Robust methods for multivariate data analysis

S. Frosch Møller1*, J. von Frese2 and R. Bro2

1Department of Seafood Research, Danish Institute for Fisheries Research, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
2Spectroscopy and Chemometrics Group, Quality and Technology, Department of Food Science, The Royal Veterinary and Agricultural University, DK-1958 Frederiksberg C, Denmark

Received 24 February 2005; Revised 06 December 2005; Accepted 27 January 2006

Outliers may hamper proper classical multivariate analysis, and lead to incorrect conclusions. To remedy the problem of outliers, robust methods are developed in statistics and chemometrics. Robust methods reduce or remove the effect of outlying data points and allow the ‘good’ data to primarily determine the result. This article reviews the most commonly used robust multivariate regression and exploratory methods that have appeared since 1996 in the field of chemometrics. Special emphasis is put on the robust versions of chemometric standard tools like PCA and PLS and the corresponding robust estimates of regression, location and scatter on which they are based. Copyright © 2006 John Wiley & Sons, Ltd.

KEYWORDS: outliers, robust estimation, PCA, PCR, PLS

1. INTRODUCTION

Outliers are observations that appear to break the pattern or grouping shown by the majority of the observations. Presence of outliers is more the rule than the exception for real world data. Many branches of chemometrics in both industry and research work with huge amounts of data, which makes visually based evaluation and screening for outliers difficult. The reasons for outliers are various, for example instrument failure, non-representative sampling, formatting errors and objects stemming from other populations. Usually, only complete objects (\(x_i\)) are considered as outliers, but it is equally relevant to look for outliers in variables (\(x_j\)) and even individual data elements (\(x_{ij}\)). Most conventional multivariate methods are sensitive to outliers due to the fact that they are based on least squares (LS) or similar criteria where even one outlier can have an arbitrarily large effect on the estimate and deteriorate the model. Therefore, it is necessary to (1) identify outliers and (2) decide whether the outliers should be accommodated or rejected in the modeling process.

The aim of any robust method is to reduce or remove the effect of outlying data points and allow the remainder to predominantly determine the results. Robust methods are helpful for both semi-automated detection of outliers by looking at the robust residuals and model building. When no outliers are present in the data set, the result from a robust method should be consistent with the result from the corresponding non-robust method. Robust methods provide a powerful methodology extending a conventional ‘manual’ analysis and elimination of outliers by using exploratory methods and ‘conventional’ outlier diagnostics.

Rousseeuw and Leroy [1] presented an overview of robust estimates in regression and outlier detection, and Maronna and Yohai [2] described recent advances in robust estimation in multivariate location and scatter estimation. Much focus has been put on making the common chemometric techniques such as principal component analysis (PCA), principal component regression (PCR) and partial least squares (PLS) regression more robust against outliers, using robust estimates to replace the non-robust LS estimate. Reference [3] holds a review of the robust methods for multivariate analysis until 1996. An overview of the recently developed methods for multivariate data analysis, based on the minimum covariance determinant and least trimmed squares estimators for location, scatter and regression, together with a detailed description of these estimators, can be found in Reference [4].

The aim of this paper is to present an overview of the most common robust chemometric methods, that is PCA, PCR and PLS, described in the literature as many new methods have emerged subsequently.

In Section 2 outliers and their effect on least squares estimation will be discussed. Section 3 introduces the robust estimates for regression, location and covariance used in the robust multivariate methods discussed in Section 4. Section 5 contains comments on software availability. Finally, Section 6 presents a discussion on the use of robust methods.
2. OUTLIERS

Different types of outliers can be discerned. Taking regression models as an example where the independent variables are denoted as \( X_{inp} \) (\( n \) stands for the number of objects \( i = 1, \ldots, n \)) and \( p \) for the original number of variables \( (j = 1, \ldots, p) \) and the \( q \) dependent variables are denoted as \( Y_{nq} \), the following categories of outliers can be considered: (1) ‘Good’ leverage points which are observations isolated from the major part of the observations in the data matrix \( X \) but following the same regression model, (2) ‘Bad’ leverage points which in addition to being isolated from the major part of \( X \) deviate strongly from the regression model defined by the other observations and (3) Outliers that are not leverage points but have large \( y \) prediction residuals in calibration and are therefore referred to as high \( y \) residual outliers or vertical outliers. In robust analysis, the good leverage points are usually not denoted as outliers as they are not detrimental to the regression model but merely reflect an ‘unfortunate design’. These three types of outliers can occur both during model fitting and during predictions with a previously established model.

Figure 1 shows a scatterplot of 10 points, \((x_1, y_1), \ldots, (x_{10}, y_{10})\), with no outliers presented. The LS solution fits the data very well, and for data with normally distributed random noise without outliers, the LS solution is in fact optimal in the maximum likelihood sense.

Why LS is not resistant to outliers follows from the properties of the objective function for LS procedures. The objective function to be minimized is the sum of the squared residuals

\[
\text{Minimize} \sum_{i=1}^{n} r_i^2 \tag{1}
\]

in which the residuals \( r_i \) are given by

\[
r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip} \tag{2}
\]

where \( y_i \) \((i, \ldots, n)\) are the corresponding values of the dependent variables, \( x_{ij} \) \((i = 1, \ldots, n; j = 1, \ldots, p)\) the values of the explanatory variables, and \( \hat{\beta}_j = (j = 1, \ldots, p) \) is the LS estimate of the parameters. This means that a large outlier will exert an inappropriately large influence on the LS-estimate as will be illustrated in the following:

Figure 2 illustrate the three outlier types where high \( y \) residual outliers are marked with a ‘1’, ‘2’ represent good leverage points and bad leverage points are marked with ‘3’. Both the high \( y \) residual outliers and the bad leverage points affect the calibration model by distorting the least squares model to a certain degree and should be eliminated. From a purely experimental point of view an observation with a high \( y \) residual does not necessarily indicate a corrupted or deviating \( y \) measurement, we only know that the corresponding \((x, y)\) pair is inconsistent with the remainder (although from a pure theoretical viewpoint the explanatory variable \( x \) is considered to be error-free). In multivariate regression models (e.g. PCR or PLS) it might be possible to assign such an outlier as originating from \( X \) in case it shows a strong deviation from the \( X \)-model. Generally, outliers are not necessarily bad measurements but could also indicate samples belonging to another group than the majority of the data.

As noted by Gnanadesikan [5], the consequence of outliers in multivariate data is intrinsically more complex than in the univariate case. A multivariate outlier can distort measures of location and scale and thereby also those of covariance structure. As a result the modeling methods may describe the shape of the majority of the data incorrectly and conclusions drawn can be misleading. An added complication is that it is much more difficult to identify multivariate outliers. A single univariate outlier may easily be detected graphically, which is not that straightforward in higher dimensions. Many multivariate methods work well for identifying single outliers, but when there are many outliers, masking and swamping effects may occur. The masking effect means that some outliers are unnoticed because the presence of other outliers masks their bad influence [6,7]. The swamping effect consists of wrongly identifying / diagnosing

\[
\text{Figure 1. Scatterplot of 10 points, } (x_1, y_1, \ldots, (x_{10}, y_{10}) \text{ and the LS regression line.}
\]

\[
\text{Figure 2. High } y \text{ residual outliers (1) and leverage points (Good leverage points are denoted ‘2’ and bad leverage points are denoted ‘3’).}
\]
an observation as an outlier because of the presence of other outliers [8].

According to Rousseeuw and Leroy [1] outlier diagnostics and robust methods have the same goal just in opposite order: In standard or non-robust diagnostic approaches the outliers are first identified and then the remaining data analyzed with a non-robust LS criterion. In robust methods, models are fitted to the majority of the data and outliers are identified as those observations with large residuals from the robust fit. A survey of diagnostic techniques can be found in Reference [1].

3. BASIC ROBUST STATISTICS

To enable the comparison of different robust methods in various situations, measures of performance are necessary. One such performance measure for robust methods is the breakdown point introduced by Hampel [1]. Another essential performance measure is the influence function. For assessing the influence function, distributional assumptions for the data have to be made. This often renders the analysis more intricate and might necessitate empirical comparisons with unknown general validity in particular for \( n < p \) (e.g. [11]).

Efficiency is another important concept for the discussion of robust estimates. The relative statistical efficiency is the ratio of the mean square error from a robust estimator to the mean square error from an ordinary LS estimator when applied to data from an uncontaminated distribution, for example with normally distributed errors [6].

Equi-variance properties are also important for understanding estimators. Equi-variance means that a systematic transformation of the data will cause a corresponding transformation of the estimator [1]. Three types of equi-variance exist for regression estimators: (1) regression-, (2) scale and (3) affine equi-variance. Regression equi-variance means that any (additional) linear dependence \( Y \to Y + Xv \) should be reflected in the regression vector accordingly \( b \to b + v \).

\(^1\)Robust estimates are marked by asterisk (*) throughout the article.

This corresponds to translation equi-variance in the case of location estimation. Scale invariance implies that the fit is essentially independent of the choice of measurement units for the response variable \( y \) and for any scaling \( y \to cy \) the regression vector scales appropriately \( b \to cb \). Affine equi-variance means that for linearly transformed data \( X \to AX \) the estimate of the regression vector transforms correspondingly \( b \to A^{-1}b \), such that the predicted values and residuals are invariant under this transformation.

A weaker condition than affine equi-variance is the orthogonal equi-variance, which means that any orthogonal transformation of the data (rotation and reflection) transforms the estimator properly. Orthogonal equi-variance is sufficient in the context of PCA or PLS since even the classical procedures are only orthogonally equi-variant [12]. For a detailed description of the different equi-variance criteria the reader is referred to Reference [1].

Multiple linear regression, as well as estimation of sample mean and covariance are the cornerstones of multivariate data analysis methods such as PCA, PCR and PLS [1,2]. The former underlying techniques are not resistant to outliers as they are based on LS techniques and robustifying them is often the basis for obtaining robust versions of the latter multivariate data analysis methods. The focus throughout this section is restricted to robust multivariate regression estimators and robust estimates of multivariate location and covariance used in the multivariate methods described afterwards in this paper.

Overviews of the important robust multidimensional estimators for regression and location and scatter are listed in Tables I and II, respectively.

3.1. Robust multivariate regression estimates

3.1.1. Multiple linear regression

In multiple linear regression (MLR), the response variable \( y_i \) is regressed on \( p \) explanatory variables \( (x_{1i}, \ldots, x_{pi}) \) in the model

\[
y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i
\]

with errors \( \varepsilon_i \). As in univariate regression, the LS estimator of \( \beta_0, \beta_1, \ldots, \beta_p \) which corresponds to minimizing the squared residuals (Equation (1)), is quite sensitive to the presence of outlying points. A robust alternative to LS estimator is therefore needed.

3.1.2. M-estimates

The methods collected under the term M-estimators (maximum likelihood type estimators), first introduced by Huber [13,14], replace the squared residuals in Equation (1) by another function of the residuals

\[
\text{Minimize} \sum_{i=1}^{n} \rho(r_i/s)
\]

The function \( \rho \) is symmetric (i.e. \( \rho(-t) = \rho(t) \)) for all \( t \) with a unique minimum at zero, \( r_i \) is the residual of the \( i \)th observation and \( s \) is a suitable estimate of the scale obtained from the residuals. For \( \rho(t) = t^2 \) and \( s = 1 \) one obtains the LS estimator. Different choices of \( \rho(t) \) correspond to assuming
## Table I. Overview of the most important robust estimators for univariate location and scatter in robust multivariate data analysis

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\epsilon$ (%)</th>
<th>Comments</th>
<th>Affine equivariant</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>0%</td>
<td>Not robust against X-outliers</td>
<td>Yes</td>
<td>[13,14]</td>
</tr>
<tr>
<td>GM</td>
<td>≤30%, decreases as $p$ increases</td>
<td>Robust with respect to outliers in X as well as outliers in Y</td>
<td>Yes</td>
<td>[1]</td>
</tr>
<tr>
<td>Siegel's repeated median</td>
<td>50%</td>
<td>No</td>
<td>[15]</td>
<td></td>
</tr>
<tr>
<td>LMS*</td>
<td>50%</td>
<td>Yes</td>
<td>[16]</td>
<td></td>
</tr>
<tr>
<td>LTS</td>
<td>50%</td>
<td>Yes</td>
<td>[16]</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>50%</td>
<td>Yes</td>
<td>[17]</td>
<td></td>
</tr>
<tr>
<td>MM</td>
<td>50%</td>
<td>Yes</td>
<td>[18,19]</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations are explained in the text.

*slow convergence ($\propto n^{-1/3}$).

## Table II. Overview of the most important robust estimators for multivariate location and scatter in robust multivariate data analysis

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\epsilon$ (%)</th>
<th>Comments</th>
<th>Equivariance properties</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>$\leq 1/(p+1)$</td>
<td>Not all M-estimates are affine equivariant, for example $L_i$</td>
<td>Method dependent</td>
<td>[20]</td>
</tr>
<tr>
<td>MVT</td>
<td>50%</td>
<td>Converge quickly</td>
<td>Affine</td>
<td>[21,22]</td>
</tr>
<tr>
<td>Stahel–Donoho</td>
<td>50% for $n$ larger than $2p+1$</td>
<td>Limitation: $n &gt; p$ after trimming</td>
<td>High computational cost</td>
<td>[23,24]</td>
</tr>
<tr>
<td>MVE</td>
<td>50%</td>
<td>Converges slowly ($n^{-1/3}$)</td>
<td>Affine</td>
<td>[16]</td>
</tr>
<tr>
<td>MCD</td>
<td>50%</td>
<td>Limitation: $n &gt; p$</td>
<td>Affine</td>
<td>[16]</td>
</tr>
<tr>
<td>S</td>
<td>50%</td>
<td>Limitation: $n &gt; p$</td>
<td>Affine</td>
<td>[25,26]</td>
</tr>
<tr>
<td>MM</td>
<td>50%</td>
<td>Limitation: $n &gt; p$</td>
<td>Affine</td>
<td>[27]</td>
</tr>
</tbody>
</table>

Abbreviations are explained in the text.

specific distributions for the errors (e.g. $\rho(t) = |t|$ for double exponentially distributed errors) [28].

M-estimator calculations can be computed by means of iteratively reweighted least squares (IRLS). The principle of IRLS is to obtain a weight for each observation depending on the size of the regression residual. Such weights make it possible to bound the effect of outliers on the final model. The scale factor is necessary to achieve affine equivariance for the estimators. A possible scale estimate for the residuals would be the standard deviation. However, the standard deviation is not robust to outliers and therefore not suitable in this context. The median absolute deviation (MAD) is a commonly used robust alternative to the standard deviation [1]. Rousseeuw and Croux [29] proposed the $Q_a$ scale estimate as an alternative to the MAD. The $Q_a$ estimator, motivated by the Hodges–Lehman estimator [30], is for a univariate data set $(z_1, \ldots, z_n)$ defined as the first quartile of the pairwise differences between the data $Q_a = 2.2219 \cdot d \cdot \left| z_i - z_j \right|_{i \leq j}$ where $d$ is a small sample correction factor (approaching 1 for increasing $n$) [29]. When comparing the efficiency of several robust scale estimators Rousseeuw and Croux [29] found that the $Q_a$ estimator yielded better results than the MAD. Unfortunately, the M-estimators can be strongly influenced by any high leverage point and thus have a break down point $\epsilon$ of 0% [1].

### 3.1.3. GM-estimates

Generalized M-estimators (GM-estimators) frequently referred to as ‘bounded influence estimators’ were developed to overcome the x-outlier problem of M-estimators [1] and thereby improve the breakdown point.

The basic purpose of these methods is to bound the influence of outlying $x_i$ by means of some weights $w_i$ that give full influence to observations assumed to come from the main body of the data, but reduced weight or influence to outlying observations. Starting with a sensible estimate the iterative procedure will continue until the sequence of estimates has converged to within the desired accuracy. The $\epsilon$ of all GM-estimators can be no better than a certain value, in general not larger than 30%, decreasing as a function of the dimension $p$ [17,31].

A number of weights have been proposed, for example Tukey’s biweight [32], Huber [13,14], Hampel [33] and Andrew’s wave [33,34]. These weights are not restricted to GM-estimators but can be used for all kind of estimates requiring a weight function. Plots of selected weight functions are illustrated in Figure 3.

### 3.1.4. Siegel’s repeated median

The first high breakdown point regression estimator was the repeated median (RM) proposed by Siegel [15]. It is based on...
calculating perfectly fitting models for all possible data subsets of size $p$ and obtaining a final regression model through a nested coordinatewise median calculation (see Reference [15]).

As the explicit calculation of the RM would involve the consideration of all possible subsets of $p$ points, where $p$ is the number of variables, the resulting computational complexity of $n^p$ [15] would mean a prohibitive amount of calculation time even for moderate $p$ [1]. Additionally, this estimator is not affine equivariant for linear transformations of $x_i$ [1].

### 3.1.5. Least median of squares

Replacing the sum of the squared residuals in Equation (1) with the robust median yields one of the most well-known instances of a high breakdown point estimator, the least median of squares (LMS) method of Hampel and Rousseeuw [16], defined by

$$\text{Minimize} \quad \text{median}_{i=1}^n r_i^2 \quad (5)$$

The LMS estimator has an $\varepsilon$ of 50%, and is robust with respect to outliers in $y$ as well as outliers in $X$. Unfortunately, the LMS has a very low efficiency (converges like $n^{-1/3}$) [16].

### 3.1.6. Least Trimmed Squares

To overcome the poor convergence rate for LMS, Rousseeuw [16] proposed the least trimmed squares (LTS) estimator

$$\text{Minimize} \quad \sum_{i=1}^h (r_i^2)_{1:n} \quad (6)$$

where $(r_i^2)_{1:n} \leq \ldots \leq (r_i^2)_{h:n}$ are the ordered squared residuals that is the sum of squared residuals is formed over a suitable lower quantile $h/n$ of the residuals. The fraction of included samples can be as low $h = n/2$. The LTS objective is equivalent to LS, with the exception that the largest residuals are not used in the sum, so that outliers are disregarded. Taking $h = n$ yields the LS estimator. The highest possible breakdown value (50%) for LTS is attained when $h = n/2$. For general $h$, the breakdown is $(n - h + 1)/n$. The LTS converges like $n^{-1/2}$ and behaves satisfactorily with respect to asymptotic efficiency [1]. The disadvantage of LTS is the sorting of the squared residuals, which blows up an already significant computation time [16]. A fast algorithm, FAST-LTS, for computing the LTS estimators was developed by Rousseeuw and van Driessen [35]. For small data sets the exact LTS is found whereas for larger data set the new algorithm gives more accurate estimates than existing LTS algorithms. A general problem of the FAST-LTS and other approximate algorithms with a given number of starting trial sets consists in a lack of consistancy for larger and larger training data sets as pointed out by Hawkins and Olive [36]. As possible solution a specific prior clustering has been suggested in [36] and FAST-LTS already makes uses of a more sophisticated empirical subsampling scheme for larger datasets but further research is needed [37].

The $\varepsilon$ for both LTS and LMS is independent of $p$, the number of variables and they also satisfy all three equivariance properties mentioned earlier [1].

Rousseeuw and Yohai [17] generalized the LMS and LTS estimators to S-estimators. The class of S-estimators corresponds to replacing the scale of the residuals in Equation (4) by a robust measure that minimizes the dispersion of the residuals. S-estimators are regression-, scale- and affine equivariant and possess a convergence rate $n^{-1}$. The $\varepsilon$ can attain 50% with a suitable choice of the constants involved, and in contrast to GM-estimators, S-estimators have a high breakdown point for any dimensionality. But S-estimators cannot simultaneously achieve high efficiency and a high breakdown point [1,17]. If a 50% breakdown point is imposed, the asymptotic Gaussian efficiency of S-estimators is at most 33% [38].

Computing the exact S-estimator is often not feasible, and may present difficult problems due to the existence of many local minima [18]. The computational cost of methods based on subsampling increases exponentially with $p$, and makes these estimates very costly for high dimensions.

To circumvent the efficiency problem of S-estimators Yohai introduced MM-estimators [18], which are basically efficient M-estimators using the result of an S-estimator as starting value and for obtaining a robust auxiliary scale estimate and thereby obtaining improved robustness with an $\varepsilon$ up to 50%. Thus, MM-estimators combine a high breakdown point with a high efficiency, requiring the same (high) computation time as S-estimators and are for example the ‘official recommendation’ for robust regression in the statistics software S-Plus [39].

The solution to the multivariate regression problem can also be reformulated in terms of the joint location $\bar{\mu} = (\bar{\mu}_x, \bar{\mu}_y)$ and scatter matrix $C = (C_{xx}, C_{xy})$ of the explanatory and dependent variables since the LS estimators of intercept and slope can be written as functions of the joint location and scatter matrix [40]. Thus, a robust regression estimate can also be obtained by applying robust estimators of location and scatter instead of the LS estimates. Following this approach, Rousseeuw et al. [40] recently derived the MCD regression as robust regression method using the MCD estimator of multivariate location and scatter (see Section 3.2. below). The resulting robust regression estimator has the appropriate equivariance properties, a bounded influence function, and inherits the breakdown value of the MCD.
estimator. In order to improve the rather low efficiency a reweighting scheme was proposed [40].

3.2. Robust estimates of multivariate location and covariance (scatter)
For a data set $X$ of $n$ points in $p$ dimensions the most well-known estimator of the multivariate location is the arithmetic mean

$$T(X) = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(7)

which can also be viewed as an LS estimator because it minimizes

$$\sum_{i=1}^{n} \|x_i - T\|^2$$

(8)

where $\|\ldots\|$ is the $L_2$ norm. The breakdown point is 0%. The maximum likelihood estimator for the population covariance matrix $C$ is defined as

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} (x_i - T)(x_i - T)^T$$

(9)

for $x_i$ the $i$th row of the $(n \times p)$ data matrix $X$, and $T$ the arithmetic $(1 \times p)$ mean vector. Robust estimates for multivariate location and covariance will be described in the following.

3.2.1. M-estimates of location and covariance
A generalization of M-estimators to multivariate location is given by

$$\text{Minimize} \sum_{i=1}^{n} \rho(||x_i - T||)$$

(10)

where $T$ can be regarded as the location estimate [1]. These estimates are not necessarily affine equivariant as the example of the $L_1$ location estimator shows

$$\text{Minimize} \sum_{i=1}^{n} ||x_i - T||$$

(11)

The $L_1$ location estimator also known as the ‘spatial median’ or ‘median center’ is a generalization of the univariate median and its breakdown point is 50% [13,41]. The $L_1$ estimator only satisfies the weaker condition of orthogonal equivariance.

Affine equivariant M-estimates of multivariate location and scatter were formally proposed by Maronna [20]. A major drawback of these is that the breakdown point of affine equivariant M-estimators is at most $1/(p+1)$, that is relatively low for even a moderately high number of variables [23,42,43]. Furthermore, Devlin et al. [44] found that M-estimators in practice could tolerate even fewer outliers than indicated by this upper bound. In addition Wisnowski et al. [45] found empirically that the usefulness of M-estimates of covariance for detecting multiple outliers is limited to low-dimensional, low-density scenarios.

3.2.2. Stahel–Donoho Estimator
The first affine equivariant multivariate location and scatter estimator with a high breakdown point was the Stahel–Donoho estimator (SDE) or ‘outlyingness-weighted median’ [23,24]. For each $x_i$ in $X$, one looks for a one-dimensional projection in which $x_i$ is most outlying in the sense defined as

$$r_i = \sup_{\|v\|^{-1}} \left| \frac{x_i v - \text{med}(x_i v)}{\text{med}(x_i v)} \right|$$

(12)

Where $\text{med}(x_i v)$ is the median of projections of all data points $x_i$ on the direction of the vector $v$. The location and scatter are then estimated by the weighted mean and the weighted covariance matrix with weights of the form $w(r)$ where $w$ is a strictly positive and decreasing weight function of $r \geq 0$. The estimator is related to projection pursuit since one principally searches over all possible projections $v$ with $r_i$ as projection index. Due to the high computational cost, when calculating the SDE in its exact form, approximate methods with subsampling procedures are commonly used [46]. The breakdown properties of the MAD (i.e. the denominator in Equation (13) basically determine the breakdown of the SDE and corresponding modifications have been suggested to obtained further improvements in this respect [47,48].

3.2.3. Multivariate trimming
Ellipsoidal multivariate trimming (MVT) was proposed by Gnanadesikan and Kettenring [21] and Devlin et al. [22]. In each step of this iterative procedure the squared Mahalanobis distance ($d^2$) of the observation vectors $x_i$ from the current robust estimate of location $\bar{x}$, are measured in the metric of $C^*$, the current robust estimate of the covariance matrix of the $X$ data. A specified percentage (the trimming percentage) of the most extreme observations (i.e. objects with the largest $d^2$) is temporarily set aside (max. 50% of the observations) and the remaining observations are used to compute $\bar{x}$ and $C^*$ exactly as $\bar{x}$ and $C$, the sample mean vector and covariance matrix. A number of samples with highest $d^2$ corresponding to the trimming percentage is again set aside and the process is repeated with the remaining samples. The iterative process terminates when both $\bar{x}$ and $C^*$ converge. The effect of trimming is that observations with large distances do not contribute to the calculations for the remaining observations. In most approaches the starting values for $\bar{x}$ and $C^*$ are taken to be $\bar{x}$ and $C$, even though robust starting values may appear as natural choices for $\bar{x}$ and $C$. Empirically, MVT has been found to converge quickly, usually in two or three steps [44,49]. Devlin et al. [44] claimed that $\epsilon$ of MVT was the same as its trimming percentage (max. 50%), and does not decrease with the number of variables. However, Donoho [24] argued that $\epsilon$ of MVT is at most about $1/p$, thus rendering this method less attractive due to its low breakdown point. In its original version the MVT procedure can be applied only if the number of objects in the $X$ matrix after trimming exceeds the number of variables. This limitation can be avoided by applying it to the score matrix $T$ from a PCA model of the data matrix $X$ [49], but in this case the result also depends on the robustness properties of the applied method for obtaining the PCA model and furthermore the affine equivariance of a conventional covariance estimate would be lost.

3.2.4. Minimum volume estimator

Another affine equivariant high breakdown point estimator of multivariate location and covariance is the minimum volume estimator (MVE) [16,50]. The objective of the MVE is

\[ T(X) = \text{Center of the minimum volume ellipsoid covering (at least) a fraction of } h \text{ points of } X \]

(13)

where \( h \) can be taken as low as \((n/2) + 1\). The location estimator is then given by the center of the ellipsoid. The corresponding covariance estimator is defined as the covariance matrix of the ellipsoid multiplied by a suitable correction factor to obtain consistency with the multivariate normal distribution. The \( e \) can be the highest possible namely 50% as \( n \to \infty \). The algorithm will have a slow convergence rate similar to the LMS estimate, \( n^{-1/3} \) [1]. Furthermore, the algorithms suffer from inefficiency and high computational complexity, making it impractical for use with large data sets [51].

3.2.5. Minimum covariance determinant

The objective of the minimum covariance determinant (MCD) estimator of multivariate location and scatter is [16,50]

\[ T(X) = \text{Mean of the } h \text{ points of } X \]

(14)

for which the determinant of the covariance matrix is minimal

The MCD seeks the \( h \) points out of the whole data set (\( n \) objects) for which the classical tolerance ellipsoid (for a given level) has a minimum volume among all possible subsets of size \( h \). Then the location and scatter estimates are given by the mean and covariance matrix for this optimal subset \( h_n \). The covariance matrix has to be scaled by a consistency factor \( c_h \) in order to obtain a consistent estimator for the multivariate Gaussian distribution. In the univariate case this corresponds to the least trimmed squares estimator where each data point receive a weight of one if it belongs to the robust confidence interval and zero otherwise.

The limitation of this method is that the number of observations should be larger than the number of variables, if \( p > h \) the covariance matrix of any \( h \)-subset has a zero determinant and thus cannot be minimized. Therefore, high dimensional data sets should first be reduced by variable selection or by using principal components. The breakdown point is the highest possible when \( h = 0.5n \). For a better compromise between efficiency and breakdown value \( h \) should be \( \approx 0.75n \) (\( \approx 25\% \)) [52]. A fast algorithm for calculating the MCD estimator (FAST–MCD) has been derived [52]. For small data sets it finds the exact MCD, whereas for larger data sets it is claimed to give more accurate results than alternative methods at the time of development [52].

If the outlier contamination can be estimated \textit{a priori}, \( h \) can be conveniently reformulated in terms of the trimming percentage \( \alpha \) (where \( 0 < \alpha \leq 0.5 \)); one can replace \( h = \lfloor n (1 - \alpha) \rfloor + 1 \) in Equation (14) and (15). The breakdown point of these estimators is equal to \( \alpha \). For \( \alpha \to 0 \) the MVE yields the center of the smallest ellipsoid covering all the data, whereas the MCD objective tends to the arithmetic mean [50].

Butler et al. [53] show that MCD has better statistical properties than MVE since MCD is asymptotically normal and further that MVE has a slower convergence rate (\( n^{-1/3} \)) [54]. Other authors have also noted the theoretical superiority of MCD to MVE [52,55]. The FAST–MCD is also an order of magnitude better than all MVE algorithms in terms of computational complexity [45].

The statistical efficiency of the MCD estimator can be increased by implementing a reweighting estimator [52,56]. After obtaining the raw MCD estimators of location (\( \hat{\mu}_{\text{MCD}} \)) and scatter (\( \hat{\Sigma}_{\text{MCD}} \)) each observation receives a weight \( w_i \); which is zero if its robust distance, \( \hat{d}(x_i, \hat{\mu}_{\text{MCD}}, \hat{\Sigma}_{\text{MCD}}) = \sqrt{(x_i - \hat{\mu}_{\text{MCD}})^T \hat{\Sigma}_{\text{MCD}} (x_i - \hat{\mu}_{\text{MCD}})} \) exceeds \( \sqrt{X_{n, 0.975}} \). The reweighted MCD estimator is then defined as the weighted mean and covariance matrix.

3.2.6. S-estimators of multivariate location and scatter

Davies [25] and Lopuhaä [26] extended multivariate regression S-estimates to multivariate location and covariance estimates. The S-estimator of multivariate location and scatter is defined as that vector \( T \) and positive definite symmetric matrix \( C \) which minimize det \( |C| \) subject to a limit on the magnitudes of the corresponding Mahalanobis distances, \( d_i = \sqrt{(x_i - T)^T C^{-1} (x_i - T)} \)

\[ \frac{1}{n} \sum_{i=1}^{n} \rho(d_i) = b_0 \]

(15)

where \( \rho(\cdot) \) is symmetric and \( \rho(0) = 0 \). The constant \( b_0 \) is often taken as the expected value of \( \rho(d_i) \) assuming a multivariate Gaussian distribution [57].

These S-estimators are affine equivariant, asymptotically normal, and for well-chosen \( \rho(\cdot) \) their breakdown points can be as high as 50% [1].

4. ROBUST MULTIVARIATE MODELS

Almost all robust multivariate models in the literature and described in the following, work under the assumption that outliers are samples (rows in the data matrix). That is all data of one sample is treated as one observation. Hence, these models aim at identifying and minimizing the influence of individual outlying samples. The situation where, for example an individual element in the data matrix \( x_{ij} \) is considered as outlying has not gained much attention in the literature. First approaches for this situation have been provided by Hawkins et al. [58], Liu et al. [59] and Croux et al. [60].

4.1. Robust principal component analysis

Principal component analysis (PCA) is often the first step of the data analysis, typically followed by a more quantitative analysis using PCR, PLS, discriminant analysis or other multivariate techniques. Classical PCA involves computing the eigenvectors and eigenvalues of the sample covariance or correlation matrix. Simple but powerful algorithms have been developed for finding the principal components with the singular value decomposition (SVD) as the most widespread [61].
In the context of PCA, an outlier can be defined as an observation/object that either lies far away from the subspace spanned by the correct k eigenvectors, and/or for which the projection into the model lies far from the remainder of the data within the subspace [62].

Several ways of robustifying principal components have been proposed. They can be grouped as follows:

(1) Techniques that replace the classical covariance matrix by a robust covariance via robust estimators of location and shape such as the MCD. Calculating the eigenvalues and eigenvectors of this robust covariance matrix provides eigenvectors that are robust to sample outliers. These approaches are limited to relatively low-dimensional data.

(2) Projection pursuit (PP) searches for structure in high dimensional data by projecting these data into a lower-dimensional space which maximizes a robust measure of spread called the projection index, \( \rho(\mathbf{X}) \), for example using the MAD. PP methods obtain robust eigenvector estimates by explicitly solving the maximization (or minimization) problem: find the direction (eigenvector) \( \mathbf{v}_p \) that maximizes \( \rho(\mathbf{Xv}) \) [63]. In subsequent steps, each new direction is constrained to be orthogonal to all previous directions. The procedure results in robust principal components and a robust covariance matrix. Classical PCA, which uses the variance as projection index is a special case of the PP algorithm. Since the principal components are computed sequentially, this approach can handle high dimensional data, \( n < p \).

(3) A Combination of (1) and (2). This approach can handle high-dimensional data.

(4) Adjustments to the internal computations of the SVD algorithm by replacing the LS criterion with a robust criterion. This approach can handle high-dimensional data [60] and elemental outliers [58].

Table III lists some of the robust PCA methods belonging to group (2) and (3) discussed in this paper.

Replacement of the classical covariance or correlation matrix by one of the abovementioned or other robust estimators is perhaps the simplest and also an intuitively appealing approach.

Applications using GM-estimators of means and covariances can be found in Campbell [34], Rivest and Plante [70] and Daigle and Rivest [71]. These methods are less suitable for high dimensionalities due to the fact that \( \varepsilon \) decreases towards zero with increasing dimensionality.

Many simulation studies, starting with Devlin et al. [44] have been carried out to find out which robust estimator should be used for estimating a covariance/correlation matrix and its principal components. More recently, Croux and Haesbroeck [72] have shown that the one-step reweighted MCD method [50] and the S-estimator of location and shape are well suited for robustifying the estimate of the covariance matrix. The theoretical results as well as the simulations favor the use of S-estimators, since they combine a high efficiency with appealing robustness properties. The results are more robust than results obtained with GM-estimators, but are unfortunately limited to small dimensions [72]. This limitation is a severe restriction in chemometrics where it is usually important to have robust PCA methods for situations with \( p > n \). A second problem is the computation of these robust estimators in high dimensions. Today’s fastest algorithms [52,73] can handle up to about 100 variables, which might pose a substantial practical limitation in many chemometric applications.

Other approaches to robustify PCA, based on PP, have been considered by Ruyymgaard [74], Li and Chen [64], Ammann [75], Galpin and Hawkins [7], Xie et al. [65], Croux and Ruiz-Gazen [66], Hubert et al. [67] and Hubert et al. [68].

The projection index in the method proposed of Li and Chen [64] was chosen as Huber’s M-estimator of scale, whereas Galpin and Hawkins [7] and Xie et al. [65] replace the variance norm by a robust measure of the spread based on optimizing the \( L_1 \) norm (e.g. MAD). Xie et al. [65] introduced the generalized simulated annealing algorithm as an optimization procedure in the process of PP calculation to pursue the global minimum. Li and Chen [64] proved that the resulting method inherits the breakdown value of the robust scale estimator and are qualitative robust. These methods can handle high dimensional data, where the number of variables exceeds the number of samples [67].

Unfortunately, the algorithm of Li and Chen [64] has a high computational cost [67]. In Croux and Ruiz-Gazen [66] a computationally more advantageous method was presented (C-R algorithm) with the \( L_1 \)—median estimator as the center of the data and the \( Q_n \) scale estimator [29] as projection index. To speed up the computation the directions to be investigated have been restricted to all directions that pass through the center and a data point. Unfortunately, the

**Table III. Robust high-dimensional PCA methods discussed in this paper**

<table>
<thead>
<tr>
<th>Name</th>
<th>Robust estimate</th>
<th>References</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP-PCA</td>
<td>Weighted mean (center of the data)</td>
<td>[64]</td>
<td>High computational complexity</td>
</tr>
<tr>
<td></td>
<td>Hubers M-estimate (projection index)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PP-PCA</td>
<td>( L_1 ) (center of the data)</td>
<td>[65]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAD (projection index)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PP-PCA</td>
<td>( L_1 ) (center of the data)</td>
<td>[12,66]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( Q_n ) (projection index)</td>
<td>[67]</td>
<td>Fast to compute</td>
</tr>
<tr>
<td>RA-PCA</td>
<td>( L_1 ) (center of the data)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( Q_n ) (projection index)</td>
<td>[67]</td>
<td></td>
</tr>
<tr>
<td>ROBPCA</td>
<td>Combined PP with reweighted MCD</td>
<td>[68]</td>
<td>A fast alternative algorithm exist, ROBPCA-( k_{max} )[69]</td>
</tr>
</tbody>
</table>

Outliers are considered as whole samples.
algorithm has a numerical accuracy problem in high dimensions (large \( p \)) due to round-off errors and is still computationally costly [61]. Hubert et al. [67] proposed an improvement of the C-R algorithm, a modified two-step version, called reflection-based algorithm for robust principal component analysis (RAPCA). According to the authors the method is more stable numerically and computationally much faster (when PCA compression to the rank of the data is performed as a first step), and can deal with both low and high dimensional data. The compression step in RAPCA can also be built in the C-R algorithm and thereby improve the speed. By doing so no difference with respect to computation speed were observed for up to 15 components [76]. An improved version of the C-R algorithm, which should not suffer from the numerical problem in high dimensions can be found in Croux and Ruiz-Gazen [12].

The most attractive advantages of the PP-procedures are that they yield covariance matrices and PCs that are of both orthogonal equivariance and show a high \( \varepsilon \), which can be as high as 50%, and does not depend on the dimensionality [64].

A different type of approach is the robust singular value decomposition method described by Ammann [51]. In the calculation of the SVD successive transposed QR decompositions are used where the corresponding regression steps have been replaced by their robust alternatives. Mallow’s GM-estimator [11] together with an iterative update of location by M-estimation has been suggested for the computations. Final estimates are obtained by an ordinary SVD on the weighted covariance matrix, where weights have been obtained within the previous QR decompositions.

Hubert et al. [68] introduce a new method for robust PCA called ROBPCA where PP ideas are combined with robust location and covariance estimation in lower dimensions. After preprocessing the data, PP is used for initial dimension reduction \((k < p)\), and the reweighted MCD estimator is then applied to this lower dimensional data space to find a robust centre and a robust covariance estimator of the projected samples. Dimension reduction is necessary because the MCD estimator is only applicable for \( p < n \) due to singularity of the covariance matrix when \( h < p \). Finally these estimates are back transformed to the original space and a robust estimate of the location of \( X \) and of its scatter are obtained. This method can handle both low and high dimensional data, and according to the authors, it produces more accurate estimates for non-contaminated data and more robust estimates for contaminated data. The ROBPCA method is orthogonal equivariant [68]. Unfortunately, this algorithm is not very fast when the results for several principal components \((k = 1, \ldots, k_{\text{max}})\) are required as it needs to run the whole procedure for each component separately [69]. A faster alternative algorithm, ROBPCA-\( k_{\text{max}} \) for moderate data sets of sizes up to 100 and \( k_{\text{max}} = 10 \) is proposed by Engelen et al. [69]. This algorithm is less precise and more time consuming if \( k_{\text{max}} \) is chosen too large because it computes the MCD estimator for \( k_{\text{max}} \) dimensions in one run. When comparing the ROBPCA and ROBPCA-\( k_{\text{max}} \) algorithms Engelen et al. [69] found that ROBPCA-\( k_{\text{max}} \) performed almost as well as the original ROBPCA. Only for very particular contaminations the ROBPCA-\( k_{\text{max}} \) estimator was unfavorable compared to ROBPCA.

All methods discussed so far consider entire samples, \( x_i \), as outliers but there are many examples where individual data elements, \( x_{ij} \), in otherwise good rows, \( x_i \), are corrupted, for example in fluorescence, microarray or proteomics data and image analysis. Methods that can handle this kind of outliers have been proposed by Hawkins et al. [58], Liu et al. [59] and Croux et al. [60]. These methods are based on the alternating least-squares algorithm proposed by Gabriel and Zamir [77]. In this algorithm the minimization problem is solved with criss-cross regressions, which involve iteratively computing dyadic (rank 1) fits using LS (similar to NIPALS). The original Gabriel-Zamir SVD algorithm is then rendered robust by using robust regression estimates such as \( L_1 \) [58], LTS [59] or weighted \( L_2 \) (weights based on MVE) [60] instead of the ordinary LS regression in the SVD algorithm.

### 4.1. Robust principal component regression

#### 4.1.1. Multiway methods

For the \( N \)-way methods Pravdova et al. [78] proposed a robust version of the three-way Tucker3 model or multimode PCA. The three-way data cube, \( X \), is decomposed into a number of components as in PCA but as opposed to PCA the number of components can be different for the three modes (i.e. the dimensions or directions). The method is based on robust initialization of the Tucker3 algorithm using MVT or MCD. The three-way data are unfolded to two-way matrices and the loadings are obtained by SVD. In this version the outliers are identified in the first mode only (\( A \)), but as all modes are treated symmetrically, one can look for outliers in any mode. First a so-called clean subset is constructed using MVT or MCD. A clean subset means that the data set contains no detected outliers. In each iteration of the ALS subroutine, the loadings \( A \), \( B \) and \( C \) for the different modes are calculated for the clean subset of objects only. The loadings \( A \) are then predicted for all objects and the data is reconstructed with the predefined number of factors. Residuals between the initial data and the reconstructed data are calculated and sorted, and 51% of objects with the smallest residuals are selected to form the clean subset for the next ALS iteration. The objective function to be minimized is the sum of squared residuals for the clean objects from the first mode. Once the robust Tucker3 model is constructed the outliers are identified by a robust estimate of the standardized residuals based on the scale estimator MAD. The final Tucker3 model is constructed as the least squares model for the data after outlier elimination. Empirically studies show that for 20% outliers in the dataset, the robust Tucker3 model converged to a good solution but for 40% outliers, the algorithm had problems for some type of outliers. According to authors MCD is a better algorithm for finding the clean subset than MVT [78], which can be understood from the deteriorating breakdown properties of MVT with increasing dimensionality [24].

#### 4.2. Robust principal component regression

Principal component regression (PCR) can be used to build regression models for rank deficient regressors \( X \) [62]. The principal components are obtained by decomposing \( X \) via PCA and subsequently the response(s) is regressed onto the score, \( T \).

Both steps of the PCR procedure can be affected by outliers because regression as well as PCA is based on least squares
fitting. The bad leverage points of the PCA model are important as they deteriorate the estimation of the T matrix, influencing in this way also the second step of PCR. The regression step of the PCR can additionally be hampered by outliers in y. It is thus important to robustify PCR both with respect to the PCA step as well as the subsequent multiple linear regression. Such robust PCR methods which make both steps robust have been presented by Walczak and Massart [49, 79], Filzmoser [80], Hubert and Verboven [81] and Zhang et al. [82].

Table IV contains an overview of the different robust PCR methods discussed in this paper.

The robust PCR (RPCR) method proposed by Walczak and Massart [49] uses robust PCA based on a robust covariance matrix (MVT), followed by LMS regression applied on all samples. The robust covariance matrix and thus the robust principal components provide the background to reveal outliers, either good or bad leverage points, in the X data while standardized residuals from the robust model are applied to detect outliers in both X and y. The breakdown point of MVT is the same as its trimming percentage (max. 50%) and does not decrease with the number of variables. However, Donoho [24] argued that the e of MVT is at most about 1/p, thus rendering this method less attractive due to its low breakdown point. The breakdown point of LMS is the highest possible, namely 50%. The lack of efficiency for LMS can be regarded as a possible disadvantage of this method, as the LMS minimizes the median of the residuals (which is equivalent with consideration of only 50% of the data) instead of the sum of the squared residual. Therefore Walczak and Massart [49] recommend that the model should be used as an outlier detection tool and not the final building method for the model.

Filzmoser [80] introduced a robust method for PCR based on the idea of PP proposed by Li and Chen [64] to obtain robust principal components, combined with LTS regression for prediction. The algorithm is the C-R algorithm [66] with the median absolute deviation from the median (MAD) estimator as projection index. Due to the fact that that first principal components (PC) with the largest variance are not necessarily the best predictors, a step-wise selection of PC’s is performed for subset (k < p) selection of PC’s for prediction. The selection starts with the PC resulting in the best prediction of the response variables (according to an appropriate association measure). The selection process continues until the quality of prediction cannot be increased significantly.

In the method of Hubert and Verboven [81] first a robust PCA method is applied on the regressors. For low-dimensional data (p < n), the MCD estimator [16] is applied as a robust estimator of the covariance matrix of X, and for high-dimensional data (p > n) the ROBPCA method [68] is used. Next a robust regression method is applied. If there is only one response variable the LTS regression estimator [16] is preferred, otherwise the MCD regression estimator [40] is utilized. All the robust estimates in this method are based on a hard thresholding, that is binary weights 0/1 are applied. The rather empirical approach of Zhang et al. [82] relies on robust regression diagnostics of the predictions for identifying outliers before building an ordinary LS-PCR model with the good data. For regression diagnostics so called ‘principal sensitive vectors’ (RPPSV) [83] are used, that is an analysis of how the predictions for a sample change when each training sample is omitted in turn. In addition, samples with significantly high prediction residuals are also omitted. The estimates computed by the procedure are affine equivariant and the convergence rate is n⁻¹ [83].

In the method proposed by Pell [79], ‘resampling by halves’ (RHM) [84] is used to detect outliers in the response matrix, X, in order to remove those samples from consideration before the PCA step. RHM is based on repeatedly estimating the Euclidean distance of each sample to the mean of a randomly drawn subset of 50% of the data. In case X consists of variables on different scales, autoscaling is first applied, using the standard deviations from the respective 50% subset. The frequency with which each sample is among the 5% most distant observations is used as the criterion for outlier detection. The PCA decomposition is performed without those extreme samples but the extreme samples are projected onto the PCA space and the scores from the extreme samples are used with the rest of the calibration samples in the regression step. Simulation studies [84] showed slightly better breakdown properties for the RHM than for the MCD 20–45% and 20–30% respectively, depending on the fractions of outliers and their multivariate distance from the rest of the data. In addition, the method can handle rank deficient data. In the regression step (step 2) the LTS estimator is used with multiple trimming parameters.

<table>
<thead>
<tr>
<th>Name</th>
<th>Principle</th>
<th>Comments</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPCR</td>
<td>PCA: MVT</td>
<td>Only suggested as outlier detection tool due</td>
<td>[49]</td>
</tr>
<tr>
<td></td>
<td>Regression: LMS</td>
<td>to lack of efficiency of LMS</td>
<td></td>
</tr>
<tr>
<td>RPCR</td>
<td>PCA: PP</td>
<td></td>
<td>[80]</td>
</tr>
<tr>
<td></td>
<td>Regression: LTS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RHM-PCR</td>
<td>PCA: RHM</td>
<td>Not orthogonal equivariant</td>
<td>[79]</td>
</tr>
<tr>
<td></td>
<td>Regression: LTS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPCR</td>
<td>PCA: MCD (low dim.) or ROBPCA</td>
<td></td>
<td>[81]</td>
</tr>
<tr>
<td></td>
<td>Regression: reweighted LTS (one response var.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>or MCD regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPPSV</td>
<td>Principal sensitivity vectors</td>
<td>Empirical</td>
<td>[82]</td>
</tr>
</tbody>
</table>
(50, 40, 30, 20 and 10%) as suggested by Ryan [6]. Unfortunately, the RHM estimator does not possess orthogonal and translation equivariance [81].

### 4.3. Robust partial least squares regression

PLS is a linear regression technique developed to deal with high-dimensional regressors and one, \( y_{n \times 1} \) (PLS1) or several response variables, \( Y_{n \times q} \) (PLS2). PLS makes use of ordinary least squares regression steps in the calculation of weights, loadings, scores and regression coefficients. Therefore outliers, either in the \( X \) or in the \( Y \) variables, pose a serious problem to PLS regression. The most common algorithms for PLS in the chemometric field are the NIPALS algorithm and SIMPLS algorithm. In cases with only one response variable \( q = 1 \) and without missing values, SIMPLS and PLS1 (NIPALS) are equivalent.

In Table V the different robust PLS methods discussed in this paper are listed.

Gil and Romera [88], Wakeling and Macfie [85] and Griep et al. [86] claimed that substituting all steps of the ordinary LS regression steps in the NIPALS PLS algorithm by a robust regression procedure will make the model completely resistant to outliers, but there are several prices to be paid: Not just a higher computational demand, but also a lower efficiency of the robust steps. Thus, an alternative approach is to replace one or two selected regression steps instead of all steps together which could still show a good performance in terms of handling outliers. Such procedures are called semi-robust [86,88].

A first robust PLS2 algorithm, (RPLS) [85] has been developed by replacing the non-robust regression steps for \( w \) (weight vector for \( X \)) and \( c \) (loading vector for \( Y \)) in the PLS2 algorithm by the robust biweight method [32]. This makes these estimates resistant to outliers by down-weighting cases with high residuals. Final values of \( f \) (score vector for \( X \)) and \( b \) (score vector for \( Y \)) are formed from the un-weighted data as in the conventional PLS2 algorithm. The price to be paid consists of a lack of orthogonality on successive \( X \) weight vectors, \( w \). The RPLS algorithm is designed to compensate independently for outliers in both \( X \) and \( Y \).

Following the idea of Wakeling and Macfie [85], Griep et al. [86] carried out a comparison among three different methods of robust regression and studied their incorporation into the PLS1 algorithm. In their study Griep et al. [86] replaced the regression step for the weight vector \( w \) with three different methods of robust regression: LMS, Siegel’s RM and IRLS. Their empirical results indicate that the best option is to use IRLS compared to LMS and Siegels RM.

According to Gil and Romera [88] this way of making the PLS models robust, by substituting some or all regressions steps with a robust alternative, does not necessarily catch ‘multivariate’ outliers. This is due to the fact that the first step of PLS is not just one regression, but is formed from the individual regressions of each variable \( x_i \) on \( y \).

Thus, the application of, for example IRLS in the first step consists of the application of IRLS to each simple regression. Therefore outliers in the projections of the data onto planes \( [x_1, y], [x_2, y], \ldots, [x_p, y] \) are taken into account, but the multivariate nature of the \( X \) data is not considered.

Another version of the IRLS algorithm for PLS has been obtained by Cummins and Andrews [87] and Pell [79] called IRPLS. The idea is the same as in ordinary IRLS, but in these cases using the cross validated residuals of the PLS regression in the sample weight function. The sample weights are initially set to one and are updated after each iteration. Both Cummins and Andrews [87] and Pell [79] tested four weight functions including bisquare, Cauchy, Fair and Huber. The only difference between the algorithms proposed by Pell [79] and Cummins and Andrews [87] is that in the algorithms of Pell [79] the number of components is fixed until the sample weights converge whereas in the algorithm proposed by Cummins and Andrews [87] the number of components is allowed to change as the sample weights change. Gil and Romera [88] claimed that a problem with the method proposed by Cummins and Andrews [87] is that the residuals for each sample would depend strongly on the number of components calculated in PLS because different criteria for choosing the number of components could cause different weights for each sample.

The breakdown point for IRPLS depends on the chosen weight function but was around 44% for the tested weight functions and moderate outliers [87].

These algorithms were only derived for a one-dimensional response variable (i.e. PLS1) and are not resistant to high

<table>
<thead>
<tr>
<th>Name</th>
<th>Principle</th>
<th>Basic algorithm</th>
<th>Comments</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPLS</td>
<td>Biweight-IRLS</td>
<td>NIPALS</td>
<td>Semi-robust</td>
<td>[85]</td>
</tr>
<tr>
<td>Robust PLS1</td>
<td>IRLS</td>
<td>NIPALS</td>
<td>Derived for PLS2</td>
<td>[86]</td>
</tr>
<tr>
<td>IRPLS</td>
<td>IRLS</td>
<td>NIPALS</td>
<td>Derived for PLS1</td>
<td>[87]</td>
</tr>
<tr>
<td>IRPLS</td>
<td>IRLS</td>
<td>NIPALS</td>
<td>Not resistant to leverage points</td>
<td>[79]</td>
</tr>
<tr>
<td>PLSMR</td>
<td>Stahel–Donoho estimator</td>
<td>NIPALS</td>
<td>Not resistant to leverage points</td>
<td>[88]</td>
</tr>
<tr>
<td>RSIMCD</td>
<td>ROBPCA for robust scores</td>
<td>SIMPLS</td>
<td>Only valid for ( n &gt; p )</td>
<td>[89]</td>
</tr>
<tr>
<td>RSIMPLS</td>
<td>ROBPCA for robust scores</td>
<td>SIMPLS</td>
<td>Derived for both PLS1 and PLS2</td>
<td>[89]</td>
</tr>
<tr>
<td>PRM</td>
<td>GM-estimators</td>
<td>SIMPLS</td>
<td>Derived for PLS1</td>
<td>[90]</td>
</tr>
</tbody>
</table>

Table V. Overview of different robust PLS methods discussed in this paper.
leverage points since the weights only depends on the residuals after each step [89]. Different weight functions as well as different tuning constant for the same weight function can give different results, which may make the methods less attractive [79,87]. Furthermore, Pell [79] obtained in some cases better prediction results with the robust methods than the estimated reference error, which may be indicative for an overfitting problem.

In Gil and Romera [88] a robust PLS1 method, PLS matrix robust (PLSMR), is obtained by robustifying the sample covariance matrix of the x-variables and the sample cross-covariance matrix between the x- and y-variables. For this the highly robust Stahel-Donoho estimator [10,23,24] is used with Huber’s weight function. To minimize the computational cost the sub-sampling scheme used to compute the estimator starts by drawing subsets of size \( p + 2 \). This means that the PLSMR method cannot be applied to high-dimensional regressors \((n \ll p)\) which is a major disadvantage. To tackle this problem Gil and Romera [88] propose to initiate with variable selection before the robust regression. It is not possible to extend the method to PLS2 [89].

In SIMPLS [91], the PLS weights are obtained as eigenvectors of the XY cross covariance matrix after successive deflation steps. Since SIMPLS is based on the sample cross-covariance \( C_{xy} \), the empirical covariance matrix \( C_{\hat{y}} \), and on linear LS regression the results are affected by abnormal observations in the data set. Hubert and Vanden Branden [89] introduced two robustified versions of the SIMPLS algorithm called RSIMCD and RSIMPLS respectively, based on replacing the cross-covariance matrix \( C_{xy} \) and the empirical covariance matrix \( C_{\hat{y}} \) by robust estimates, and by performing a robust regression method instead of MLR. The robust algorithms are built on two main stages. First, robust scores \( t_i \) are constructed based on a robust criterion (ROBPCA) on \( Z_{nm} = (X_{nm}, Y_{nm}) \), and secondly a robust linear regression is performed based on the robust scores from the ROBPCA. The first stage is similar for both methods but the regression step differs: RSIMCD is based on MCD regression [40] while RSIMPLS uses additional information from the previous ROBPCA step for a reweighted MLR. The proposed algorithms are fast compared with previously developed robust methods and can handle cases where \( n \ll p \) and \( q = 1 \) [11]. Hubert and Vanden Branden [89] recommend RSIMPLS because it is roughly twice as fast as RSIMCD. The breakdown value for RSIMPLS is roughly 1–2.5\% (where \( \alpha \) is the assumed minimal fraction of regular observations) as for the MCD estimator [11]. Both RSIMPLS and RSIMCD are equivariant for translation and orthogonal transformations in \( x \) and \( y \) [89].

Recently, Serneels et al. [90] proposed a method, partial robust M-regression (PRM), for robust regression based on GM-estimators. PRM uses continuous weights, resulting in a gradual down-weighting of outliers according to their degree of outlyingness. This is in contrast to, for example RSIMPLS where a weight of zero is given to all observations with residuals above a certain cut-off value and unity to all others. As GM-estimator the weights are computed from both the residuals as well as the leverage for a sample. This can be performed in a way such that the orthogonal and scale equivariance of the PLS estimator is retained. The weighting is used both in the SIMPLS step of computing the PLS scores as well as in the regression of \( y \) on these scores. For \( p \gg n \) the computation time is sped up by carrying out a prior SVD on \( X(n \times p), X' = US \). The iteration procedure is then applied to the reduced data matrix, \( X = US \times (n \times n) \) and the resulting PRM regression estimate \( \hat{\beta} \) needs then to be back transformed into \( \hat{\beta} = U\hat{\beta} \).

A simulation study testing various distributions of the error terms, different samples sizes and dimensionalities showed that in terms of stability and numerical efficiency, PRM generally outperformed PLS and RSIMPLS. Only for the normal error terms PLS was more efficient than PRM. In situations with 10% bad leverage points PRM and RSIMPLS performed almost similar and clearly outperformed PLS. When comparing the computation time for PRM and SIMPLS it turned out that the computation time for RSIMPLS was consistently substantially higher than for PRM both for increasing number of observations and increasing number of predictor variables [90].

The PRM method is computationally possible for high dimensional data sets and thus provides an appealing alternative to RSIMPLS in particular when a gradual outlier behavior can be expected. But the method is currently only derived for univariate \( y \) (i.e. PLS1) and the highest possible breakdown point of all GM-estimators is in general not larger than 30% and decreases as a function of the dimensionality \( p \) [17,31].

5. SOFTWARE

The common basic methods for robust estimation of location and scatter (i.e. MCD) and robust regression (i.e. M-, LMS-, LTS-, and MM-estimators) are all available within the standard statistical software packages SAS (release > 6.12) [92], S-Plus [39,93] and R [94]. An implementation for robust PCA is also available for S-Plus [95].

Recently, a comprehensive MATLAB toolbox, ‘LIBRA’, for robust estimates and multivariate methods has appeared [96]. Apart from MCD and LTS it also contains implementations of other methods that have been developed at the research groups at the University of Antwerp and the Katholieke Universiteit Leuven, in particular for robust PCA (ROBPCA) and PLS (RSIMPLS). The toolbox also includes many graphical tools for model checking and outlier detection. Additionally, an incorporation of several of these methods into the widely used PLS Toolbox for Matlab is in preparation (Eigenvector, Inc., Pers. Comm.). The partial robust M-regression is also available as Matlab implementation [97].

6. DISCUSSION

Real world data is often contaminated with gross outliers, which can lead conventional data analysis methods based on least squares completely a-stray. Thus, outlier detection and/or the use of robust methods are of paramount importance for applied multivariate data analysis in general and chemometrics in particular.

Although simple robust methods such as the median, median absolute deviation or interquartile range are easy to
compute and have been applied for a long time, the use of robust methods for multiple regression or latent variable models has been computationally prohibitive for multivariate datasets of even moderate size. This is because theoretically rigorous methods usually require the consideration of all possible sample subsets of a given size, which means that their computationally complexity shows a combinatorial growth. Thus, although the exponentially increasing computational power (Moore’s law) has contributed a lot to the advancement of computationally demanding data analysis methods, the progress in the applicability of robust methods has largely been achieved by improved fast approximate methods.

When considering robust methods for applied multivariate data analysis, practical considerations might deviate from a pure theoretical viewpoint. Data sets with more than 10–20% outliers would probably be rejected completely and the generation of new data with a higher quality required. Hence, a theoretical breakdown point of 50% might be helpful for initially assessing the data quality and detecting outliers. But on the other hand it should be kept in mind that there is a price to pay for such an extreme robustness. Apart from the higher computational complexity, robust methods usually also show a lower statistical efficiency and convergence rate. For example the median only shows 64% asymptotic efficiency for normally distributed data in comparison to the mean [98]. Thus, although robust methods are able to obtain reasonable estimates in the presence of gross outliers in the original data, their estimates for the ‘good’ data are usually subject to a higher uncertainty and they require more training samples than conventional least squares methods. For methods with adjustable breakdown properties, therefore a good compromise between robustness and efficiency should be aimed for.

In addition it might be remarked that breakdown points are derived for arbitrarily large, that is ‘infinite’ outliers, whereas in practice outliers are usually bounded due to a finite measurement range. Extreme outliers might also be easy to detect individually by applying appropriate outlier diagnostics. With state-of-the-art robust methods, for example based on projection pursuit or fast MCD estimators, efficient tools for robust multivariate data analysis are available and convenient software implementations, for example in Matlab exist.

Acknowledgements
First author acknowledges support provided by the Danish Ministry of Food, Agriculture and Fisheries. We thank the two referees for their helpful comments.

REFERENCES


