1.23(0.14)	0.08(0.04)	0.02(0.06)	0.02(0.02)
MSE_a	2.08(0.23)	0.48(0.16)	0.19(0.10)	0.19(0.09)

Table 4: Estimation results in 1000 replications. The sample size is n=100, p=5 and the proportion of outliers is 30%.

Algorithm	PLS	PLS-SD	PLS-KurSD	RSIMPLS
No Contamination				
Mean(Angle)	0.06(0.03)	0.07(0.03)	0.07(0.03)	0.08(0.03)
Norm(β)	0.01(0.01)	0.01(0.01)	0.01(0.01)	0.02(0.01)
MSE_a	0.16(0.08)	0.18(0.09)	0.18(0.09)	0.18(0.09)
30% Bad leverage points				
Mean(Angle)	1.36(0.18)	0.61(0.21)	0.10(0.10)	1.29(0.26)
Norm(β)	1.39(0.13)	0.75(0.20)	0.04(0.11)	1.37(0.22)
MSE_a	2.23(0.24)	1.58(0.22)	0.24(0.22)	2.19(0.25)
30% Vertical outliers				
Mean(Angle)	1.36(0.19)	0.62(0.21)	0.11(0.12)	1.30(0.27)
Norm(β)	1.40(0.14)	0.75(0.19)	0.04(0.13)	1.37(0.19)
MSE_a	2.25(0.24)	1.58(0.22)	0.26(0.27)	2.20(0.26)
30% Orthogonal outliers				
Mean(Angle)	1.36(0.17)	0.61(0.21)	0.10(0.11)	1.31(0.25)
Norm(β)	1.40(0.16)	0.75(0.19)	0.04(0.13)	1.37(0.17)
$\text{MSE}(\sigma_e)$	2.26(0.24)	1.59(0.22)	0.25(0.23)	2.22(0.26)
30% Concentrated outliers				
Mean(Angle)	1.36(0.18)	0.61(0.20)	0.10(0.10)	1.29(0.26)
Norm(β)	1.39(0.21)	0.74(0.20)	0.04(0.11)	1.37(0.20)
MSE_a	2.26(0.23)	1.59(0.21)	0.24(0.23)	2.21(0.24)

Table 5: Some descriptive statistics for Fish data

Regressors	Mean	Median	Std	$Corr\ Coef\ (Wi, y)$
W1	1.55	1.49	0.16	0.63
W2	1.43	1.37	0.16	0.62
W3	1.24	1.18	0.20	0.62
W4	1.20	1.15	0.19	0.63
W5	1.00	0.95	0.16	0.65
W6	0.85	0.82	0.15	0.71
W7	0.91	0.87	0.16	0.71
W8	0.94	0.90	0.16	0.70
W9	0.55	0.53	0.09	0.73

The values for the parameters chosen in our experiment are given in Table 2. Tables 3 and 4 report averages and standard deviations of the three performance measures for the four algorithms compared. Both tables show in the first four rows, corresponding to the no contamination case, that the three robust algorithms are slightly worse than PLS when the data is not contaminated. The loss of efficiency is very small, and similar in the three robust algorithms. However, in the case of contamination the improvement over PLS is high. Table 3 shows that with small percentage of contamination the RSIMPLS and PLSKurSD methods improve the one proposed by Gil and Romera (1998) and Table 4 shows that when the amount of contamination increases (30%) both PLS-SD and RSIMPLS breakdown. Note that the proposed method provides robust estimates even for concentrated contamination, which is a very difficult case in outlier detection. From the computational point of view PLSKurSD is much faster than RSIMPLS, as shown in Table 9 where computational times for a real data example are presented. All codes compared were implemented in Matlab. The algorithm RSIMPLS was taken from library for robust analysis LIBRA(see http://wis.kuleuven.be/stat/robust/LIBRA.html).

5 Real Data Study

In this section we present an application in which PLS can be useful in analyzing a set of data. This data was primarily introduced by Naes (1985) and it has been analyzed by Gil and Romera (1985) and Hubert and Vanden Branden (2003). It contains observations on 45 samples of Fish (rainbow trout). For each sample, fat concentration is determined. The spectra are obtained on a NIR instrument and consist of nine wavelengths. The objective of the analysis is to search for the relation between fat concentration and these spectra. Table 5 presents some descriptive statistics for this data set.

In order to select the number of PLS components we use a *leave-one-out* cross-validation method. Let $\hat{y}_{(i),a}$ be the estimate of y_i which comes from a

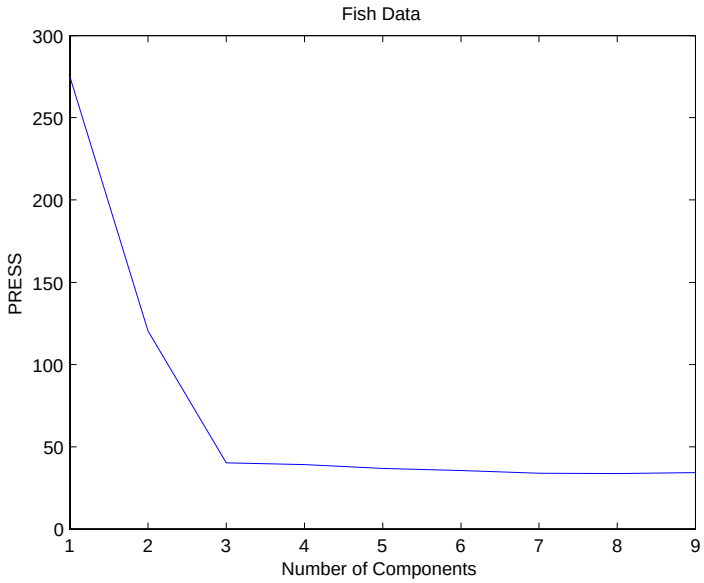

Figure 1: PRESS values for the PLS algorithm, $a = 1, ..., 10$

PLS regression with a components when we have eliminated the ith observation. The value of the predictive residual error sum of squares (PRESS) is

$$PRESS(a) = \sum_{i=1}^{n} \frac{(y_i - (\hat{y}_{(i),a})_i)^2}{n} \tag{14}$$

The PRESS values decrease very slowly from $a = 3$ onwards, as it is shown in Figure 1. The same result has been shown by Gil and Romera (1998) and Hardy et al (1996).

It was reported by Naes (1985) that observations 39-45 are outlying. We have verified that in fact these observations are multivariate outliers using the Mahalanobis distance with respect to the 1-38 observations. We have applied the proposed PLSKurSD procedure to the Fish data and we have found 17 suspicious outlying observations: 1, 3, 10, 12, 16, 17, 18, 27, 30, 35, 37, 39, 40, 41, 43, 44 and 45. The corresponding 17 squared standardized Mahalanobis distances (in the log-scale) with respect to the good observations, according to the PLSKurSD procedure, are plotted in Figure 2. In this figure we have also plotted the two standard χ^2-percentiles (in logarithms) which are usually selected as cutoff values. We observe that all the scores exceed by far the cut-off values $\log(\chi^2_{10;0.95}) = 2.907$ and $\log(\chi^2_{10;0.99}) = 3.144$.

Table 6 compares different regression estimates for this data set. The first column is the ordinary least squares regression estimate, (OLS), the second one is the standard PLS algorithm, the third is the RSIMPLS robust algorithm

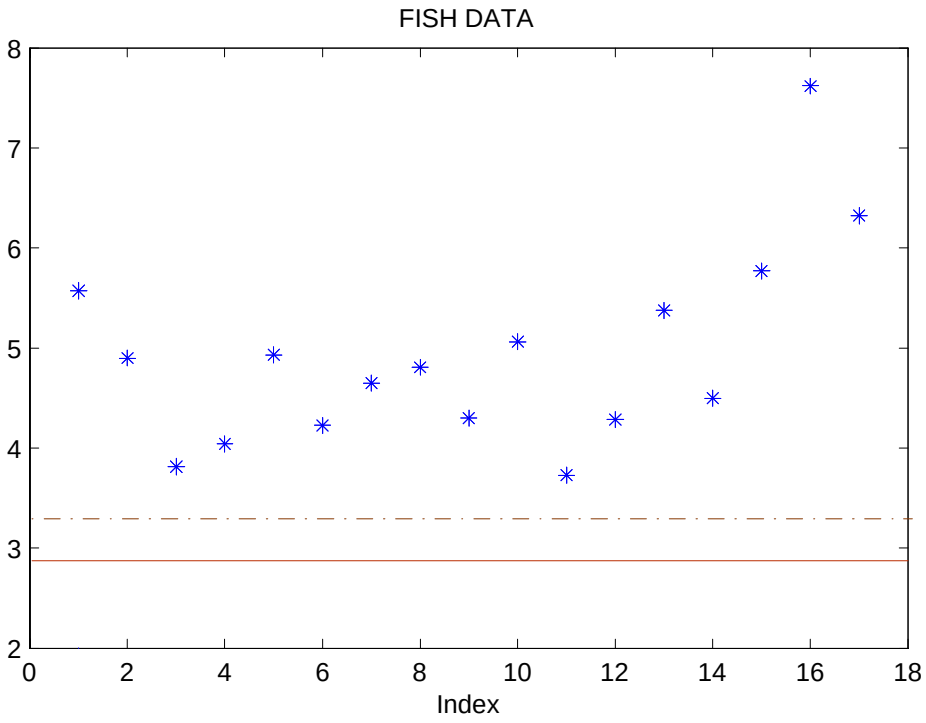

Figure 2: Squared Standardized Mahalanobis distances of the 17 outliers found by PLSKurSD in the log-scale. The cut-off values $\log(\chi^2_{10;0.95}) = 2.907$ and $\log(\chi^2_{10;0.99}) = 3.144$ are represented by the horizontal lines.

Table 6: Regression coefficients with four procedures for FISH data. The * in the OLS regression indicates a t statistic larger than 2.

	OLS	PLS	RSIMPLS	PLSKurSD
W1	-186.37	149.23	64.7557	55.1565
W2	425.81	60.491	-18.5719	-17.4461
W3	-61.52	-107.91	-88.0727	-82.0153
W4	-376.54	-114.25	-67.9800	-63.1873
W5	-33.46	-97.353	-47.6929	-63.2612
W6	-1188.83(*)	-25.934	52.5208	57.8444
W7	-20.08	-29.735	65.3958	71.0942
W8	1223.78(*)	-0.5704	55.7535	62.9704
W9	303.24(*)	258.93	9.9509	3.7658

proposed by Hubert and Vanden Branden (2003) and the last one is the proposed algorithm, which is called PLSKurSD. The (*) in the OLS regression indicates a t statistic larger than 2. Both robust regression methods give similar weights to the same wavelengths. The differences between the OLS and the robust estimates are remarkable as can be expected by the large number of outliers found by the robust methods.

Our procedure found a larger set of outliers, 17, than the RSIMPLS algorithm, which finds 13. As we have shown in Figure 2 these 17 observations are clearly outliers with respect to the rest of the data. It is interesting to show that our procedure leads to a better fit to the data. Table 7 reports the standard deviation and the MAD (median of the absolute deviations to the median) of the regression residuals in the set of clean data and in the set of outliers found by the two robust procedures. It can be seen that the PLSKurSD leads to a better fit. We have also compared the correlation structure of the regressors in the two groups and Table 8 gives the correlations between the response and the spectra in the group of clean data and in the group of outliers for the two robust procedures. It can be seen that the differences between the correlation in the group of clean data and the group of outliers are larger with the proposed procedure PLSKurSD. For instance, the euclidean distance between columns 2 and 3 in Table 8 is 0.1025 whereas the euclidean distance between the columns 4 and 5 is 0.2764. This result suggests that PLSKurSD finds a group of outliers with a correlation structure different from the one in the group of good observations. This effect is less clear with the RSIMPLS procedure.

Table 9 presents the computational times for the three PLS methodologies used in our Fish data analysis. The proposed PLSKurSD algorithm is remarkably much faster than the RSIMPLS.

Figure 3 plots the two groups of observations in the plane generated by the two first PLS components. Most of the data correponding to the group of outliers appear as extremes in this plot. Figure 4 shows the projection of the

Table 7: Dispersion measures (Standard deviation and MAD) in four cases: the clean group according to RSIMPLS$_g$; the group of outliers found by RSIMPLS$_o$; the clean group according to PLSKurSD$_g$; and the group of outliers found by PLSKurSD$_o$

	RSIMPLS$_g$	RSIMPLS$_o$	PLSKURSD$_g$	PLSKURSD$_o$
STD	0.8175	2.0965	0.6231	2.5152
MAD	0.6518	1.5318	0.5782	1.6286

Table 8: Correlation coefficients between the response and the spectra in five cases: All the data (Total); the clean group according RSIMPLS$_g$; the group of outliers found by RSIMPLS$_o$; the clean group according to PLSKurSD$_g$; and the group of outliers found by PLSKurSD$_o$

	TOTAL	RSIMPLS$_g$	RSIMPLS$_o$	PLSKURSD$_g$	PLSKURSD$_o$
W1	0.6357	0.7304	0.6022	0.7499	0.5200
W2	0.6246	0.7077	0.5951	0.7188	0.5127
W3	0.6198	0.6918	0.5937	0.6865	0.5134
W4	0.6311	0.7078	0.6047	0.7028	0.5256
W5	0.6579	0.7340	0.6277	0.7117	0.5544
W6	0.7168	0.8002	0.6927	0.7922	0.6290
W7	0.7151	0.8021	0.6901	0.7962	0.6251
W8	0.7088	0.7951	0.6842	0.7897	0.6175
W9	0.7381	0.7899	0.7162	0.7591	0.6621

Table 9: Computational times

Algorithm	SIMPLS	RSIMPLS	PLS-KurSD
Time(seg.)	0.000	3.120	0.062

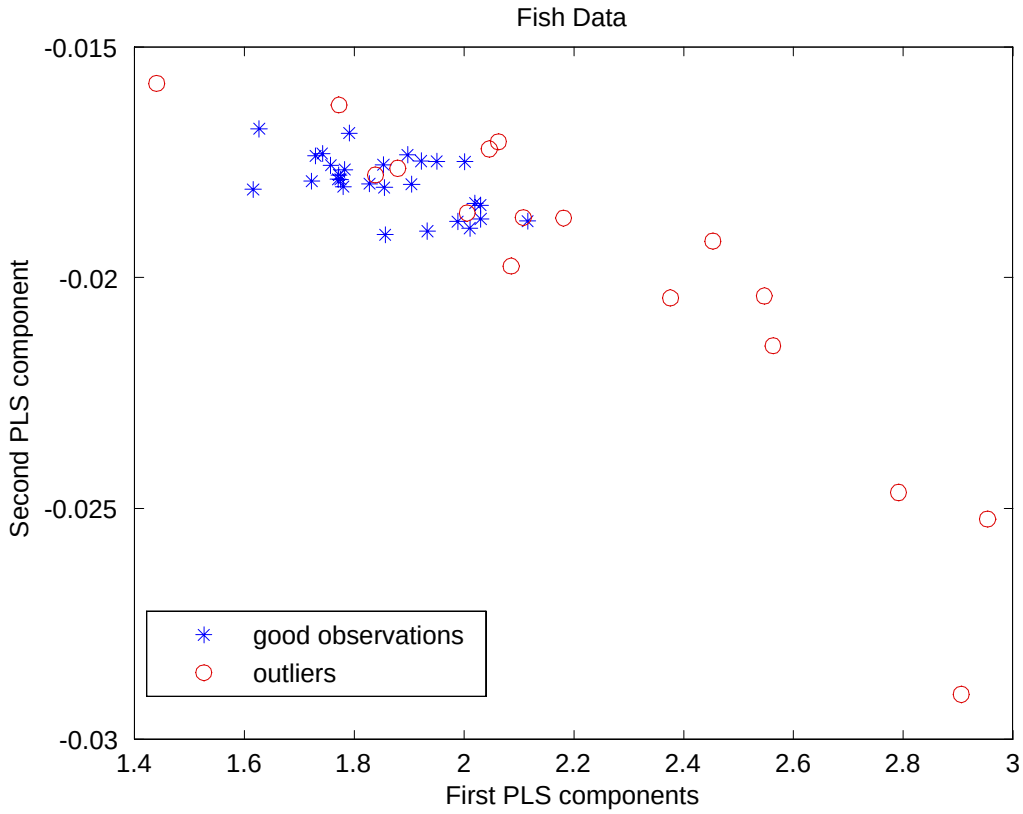

Figure 3: Projection of the Fish data on the plane of the two first PLS components according to PLSKurSD procedure

data on the second and the third PLS components. Again, most of the points identified as outliers are extreme data in this plot.

Finally, we found surprising the large group of outliers found by our procedure. One possibility is that this set of points is indicating a wrong specification of a linear model, as outliers are extreme points but with respect to the given model. We have found that there is a large evidence of a nonlinear relationship between the response and the predictors. For instance, Figure 5 shows a plot of the response with respect to Wavelength 9. We have chosen this predictor because this variables has a significative weight in the OLS regression model, according to Table 6, and high correlation with the response, both in the overall sample and in the subsets of clean data, as is shown in Tables 5 and 8. It can be seen in this Figure that the relationship seems to be non linear, in the whole sample and also when points 39-45 are deleted. This may explain why so many

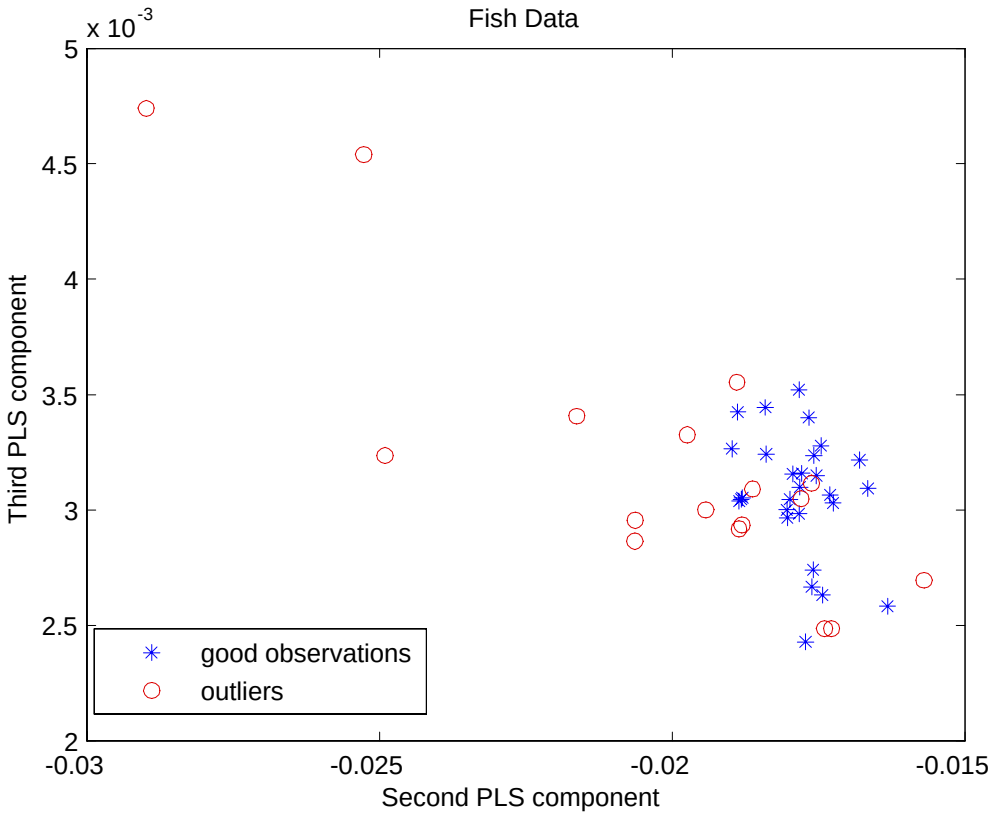

Figure 4: Projection of the Fish data on the plane of the second and the third PLS components according to PLSKurSD procedure

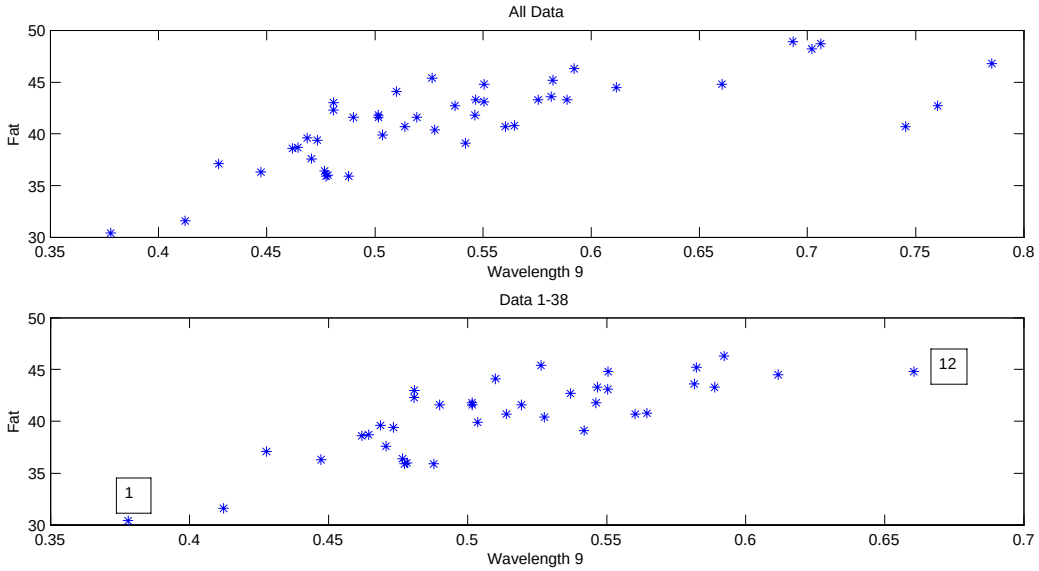

Figure 5: Plot of the Fat variable versus the nineth wavelength. Upper panel whole sample, lower panel observations 1-38.

point are found as outlying in the multivariate regression. For instance, the two extreme points in the bottom panel in Figure 5, which deviate strongly from the linear relationship between both variables, are found as outliers by our procedure.

6 Acknowledgments

This research has been supported by Spanish MEC grant SEJ2004-03303.

References

[1] M. Barker & W. S Rayens (2003), 'Partial least squares for discrimination,' *Journal of Chemometrics*, **17**, 166-173.

[2] Gil, J.A. & Romera, R. (1998), 'On Robust Partial Least Squares (PLS) Methods,' *Journal of Chemometrics*, **12**, 365-378.

[3] Griep, M., Wakeling, P. Vankeerberghen & P. Massart, D. (1995), 'Comparison of semirobust and robust partial least squares procedures,' *Chemometrics and Intelligent Laboratory Systems*, **29**, 1, 37-50.

[4] Hardy, A.J., MacLaurin, P., Haswell, S.J. de Jong, S. & Vandeginste, B.G., (1996), 'Double-case diagnostic for outliers identification', *Chemometrics and Intelligent Laboratory Systems*, **34**, 117-129.

[5] Helland, I. (1992), 'Maximum Likelihood Regression on Relevant Components,' *Journal of the Royal Statitical Society*, **54, 2**, 637-647.

[6] Hubert, M. & Vanden Branden, K. (2003), 'Robust methods for partial least squares regression,' *Journal of Chemometrics*, **17**, 537-549.

[7] Maronna, R.A. & Yohai, V.(1995), 'The Behavior of the Stahel-Donoho Robust Multivariate Estimator,' *Journal of the American Statistical Association*, **90**, 330-341.

[8] Moller, S.F., von Frese, J. and Bro, R. (2005), 'Robust Methods for Multivariate Data Analysis', *Journal of Chemometrics,* **19**, 549-563.

[9] Nguyen D. N & Rocke D. M. (2002), 'Tumor classification by partial least squares using microarray gene expresion data'. *Bioinformatics*, **18**, 1, 39-50.

[10] Peña, D. & Prieto, J. (2001), 'Robust Covariance Matrix Estimation and Multivariate Outlier Detection,' *Technometrics*, **3**, 286-310.

[11] Peña, D. & Prieto, J. (2006), 'Combining Random and Specific Directions for Robust Estimation of High-Dimensional Multivariate Data,' *Journal of Computational and Graphical Statistics* (in press).

[12] Wakeling, I. & Macfie, H. (1992), 'A robust PLS procedure,' *Journal of Chemometrics*, **6**, 189-198.