Robust Outlier Detection in Partial Least Squares Regression

Aylin Alin*
Department of Statistics, Dokuz Eylul University, Izmir, Turkey - aylin.alin@deu.edu.tr

Claudio Agostinelli
Dipartimento di Matematica, Università di Trento, Trento, Italy - claudio.agostinelli@unitn.it

Abstract

Partial Least Squares Regression (PLSR) is a very popular multivariate modeling method used to model multicollinear data as well as data sets where the number of explanatory variables exceed the number of samples. Outliers are very common in multivariate data sets and may have a significant effect on the quality of a PLSR model. Hence, their identification is a critical part of modeling process. The aim of this study is to introduce a robust method to detect possible outliers and influential observations in a PLSR model. We propose robust leverage values and robust Cook's Distance statistic estimated from a robust weighted PLSR (RWPLSR) method whose weights can also serve as a diagnostic tool to detect outliers in response and explanatory variable spaces. We compare the proposed diagnostics with classical outlier detection techniques from least squares based PLSR on a real data set.

Keywords: Robust Diagnostics; Robust SIMPLS; SIMPLS; Weighted Likelihood Estimation.

1. Introduction

Partial Least Squares Regression (PLSR) is a very popular multivariate technique. It is used to build a regression model with multicollinear data or with data where number of explanatory variables exceeds the number of samples. The idea of PLSR is to extract uncorrelated latent variables (components) iteratively using algorithms among which SIMPLS (DeJong (1993)) is one of the popular ones. Outliers are very common in multivariate data sets and may have a significant effect on the quality of the model. They may be caused by the instrument used to measure the sample, operation and sample preparation. Despite its popularity SIMPLS is very sensitive against outlying observations since it uses least squares estimation and a nonrobust covariance matrix. When the data set contains outliers, the estimates from SIMPLS differ from the estimates that we would obtain without outliers. Furthermore, SIMPLS might not able to correctly detect outliers by means of classical diagnostics tools based on residuals, leverage values, or leave-one-out diagnostics, since it is sensitive to masking effects in presence of multiple outliers. In fact, a group of outliers would effect, the mean, the covariance matrix and the regression parameter estimates and they would pull the regression line in their direction so that they will exhibit small residuals and leverage values. To overcome this problem we suggest to use a robust SIMPLS procedure in order to build a robust model that would also detect outliers; based on this estimates, diagnostic tools using residuals, leverage values and Cook's distances do not suffer from masking effects. Our proposal is based on a Robust Iteratively Reweighted SIMPLS (RWSIMPLS) introduced by Alin and Agostinelli (2017).

In Section 2, we give some details of the RWSIMPLS method and proposed diagnostics (For details on SIMPLS refer DeJong (1993)). In Section 3, we compare the performance of proposed method with the diagnostics from ordinary SIMPLS on a real data set. We finalize the paper with concluding remarks.

2. Robust Iteratively Reweighted SIMPLS (RWSIMPLS) Method

In a PLSR problem, our aim is to build the linear regression model (1) where $\boldsymbol{X}$ is $n \times k$ mean centered explanatory variables matrix and $\boldsymbol{Y}$ is $n \times m$ mean centered response variable matrix.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1}$$

We assume $\boldsymbol{X}$ and $\boldsymbol{Y}$ are modeled by the linear components as in Equation (2).

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^\top + \boldsymbol{E} \qquad \boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^\top + \boldsymbol{F} \tag{2}$$

where $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_A)$ and $\boldsymbol{U} = \boldsymbol{Y}\boldsymbol{C} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_A)$ are the $n \times A$ component score matrices for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively with A representing the number of components which is less than or equal k. $\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_A)$ and $\boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_A)$ are the $k \times A$ and $m \times A$ loading matrices for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The $k \times A$ matrix $\boldsymbol{W}$ and the $m \times A$ matrix $\boldsymbol{C}$ are the weight matrices. The first score vectors $\boldsymbol{t}_1$ and $\boldsymbol{u}_1$ are the solutions to the following maximization problem. Under the constraints $\|\boldsymbol{t}\| = \|\boldsymbol{u}\| = 1$ we have

$$\max_{\boldsymbol{t},\boldsymbol{u}} \mathbb{C}\mathrm{ov}(\boldsymbol{t}, \boldsymbol{u}) = \max_{\boldsymbol{w},\boldsymbol{c}} \mathbb{C}\mathrm{ov}(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{Y}\boldsymbol{c}) = \max_{\boldsymbol{t},\boldsymbol{u}} \boldsymbol{t}^\top \boldsymbol{u} = \max_{\boldsymbol{w},\boldsymbol{c}} (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{Y}\boldsymbol{c} = \max_{\boldsymbol{w},\boldsymbol{c}} \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{c}. \tag{3}$$

The weight vectors $\boldsymbol{w}$ and $\boldsymbol{c}$ can be found by the Singular Value Decomposition (SVD). According to that, among all possible directions $\boldsymbol{w}$ and $\boldsymbol{c}$, the maximum of Equation (3) is attained for the vectors $\boldsymbol{w_1}$ and $\boldsymbol{c_1}$ corresponding to the largest singular value of $\boldsymbol{X}^\top \boldsymbol{Y}$ (Varmuza and Filzmoser (2009)). An additional constraint is needed for the subsequent score vectors. It is generally taken as the orthogonality of the previous score vectors; i.e. $\boldsymbol{t_a}\boldsymbol{t_j} = 0$ and $\boldsymbol{u_a}\boldsymbol{u_j} = 0$ for $1 \leq a < j \leq A$. The SIMPLS algorithm is based on deflating the empirical covariance matrix of the mean centered data which is highly sensitive against outlying data points, and the loads as the ordinary least squares regression coefficients of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with respect to their score vectors. Hence, the loading vector estimates are very sensitive to outliers and they might be less efficient with the non-normal errors. As a solution Alin and Agostinelli (2017) proposed the Robust Iteratively Reweighted SIMPLS (RWSIMPLS) which is a modified version of SIMPLS where the following weights are used to reduce the influence of outlier observations as it is described below

$$\omega_{r_i} = \omega(r_i(\hat{\boldsymbol{\beta}}); m(\cdot; \hat{\sigma}), \hat{F_n}) = \min\{1, \frac{\Big[A(\delta(r_i(\hat{\boldsymbol{\beta}}))) + 1\Big]^+}{\delta(r_i(\hat{\boldsymbol{\beta}})) + 1}\}. \tag{4}$$

The $r_i(\hat{\boldsymbol{\beta}})$ is the residual corresponding to the estimate $\hat{\boldsymbol{\beta}}$. $\delta(r_i(\hat{\boldsymbol{\beta}})) = \frac{f^*(r_i(\hat{\boldsymbol{\beta}}))}{m^*(r_i(\hat{\boldsymbol{\beta}}); \hat{\sigma})} - 1$ define the Pearson residual which express the agreement between the residuals and their assumed probability model. $f^*(r_i(\hat{\boldsymbol{\beta}})) = \int k(r_i(\hat{\boldsymbol{\beta}}); t, h)\ d\hat{F}_n(t)$ represent a kernel density estimator obtained with the observed values of the residuals based on $\hat{\boldsymbol{\beta}}$ where $\hat{F}_n$ is the empirical distribution of the residual vector $r_i(\hat{\boldsymbol{\beta}})$, $i = 1, \cdots, n$. $m^*(r_i(\hat{\boldsymbol{\beta}}); \hat{\sigma}) = \int k(r_i(\hat{\boldsymbol{\beta}}); t, h)\ dM(t; \hat{\sigma})$ is the smoothed model density. In this study, we use normal kernel density $k(r; t, h) = \exp(-(r-t)^2/(2h^2))/(\sqrt{2\pi}h)$ with $h = \sqrt{k\sigma^2}$, where k is a constant independent of the scale of the data, so that very small weight is assigned to an outlying observation. The function $A(\cdot)$ (Lindsay, 1994) is a Residual Adjustment Function (RAF). $A(\delta) = \delta$ leads the weight value of 1 corresponding the maximum likelihood estimates (MLE) and $A(\delta) = 2(\delta+1)^{1/2} - 1$ corresponds to the Hellinger distance estimate that is more stable than MLE when some assumptions fail. In this study the weights are calculated using the Hellinger RAF. Unlike the SIMPLS method, the RWSIMPLS algorithm is based on deflating weighted covariance matrix $\boldsymbol{S}^w = \boldsymbol{X}^{w\top} \boldsymbol{Y}^w$ where $\boldsymbol{X}^w$ and $\boldsymbol{Y}^w$ are obtained by multiplying each ith row by the squared root of

$$\omega_{r_i} = \underset{s}{median}\{\omega_{r_{is}}\} \quad for \quad s = 1, \ldots, m; \quad i = 1, \ldots, n. \tag{5}$$

$\omega_{r_{is}}$ is the weight of the ith residual corresponding to the sth response. Below are the steps of the RWSIMPLS algorithm.

Step 1: Start with robustly centered $\boldsymbol{X}$ and $\boldsymbol{Y}$ as

$$\begin{aligned} \tilde{x}_{ij} &= x_{ij} - \mathrm{med}_{L_1}(\boldsymbol{X}) \quad j = 1, \ldots, k \\ \tilde{y}_{is} &= y_{is} - \mathrm{med}_{L_1}(\boldsymbol{Y}) \quad s = 1, \ldots, m \end{aligned} \tag{6}$$

where $\mathrm{med}_{L_1}(\boldsymbol{X})$ is the L_1-median computed from the collection of the vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. It is a robust estimator of the center of the data cloud of the h-dimensional vectors (Serrneels et al., 2005).

Step 2: Scale the centered response variable(s) by the robust scale parameter obtained by the Median Absolute Deviation (MAD)

$$y_{is}^{*} = \frac{y_{is}}{\text{MAD}(y_{1s}, \ldots, y_{ns})} \quad s = 1, \ldots, m \tag{7}$$

where $\text{MAD}(y_{1s}, \ldots, y_{ns}) = \frac{\underset{i}{\text{median}}|y_{is} - \text{med}_{L_1}(\boldsymbol{Y})|}{0.6745}$. Those scaled response values serve as the starting residuals.

Step 3: Calculate the weights $\omega_{y_i^*}$ for each y_i^* as given in (4) and (5) with r_{is} replaced by y_{is}^*.

Step 4: Calculate the weighted explanatory variables $\boldsymbol{X}^w$ and response variables $\boldsymbol{Y}^w$ by multiplying each row by the squared root of weights.

$$\begin{aligned} x_{ij}^w &= x_{ij}\sqrt{\omega_{y_i^*}} \quad \text{for each} \quad j = 1, \ldots, k \\ y_{is}^w &= y_{is}\sqrt{\omega_{y_i^*}} \quad \text{for each} \quad s = 1, \ldots, m \end{aligned} \tag{8}$$

Step 5: Select a small sample without replacement (say with the size of 5) $\boldsymbol{X}_z^w$ and $\boldsymbol{Y}_z^w$ out of $\boldsymbol{X}^w$ and $\boldsymbol{Y}^w$, then follow the steps of ordinary SIMPLS on $\boldsymbol{X}_z^w$ and $\boldsymbol{Y}_z^w$ to find the starting coefficient estimates $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_m)$.

Step 6: Using $\hat{\boldsymbol{\beta}}_s = (\hat{\beta}_{s0}, \ldots, \hat{\beta}_{sk})$, calculate $r_{is}(\hat{\boldsymbol{\beta}}_s) = y_{is} - \boldsymbol{x}_i^w \hat{\boldsymbol{\beta}}_s$ with $\boldsymbol{x}_i^w = (x_{i1}^w, \ldots, x_{ik}^w)$ for $s = 1, \ldots, m$ and $i = 1, \ldots, n$.

Step 7: Calculate the robustly centralized and scaled residuals r_{is}^* as

$$r_{is}^{*} = \frac{r_{is} - \text{med}_{L1}(\boldsymbol{r})}{\text{MAD}(r_{1s}, \ldots, r_{ns})} \tag{9}$$

with $\text{MAD}(r_{1s}, \ldots, r_{ns}) = \frac{\underset{i}{\text{median}}|r_{is} - \text{med}_{L_1}(\boldsymbol{r})|}{0.6745}$.

Step 8: Calculate the weights for each r_i^*, then reweight the explanatory and response variables as in Equation

$$\begin{aligned} x_{ij}^w &= x_{ij}\sqrt{\omega_{r_i^*}} \quad \text{for each} \quad j = 1, \ldots, k \\ y_{is}^w &= y_{is}\sqrt{\omega_{r_i^*}} \quad \text{for each} \quad s = 1, \ldots, m \end{aligned} \tag{10}$$

Step 9: Calculate the reweighted covariance matrix $\boldsymbol{S}^w = \boldsymbol{X}^{w\top}\boldsymbol{Y}^w$.

Step 10: Perform the ordinary SIMPLS on $\boldsymbol{S}^w$ to estimate the coefficients.

Step 11: With the new estimated coefficients, repeat the steps 6-10 until the specified threshold (say $1e^{-4}$) is reached for the maximum of the absolute deviations for two consecutive $\hat{\boldsymbol{\beta}}$s. Once convergence is reached store the final coefficients and final value of the absolute deviations.

Step 12: Repeat the steps 5-11 for few, say $c = 5$, times. Then, the coefficient set with the minimum of those c absolute deviations will be the final $\hat{\boldsymbol{\beta}}$.

3. Outlier Diagnostics

Leverage points which are the diagonal elements of the hat matrix are useful to detect outliers in explanatory variable space. However, leverage points based on ordinary least square estimates may suffer from masking effect so that corresponding leverage values appear normal. In this paper, we propose a robust hat matrix (11) leading robust leverage points. Those robust measures are based on the estimates from RWSIMPLS algorithm.

$$\boldsymbol{H}^w = \boldsymbol{T}^w(\boldsymbol{T}^{w\top}\boldsymbol{T}^w)^{-1}\boldsymbol{T}^{w\top} \tag{11}$$

where $\boldsymbol{T}^w$ is the $n \times A$ component matrix calculated as $\boldsymbol{T}^w = \boldsymbol{X}\boldsymbol{W}^{\boldsymbol{w}}$ where $\boldsymbol{W}^{\boldsymbol{w}}$ is the robust weight matrix of $\boldsymbol{X}$ from RWSIMPLS method and $\boldsymbol{X}$ is the explanatory variable matrix robustly centered as in (6). The leverage value h_{ii}^r is the ith diagonal element of $\boldsymbol{H}^w$ measuring the effect of ith response point on its own

prediction when a linear regression model is fit on component matrix. Large leverage value means data point has an unusually large influence on its predicted value. The trace of $\boldsymbol{H}^w$ equals to the number of components A, and any leverage point greater that 2 times the average of leverage values may be considered as an outlier in component space which is also an outlier in $\boldsymbol{X}$. Leverage points do not consider the response variables. To be able to detect unusual points in response space, we propose robust standardized residuals in (12) to disclose outliers in response variable and robust Cook's Distance (13) to disclose observations that are influential on $\hat{\boldsymbol{\beta}}$.

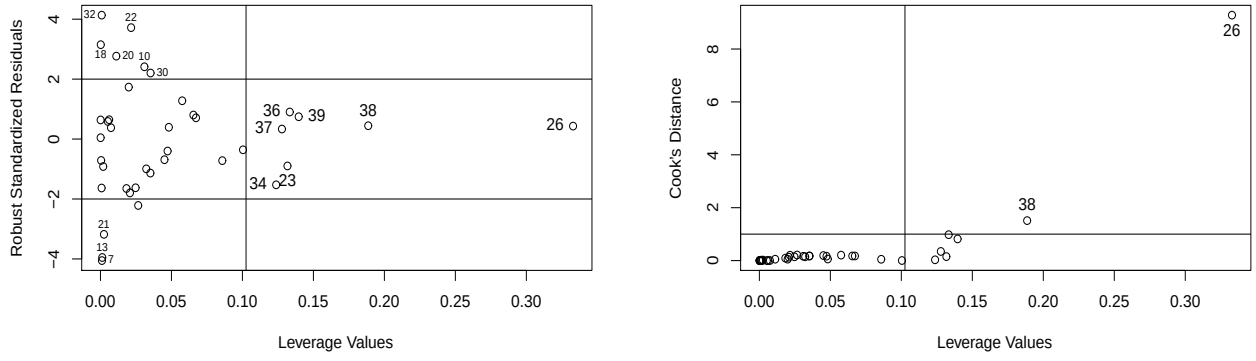
$$r_{is}^w(\hat{\boldsymbol{\beta}}) = \frac{\tilde{r}_{is}}{\hat{\sigma}_s^w} \tag{12}$$

$$D_{is}^w = \left(\frac{y_{is} - \hat{y}_{is}^w}{\hat{\sigma}_s^w \sqrt{1 - h_{ii}}} \right)^2 \frac{h_{ii}}{A(1 - h_{ii})} \tag{13}$$

where $\tilde{r}_{is} = r_{is}(\hat{\boldsymbol{\beta}}) - \bar{r}(\hat{\boldsymbol{\beta}})$, $r_{is}(\hat{\boldsymbol{\beta}}) = y_{is} - \boldsymbol{x_i}\hat{\boldsymbol{\beta}}_s$ and $\hat{\sigma}_s^w = \sqrt{\frac{\sum_{i=1}^n \omega_{r_{is}} \tilde{r}_{is}^2}{\sum_{i=1}^n \omega_{r_{is}} - A}}$. Latter is the robust scale estimate of the model for sth response. $\hat{y}_{is}^w$ is the fitted value at the ith row and sth column of the matrix $\boldsymbol{T}^w \boldsymbol{C}^{w\top}$ where $\boldsymbol{C}^w$ is the robust weight matrix of $\boldsymbol{Y}$ from RWSIMPLS. Any observations such that D_{is}^w is greater than 1 can be considered as a point having influence on $\hat{\boldsymbol{\beta}}$. The following section includes an example where we compare the proposed robust diagnostics with their counterparts from ordinary SIMPLS based PLSR.

4. Example-Octane Data

This data set also studied by Hubert et al. (2005) consists of NIR absorbance spectra over $k = 226$ wavelengths ranging from 1102nm to 1552nm with measurements every two nm. For each of the $n = 39$ production gasoline samples the octane number y was measured, so $m = 1$. It is known that this data set contains six outliers $25, 26, 36, 37, 38, 39$ to which alcohol was added. The robust diagnostics from RWSIMPLS method with $A = 2$ components are shown in Figure 1. The first plot is for robust leverage vs robust standardized residuals. Spectrum points $23, 26, 34, 36 - 39$ seem to be outliers in spectrum space with large leverage values. But, they do not have large standardized residuals. Among those seven data points only points 38 and 26 seem to be influential on the coefficient estimates.



(a) (b)

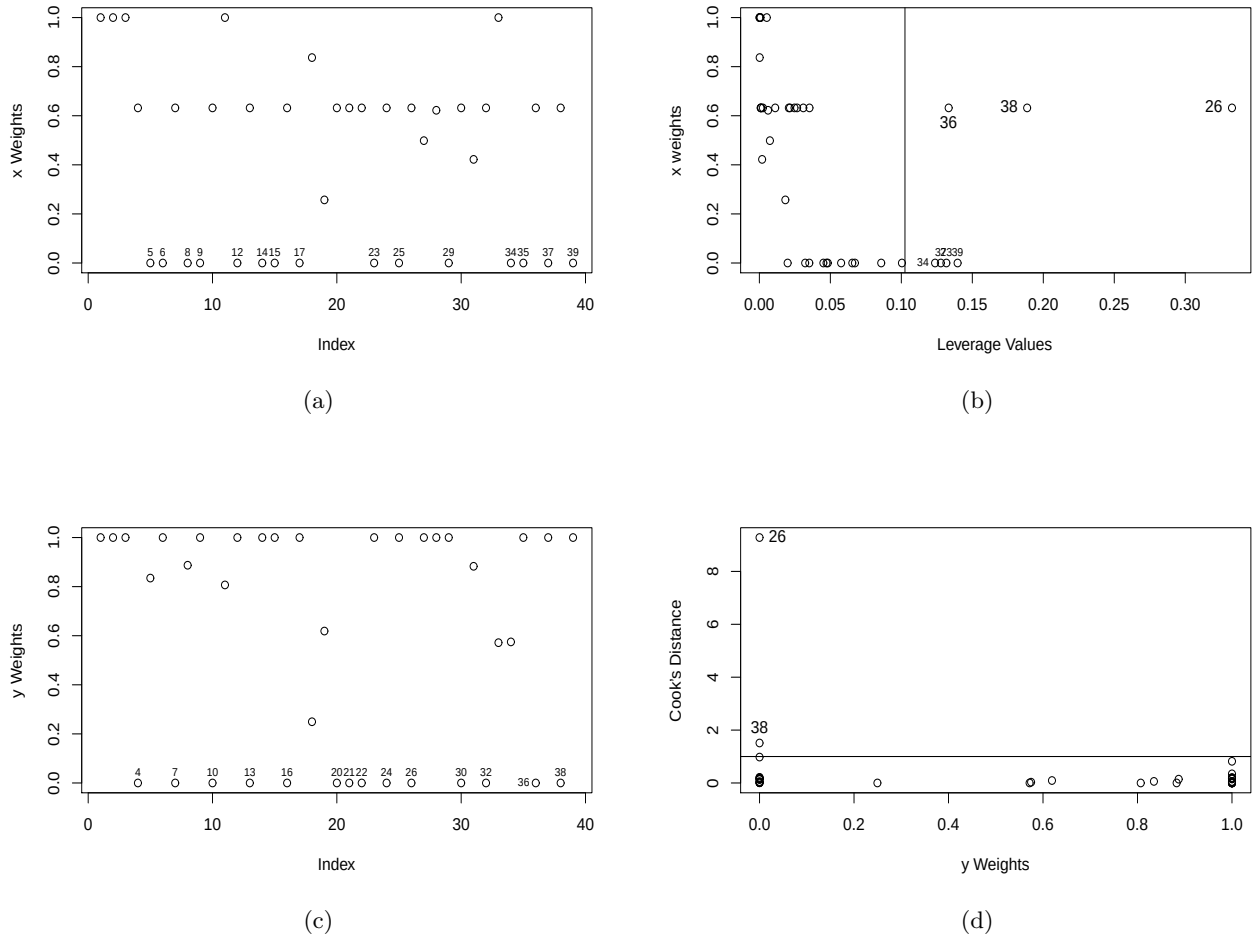
Figure 1: Diagnostics from RWSIMPLS for Octane Data

The weights of the RWSIMPLS method can also be used as diagnostics to detect data points deviating from normal distribution. These points are illustrated with Figure 2. The data points with the final weights of y equal to zero ($\omega_{y_i} = 0$) are the points causing non-normality for response. There are 14 octane points with

zero weights but only points 26 and 38 seem to be influential. The first plot in Figure 2 is for the weights on explonatory variables (spectrums) calculated as in (14).

$$\omega_{x_i} = \underset{j}{\text{median}}(\omega_{x_{ij}}) \quad for \quad j = 1, \ldots, k; \quad i = 1, \ldots, n \tag{14}$$

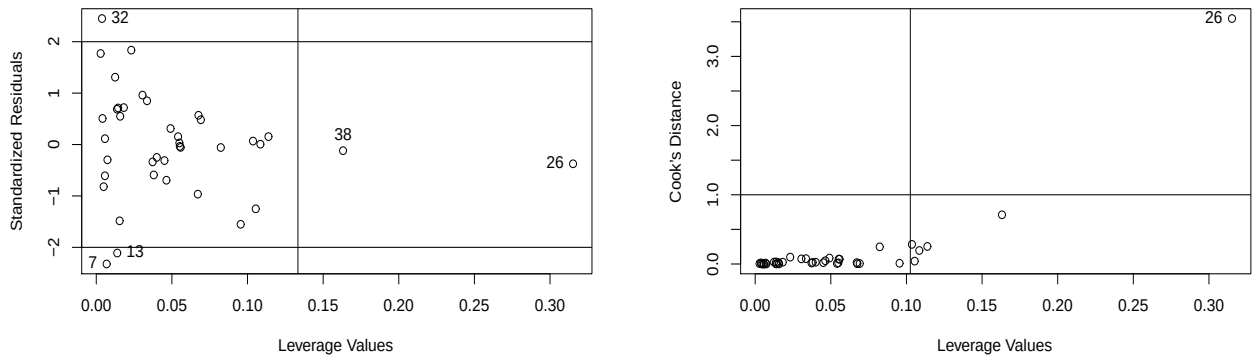
where $\omega_{x_{ij}}$ is obtained as in (4). The points with zero ω_{x_i} cause non-normality for $\boldsymbol{X}$ space. There are 14 points with zero ω_{x_i} but only spectrum points $23, 26, 34, 36-39$ both cause non-normality and large effect on response fits. Even though $25, 26, 36-39$ reported as outliers (Hubert et al. (2005)), only $26, 36-39$ seem to be outliers in explanatory variable space with only two of them being influential on coefficient estimates $\hat{\boldsymbol{\beta}}$. Point 25 seems to cause deviation from normality with no harm on coefficient estimates and fitted values. The diagnostics from ordinary SIMPLS algorithm are presented with Figure 3. In the ordinary SIMPLS method, we can only detect two spectrum points 26 and 38 as large leverage points with 26 being only influential point. As mentioned above this data set has also been studied by Hubert et al. (2005) who only detect point 26 as being a bad leverage point but giving no information about if it is influential or if the point causes non-normality for response or explanatory variable space.



(a) (b)

(c) (d)

Figure 2: Weights from RWSIMPLS for Octane Data



(a) (b)

Figure 3: Diagnostics from SIMPLS for Octane Data

5. Conclusions

We proposed robust diagnostics to detect outliers and influential observations in a PLSR model. The proposed diagnostic measures are calculated by robust weighted partial least squares algorithm (RWSIMPLS). We compared the proposed measures with their classical counterparts on a data set which has also been studied by various authors. Apart from the classical diagnostic measures such as residuals, leverage points and Cook's distances, the weights of the RWSIMPLS method also serve as diagnostic tool to detect points deviating from normality.

References

Alin, A. and Agostinelli, C. (2017). Robust iteratively reweighted SIMPLS. *Journal of Chemometrics*, 31(3):e2881.

DeJong, S. (1993). SIMPLS: An alternative approach tp partial least squares regression. *Chemom. Intell. Lab. Syst.*, 18:251–263.

Hubert, M., Rousseeuv, P., and Aelst, S. (2005). *Multivariate Outlier Detection and Robustness*, pages 263–302. Elsevier. In: Handbook of Statistics, Volume 23: Data Mining and Computation in Statistics.

Lindsay, B. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Stat.*, 22:1018–1114.

Serrneels, S., Croux, C., Filzmoser, P., and P.J., V. E. (2005). Partial robust M-regression. *Chemom. Intell. Lab. Syst.*, 79:55–64.

Varmuza, K. and Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press.