A New Approach to Robust Partial Least Squares Regression Analysis

Esra Polat[#1], Suleyman Gunay[*2]

[#] *Department of Statistics, Faculty of Science, Hacettepe University, 06800, Ankara, Turkey*

***Abstract*— Partial Least Squares Regression (PLSR) is a linear regression technique developed to relate many independent variables to one or several dependent variables. Robust methods are introduced to reduce or remove the effects of outlying data points. In the previous studies in robust PLSR field it has been mentioned that if the sample covariance matrix is properly robustified further robustification of the linear regression steps of the PLS1 algorithm (PLSR with univariate dependent variable) becomes unnecessary. Therefore, the purpose of this study is to propose a new approach to robust PLSR based on statistical procedures for covariance matrix robustification by selecting the well-known S-estimators. Both simulation results and an analysis on a real data set, which is used in robust PLSR literature frequently, showing the effectiveness, success in fitting to regular data points and predictive power of the new proposed robust PLSR method.**

Keywords*— *robust partial least squares regression, robust covariance matrix, S-estimators, goodness-of-fit, prediction, efficiency.

I. Introduction

Partial Least Squares (PLS) is a useful procedure for relating a set of dependent variables to many independent variables. It could be seen as a general dimension reduction technique which takes into account the linear relationship between the dependent variables and the independent variables. It is well known that the popular algorithms for PLS regression (NIPALS and SIMPLS) are very sensitive to outliers in the dataset. For univariate or multivariate dependent variables, several robustified versions have already been proposed. Wakeling and Macfie (1992) worked with the PLS with multivariate dependent variables (which was 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 dependent variable (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), Hubert and Vanden Branden (2003). González et al. (2009) also concentrated in the case of univariate response (PLS1) and showed that if the sample covariance matrix is properly robustified the PLS1 algorithm will be robust and, therefore, further robustification of the linear regression steps of the PLS1 algorithm is unnecessary [2, 3, 5].

In this study, similar to Gil and Romera (1998) and González et al. (2009) studies, we also concentrate in the case of univariate response (PLS1) and we present a procedure which applies the standard PLS1 algorithm to a robust covariance matrix. In our study, we estimate the covariance matrix used in PLS1 algorithm robustly by using well-known S-estimators.

The rest of the paper is organized as follows. Section 2 reviews briefly the PLS1 algorithm for an one-dimensional dependent variable and analyzes the implication of the robustification of the covariance matrix for the regression steps. Section 3 presents the new approach to robust PLSR analysis. Section 4 reports a simulation study where the performance of the new robust method is compared to classical method and other four robust methods existing in robust PLSR literature. Section 5 illustrates the performance of the proposed method on a well-known set of a real data in literature. Conclusions are reported in Section 6.

II. The Classical PLS1 Algorithm

It is supposed that we have a sample of size n of a 1+p dimensional vector $\mathbf{z} = (\mathbf{y}, \mathbf{X})'$, which could be decomposed as a set of p independent variables, x and a univariate dependent variable y. Throughout this paper, matrices are denoted by bold capital letters and vectors are denoted by bold lowercase letters. Let $\mathbf{S_z}$, be the sample covariance matrix of **z**, consisting of the elements $\mathbf{S_z} = \begin{pmatrix} \mathbf{s}_y^2 & \mathbf{s}'_{y,X} \\ \mathbf{s}_{y,X} & \mathbf{S_X} \end{pmatrix}$, where $\mathbf{s}_{y,X}$ is the $p \times 1$ vector of covariances between y and the x variables. The aim of this study is to estimate the linear regression $\hat{\mathbf{y}} = \hat{\boldsymbol{\beta}}'\mathbf{x}$, and it is assumed that the dependent variable can be linearly explained by a set of a components $(\mathbf{t}_1, \cdots, \mathbf{t}_k)$, with k<<p, which are linear functions of the x variables. Hence, calling **X** the $n \times p$ data matrix of the independent variables, and $\mathbf{x}'_i$ to its *i*th row, the following model showed by Eq. (2.1) and Eq. (2.2) holds [3].

$$\mathbf{x}_i = \mathbf{P}\mathbf{t}_i + \boldsymbol{\varepsilon}_i \tag{2.1}$$

$$\mathbf{y}_i = \mathbf{q}'\mathbf{t}_i + \eta_i \tag{2.2}$$

Here, **P** is the $p \times k$ matrix of the loadings of the vector $\mathbf{t}_i = (t_{i1}, \cdots, t_{ik})'$ and **q** is the k-dimensional vector of the y-loadings. The vectors $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\eta}_i$ have zero mean, follow normal distributions and are uncorrelated. The component matrix $\mathbf{T} = (\mathbf{t}_1, \cdots, \mathbf{t}_k)'$ is not directly observed and should be estimated. Then, it can be shown that the maximum likelihood estimation of the **T** matrix is given as in Eq. (2.3) [3].

$$\mathbf{T} = \mathbf{X}\mathbf{W}_k \tag{2.3}$$

Here the loading matrix $\mathbf{W}_k = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_k]$ is the $p \times k$ matrix of coefficients and the vectors $\mathbf{w}_i, 1 \le i < k$, are the solution of Eq. (2.4) under the constraint in Eq. (2.5) with $\mathbf{w}_1 \; \alpha \; \mathbf{s}_{y,x}$. Consequently, we can conclude that components $(\mathbf{t}_1, \cdots, \mathbf{t}_k)$ are orthogonal [3].

$$\mathbf{w}_i = \arg\max_{\mathbf{w}} \operatorname{cov}^2 (\mathbf{X}\mathbf{w}, \mathbf{y}) \tag{2.4}$$

$$\mathbf{w}'\mathbf{w} = 1 \text{ and } \mathbf{w}_i' \mathbf{S}_x \mathbf{w}_j = 0 \quad \text{for } 1 \le j < i \tag{2.5}$$

It can be shown that vectors $\mathbf{w}_i$ are found as the eigenvectors linked to the largest eigenvalues of the matrix is given as in Eq. (2.6).

$$(\mathbf{I} - \mathbf{P}_x(i))\mathbf{s}_{y,x}\mathbf{s}'_{y,x} \tag{2.6}$$

$\mathbf{P}_x(i)$ is the projection matrix on the space spanned by $\mathbf{S}_X\mathbf{W}_i$, given by $\mathbf{P}_x(i) = (\mathbf{S}_x\mathbf{W}_i)\left[(\mathbf{S}_x\mathbf{W}_i)'(\mathbf{S}_x\mathbf{W}_i)\right]^{-1}(\mathbf{S}_x\mathbf{W}_i)'$. From these results it is easy to see that the vectors $\mathbf{w}_i$ can be computed recursively as in below.

$$\mathbf{w}_1 \; \alpha \; \mathbf{s}_{y,x} \tag{2.7}$$

$$\mathbf{w}_{i+1} \; \alpha \; \mathbf{s}_{y,x} - \mathbf{S}_x\mathbf{W}_i(\mathbf{W}_i'\mathbf{S}_x\mathbf{W}_i)^{-1}\mathbf{W}_i'\mathbf{s}_{y,x}, \; 1 \le i < k \tag{2.8}$$

It could be mentioned that by using the expressions given by Equations (2.7) and (2.8), it is not necessary to calculate the PLS components $\mathbf{t}_i$. In each step of the algorithm, $\mathbf{w}_{i+1}$ only depends on the value of the *i* previous vectors $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_i$, on $\mathbf{S}_x$ and on $\mathbf{s}_{y,x}$. Moreover, as $\mathbf{w}_1$ only depends on $\mathbf{s}_{y,x}$, the calculation of **W** completely fixed by the values of $\mathbf{S}_x$ and $\mathbf{s}_{y,x}$. Finally, as the regression coefficients in Eq. (2.2) are uncorrelated, due to the uncorrelation of the t variables, it is easy to see that the regression coefficients $\hat{\boldsymbol{\beta}}_k^{PLS}$ are given by Eq. (2.9) [3].

$$\hat{\boldsymbol{\beta}}_k^{PLS} = \mathbf{W}_k(\mathbf{W}_k'\mathbf{S}_x\mathbf{W}_k)^{-1}\mathbf{W}_k'\mathbf{s}_{y,x} \tag{2.9}$$

The application of this algorithm can be seen as a two step procedure: (1) the weights $\mathbf{w}_i$ that define the new orthogonal regressor t_i, are computed with Equations (2.7) and (2.8) by using the covariance matrix of the observations; (2) the regression coefficients $\mathbf{q}_i$ are computed from a simple regression between the response, y and the regressor t_i. As it is shown in Equation (2.9), these two steps depend only on the covariance matrix of the observations and it may be thought that if this matrix is properly robustified the procedure will be robust [3].

III. A New Approach to Robust Partial Least Squares Regression Analysis

In this section, following the idea of the methods proposed by Gil and Romera (1998) and González et al. (2009) we propose a new approach to robust PLSR by using S-estimators in order to robustify the sample covariance matrix, $\mathbf{S}_z$, in the PLS1 algorithm. Thus, firstly, we will briefly recall the definition of an S-estimator of multivariate location and scatter. Then, we will give detailed information about FastS algorithm used for calculating multivariate S-estimators for location and scatter. Finally, we will give the three steps of our new proposed robust PLSR method which we named as 'PLS-Smult' [8].

S-estimators for multivariate location and scatter have been studied by Davies (1987), Rousseeuw and Leroy (1987) and Lopuhaä (1989). S-estimators for multivariate location and scatter are highly robust with breakdown value (BDP) up to 50%.

We briefly recall the definition of an S-estimator of multivariate location and scatter. For a sample $\mathbf{z}_1,\cdots,\mathbf{z}_n \in \Re^{p'}$, an S-estimator is defined as the couple $(\tilde{\boldsymbol{\mu}}_z, \tilde{\boldsymbol{\Sigma}}_z)$ which minimizes $|\mathbf{C}_z|$ under the condition in Eq. (3.1)

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\sqrt{(\mathbf{z}_i-\mathbf{m}_z)'\mathbf{C}_z^{-1}(\mathbf{z}_i-\mathbf{m}_z)}\right)=b \tag{3.1}$$

over all $(\mathbf{m}_z, \boldsymbol{C}_z)$ where $\mathbf{m}_z \in \Re^{p'}$ and $\mathbf{C}_z$, is a $p'\times p'$ symmetric positive definite (SPD) matrix. In order to obtain positive breakdown estimates, ρ-function should satisfy the following conditions [6, 8, 10]:

1. ρ-function is symmetric around zero and twice continuously differentiable
2. ρ-function is strictly increasing on $[0,c]$ for some $c>0$, constant on $[c,\infty)$ and $\rho(0)=0$.

For ρ-function one often chooses the function is given in Eq. (3.2). Here c>0 is an appropriate and a user-chosen tuning constant. The derivative of this function is known as Tukey's bisquare function is shown as in Eq. (3.3) [6, 8, 10]:

$$\rho(z)=\begin{cases}\dfrac{z^2}{2}-\dfrac{z^4}{2c^2}+\dfrac{z^6}{6c^4}, & |z|\le c\\ \dfrac{c^2}{6}, & |z|>c\end{cases} \tag{3.2}$$

$$\rho'(z)=\psi(z)=\begin{cases}z\left(1-\left(\dfrac{z}{c}\right)^2\right)^2, & |z|\le c\\ 0, & |z|>c\end{cases} \tag{3.3}$$

Lopuhaä and Rousseeuw (1991) showed that the BDP of a multivariate S-estimator is $\frac{b}{\rho(c)}$. The constant b could be computed as $E_{F_0}[\rho(\|z\|)]$ where $F_0 = N(\mathbf{0}, \mathbf{I_p})$ to ensure consistency at the normal model. Therefore, under the normal model b can be computed as in Eq. (3.4).

$$b=\frac{(p')}{2}\chi^2_{p'+2}(c^2)-\frac{p'(p'+2)}{2c^2}\chi^2_{p'+4}(c^2)+\frac{p'(p'+2)(p'+4)}{6c^4}\chi^2_{p'+6}(c^2)+\frac{c^2}{6}\left(1-\chi^2_{p'}(c^2)\right) \tag{3.4}$$

Here $\chi^2_{p'}$ is the cdf of the χ^2 with p' degrees of freedom. The value of the corresponding tuning parameters for a given BDPs between 0% and 50% could be found in Rousseeuw and Yohai (1984). For example, for the BDPs 0.50, 0.25, 0.20, 0.15, the corresponding c values are 1.5476, 2.937, 3.42, 4.00, respectively [6, 8, 10].

The FastS algorithm, which was developed by Salibian-Barrera and Yohai (2006) for regression S-estimators, was extended to multivariate S-estimators for location and scatter by Salibian-Barrera et al. (2006) [6].

A. The FastS Algorithm for Multivariate Location and Scatter

In this section, the main idea of the FastS algorithm will be layed out. Firstly, the $\mathbf{C}_z$ in Eq. (3.1) is written as $\sigma_z^2\boldsymbol{\Gamma}_z$ with $|\boldsymbol{\Gamma}_z|=1$ and $\sigma_z=|\boldsymbol{\Sigma}_z|^{1/2p'}$, so that the equivalent objective is to find the triplet that minimizes s under the restriction

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{\sqrt{(\mathbf{z}_i-\mathbf{m}_z)'\mathbf{G}_z^{-1}(\mathbf{z}_i-\mathbf{m}_z)}}{s}\right)=b \tag{3.5}$$

over all $(\mathbf{m}_z, \mathbf{G}_z, s)$ where $\mathbf{m}_z \in \Re^{p'}$, $\mathbf{G}_z$ is a $p'\times p'$ SPD matrix with $|\mathbf{G}_z|=1$ and s is a positive scalar. The location and scatter estimates are then $(\tilde{\boldsymbol{\mu}}_z, \tilde{\sigma}_z^2\tilde{\boldsymbol{\Gamma}}_z)$ [6, 8].

The algorithm starts with N initial estimates $\left(\tilde{\boldsymbol{\mu}}_1^{(0)},\tilde{\boldsymbol{\Gamma}}_1^{(0)},\tilde{\sigma}_1^{(0)}\right),\cdots,\left(\tilde{\boldsymbol{\mu}}_N^{(0)},\tilde{\boldsymbol{\Gamma}}_N^{(0)},\tilde{\sigma}_N^{(0)}\right)$ obtained by drawing N random subsets of size $p'+1$ that have a covariance matrix with non-zero determinant, and calculating the classical mean $\tilde{\boldsymbol{\mu}}_l^{(0)}$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}_l^{(0)}$ of the *l*th subset [6, 8].

Then we set $\tilde{\boldsymbol{\Gamma}}_l^{(0)}=\left|\tilde{\boldsymbol{\Sigma}}_l^{(0)}\right|^{-1/p'}\tilde{\boldsymbol{\Sigma}}_l^{(0)}$ and $\tilde{\sigma}_l^{(0)}=\text{med}_{i=1}^{n}\sqrt{\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(0)}\right)'\left(\tilde{\boldsymbol{\Gamma}}_l^{(0)}\right)^{-1}\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(0)}\right)}$ for all $l=1,\cdots,N$. Next those estimates are refined by performing k so-called I-steps, resulting in [6, 8].

$$\left(\tilde{\boldsymbol{\mu}}_1^{(k)},\tilde{\boldsymbol{\Gamma}}_1^{(k)},\tilde{\sigma}_1^{(k)}\right),\cdots,\left(\tilde{\boldsymbol{\mu}}_N^{(k)},\tilde{\boldsymbol{\Gamma}}_N^{(k)},\tilde{\sigma}_N^{(k)}\right) \tag{3.6}$$

The jth I-step to refine the estimate $\left(\tilde{\boldsymbol{\mu}}_l^{(j-1)},\tilde{\boldsymbol{\Gamma}}_l^{(j-1)},\tilde{\sigma}_l^{(j-1)}\right)$ goes as follows [6, 8]:

1. Refine the scale: $\tilde{\sigma}_l^{(j)}=\tilde{\sigma}_l^{(j-1)}\sqrt{\frac{1}{nb}\sum_{i=1}^{n}\rho\left(\frac{\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(j-1)}\right)'\left(\tilde{\boldsymbol{\Gamma}}_l^{(j-1)}\right)^{-1}\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(j-1)}\right)}{\tilde{\sigma}_l^{(j-1)}}\right)}$.

2. Use $\tilde{\sigma}_l^{(j)}$ to compute weights $w_i^{(j)}=\frac{\rho'(u)}{u}$ with $u=\frac{\sqrt{\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(j-1)}\right)'\left(\tilde{\boldsymbol{\Gamma}}_l^{(j-1)}\right)^{-1}\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(j-1)}\right)}}{\tilde{\sigma}_l^{(j)}}$.

3. Compute the weighted mean $\tilde{\boldsymbol{\mu}}_l^{(j)}$ and the weighted covariance $\tilde{\boldsymbol{\Sigma}}_l^{(j)}$, which leads to the refinement $\tilde{\boldsymbol{\Gamma}}_l^{(j)}=\left|\tilde{\boldsymbol{\Sigma}}_l^{(j)}\right|^{-1/p'}\tilde{\boldsymbol{\Sigma}}_l^{(j)}$.

After performing k I-steps, the scale $\tilde{\sigma}_l^{(k)}$ is improved for each $\left(\tilde{\boldsymbol{\mu}}_l^{(k)},\tilde{\boldsymbol{\Gamma}}_l^{(k)},\tilde{\sigma}_l^{(k)}\right)$ by iteratively solving Eq. (3.7)

$$\tilde{\sigma}_l^{(k+1)}=\tilde{\sigma}_l^{(k)}\sqrt{\frac{1}{nb}\sum_{i=1}^{n}\rho\left(\frac{\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(k)}\right)'\left(\tilde{\boldsymbol{\Gamma}}_l^{(k)}\right)^{-1}\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(k)}\right)}{\tilde{\sigma}_l^{(k)}}\right)} \tag{3.7}$$

until convergence while keeping $\tilde{\boldsymbol{\mu}}_l^{(k)}$ and $\tilde{\boldsymbol{\Gamma}}_l^{(k)}$ fixed. The refined estimates $\left(\tilde{\boldsymbol{\mu}}_1^{(B)},\tilde{\boldsymbol{\Gamma}}_1^{(B)},\tilde{\sigma}_1^{(B)}\right),\cdots,\left(\tilde{\boldsymbol{\mu}}_\nu^{(B)},\tilde{\boldsymbol{\Gamma}}_\nu^{(B)},\tilde{\sigma}_\nu^{(B)}\right)$ with the smallest fully iterated scales are kept. However, it must be mentioned that not all scales $\tilde{\sigma}_l^{(k)},l=1,\cdots,N$ need to be computed by solving Eq. (3.7). The first *ν* scales $\tilde{\sigma}_l^{(k)},l=1,\cdots,\nu$ are always computed, but for l > ν the lth scale is only computed if Eq. (3.8) holds.

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{\sqrt{\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(k)}\right)'\left(\tilde{\boldsymbol{\Gamma}}_l^{(k)}\right)^{-1}\left(\mathbf{z}_i-\tilde{\boldsymbol{\mu}}_l^{(k)}\right)}}{A}\right)<b \tag{3.8}$$

Here A is the maximum of the ν best scales that were fully iterated so far. This idea was first developed by Yohai and Zamar (1991). The ν estimates $\left(\tilde{\boldsymbol{\mu}}_1^{(B)},\tilde{\boldsymbol{\Gamma}}_1^{(B)},\tilde{\sigma}_1^{(B)}\right),\cdots,\left(\tilde{\boldsymbol{\mu}}_\nu^{(B)},\tilde{\boldsymbol{\Gamma}}_\nu^{(B)},\tilde{\sigma}_\nu^{(B)}\right)$ with the smallest scales need to be refined until convergence using I-steps as described above, and the final estimate $\left(\tilde{\boldsymbol{\mu}}^{(F)},\tilde{\boldsymbol{\Gamma}}^{(F)},\tilde{\sigma}^{(F)}\right)$ is the one with the smallest scale after full refinement. The final estimate for the covariance matrix $\boldsymbol{\Sigma}_z$ is then $\tilde{\boldsymbol{\Sigma}}^{(F)}=\left(\tilde{\sigma}^{(F)}\right)^2\tilde{\boldsymbol{\Gamma}}^{(F)}$ [6, 8].The FastS algorithm code, which was written by Riani et al. (2012), could be found in MATLAB FSDA Toolbox and it was named as 'Smult' [9].

In this study, firstly, by using robust covariance estimator obtained by using FastS algorithm, the robust covariance estimator $\tilde{\mathbf{S}}_z$ of $\mathbf{S}_z=\begin{pmatrix}s_y^2 & \mathbf{s}'_{y,x}\\ \mathbf{s}_{y,x} & \mathbf{S}_x\end{pmatrix}$ is obtained. Then, by using robust covariance estimator $\tilde{\mathbf{S}}_z$ in the alternative definition of PLS1 algorithm given between Equations (2.7)-(2.9), a new robust PLSR method named as 'PLS-Smult' is proposed. The steps of the PLS-Smult algorithm could be given as in Eq. (3.9) [8].

$$\begin{aligned}
&\mathbf{w}_1 \propto \tilde{\mathbf{s}}_{y,x} \\
&\mathbf{w}_{i+1} \propto \tilde{\mathbf{s}}_{y,x} - \tilde{\mathbf{S}}_x \mathbf{W}_i \left(\mathbf{W}_i' \tilde{\mathbf{S}}_x \mathbf{W}_i\right)^{-1} \mathbf{W}_i' \tilde{\mathbf{s}}_{y,x}, 1 \le i < k \\
&\hat{\boldsymbol{\beta}}_k^{PLS-Smult} = \mathbf{W}_k \left(\mathbf{W}_k' \tilde{\mathbf{S}}_x \mathbf{W}_k\right)^{-1} \mathbf{W}_k' \tilde{\mathbf{s}}_{y,x}
\end{aligned} \tag{3.9}$$

The $\tilde{\mathbf{s}}_{y,x}$ and $\tilde{\mathbf{S}}_X$ robust covariance estimations could be obtained by decomposing the robust covariance estimation of combined data set $\mathbf{z}_i' = (y_i, \mathbf{x}_i)', i = 1, \cdots, n$ obtained by S-estimator as in the form $\tilde{\mathbf{S}}_z = \begin{pmatrix} \tilde{s}_y^2 & \tilde{\mathbf{s}}_{y,x}' \\ \tilde{\mathbf{s}}_{y,x} & \tilde{\mathbf{S}}_x \end{pmatrix}$ [8].

IV. Simulation Study

In the previous section, the new proposed robust PLSR method 'PLS-Smult' is explained in detail. In this section, the comparison of PLS-Smult with other four robust PLSR methods existing in literature is shown in order to validate the good properties of the new PLS robusification. Hence, in this study, five robust PLSR procedures are compared to the classical PLSR method. The first one, RSIMPLS, is the algorithm proposed by Hubert and Vanden Branden (2003) [5].The second one, PRM, is the partial robust M-estimator proposed by Serneels et al. (2006) [11]. The third one, PLS-SD, is the one proposed by Gil and Romera (1998) [2]. The fourth one, PLS-KurSD, is the one proposed by González et al. (2009) [3]. The last robust PLSR method is PLS-Smult, is the one proposed in this paper. In this study, following the study of Hubert et al. (2012) the number of subsets is chosen as N=500 for 'Smult' function used in PLS-Smult algorithm.

We compare efficiency, goodness-of-fit (GOF) and predictive ability of classical PLSR, robust RSIMPLS, PRM, PLS-SD, PLS-KurSD and new proposed robust PLS-Smult methods by performing a simulation study on uncontaminated and contaminated data sets.

According to the initial models given in Equations (2.1) and (2.2), and following a simulation design similar as the one described in Engelen et al. (2003), we have generated the data sets as in Eq. (4.1).

$$\begin{aligned}
&\mathbf{T} \sim N_2\left(\mathbf{0}_2, \boldsymbol{\Sigma}_t\right) \\
&\mathbf{X} = \mathbf{T}\mathbf{I}_{2,p} + N_p\left(\mathbf{0}_p, 0.1\,\mathbf{I}_p\right) \\
&\mathbf{y} = \mathbf{T}\mathbf{A}_{2,1} + N(0,1)
\end{aligned} \tag{4.1}$$

Here, $\left(\mathbf{I}_{k,p}\right)_{i,j} = 1$ for $i = j$ and $\left(\mathbf{I}_{k,p}\right)_{i,j} = 0$, otherwise; $\mathbf{I}_p$ is the $p \times p$ dimensional identity matrix; $\mathbf{0}_2 = (0,0)'$ is a two-dimensional vector of zeros and $\mathbf{A}_{2,1} = (1,1)'$ is a two-dimensional vector of ones and $\mathbf{T}$ is the $n \times 2$ dimensional component matrix. Furthermore, we select n=50, p=10 and we set $\boldsymbol{\Sigma}_t = \begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix}$.

Next, contamination is added by replacing 10% of the observations by different types of outliers. The contaminated parts of the data are denoted as $\mathbf{T}_\epsilon$, $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$.

1. Bad leverage points were constructed by substituting $\mathbf{T}_\epsilon \sim N_2\left((15,15)', \boldsymbol{\Sigma}_t\right)$: $\mathbf{X}_\epsilon = \mathbf{T}_\epsilon \mathbf{I}_{2,p} + N_p\left(\mathbf{0}_p, 0.1\mathbf{I}_p\right)$. However, the corresponding y-values did not change.

2. Vertical outliers have uncontaminated x-values, but their y-values were changed by adjusting the error term: $\mathbf{Y}_\epsilon = \mathbf{T}\mathbf{A}_{2,1} + N(15, 0.1)$.

For each situation, m=1000 data sets were generated and they were analyzed with k=1; 2 and 3 components. The efficiency of the considered methods is evaluated by means of the MSE of the estimated regression parameters $\hat{\boldsymbol{\beta}}$ that is defined as in Eq. (4.2). Moreover, it is clear that the true parameter vector is determined as $\boldsymbol{\beta}_{p,1} = \mathbf{I}_{p,2}'\mathbf{A}_{2,1}$ [1].

$$MSE_k\left(\hat{\boldsymbol{\beta}}\right) = \frac{1}{m}\sum_{l=1}^{m}\left\|\hat{\boldsymbol{\beta}}_k^{(l)} - \boldsymbol{\beta}\right\|^2 \tag{4.2}$$

Here $\hat{\boldsymbol{\beta}}_k^{(l)}$ denotes the estimated parameter based on k components in the lth simulation. The MSE indicates to what extent the slope and intercept are correctly estimated. Therefore, the aim is to obtain a MSE value close to zero. Furthermore, we are

interested on how well the methods fit the regular data points. Because of the simulation settings, we know exactly their indices as we store in the set G_r. Then, the GOF criterion is defined as in Eq. (4.3). Here $r_{i,k}$ is the residual of the ith observation when k components are computed. The objective is to obtain a GOF value close to 1 [1].

$$GOF_k = 1 - \frac{\underset{i \in G_r}{var}(r_i, k)}{\underset{i \in G_r}{var}(y_i)} \tag{4.3}$$

The predictive ability of the methods could be measured by means of the Root Mean Squared Error (RMSE). First a test set G_t of uncontaminated data points with size n_t=50 is generated and then Eq. (4.4) is computed. Here, $\hat{y}_{i,k}$ is the predicted y-value of observation i from the test set when the regression parameter estimates are based on the training set (X, Y) of size n and k components are retained in the model. The optimal number of components is often selected as that k for which this RMSE value is minimal [1].

$$RMSE_k = \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}(y_i - \hat{y}_{i,k})^2} \tag{4.4}$$

The results of the simulations are shown in Tables I-III. Table I shows that in case of no contamination is added and when there is only k=1 component is selected, although RSIMPLS method performs better than classical PLSR method in terms of efficiency and predictive ability, the other four robust PLSR methods (including the new proposed robust PLS-Smult method) have nearly close performance to classical method in terms of efficiency, fitting to data and predictive ability. When only k=2 components are retained in the model, PRM and new proposed robust PLS-Smult methods have nearly close performance to classical method in terms of efficiency and predictive ability. When the model with k=3 is examined, though robust RSIMPLS and PRM methods are better than classical PLSR method in terms of efficiency and showing a close performance to classical method in terms of predictive ability, it could be mentioned that the classical PLSR method outperforms the other three robust PLSR methods (including PLS-Smult) in terms of efficiency and predictive ability. Overall, when no contamination is added, the classical method performs somewhat better than their robust versions, as it would be expected.

Table I The Sample Size is n=50 and p=10, No Contamination.

Number of Components		PLSR	RSIMPLS	PRM	PLS-SD	PLS-KurSD	***PLS-Smult***
k=1	MSE	0.5526	0.5352	0.5597	0.5532	0.5645	0.5526
	GOF	0.7982	0.7997	0.7960	0.7970	0.7930	0.7977
	RMSE	1.4925	1.4845	1.4977	1.4949	1.5085	1.4935
k=2	MSE	0.0338	0.0519	0.0357	0.0432	0.0730	0.0379
	GOF	0.8935	0.8916	0.8921	0.8922	0.8883	0.8930
	RMSE	1.1135	1.1242	1.1165	1.1200	1.1380	1.1162
k=3	MSE	1.7041	1.4097	1.5228	1.9331	3.1495	1.8428
	GOF	0.9082	0.9007	0.9038	0.9029	0.8832	0.9057
	RMSE	1.1883	1.1860	1.1829	1.2021	1.2760	1.1973

It could be seen from both Table II and Table III that when the data set is contaminated, classical PLSR method clearly break downs. The MSE of the regression parameter estimates for classical method increases drastically and even attains their minimum at k=1. The GOF values for classical PLSR method are very low, especially when the data contain bad leverage points. The low GOF values mean that the regular data points are badly fitted. The high RMSE values indicate the low predictive ability of the classical method.

Table II shows that when the data contain bad leverage points, the performance of classical PLSR method in terms of efficiency, fitting to data and predictive ability decrease drastically against robust methods for k=1, k=2 and k=3. The MSE of the regression parameter estimates for classical method increase drastically and even attain their minimum at k=1. When the GOF values of methods are examined for each of the number of components (k=1, 2 3), it is seen that the values related to classical method is lower than the robust methods as it is expected. This shows that the regular data points are badly fitted for classical PLSR method.

It is obvious from Table II that when k=1, k=2 or k=3, the new proposed robust PLS-Smult method outperforms robust PRM, PLS-SD and PLS-KurSD methods in terms of efficiency, fitting to data and predictive ability. Consequently, for this simulation setting in the presence of 10% bad leverage points in the data set, the superiority of new proposed robust PLS-Smult method

against robust PRM, PLS-SD and PLS-KurSD methods in terms of efficiency, fitting to data and predictive ability could be seen clearly.

Table II The Sample Size is n=50 and p=10, 10% Bad Leverage Points.

Number of Components		PLSR	RSIMPLS	PRM	PLS -SD	PLS-KurSD	***PLS -Smult***
k=1	MSE	1.5875	0.5107	0.6115	0.5867	0.5694	0.5561
	GOF	0.2874	0.8087	0.7894	0.7955	0.7967	0.8023
	RMSE	2.7942	1.4646	1.5389	1.5145	1.5079	1.4931
k=2	MSE	2.3084	0.0508	0.1136	0.2354	0.0891	0.0404
	GOF	0.3992	0.8937	0.8827	0.8639	0.8872	0.8944
	RMSE	2.6040	1.1226	1.1688	1.2607	1.1470	1.1162
k=3	MSE	12.5550	1.6954	5.7283	3.1187	3.0885	1.9656
	GOF	0.4713	0.9046	0.7351	0.8673	0.8882	0.9087
	RMSE	2.7935	1.1932	1.9097	1.4121	1.2794	1.2008

Table III shows that the classical PLSR method has a very low efficiency and predictive ability, much badly fitting to data than the five robust PLSR methods (including PLS-Smult) in the presence of vertical outliers in the model with k=1, 2 or 3 components. Especially when the model with k=3 is examined, it is seen that the MSE value of classical method (29.0442) is higher than the MSE values of the five robust methods. RSIMPLS method is the forefront robust method in terms of efficiency and predictive ability for k=1, however, the new proposed robust PLSR method shows a close performance to the robust PRM, PLS-SD and PLS-KurSD methods. When k=2 components are retained in the model, PLS-Smult method is more efficient, fitting to data better and it has a higher predictive ability than PLS-KurSD method. Furthermore, for the model with k=2 components, the new proposed robust PLS-Smult, robust RSIMPLS and PRM are forefront methods especially in terms of efficiency and predictive ability. For the model with k=3 components, it is seen that the new proposed robust PLS-Smult method is more efficient and it has a higher predictive ability than robust PLS-SD and PLS-KurSD methods existing in literature.

Table III The Sample Size is n=50 and p=10, 10% Vertical Outliers.

Number of Components		PLSR	RSIMPLS	PRM	PLS -SD	PLS-KurSD	***PLS -Smult***
k=1	MSE	0.6133	0.4910	0.5419	0.5487	0.5543	0.5465
	GOF	0.7613	0.8126	0.8027	0.8019	0.7984	0.8034
	RMSE	1.6175	1.4485	1.4860	1.4886	1.4956	1.4840
k=2	MSE	1.1477	0.0517	0.0428	0.0549	0.2505	0.0406
	GOF	0.8132	0.8942	0.8932	0.8926	0.8868	0.8950
	RMSE	1.4543	1.1198	1.1189	1.1270	1.1460	1.1135
k=3	MSE	29.0442	1.6097	1.8694	2.6334	4.6291	1.8839
	GOF	0.6241	0.9050	0.9043	0.9000	0.8723	0.9089
	RMSE	2.2290	1.1933	1.2033	1.2438	1.3276	1.1998

Both GOF and RMSE appear to be good criteria to select the optimal number of components 'k_{opt}'. In this study, generally, it is clearly seen from Tables I-III that the differences GOF_3-GOF_2 are very small compared to GOF_2-GOF_1, however, as it is mentioned in Engelen et al. (2003) study it could not be concluded that k_{opt} should be chosen for which GOF_k is maximal. On the other hand, the minimal value of RMSE is always reached at the k=2. This suggesting to select k such that $RMSE_k$ is minimal, therefore, k_{opt}=2 is selected [1]. If Table II and Table III are examined together, it is concluded that in case of the data set is contaminated by 10% of bad leverage points or vertical outliers, it is clear that the new proposed robust PLS-Smult is one of the most efficient methods for k_{opt}=2. The new proposed robust PLS-Smult method is more efficient and it has a higher predictive ability than robust RSIMPLS, PRM, PLS-SD and PLS-KurSD methods in case of the data set is contaminated by bad leverage points and k_{opt}=2 is selected. Moreover, when the data set is contaminated by vertical outliers and k_{opt}=2 is selected, PLS-Smult and PRM are the forefront methods with their performance in terms of efficiency and predictive ability.

V. An Example: Fish Data

In this section, the new proposed robust PLSR method and four robust PLSR methods existing in the literature will be compared on a real data including outliers in terms of goodness-of-fit and predictive ability by using Eq. (4.3) and Eq. (4.4). For this purpose, the fish data which was given in Naes (1985) will be used. The fish data comprise 45 observations and the last 7 are outliers (in the words of Næs, 'abnormal samples'). In this example, fat concentration (percentage, %) of 45 fish samples (rainbow trout) and independent variables of the absorbance at 9 Near Infrared Reflectance (NIR) wavelengths measured after sample homogenisation. The aim of the analysis made on this data set is to model the relationships between the fat concentration (one dependent variable) and these nine spectrums (independent variables). In this study, the data set is divided into two parts. The first 5 observations are the test set and the other remained 40 samples are the training set [2, 4, 7].

Firstly, similar to the our simulation studies, while computing the GOF values 7 outliers are removed from training set that occurs of 40 samples. However, while computing the RMSE value the models are constituted using the training set including the 7 outliers. Then, by using the regression coefficients obtained from these models, the predictions are made from clean test set that occurs of 5 samples. Hence, the predictive ability of the new robust PLSR method especially against the classical PLSR method and the other four robust methods is examined.

The GOF or RMSE values could be considered while selecting the number components that will be retained in the model. The optimal number of components could be selected as the k for which the GOF values are no more change. However, as it is mentioned before, it is more convenient to consider the RMSE values while selecting the optimal number of components. The significant point while selecting the optimal number of components that will retain in the model is that adding one more component whether cause an important decrease or not in RMSE value. Hence, both the aim of data reduction is not deviated and an unnecessary component is not added to model. From Table IV, it is seen clearly that the optimal number of components should be selected as $k_{opt}=3$ for this data set, as adding the third component to the model cause an important decrease in the RMSE values for all the robust methods and classical method. Furthermore, it is clear that the fitting to data also improves for all the methods after adding the third component. Table IV shows that PLS-Smult has a higher predictive ability than both classical method and robust PLS-SD and PLS-KurSD methods for $k_{opt}=3$. Moreover, PLS-Smult method fits the data better than the robust PRM method existing in the literature for $k_{opt}=3$.

Table IV The GOF and RMSE Values for Fish Data in Case of the First 5 Observations are the Test Set and the Remaining 40 Samples are the Training Set.

Number of Components		PLSR	RSIMPLS	PRM	PLS -SD	PLS -KurSD	***PLS -Smult***
k=1	GOF	0.4012	0.5407	0.4582	0.5420	0.5388	0.5263
	RMSE	3.5624	2.4939	1.6635	2.3197	2.1207	1.9402
k=2	GOF	0.7733	0.8333	0.4228	0.8556	0.8652	0.8431
	RMSE	3.0175	2.8755	1.9905	2.7618	2.8415	2.7439
k=3	GOF	0.9240	0.9624	0.6813	0.9603	0.9502	0.9608
	RMSE	2.2604	1.8794	1.2443	2.0029	2.1007	1.9382
k=4	GOF	0.9291	0.9621	0.6787	0.9583	0.9598	0.9597
	RMSE	2.1734	1.8312	1.1445	1.9081	1.9179	1.9307
k=5	GOF	0.9337	0.9668	0.6793	0.9654	0.9685	0.9669
	RMSE	2.2326	1.8506	1.0011	1.9618	1.8130	1.8737
k=6	GOF	0.9377	0.9633	0.8048	0.9695	0.9628	0.9666
	RMSE	1.9128	1.8871	1.1785	1.7740	1.9087	1.8558
k=7	GOF	0.9407	0.9679	0.8135	0.9670	0.9405	0.9674
	RMSE	1.8834	1.7305	1.1420	1.8190	1.7419	1.7700
k=8	GOF	0.9432	0.9600	0.8187	0.9686	0.9370	0.9646
	RMSE	1.9318	1.8257	1.2743	1.8119	1.7429	1.8099
k=9	GOF	0.9424	0.9662	0.8186	0.9690	0.9478	0.9654
	RMSE	1.9935	1.7920	1.1997	1.7623	1.6864	1.8023

VI. Conclusions

In this study, we propose a new robust PLSR method for the PLSR model with one dependent variable, called as 'PLS-Smult', in order to obtain robust predictions in case of outliers existing in the data set.

The simulation study shows that when no contamination is added to the data set, the new proposed PLS-Smult gives almost give identical results to classical PLSR method for k_{opt}=2. However, when the data set is contaminated with bad leverage points or vertical outliers, it is seen that the new proposed robust PLS-Smult method outperforms especially the classical PLSR method but also the robust RSIMPLS, PRM, PLS-SD and PLS-KurSD methods with more or less differences in terms of efficiency, fitting to data and predictive ability. In case of the data set is contaminated with bad leverage points, it is seen that most efficient methods are PLS-Smult and RSIMPLS for k_{opt}=2, respectively. Moreover, when the data contain bad leverage points and k_{opt}=2, the new proposed robust PLS-Smult method shows a better predictive ability than robust RSIMPLS, PRM, PLS-SD and PLS-KurSD methods existing in the literature. In case of the data containing vertical outliers and k_{opt}=2, the most efficient methods are the new proposed robust PLS-Smult and PRM methods. Furthermore, these two robust methods are also come forefront with their predictive ability performance.

The results obtained from real data analysis show that the optimal number of components is selected k_{opt}=3, as adding the third component to the model causes a considerably decrease in the RMSE values of robust methods. The results for the model containing k_{opt}=3 components show that the GOF values for the new proposed robust PLS-Smult method are higher than especially both robust PRM method existing in the literature and classical PLSR method. However, when k_{opt}=3 is selected, the RMSE value for PLS-Smult is lower than classical PLSR method. Generally, for the real data analysis it could be mentioned that whatever the number of the components in the model, the new proposed robust method gives much better models than classical PLSR method in terms of fitting to data and predictive ability.

Consequently, it could be mentioned that when the data contaminated by a reasonable amount of outliers the new proposed robust PLS-Smult method outperforms the classical PLSR method in terms of efficiency, fitting to data and predictive ability. Moreover, PLS-Smult is a good alternative to robust RSIMPLS, PRM, PLS-SD and PLS-KurSD methods existing in the robust PLSR literature that in some cases it outperforms or shows a similar performance with these four robust PLSR methods.

References

[1] S. Engelen, M. Hubert, K. Vanden Branden, and S. Verboven, "Robust PCR and Robust PLSR: A comparative study", *In Theory and Applications of Recent Robust Methods* (M. Hubert, G. Pison, A. Struyf and S. V. Aelst, eds.), Birkhäuser, Basel, pp. 105–117, 2004.

[2] J. A. Gil and R. Romera, "On robust partial least squares (PLS) methods", *Journal of Chemometrics*, vol. 12, pp. 365-378, 1998.

[3] J. González, D. Peña, and R. Romera, "A robust partial least squares regression method with applications", *Journal of Chemometrics*, vol. 23, pp. 78–90, 2009.

[4] A. J. Hardy, P. MacLaurin, S. J. Haswell, S. De Jong, and B. G. M. Vandeginste, "Double-case diagnostic for outliers identification", *Chemometrics and Intelligent Laboratory Systems*, vol. 34, pp. 117-129, 1996.

[5] M. Hubert and K. Vanden Branden, "Robust methods for Partial Least Squares Regression", *Journal of Chemometrics*, vol. 17, pp. 537-549, 2003.

[6] M. Hubert, P.J. Rousseeuw, and T. Verdonck, "A deterministic algorithm for robust location and scatter", *Journal of Computational and Graphical Statistics*, vol. 21, pp. 618-637, 2012.

[7] T. Naes, "Multivariate calibration when the error covariance matrix is structured", *Technometrics*, 27:3, pp. 301-311, 1985.

[8] E. Polat, "New Approaches in Robust Partial Least Squares Regression Analysis", Ph.D Turkish Thesis, Hacettepe University Department of Statistics, Ankara, Turkey, March 2014.

[9] M. Riani, D. Perrotta, and F. Torti, "FSDA: A MATLAB toolbox for robust analysis and interactive data exploration", *Chemometrics and Intelligent Laboratory Systems*, vol. 116, pp. 17–32, 2012.

[10] M. Salibian-Barrera, S. Van Aelst, and G. Willems, "PCA based on multivariate MM-estimators with fast and robust bootstrap", *Journal of the American Statistical Association*, vol. 101, pp. 1198-1211, 2006.

[11] S. Serneels, C. Croux, P. Filzmoser, and P. J. Van Espen, "Partial Robust M-regression", *Chemometrics and Intelligent Laboratory Systems*, vol. 79, pp. 55-64, 2005.