Robust and Classical PLS Regression Compared

Bettina Liebmann[a]*, Peter Filzmoser[b] and Kurt Varmuza[a]

---

* *Correspondence to: B. Liebmann, Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9/166, A-1060 Vienna, Austria*
E-mail: Bettina.Liebmann@tuwien.ac.at

[b] *P. Filzmoser*
Institute of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria
E-mail: P.Filzmoser@tuwien.ac.at

[a] *B. Liebmann, K. Varmuza*
Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9/166, A-1060 Vienna, Austria
E-mail: kvarmuza@email.tuwien.ac.at

Abstract.

Classical PLS regression is a well-established technique in multivariate data analysis. Since classical PLS is known to be severely affected by the presence of outliers in the data or deviations from normality, several PLS regression methods with robust behavior towards data contamination have been proposed. We compare the performance of the classical SIMPLS approach with the partial robust M regression (PRM). Both methods are applied to three different data sets including outliers intentionally created. A simulated data set with known true model parameters allows insight in the modeling performance with increasing data contamination. QSPR data are modified with a cluster of outlying observations. A third data set from near infrared (NIR) spectroscopy is likely to include noise and experimental errors already in the original variables, and is further contaminated with outliers. To provide a sound comparison of the considered methods we apply repeated double cross validation. This validation procedure judiciously optimizes the model complexity (number of PLS components) and estimates the models' prediction performance based on test set predicted errors. All studied robust regression models outperform the classical PLS models when outlying observations are present in the data. For uncontaminated data the prediction performance of both the classical and the robust model are in the same range.

Keywords: partial robust M-regression (PRM), PLS regression; repeated double cross validation (rdCV), outliers, R

1. INTRODUCTION

Partial Least Squares (PLS) regression is a well-known and often successfully applied technique in multivariate data analysis. The typical task of regression is to model a response *y* by means of a set of explanatory variables ("features") $x_1, \ldots, x_m$.

Many different PLS algorithms have been developed over the last 25 years, and have been implemented in various software products. For the "classical PLS" method as referred to in this work, we choose the SIMPLS algorithm introduced by de Jong [1]. Essential advantages of the PLS approach are its ability to deal with collinear variables and numerous *x*-variables, and it allows to optimize the model's complexity [2]. These properties are especially useful with modern analytical instruments such as spectrometers, where many and strongly correlated *x*-variables are recorded.

However, the classical PLS procedures are known to be severely affected by the presence of outliers in the data or deviations from normality [3]. The non-robustness of PLS was justified theoretically in [4]. Outliers are different from the majority of the data, but they are not necessarily incorrect. Often the outlying observations were made under exceptional circumstances or they belong to another statistical population. In general, classical methods usually fail in identifying the outliers. Consequently, the resulting model may be fitting the outlying observations thus "masking" their erroneous nature (masking

effect). By contrast, some good data points might show up as outliers (swamping effect).

Several robust alternatives to classical PLS have been proposed. Their common goal is to detect data contamination and estimate a regression model that primarily fits the “good” data. Outliers can then be identified easily by their residuals from this robust fit.

The two main strategies for robust PLS regression are (1) downweighting of outliers and (2) robust estimation of the covariance matrix. The early approaches for robust regression by downweighting of outliers are considered semi-robust: they had, for instance, non-robust initial weights [5] or the weights were not resistant to leverage points [6]. Based on the second strategy, a robust covariance estimation, the robust SIMPLS method [7] provides resistance to all types of outliers including leverage points. The latter also applies to the “Partial Robust M Regression” (PRM), introduced in 2005 by Serneels et al [8]. As the name suggests, it is a partial version of the robust M-regression. In an iterative scheme, weights ranging between zero and one are calculated to reduce the influence of deviating observations in the *y* space as well as in the space of the regressor variables. PRM is very efficient in terms of computational cost and statistical properties, and therefore the robust method of choice in this paper.

The objective is to compare the predictive performance of classical and robust PLS regression by a judicious validation method. We apply the repeated double

cross validation (rdCV) procedure to both regression types, and study the models resulting from three data sets with different characteristics.

2. METHODS

2.1. Partial Robust M-Regression

The PRM approach is “partial” because it follows the idea of dimensionality reduction by using a few latent variables. The original regressor variables are replaced with orthogonal latent variables with maximum covariance with $\boldsymbol{y}$, as in classical PLS regression. Suppose that observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$ and y_i, for $i = 1, \ldots, n$, are available, forming the $(n \times m)$ matrix $\boldsymbol{X}$ and the vector $\boldsymbol{y}$, respectively. For simplicity, we assume mean-centered $\boldsymbol{y}$. Then the original regression problem

$$y_i = \boldsymbol{x}_i \boldsymbol{\beta} + \varepsilon_i \qquad (1)$$

with the coefficients $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_m)^T$ and the error terms ε_i is reduced to the latent variables regression model

$$y_i = \boldsymbol{t}_i \boldsymbol{g} + \delta_i \qquad (2)$$

with the new regression coefficients $\boldsymbol{g} = (g_1,\ldots,g_a)^T$ and the error terms δ_i. The new model is of lower dimension $a < m$, and it is in fact a regression on the score vectors $\boldsymbol{t}_i$, which are to be determined.

In general, two types of outliers can be influential to the estimation of the regression coefficients: Leverage points, which are multivariate outliers in the space of the regressor variables, and vertical outliers, which are not atypical in

the regressor space but have large residuals. PRM offers good robustness properties by taking into account both types of outliers: a weight w^x_i is responsible for dealing with leverage points, while a weight w^r_i is relevant for vertical outliers. For each observation, these continuous weights are iteratively adjusted, in order to diminish the negative influence of outlying objects on the regression model. The total weight to multiply each object with is then defined as

$$w_i = \sqrt{w_i^x w_i^r} \qquad (3)$$

Note that all PLS regressions are performed on weighted observations $w_i \boldsymbol{x}_i$ and $w_i y_i$.

A brief description of the PRM algorithm:

Step 1: Compute robust starting values for weights from data $\boldsymbol{x}_i$ and y_i.

Step 2: Perform classical PLS on weighted observations $w_i \boldsymbol{x}_i$ and $w_i y_i$.

Step 3: Recompute weights w^r_i (from residuals), w^x_i (from PLS scores), and total weights w_i

Step 4: Iterate step 2 and 3 until convergence of the regression coefficients

Step 5: Obtain final regression coefficients $\boldsymbol{b}_{\mathrm{PRM}}$ directly from last PLS step

In the first and crucial step, the weights are initialized – in a robust manner. Therefore, “robust autoscaling” is applied to the $\boldsymbol{X}$ matrix as well as the $\boldsymbol{y}$ vector. Instead of the usually applied measures mean and standard deviation, their robust counterparts median and median absolute deviation (MAD) are used [3]. The data is centered to the median, and then divided by MAD.

Robust autoscaling in $\boldsymbol{y}$ thus results in the intermediate distances h_i

$$h_i = \frac{y_i - y_{median}}{\underset{i}{median}\left|y_i - y_{median}\right|} \tag{4}$$

The robust center of $\boldsymbol{X}$ can be calculated by a multidimensional median estimator such as the column-wise median or the L1-median $\tilde{\boldsymbol{x}}$ [9]. Then we yield each objects' Euclidean distance $\left\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}\right\|$ to the robust center $\tilde{\boldsymbol{x}}$. The robust autoscaling in $\boldsymbol{X}$ results in the intermediate distances g_i

$$g_i = \frac{\left\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}\right\|}{\underset{i}{median}\left\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}\right\|} \tag{5}$$

Note that in all subsequent steps of the algorithm, the distances g_i are computed in the score space, i.e. according to the scores $\boldsymbol{t}_i$.

By passing the intermediate distances h_i and g_i to a weight function, they are transformed to values between 0 and 1. Observations with large distances to the data majority receive a weight close to zero, so to have reduced influence on the regression model. Observations among the data majority get a weight close to one. We choose the "Fair" weight function [6] with a tuning constant set to four, which is reported to have good performance properties. The residual weight $w^r{}_i$ and the leverage weight $w^x{}_i$ for object i is then calculated as:

$$w^r{}_i = \frac{1}{\left(1 + \left|\frac{h_i}{4}\right|\right)^2} \quad \text{and} \quad w^x{}_i = \frac{1}{\left(1 + \left|\frac{g_i}{4}\right|\right)^2} \tag{6}$$

The transforming character of the Fair function is shown in Figure 1. By choosing continuous weights the dilemma of an all-or-nothing decision – the

object is an outlier: yes or no – can be avoided. The weight given to each object is corresponding to its degree of outlyingness.

Once the robust starting weights are computed, we construct a model using classical SIMPLS on the weighted rows of $\boldsymbol{X}$ and weighted $\boldsymbol{y}$. This analysis yields a first estimate of the regression coefficients $\boldsymbol{g}$ and the PLS scores $\boldsymbol{t}_i$. Note that the resulting scores have to be corrected by division by the total weight w_i. At this point the residuals r_i are computed:

$$r_i = y_i - \boldsymbol{t}_i \boldsymbol{g} \qquad (7)$$

The residual weights w^r_i are then updated according to equation (4) and (6), by substituting y_i with the residuals r_i. For an update of the leverage weights w^x_i, we replace the original x-variables in equation (5) by the current set of score vectors $\boldsymbol{t}_i$ and apply the Fair function given in (6). The original data matrix $\boldsymbol{X}$ as well as vector $\boldsymbol{y}$ is reweighted with the updated total weights, and the next classical SIMPLS regression step is performed until convergence of the regression coefficients $\boldsymbol{g}$. If the difference between the regression coefficients of two consecutive PLS steps is smaller than a certain threshold value, here 10^{-2}, the iterative procedure is terminated. From the last regression step, the robust PLS model is obtained.

2.2. Performance criteria

SEP. We estimate the prediction performance of the models based on many test set predicted errors (residuals), that is the difference between the experimental value y_i and the predicted (modeled) value $\hat{y}_i$ for an object *i*. The standard deviation of these prediction errors - usually abbreviated to standard error of prediction - SEP, is defined by

$$\mathrm{SEP} = \sqrt{\frac{1}{n_{SEP} - 1} \sum_{i=1}^{n_{SEP}} (y_i - \hat{y}_i - \mathrm{bias})^2} \tag{8}$$

$$\mathrm{bias} = \frac{1}{n_{SEP}} \sum_{i=1}^{n_{SEP}} (y_i - \hat{y}_i) \tag{9}$$

The SEP used within this work is equivalent to $\mathrm{SEP}_{\mathrm{TEST}}$, because all predicted $\hat{y}_i$ values are derived from test set objects. Applying the rdCV approach (see Section 2.3) the number of available $\hat{y}$-values, n_{SEP}, is the number of objects times the number of repetitions. The bias is the arithmetic mean of the prediction errors; especially for large n_{SEP} it is near zero.

$\mathrm{SEP}_{\mathrm{TRIM}}$. The SEP criterion becomes illusive when applied to robust models fitted to contaminated data. A good robust fit leads to large residuals for the outlying objects, whilst a classical model tends to describe outliers better - sometimes even better than the regular observations. Since we intend to assess the robust model's performance in fitting the good data but not the outliers, a robust SEP measure is necessary [3]. The exclusion of a certain

percentage of unusually large (absolute) residuals leads to an acceptable robust performance criterion SEP_{TRIM}. We choose a trimming constant of 20 %. This choice was also made in other papers [10], as with real data the percentage of outliers is unknown. Note that for data sets where only few outliers are expected, a smaller trimming constant can prevent too optimistic estimates of the prediction performance.

MSE. Another statistical error criterion based on residuals computed from (repeated double) cross validation is the mean squared error (MSE).

$$\mathrm{MSE} = \frac{1}{n_{MSE}} \sum_{i=1}^{n_{MSE}} (y_i - \hat{y}_i)^2 \qquad (10)$$

A trimmed version of this measure, MSE_{TRIM}, is easily computed by excluding 20 % of the largest squared residuals. In the robust repeated double cross validation algorithm, MSE_{TRIM} is used for an estimation of the optimum number of PLS components.

RMSE of regression coefficients. We expect a good regression method to find the true underlying linear function relating ***X*** and ***y***. Given a defined set of true model coefficients ***β*** and a random data set ***X***, the calculation of a perfectly corresponding response ***y*** is straightforward. It is an easy task to solve the linear regression problem of perfectly related ***X*** and ***y*** data, obtaining estimated regression coefficients ***b*** that are (almost) identical with ***β***. In case of data contamination or noise, however, the estimated coefficients will deviate from ***β***. The RMSE indicates to what extent the predefined coefficients ***β*** are correctly estimated by the considered method.

The RMSE of the estimated regression parameters $\boldsymbol{b} = (b_1,\ldots,b_m)^T$ is introduced as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\beta_j - b_j\right)^2} \qquad (11)$$

where $\boldsymbol{\beta}$ denotes the true model coefficients. Ideally, the RMSE value is close to zero.

2.3. Repeated double cross validation (rdCV)

Model evaluation is of high importance in chemometrics. For that purpose, we use repeated double cross validation [11]. This procedure allows a reasonable estimation of the optimum model complexity (number of PLS components) as well as the prediction performance. A randomly chosen subset of data – the calibration set – is subjected to a *k*-fold cross validation loop, yielding a first suggestion for the optimum model complexity. Subsequently, a model for the entire calibration set is constructed and applied to the left out test data. Due to the repetitive nature of rdCV, the variability of optimum model complexity as well as the variability of test set predicted errors with different data subsets is accessible. The rdCV procedure combined with classical PLS is published in detail in [11], and an application is presented in [12]. A fair comparison of classical and robust PLS models requires a comparable validation technique. Therefore, we implemented the robust PRM method into the three nested loops of the rdCV procedure. The pseudo-code is as follows:

Repetition loop: **FOR** $\rho = 1$ TO n_{REP}

(1) Split all n objects randomly into SEG_{TEST} segments (typically 3-10) of approximately equal size.

(2) Outer loop: **FOR** $\tau = 1$ TO SEG_{TEST}

(a) Test set = segment with number τ (n_{TEST} objects)

(b) Calibration set = other SEG_{TEST} - 1 segments (n_{CALIB} objects)

(c) Split calibration set into SEG_{CALIB} segments (typ. 4-10) of approximately equal size.

(d) Inner loop: **FOR** $\kappa = 1$ TO SEG_{CALIB}

(i) Validation set = segment with number κ (n_{VAL} objects)

(ii) Training set = other SEG_{CALIB} - 1 segments (n_{TRAIN} objects)

(iii) Make PRM models from the training set, with $a = 1, ..., a_{MAX}$ components

(iiii) Apply the PRM models to the validation set, resulting in $\hat{\boldsymbol{y}}_{CV}$ for the objects in segment κ for $a = 1, ..., a_{MAX}$

NEXT κ

(e) Estimate optimum number of components, a_{OPT}, from $\hat{\boldsymbol{y}}_{CV}$ of the calibration set by the “one-standard error” method (see below), giving $a_{OPT}(\tau)$ for this outer loop.

(f) Make PRM models from the whole calibration set for $a = 1, ..., a_{MAX}$ components

(g) Apply the models to the current test set, resulting in test-set

predicted $\hat{\boldsymbol{y}}$ for n_{TEST} test set objects and $a = 1, ..., a_{MAX}$ components.

NEXT τ

(3) After completing the outer loop, we have one test-set predicted $\hat{y}$ for each of the n objects for a_{MAX} different model complexities

NEXT ρ

(A.) After completing the repetition loops, a 3-dimensional data array consisting of test-set predicted $\hat{y}$ for each object, every repetition, and all considered numbers of components is available. The calculation of the corresponding prediction errors is straightforward.

(B.) The choice of a final optimum number of PLS components, a_{FINAL}, is based on all (SEG_{TEST} times n_{REP}) available values for a_{OPT} and picks the value with the highest frequency. Note that a_{FINAL} is determined without using test set predicted values.

(C.) Eventually, the residuals at a_{FINAL} for all the repetitions are summarized in the performance criterion SEP_{TRIM}.

Steps (f) and (g) are primarily important for model diagnostic plots, and could be omitted to speed up calculations. Having finished the rdCV procedure, a final regression model for future use is built from all objects with a_{FINAL} components. Unless this robust model is applied to new samples that are from a very different data population, the future prediction errors can be expected within the range of $\pm$ 2 SEP_{TRIM}.

Note: The optimum number of PLS components is based on a „one-standard error“ rule [3,13]. The error criterion used to estimate the optimum number of components, a_{OPT}, is the MSE_{TRIM}. With, for instance, seven segments in the inner rdCV loop (SEG_{CALIB}) there are seven MSE_{TRIM} values available. For each model complexity *a*, the mean as well as the standard deviation of MSE_{TRIM} is computed. The least complex model within one standard error of the best is chosen as optimum. The pseudo-code given for PRM also applies to PLS; except that non-trimmed MSE values are used within the one-standard error rule for PLS models.

3. SOFTWARE AND DATA

3.1. Software

The free software R offers an environment with focus on statistical data analysis and graphical representation [14]. It is licensed under the GNU General Public License (GPL) and available from the Comprehensive R Archive Network (CRAN). R is an open source programming language that is easily extended by freely available collections of functions (“packages”). The package “pls”, for instance, provides routines for principal component regression (PCR) and the partial least-squares regression (PLS) used within our rdCV functions [15]. We employ the rdCV function “mvr_dcv” for classical PLS as available in the R-package “chemometrics” [16]. The function manages the data and passes the settings for the nested loops in rdCV, e. g. number of segments for creation of test, calibration and validation sets. The package “chemometrics” also provides the partial robust M regression algorithm in the function “prm”. Additionally, a single k-fold cross-validation procedure for PRM is implemented in “prm_cv”. We developed “prm_dcv”, the robust counterpart to “mvr_dcv” employing the above-mentioned subfunctions for the robust method in the repetitive validation scheme. A typical call of both the classical and the robust PLS evaluated by rdCV is:

```
library(chemometrics)       # load package "chemometrics"
data(PAC)                   # load PAC dataset
```

```
class.result <- mvr_dcv(y~X, data=PAC, ncomp=20,
method="simpls")
rob.result <- prm_dcv(y~X, data=PAC, ncomp=20)
```

By default, no scaling of the data is provided; the number of repetitions is 100; the data set is split into four segments in the outer and seven segments in the inner loop. This call will consider models up to 20 PLS components. Further parameters of "mvr_dcv" and “prm_dcv” are explained in their help files.

In this work, all classical PLS models are run with 100 repetition loops. To reduce computational cost, the repetitions for all robust PRM models are reduced to 25. A comparison of computation time will be given in the Results section.

3.2. Data

The performance of robust and classical PLS models is compared by three different types of data that are intentionally contaminated with outliers.

ART. This artificial data set is intended to compare the modeling capability of PLS and PRM. The data set contains 500 samples described by 10 *x*-variables and a defined underlying model with fixed coefficients $\boldsymbol{\beta}$. The *x*-variables contain random numbers of a uniform distribution U(0,10) in the range of 0 to 10. The model coefficients are randomly drawn from the uniform distribution U(-

50,50). The corresponding response ***y*** results from a linear relationship defined by ***β*** (without intercept) and yields values between 0 and 1539.

PAC. The second data set is associated with quantitative structure-property relationship (QSPR). It is available in the R-package "chemometrics" by calling `data(PAC)`, and contains data for 209 polycyclic aromatic compounds. Each compound is characterized by 467 numerical descriptors of its approximated 3-dimensional molecular structure (*x*-variables) calculated by software Dragon [17,18]. The goal is to model the gas chromatographic retention index as dependent *y*-variable [19]. As the molecular descriptors are calculated from the molecular structures, they are not prone to experimental error. Outliers are more likely to appear in the response variable ***y***, the experimentally determined retention index, ranging from about 200 to 500.

NIR. The third data set is from 166 mash samples withdrawn from bioethanol fermentation experiments that varied with respect to enzymatic pretreatment and type of feedstock (wheat, corn or rye). The samples span the range of 22 to 88 g/L ethanol concentration, which was determined by HPLC and serves as the property of interest (*y*-variable). The first derivatives of near infrared (NIR) absorbance spectra in the wavelength range of 1100 to 2300 nm provide 235 *x*-variables for each sample [12]. In this data experimental errors are possible in both ***X*** and ***y***.

3.3. Creating Outliers

Preliminary regression tests showed the absence of strong outliers in the above-mentioned data sets. Hence, some perturbing observations are artificially generated.

Leverage points are constructed by substituting a fraction of the original *x*-variables with a considerably higher value ("***X*** outlier"). Therefore, the maximum of an *x*-variable x_j is calculated. According to equation (12), $x_{j,max}$ is then multiplied with a random value chosen from the uniform distribution U(3,10).

$$x_{ij,out} = x_{j,\max} \cdot \mathrm{U}(3,10) \qquad (12)$$

Typically, the first 3 % of the objects are contaminated; for instance, the first 15 samples in the artificial data set ART, having in total 500 objects. For generating the outlying observations, always and only the first three *x*-variables (j = 1, 2, 3) were contaminated. This choice is rather arbitrary for the PAC and NIR data set; however, with the ART data it gives some interesting insights for assessing the final model.

Vertical outliers are not contaminated in the ***X***-space, but their original values in ***y*** are changed ("***y*** outlier") according to equation (13), which worked well for the used data.

$$y_{i,out} = y_i \pm \mathrm{U}(y_{\max}, 2y_{mean}) \qquad (13)$$

An outlying observation $y_{i,out}$ is calculated by adding or substracting (randomly) a value drawn from a uniform distribution of values ranging between the

maximum, y_{max}, and twice the mean, y_{mean}. In case of a resulting negative value of $y_{i,out}$, it is set to zero instead.

A third type of outlier, belonging to the class of leverage points, is introduced by combination of data contamination in $\boldsymbol{X}$ and $\boldsymbol{y}$. The creation of these three types of outliers is shown schematically in Figure 2 for ART data.

4. RESULTS

4.1. ART

The artificial data set consists of 500 samples described by 10 regressor variables and one response. As long as no outlying observations contaminate the data, the classical PLS model performs equally to the robust PRM model. This is primarily reflected in the RMSE value for the estimated regression coefficients, which is zero for the classical PLS model, and close to zero for PRM. With respect to the prediction performance the classical model performs perfectly (SEP = 0), while the robust model is slightly worse with a SEP_{TRIM} of 2 (Table I).

If we add leverage points with errors in the first three *x*-variables, x_{i1}, x_{i2}, and x_{i3}, the classical PLS method estimates regression coefficients with considerable deviations from the real model parameters $\boldsymbol{\beta}$ (see Figure 3). Evidently, the regression coefficients having the largest deviations are associated with the first three *x*-variables. The RMSE of the estimated regression coefficients results in 10.1 for classical PLS, and yields 1.5 for the robust PRM method (Table I). Hence, the robust method succeeds in downweighting the influence of erroneous *x* data, even if the outliers are not completely excluded from the regression.

Once we introduce not only leverage points but also vertical outliers, the prediction performance of classical PLS deteriorates drastically with a SEP of

247. Comparing the trimmed SEP values of both considered methods, the classical model yields about ten times higher values (SEP_{TRIM} = 76) than the robust model (SEP_{TRIM} = 8). In the investigated data set, 9 % of the samples are outliers – 15 samples with errors in *x*, another 15 with errors in *y*, and additional 15 samples with errors in both *x* and *y*. In Figure 4 test set predicted *y* versus experimental (simulated) *y*-values for classical PLS (4a, b) and robust PRM (4c, d) are shown. All models discussed result from a repeated double cross validation procedure with an optimum model complexity determined by applying the one-standard error rule (see Section 2.3). For the classical PLS models, the validation yields 100 test set predicted *y*-values for every object. Due to the iterative reweighting loops in the robust method, the computational effort of PRM within the rdCV algorithm is elevated. Therefore, the number of repetitions for validation of the robust PLS models is reduced to 25, giving 25 test set predicted *y*-values for each sample. With the given rdCV settings the classical method takes 30 seconds, whilst the robust method takes 6 minutes for the ART data.

The predicted values $\hat{y}_i$ for all repetitions are included in Figure 4a and 4c as gray crosses, and give a picture of the variability of the predicted responses. The mean of all predicted values for each object is denoted by a black cross. In Figure 4a, the data points are notably spread, and the cloud containing the majority of data is systematically distorted from the 45° line. It would be difficult to select candidates for outliers based on this regression result. In contrast, the robust model in Figure 4c gives superior results with a distinct fit of the majority

of data to the optimum 45° line. Furthermore, the variation of predicted values with every repetition is smaller. In Figure 4b (PLS) and 4d (PRM) only the mean of the predicted values for each object is shown, and marked individually according to the type of outlier. It can be seen that the robust method computes exceptionally large residuals for some outlying observations, while the good data are fitted almost perfectly. The dangerous influence of outliers on classical PLS models, namely “pulling” a model towards their direction, can be observed in Figure 4b.

4.2. PAC

The second data set is composed of 467 molecular descriptors (*x*-variables) calculated from the 3-dimensional structure of 209 polycyclic aromatic compounds. As molecular descriptors will never be prone to experimental error, we assume error sources such as data manipulation errors, the inclusion of a partly wrong molecular structure, or modeling errors caused by choosing the wrong descriptor model. All outlying observations are computed according to the concept presented in Section 3.3. The outliers are allowed to be physical impossibilities, because in this work the focus is on observing effects of outliers rather than interpreting their physical meaning.

The contamination of this data set is exceptional in that it is designed to affect only samples with low values of the gas chromatographic retention index, here used as dependent *y*-variable.

Evidently, the prediction performance of the classical PLS models - measured by SEP - deteriorates with increasing number of outliers (Table II). While the original data set gives a SEP of 11, the strongly contaminated data give a ten times higher value. The robust model becomes slightly worse by adding more outliers too, but it is still in the same range of prediction quality as the classical model without outlying observations.

It is notable that the optimum number of PLS components decreases from 11 to 1 for the classical models, once outliers are present. One might claim a higher model complexity for a better prediction performance; for fairness, even a value comparable to the robust PRM models' optimum complexity (a_{FINAL} = 5 and 7, respectively). To confirm, we consult the available diagnostic plot on SEP as a function of the number of PLS components (Figure 5). The optimum complexity is marked with the vertical dashed line at one PLS component. With higher model complexity, e.g. a = 4, the mean value of SEP (black line) of all 100 repetitions (gray lines) is slightly lower. The price to be paid is having drastically larger variations in SEP for different test sets (repetitions), which indicates overfitting.

The included 12 leverage objects as well as 6 vertical outliers strongly affect the classical PLS model. First, data points denoted as ***X***-outliers in Figure 6a seem well fitted to the rest of the data. However, the regression line is definitely twisted by the leverage objects, especially because they form a strong cluster in the low value range of y and mask each other. A further effect of the outliers is

that the relationship between estimated and predicted *y*-values even indicates non-linearity in the data. The best achievable classical model has one PLS component. 95 % of the prediction errors for the gas chromatographic retention index are expected in the range of $\pm$ 2 SEP = 230, which is about 70 % of the mean retention index.

The robust method reveals all leverage objects and downweights the flawed observations with weights below 0.3 (Figure 7). These samples have only a small influence on the modeling process, while the “good” data prevail and allow for a good regression model. Consequently, the test set predicted values for outliers have large residuals (Figure 6b). Excluding these large absolute residuals from the performance criterion, we get a 95 % error range for future prediction errors of $\pm$ 2 SEP_{TRIM} = 30, which equals 8 % of the mean retention index *y*.

4.3. NIR

The third data set used comes from near-infrared spectroscopy measurements in 166 different liquid fermentation samples. Apart from a large range of ethanol concentration covered by the samples, they differ from each other with respect to the feedstock used and the enzymatic pre-treatment applied in the production process. Consequently, the original data may include observations from different statistical populations, which might show outlying behavior in the regression. Additionally, 15 outliers are created on purpose. As for the PAC data, the optimum model complexity decreases with increasing number of

outliers, in particular for the classical PLS model (Table III). Using more than two PLS components for the classical method promotes over-fitting of the outliers present in the calibration data, and gives even worse results for the prediction quality. In Figure 8a the ***X***-outliers are well fitted to the data majority, while observations with errors in ***y*** and errors in both ***X*** and ***y*** are easily detectable. A systematic deviation of points being perpendicular to the 45° line can be observed for “good” samples with high values in experimental ***y***. As they are not conspicuous in regression results of the original data (without contamination added), we encounter the so-called swamping effect. Because of the presence of outliers, good data is incorrectly fitted.

The total weights assigned to each NIR sample are displayed in Figure 9. The robust method detects most of the introduced outliers (sample 1 to 15). In contrast to the results presented for the other two data sets, the influence of ***X***-outliers in NIR data is reduced only moderately. These outliers are influenced by a group of samples with slightly different multivariate data structure (samples 16 to 52), which are withdrawn from experiments with the particular feedstock rye.

4.4. Summary

We compared a classical PLS regression method with the partial robust M regression method by application of repeated double cross validation. The rdCV procedure is published in a previous work with focus on classical PLS (SIMPLS) [11], and it is freely available in the package “chemometrics” for the R programming environment. For this study, we extended the rdCV algorithm to robust PLS regression (PRM) to provide a common ground for a fair and careful comparison of both the considered methods.

Even if the main settings for the rdCV procedure (e.g., number of segments for creating test, calibration, and validation sets; maximum number of considered PLS components) are the same for both regression methods, the robust method PRM needs more time for computation due to the iterative adjustment of weights. The compromise chosen in this work is to reduce the number of repetitions from 100 to 25 for PRM. The computational effort for a typical data set in chemometrics, such as the NIR data set, is two minutes for the classical PLS model and 30 minutes for the robust PLS model. Since PRM yields better prediction results in all investigated data sets with outliers being present, the higher computation time is justified. However, the rdCV procedure can be accelerated by omitting some calculation steps only necessary for model diagnostics plots.

Repeated double cross validation is a reliable validation technique that provides a realistic estimation of the models’ prediction performance, and allows

optimizing the model complexity. Apart from profound model diagnostic plots made available by rdCV, the weights plot calculated by PRM is useful for data inspection, in particular for the detection of outliers. We present three data sets with different characteristics. The artificial data ART is simulated following a perfect linear relationship between *x* and *y* variables, and contain neither errors nor noise. The PAC data is likely to contain experimental errors and/or noise in the *y*-variables only. The most realistic chemical data set NIR is prone to experimental error and noise for both *x*- and *y*-variables.

5. CONCLUSIONS

Whenever outliers are probable in the data the application of a robust method should be considered. The main advantage of the presented robust PRM method including repeated double cross validation is that no outlier detection is necessary prior to model creation, and a realistic estimation of the model's future performance is made available. If no outliers are present in the data, the robust method is practically as good as the classical method. It is shown for artificial data that the true underlying model parameters are estimated correctly by PRM; in particular with aberrant observations being present in the calibration data the robust methods clearly outperforms classical PLS. Consequently, the robust models give better prediction results for non-outliers than the classical models. Nevertheless the problem of detecting outliers in new data remains. A straightforward way is to perform robust autoscaling in $\boldsymbol{X}$ (equation (5)) for both the new data and data used for model creation, and then calculate the robust weights w_i^x (equation (6)). Other more sophisticated robust outlier detection methods are available, see for example [20].

Acknowledgements

This work was partly funded by the Austrian Research Promotion Agency (FFG), BRIDGE program, project no. 812097/11126. We thank Anton Friedl (Vienna University of Technology, Institute of Chemical Engineering) for encouragement and support.

REFERENCES

1. de Jong S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* 1993; **18** (3): 251-253.
2. Martens H, Naes T. *Multivariate calibration*. Chichester, United Kingdom: Wiley; 1989.
3. Varmuza K, Filzmoser P. *Introduction to multivariate statistical analysis in chemometrics*. Boca Raton, FL: CRC Press; 2009.
4. Serneels S, Croux C, Van Espen PJ. Influence properties of partial least squares regression. *Chemom. Intell. Lab. Syst.* 2004; **71**: 13-20.
5. Wakeling IN, Macfie HJH. A robust PLS procedure. *J. Chemom.* 1992; **6** (4): 189-198.
6. Cummins DJ, Andrews CW. Iteratively reweighted partial least squares: A performance analysis by Monte Carlo simulation. *J. Chemom.* 1995; **9** (6): 489-507.
7. Hubert M, Vanden Branden K. Robust methods for partial least squares regression. *J. Chemom.* 2003; **17** (10): 537-549.
8. Serneels S, Croux C, Filzmoser P, Van Espen PJ. Partial robust M-regression. *Chemom. Intell. Lab. Syst.* 2005; **79**: 55-64.
9. Hössjer O, Croux C. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *J. Nonparametr. Stat.* 1995; **4**: 293-308.
10. Serneels S, Filzmoser P, Croux C, Van Espen PJ. Robust continuum regression. *Chemom. Intell. Lab. Syst.* 2005; **76** (2): 197-204.
11. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. *J. Chemom.* 2009; **23** (4): 160-171.
12. Liebmann B, Friedl A, Varmuza K. Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Anal. Chim. Acta* 2009; **642**: 171-178.
13. Hastie T, Tibshirani RJ, Friedman J. *The elements of statistical learning*. New York, NY: Springer; 2009.

14. R. *A language and environment for statistical computing*. Vienna, Austria: R Development Core Team, Foundation for Statistical Computing, www.r-project.org; 2009.
15. Mevik BH, Wehrens R. The pls package: Principal component and partial least squares regression in R. *J. Stat. Software* 2007; **18** (2): 1-24.
16. Filzmoser P, Varmuza K. chemometrics: Multivariate statistical analysis in chemometrics. R package version 0.5. Vienna, Austria: http://cran.r-project.org; 2009.
17. Todeschini R, Consonni V. *Handbook of molecular descriptors*. Weinheim, Germany: Wiley-VCH; 2000.
18. Todeschini R, Consonni V, Mauri A, Pavan M. *Dragon. Software for calculation of molecular descriptors, by Todeschini R, Consonni V, Mauri A, Pavan M.* Milan, Italy: Talete srl, www.talete.mi.it; 2004.
19. Lee M, Vassilaros D, White C, Novotny M. Retention indices for programmed-temperature capillary-column gas chromatography of polycyclic aromatic hydrocarbons. *Anal. Chem.* 1979; **51**: 768-773.
20. Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. *Comput. Stat. Data Anal.* 2008; **52** (3): 1694-1711.

Table I.

Regression results for ART data (n = 500, m = 10) with classical PLS and robust PRM regression. The data is contaminated with varying number of outliers in $\boldsymbol{X}$, $\boldsymbol{y}$ and both $\boldsymbol{X}$ and $\boldsymbol{y}$. RMSE indicates to what extent the considered method estimates the real model parameters correctly. Ideally, RMSE is 0. The models' prediction performance is assessed by SEP and SEP_{TRIM}, respectively, based on test set predicted errors. The optimum model complexity, a_{FINAL}, is determined by rdCV.

no. of outliers in			**RMSE**		$\boldsymbol{a_{FINAL}}$		**SEP**	$\boldsymbol{SEP_{TRIM}}$	
X	*y*	*X* & *y*	PLS	PRM	PLS	PRM	PLS	PLS	PRM
0	0	0	0.0	0.1	8	4	0	0	2
25	0	0	10.1	1.5	4	3	124	67	11
0	25	0	7.2	2.0	3	3	190	19	6
25	25	0	10.3	1.9	3	3	231	71	16
15	15	15	11.0	0.7	3	3	247	76	8

Table II.

Regression results for PAC data (n = 209, m = 467) with classical PLS and robust PRM regression using original data as well as data including vertical outliers and leverage points. The optimum model complexity, a_{FINAL}, is determined by rdCV. The models' prediction performance is assessed by SEP and SEP_{TRIM}, respectively, based on test set predicted errors.

no. of outliers in			a_{FINAL}		**SEP**	**SEP_{TRIM}**	
X	*y*	*X* & *y*	PLS	PRM	PLS	PLS	PRM
0	0	0	11	14	11	6	6
0	10	0	1	5	108	29	14
6	6	6	1	7	115	28	15

Table III:

Regression results for NIR data (n = 166, m = 235) with classical PLS and robust PRM regression. The original data might contain experimental error and noise, and is further contaminated with each 3 % of outliers in ***X***, ***y*** and both ***X*** and ***y***. The optimum model complexity, a_{FINAL}, is determined by rdCV. The models' prediction performance is assessed by SEP and SEP_{TRIM}, respectively, based on test set predicted errors.

no. of outliers in			a_{FINAL}		**SEP**	SEP_{TRIM}	
X	*y*	*X* & *y*	PLS	PRM	PLS	PLS	PRM
0	0	0	14	15	2	1	1
5	5	5	2	5	20	5	4

Figures

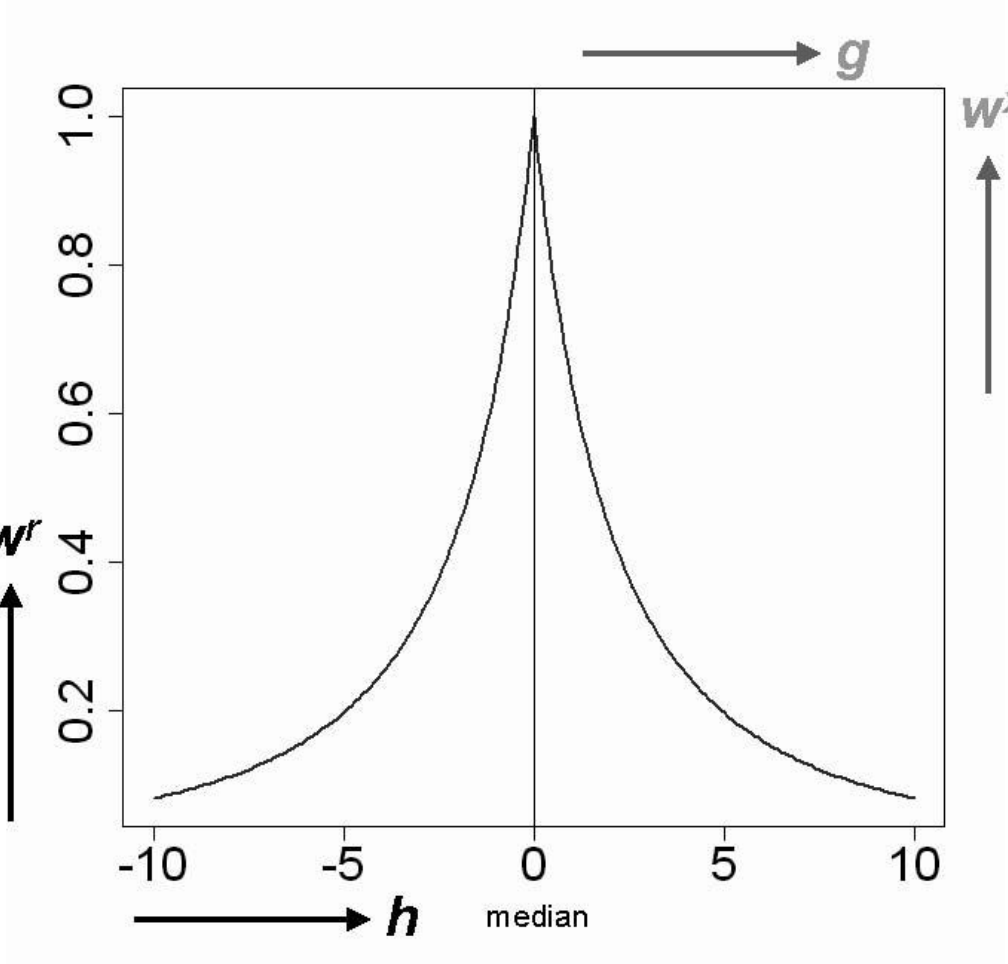

Figure 1.

Weight w^x (or w^r) to account for outliers - calculated by the Fair function. The origin can be considered the robust data center (median). Observations far from the origin are downweighted by weights much smaller than 1.

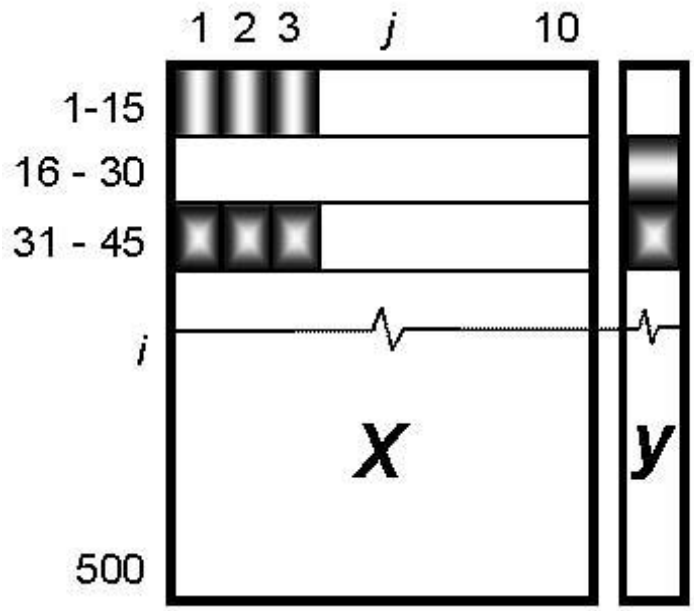

Figure 2.

Scheme for creating outliers, shown for the ART data set. The first three *x*-variables are changed for creation of leverage points, while the *y*-value is unchanged (sample 1-15). Additional 15 observations are contaminated in the *y*-value (vertical outliers, sample 16-30). Samples 31-45 have outliers in both x_i and y_i.

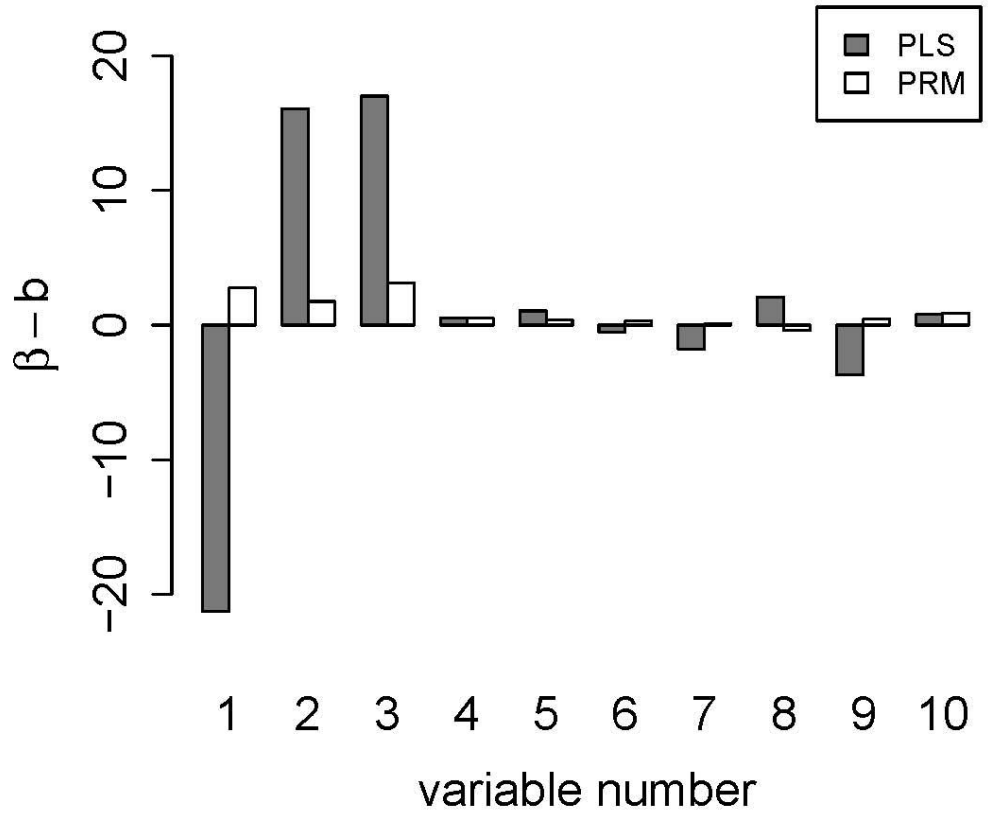

Figure 3.

5 % of outlying observations in the first three *x*-variables of the simulated data set ART severely influence the estimated regression coefficients (b_1, b_2, b_3) of the classical PLS model. Deviations to the real model parameters $\boldsymbol{\beta}$ are considerably lower for the robust PRM model.

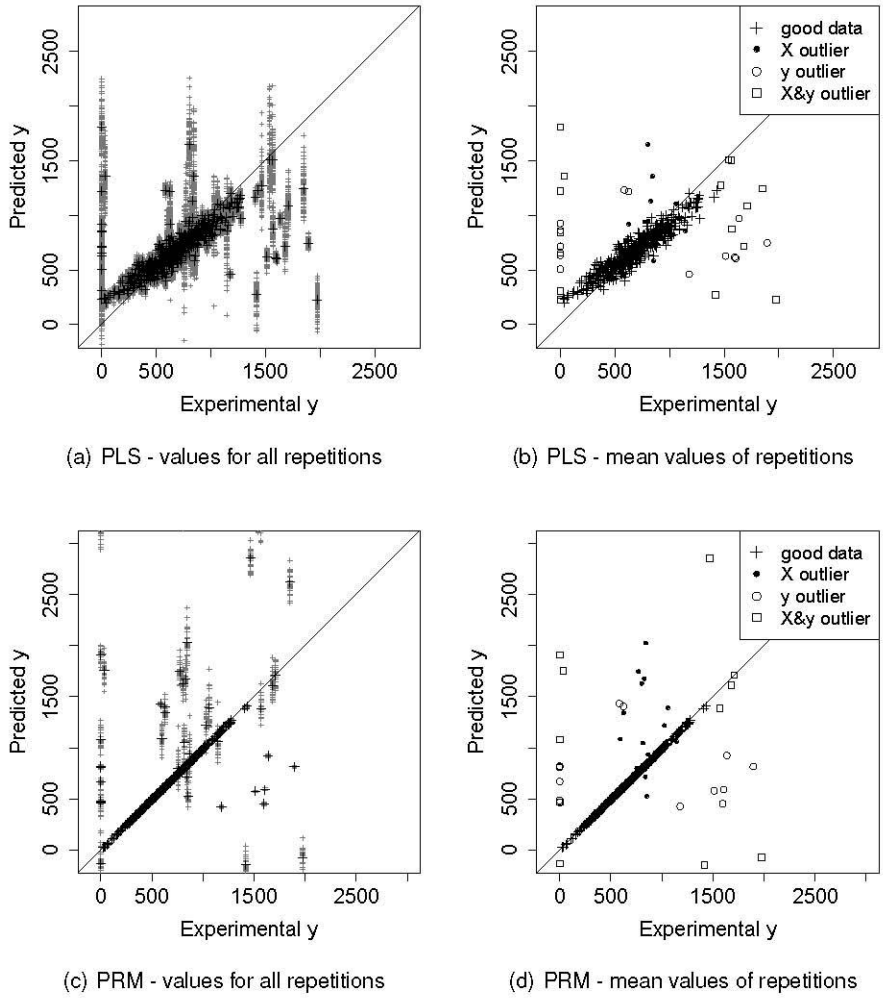

Figure 4.

Test set predicted *y* versus experimental (simulated) *y*-values for classical PLS (a, b) and robust PRM (c, d) for the simulated data set ART. The predicted *y* for every repetition is included in (a) (100 repetitions) and (c) (25 repetitions) as gray cross; the mean of all repetitions is denoted by black crosses. Equivalently, these mean values with good data and outliers marked individually are shown in (b, d). In all plots, a 45° line is included as target line.

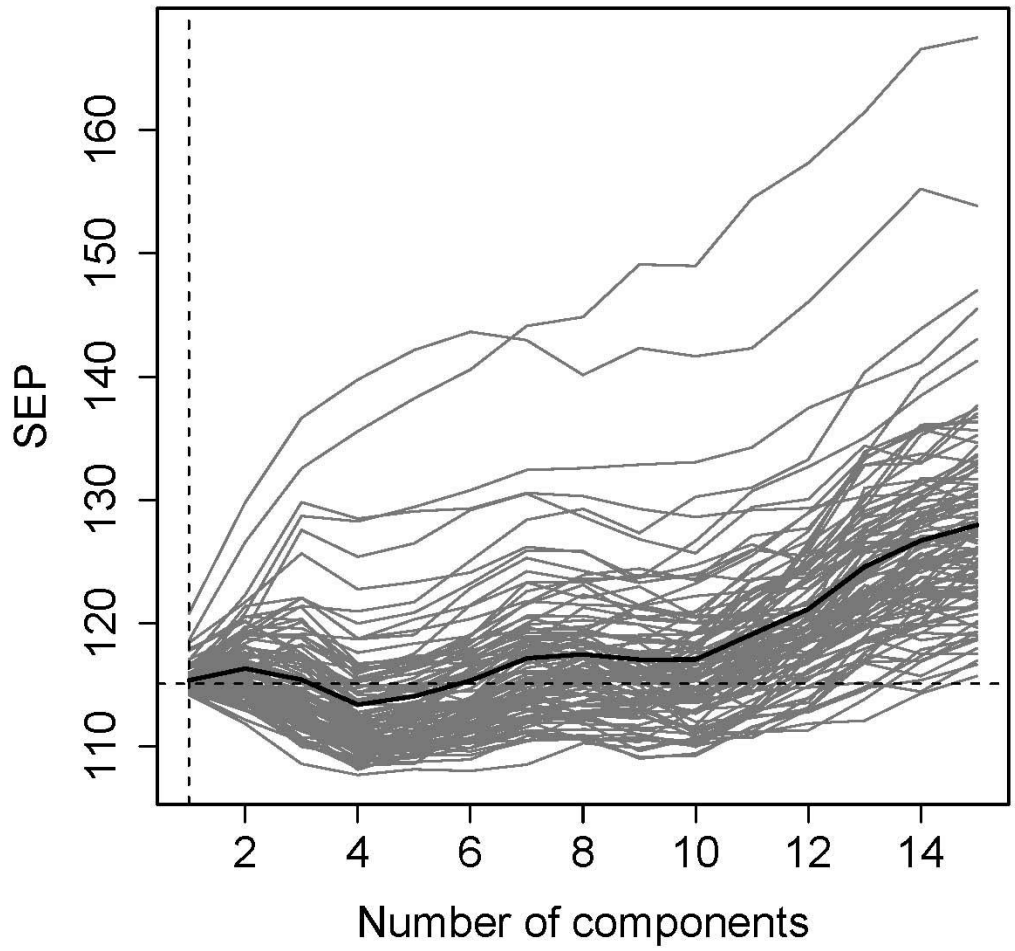

Figure 5.

SEP of the classical PLS model as a function of the number ob PLS components for PAC data with 18 outlying observations. Gray lines are for each of 100 repetitions. The black line is the mean of the 100 repetitions. The dashed lines indicate the computed optimum at one PLS component with SEP = 115. Higher number of components yields much larger variation in the prediction errors for different repetitions.

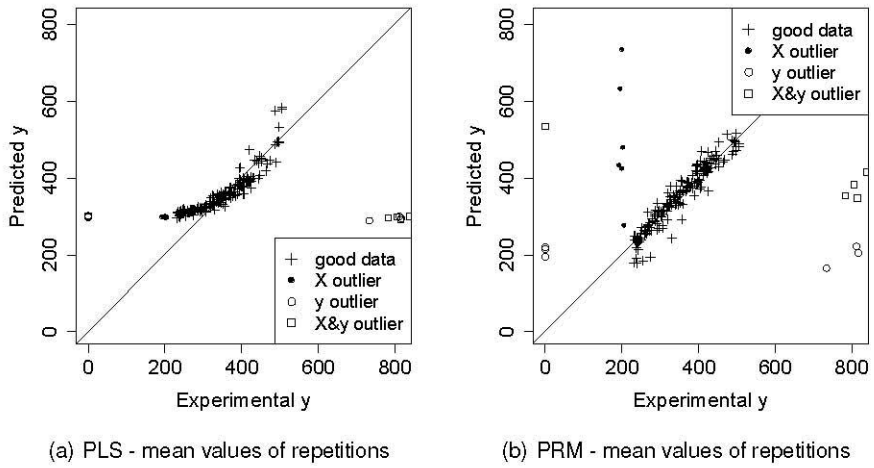

Figure 6.

PAC data with 18 outliers: Test set predicted *y* versus experimental *y*-values for classical PLS (a) and robust PRM (b). The predicted *y*-values are means of 100 repetitions for PLS and 25 repetitions for PRM, respectively. Good data and different types of outliers are marked individually.

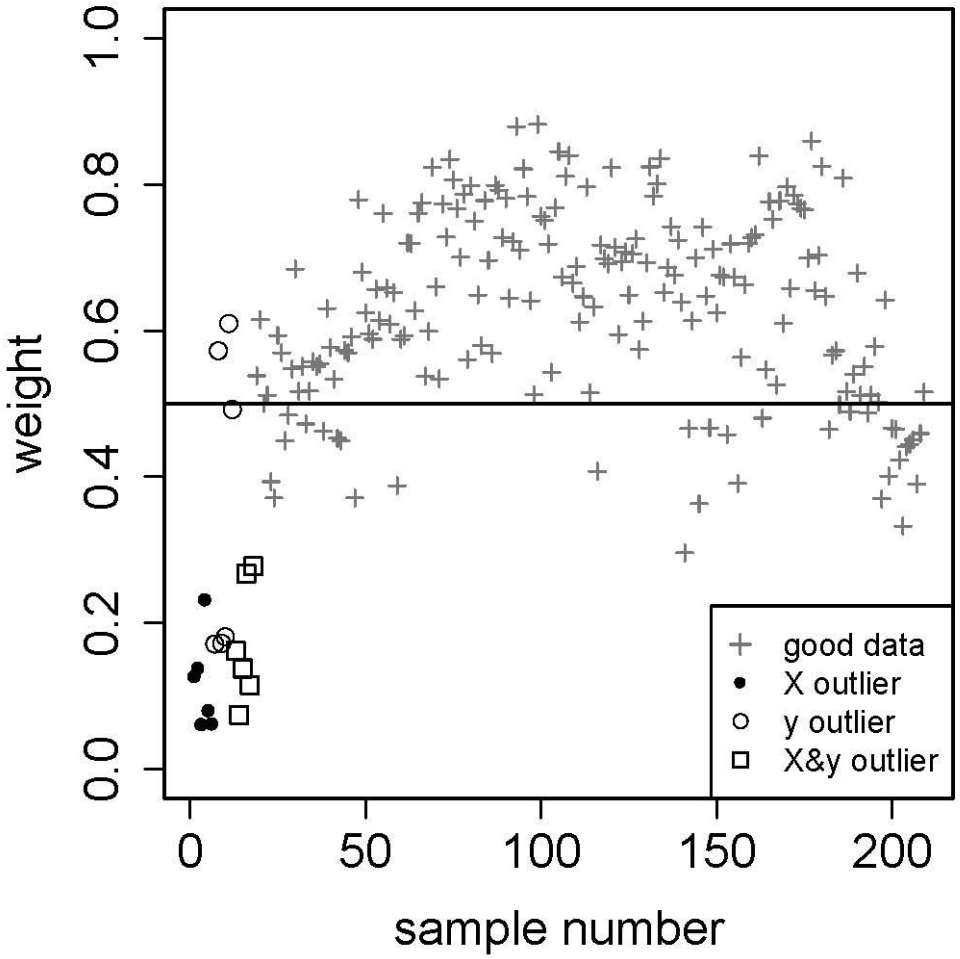

Figure 7.

Total weight w_i assigned to each PAC sample by the robust PRM method. The first 18 samples are outliers created intentionally; they are unveiled and downweighted by PRM. Three *y* outliers, however, are found close to the data majority.

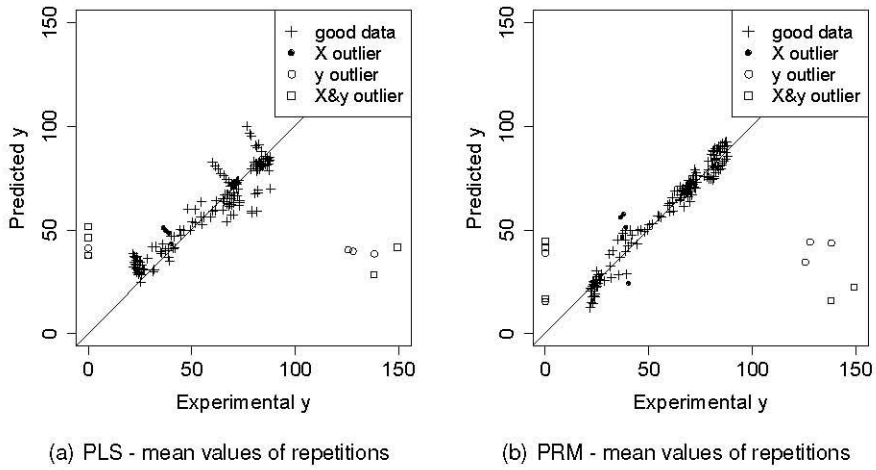

Figure 8.

NIR with 15 outliers: Test set predicted *y* versus experimental *y*-values for classical PLS (a) and robust PRM (b). The predicted y-values are means of 100 repetitions for PLS and 25 repetitions for PRM, respectively. Good data and different types of outliers are marked individually.

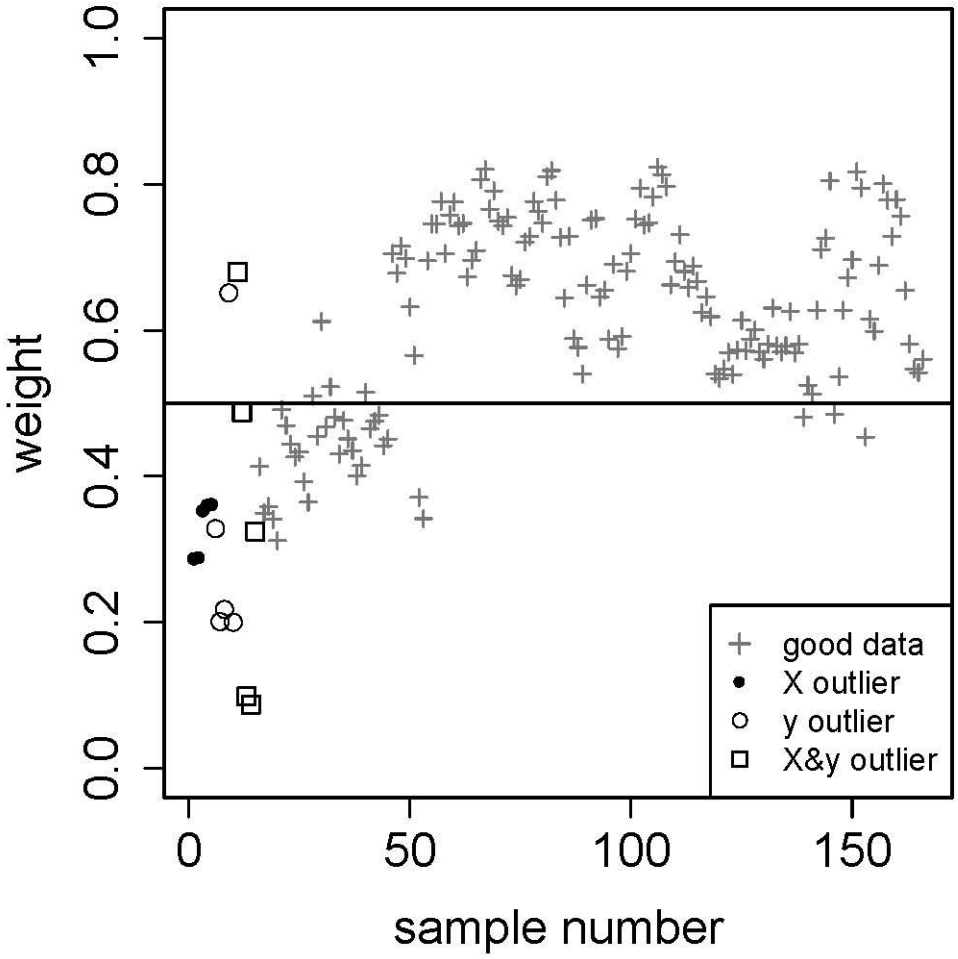

Figure 9.

Total weight w_i assigned to each NIR sample by the robust PRM method. The first 15 samples are outliers created intentionally that are mostly unveiled and downweighted. In addition, the regular samples 16 to 52 appear as outliers, too. Indeed, all these samples are withdrawn from experiments with a particular feedstock.