

# Robustness and Outlier Detection in Chemometrics

Michiel Debruyne, Sanne Engelen, Mia Hubert, and Peter J. Rousseeuw

April 10, 2006

## Abstract

In analytical chemistry, experimental data often contain outliers of one type or another. The most often used chemometrical/statistical techniques are sensitive to such outliers, and the results may be adversely affected by them. This paper presents an overview of *robust* chemometrical/statistical methods which search for the model fitted by the majority of the data, and hence are far less affected by outliers. As an extra benefit, we can then *detect* the outliers by their large deviation from the robust fit. We discuss robust procedures for estimating location and scatter, and for performing multiple linear regression, PCA, PCR, PLS, and classification. We also describe recent results concerning the robustness of Support Vector Machines, which are kernel-based methods for fitting non-linear models. Finally, we present robust approaches for the analysis of multiway data.

## 1 Introduction

When analyzing real data, it often occurs that some observations are different from the majority. Such observations are called *outliers*. Sometimes they are due to recording or copying mistakes (yielding e.g. a misplaced decimal point or the permutation of two digits). Often the outlying observations are not incorrect but they were made under exceptional circumstances, or they belong to another population (e.g. it may have been the concentration of a different compound), and consequently they do not fit the model well. It is very important to be able to detect these outliers. For instance, they can pinpoint a change in the production process or in the experimental conditions.

In practice one often tries to detect outliers using diagnostics starting from a classical fitting method. However, classical methods can be affected by outliers so strongly that the resulting fitted model does not allow to detect the deviating observations. This is called the *masking* effect. Additionally, some good data points might even appear to be outliers, which is known as *swamping*. To avoid these effects, the goal of *robust statistics* is to find a fit which is similar to the fit we would have found without the outliers. We can then identify the outliers by their large residuals from that robust fit.

In Section 2 we briefly describe some robust procedures for estimating univariate location, scale, and skewness, as well as low-dimensional multivariate location and scatter. Apart from the traditional elliptical distributions, we also consider outlier detection at asymmetric (e.g. skewed) multivariate distributions. Section 3 considers linear regression in low dimensions.

In chemometrics high-dimensional data frequently occurs, often with the number of variables exceeding the number of observations. Robust methods for analyzing such complex data structures were developed in recent years. We discuss robust versions of PCA (Section 4), PCR, and PLS (Section 5), with their accompanying outlier detection tools. Section 6 is devoted to linear classification.

Although linear models are easy to interpret, a non-linear model can sometimes provide more accurate prediction results. Kernel-based methods for fitting such non-linear models have been introduced in the machine learning community and are becoming more popular in chemometrics and bioinformatics. In Section 7 we describe some recent results concerning the robustness of Support Vector Machines, a kernel method for classification and regression.

Outlier detection and robustness in multiway data is discussed in Section 8. Finally, Section 9 discusses software availability.

## 2 Location and covariance estimation in low dimensions

In this section we assume that the data are stored in an  $n \times p$  data matrix  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$  with  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$  the  $i$ th observation. Hence  $n$  stands for the number of objects and  $p$  for the number of variables. First we consider the univariate case  $p = 1$ .

### 2.1 Univariate location and scale estimation

The location-scale model states that the  $n$  univariate observations  $x_i$  are independent and identically distributed (i.i.d.) with distribution function  $F((x - \mu)/\sigma)$  where  $F$  is known. Typically  $F$  is the standard gaussian distribution function  $\Phi$ . We then want to find estimates for the center  $\mu$  and the scale parameter  $\sigma$  (or for  $\sigma^2$ ).

The classical estimate of location is the *sample mean*  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ . However, the mean is very sensitive to aberrant values. Replacing only 1 out of  $n$  observations by a very large value can change the estimate completely. We say that the *breakdown value* of the sample mean is  $1/n$ , so it is 0% for large  $n$ . More precisely, the breakdown value is the smallest proportion of observations in the data set that need to be replaced to carry the estimate arbitrarily far away. The robustness of an estimator can also be quantified by its *influence function*<sup>1</sup> which measures the effect of a small number of outliers. The influence function of the mean is unbounded, which reflects its non-robust behavior.

Another well-known location estimator is the *median*. Denote the  $i$ th ordered observation as  $x_{(i)}$ . Then the median is defined as  $x_{((n+1)/2)}$  if  $n$  is odd and  $(x_{(n/2)} + x_{((n+1)/2)})/2$  if  $n$  is even. Replacing only 1 observation by an arbitrary value will not change the median much. Its breakdown point is about 50%, meaning that the median can resist up to 50% of outliers, and its influence function is bounded. The robustness of the median comes at a cost: at the normal model it is less efficient than the mean. To find a better balance between robustness and efficiency, many other robust procedures have been proposed, such as trimmed means and M-estimators<sup>2</sup>.

The situation is very similar for the scale parameter  $\sigma$ . The classical estimator is the *standard deviation*  $s = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)}$ . Since one outlier can make  $s$  arbitrarily large, its breakdown value is 0%. A robust measure of scale is the *median absolute deviation* given by the median of all absolute distances from the sample median:

$$\text{MAD} = 1.483 \operatorname{median}_{j=1, \dots, n} |x_j - \operatorname{median}_{i=1, \dots, n}(x_i)|. \quad (1)$$

The constant 1.483 is a correction factor which makes the MAD unbiased at the normal distribution. Another alternative is the  $Q_n$  estimator<sup>3</sup>, defined as

$$Q_n = 2.2219 c_n \{ |x_i - x_j|; i < j \}_{(k)} \quad (2)$$

with  $k = \binom{h}{2} \approx \binom{n}{2} / 4$  and  $h = \lfloor \frac{n}{2} \rfloor + 1$ . Here  $\lfloor z \rfloor$  denotes the largest integer smaller than or equal to  $z$ . So, we compute the  $k$ th order statistic out of the  $\binom{n}{2} = \frac{n(n-1)}{2}$  possible distances  $|x_i - x_j|$ . This scale estimator is thus essentially the first quartile of all pairwise differences between two data points. The constant  $c_n$  is a small-sample correction factor, which makes  $Q_n$  an unbiased estimator. The breakdown value of both the MAD and the  $Q_n$  estimator is 50%.

Also popular is the interquartile range (IQR) defined as the difference between the third and first quartiles, so roughly  $\text{IQR} = x_{(3n/4)} - x_{(n/4)}$ . Its breakdown value is only 25% but it has an easy interpretation and it is commonly used to construct the boxplot (Section 2.2).

In order to detect outliers it is important to have an *outlier detection rule*. A classical rule is based on the standardized residuals of the observations. More precisely, it flags  $x_i$  as outlying if

$$\frac{|x_i - \bar{x}|}{s} \quad (3)$$

exceeds e.g. 2.5. But this outlier rule uses non-robust estimates itself, namely the mean  $\bar{x}$  and the standard deviation  $s$ . Therefore it is very well possible that some outliers are not detected and/or some regular observations are incorrectly flagged as outliers. Plugging in robust estimators of location and scale such as the median and the MAD yields

$$\frac{|x_i - \text{median}_{j=1, \dots, n}(x_j)|}{\text{MAD}} \quad (4)$$

which is a much more reliable outlier detection tool.

## 2.2 Skewness and the adjusted boxplot

Tukey's boxplot is a graphical tool to visualize the distribution of a univariate data set, and to pinpoint possible outliers. In this plot a box is drawn from the first quartile  $Q_1 \approx x_{(n/4)}$  to the third quartile  $Q_3 \approx x_{(3n/4)}$  of the data. Points outside the interval  $[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}]$ , called the fence, are traditionally marked as outliers. If the data are sampled from the normal distribution, only about 0.7% of them will lie outside the fence. The boxplot clearly assumes symmetry, since we add the same amount to  $Q_3$  as what we subtract from  $Q_1$ . At asymmetric distributions this approach may flag much more than 0.7% of the data. For example, consider the boxplot in Figure 1(a) of the Cu (copper) variable of the Kola ecogeochemistry project which was a large multi-element geochemical mapping project (see e.g.<sup>4</sup>). We see that the usual boxplot flags many data points as outlying while they probably aren't since the underlying distribution is right-skewed. Figure 1(b) shows a *skewness-adjusted boxplot* of the same variable<sup>5</sup>. It only flags four suspiciously large observations, as well as cases with an abnormally small Cu concentration.

The skewness-adjusted boxplot labels observations as outliers if they lie outside the fence

$$[Q_1 - 1.5e^{-3.5MC} \text{ IQR}, Q_3 + 1.5e^{3.5MC} \text{ IQR}]. \quad (5)$$

Here, MC stands for the *medcouple* which is a robust measure of skewness<sup>6</sup>. It is defined as

$$MC = \text{median}_{i,j} \left\{ \frac{(x_j - \text{median}) - (\text{median} - x_i)}{x_j - x_i} \right\} \quad (6)$$

where  $i$  and  $j$  have to satisfy  $x_i \leq \text{median} \leq x_j$  and  $x_i < x_j$ . From its definition it follows that the medcouple is a number between  $-1$  and  $1$ . If the data is symmetrically distributed, the medcouple is zero and the adjusted boxplot thus reduces to the standard boxplot. A positive (negative) medcouple corresponds to a right-skewed (left-skewed) distribution.

## 2.3 Multivariate location and covariance estimation

### 2.3.1 Empirical mean and covariance matrix

In the multivariate situation we assume that the observations  $\mathbf{x}_i$  in  $\mathbb{R}^p$  are  $p$ -dimensional. Classical measures of location and scatter are given by the empirical mean  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$  and the empirical

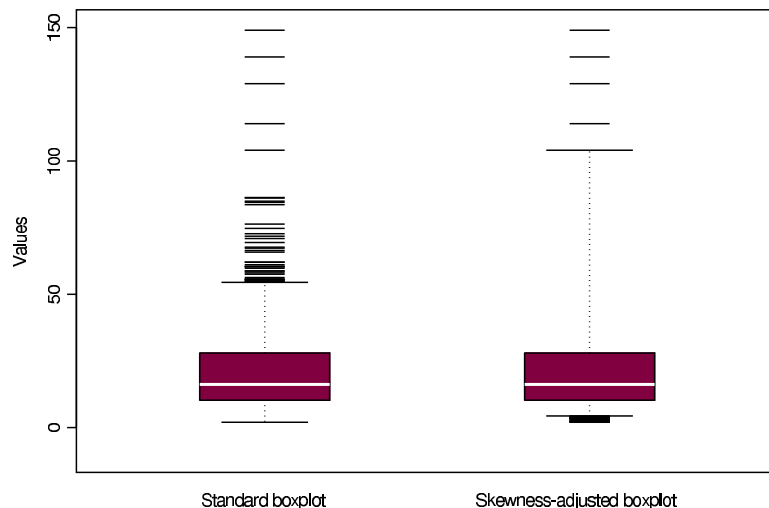


Figure 1: Difference between the standard and the skewness-adjusted boxplot of the copper data set.

covariance matrix  $\mathbf{S}_x = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T / (n - 1)$ . As in the univariate case, both classical estimators have a breakdown value of 0%, that is, a small fraction of outliers can completely ruin them. To illustrate this, consider the simple example in Figure 2. It shows the concentration of inorganic phosphorus and organic phosphorus in soil<sup>7</sup>. On this plot the classical tolerance ellipse is superimposed, defined as the set of  $p$ -dimensional points  $\mathbf{x}$  whose *Mahalanobis distance*

$$\text{MD}(\mathbf{x}) = \sqrt{(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}_x^{-1} (\mathbf{x} - \bar{\mathbf{x}})} \quad (7)$$

equals  $\sqrt{\chi_{2,0.975}^2}$ , the square root of the 0.975 quantile of the chi-square distribution with 2 degrees of freedom. (Note that (7) becomes (3) for  $p = 1$ .) If the data are normally distributed, 97.5% of the observations should fall inside this ellipse. Observations outside this ellipse are suspected to be outliers. We see however that all data points lie inside the classical tolerance ellipse.

### 2.3.2 The robust MCD estimator

Contrary to the classical mean and covariance matrix, a robust method yields a tolerance ellipse which captures the covariance structure of the majority of the data points. Starting from such highly robust estimates  $\hat{\boldsymbol{\mu}}_R$  of multivariate location and  $\hat{\boldsymbol{\Sigma}}_R$  of scatter, we plot the points  $\mathbf{x}$  whose *robust distance*

$$\text{RD}(\mathbf{x}) = \sqrt{(\mathbf{x} - \hat{\boldsymbol{\mu}}_R)^T \hat{\boldsymbol{\Sigma}}_R^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_R)} \quad (8)$$

is equal to  $\sqrt{\chi_{2,0.975}^2}$ . In Figure 2, this robust tolerance ellipse is much narrower than the classical one. Observations 1, 6 and 10 lie outside the ellipse, and are flagged as suspicious observations. The classical tolerance ellipse on the other hand was clearly stretched in the direction of the three outliers, so strongly that the latter even fell inside the ellipse and were not flagged. This is an example of the masking effect.

The robust estimates of location and scatter used in Figure 2 were obtained by the Minimum Covariance Determinant (MCD) method<sup>8</sup>. The MCD looks for those  $h$  observations in the data set

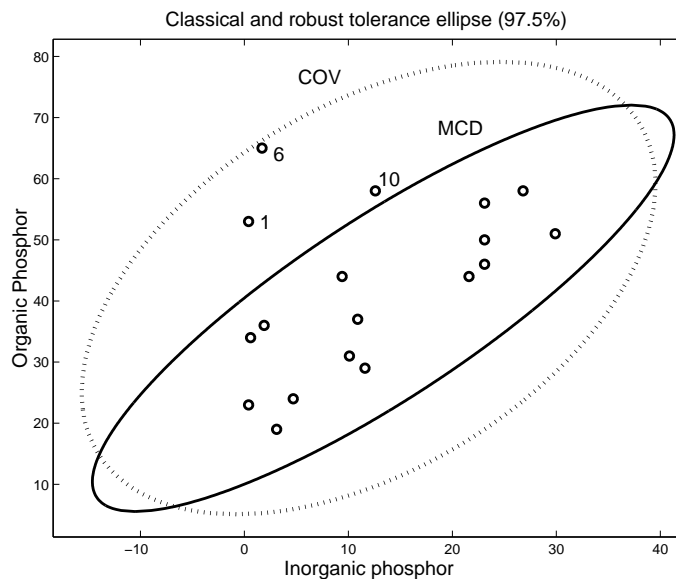


Figure 2: Classical (COV) and robust (MCD) tolerance ellipse of the phosphorus data set.

(where the number  $h$  is given by the user) whose classical covariance matrix has the lowest possible determinant. The MCD estimate of location  $\hat{\boldsymbol{\mu}}_0$  is then the average of these  $h$  points, whereas the MCD estimate of scatter  $\hat{\boldsymbol{\Sigma}}_0$  is their covariance matrix, multiplied with a consistency factor. Based on the raw MCD estimates, a reweighing step can be added which increases the finite-sample efficiency considerably. In general, we can give each  $\mathbf{x}_i$  some weight  $w_i$ , for instance by putting  $w_i = 1$  if  $(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0)^T \hat{\boldsymbol{\Sigma}}_0^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0) \leq \chi_{p,0.975}^2$  and  $w_i = 0$  otherwise. The resulting reweighed mean and covariance matrix are then defined as

$$\hat{\boldsymbol{\mu}}_R(\mathbf{X}) = \left( \sum_{i=1}^n w_i \mathbf{x}_i \right) / \left( \sum_{i=1}^n w_i \right) \quad (9)$$

$$\hat{\boldsymbol{\Sigma}}_R(\mathbf{X}) = \left( \sum_{i=1}^n w_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_R)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_R)^T \right) / \left( \sum_{i=1}^n w_i - 1 \right). \quad (10)$$

The final robust distances  $\text{RD}(\mathbf{x}_i)$  were obtained by inserting  $\hat{\boldsymbol{\mu}}_R(\mathbf{X})$  and  $\hat{\boldsymbol{\Sigma}}_R(\mathbf{X})$  into (8).

The MCD estimator has a bounded influence function<sup>9</sup> and breakdown value  $(n - h + 1)/n$ , hence the number  $h$  determines the robustness of the estimator. The MCD has its highest possible breakdown value when  $h = \lceil (n + p + 1)/2 \rceil$ . When a large proportion of contamination is presumed,  $h$  should thus be chosen close to  $0.5n$ . Otherwise an intermediate value for  $h$ , such as  $0.75n$ , is recommended to obtain a higher finite-sample efficiency.

The computation of the MCD estimator is non-trivial and naively requires an exhaustive investigation of all  $h$ -subsets out of  $n$ . Rousseeuw and Van Driessen<sup>10</sup> constructed a much faster algorithm called FAST-MCD.

### 2.3.3 Other robust estimators of location and scatter

Many other robust estimators of location and scatter have been presented in the literature. The first such estimator was proposed independently by Stahel<sup>11</sup> and Donoho<sup>12</sup>. They defined the

so-called Stahel-Donoho outlyingness of a data point  $\mathbf{x}_i$  as

$$\text{outl}(\mathbf{x}_i) = \max_{\mathbf{d}} \frac{|\mathbf{x}_i^T \mathbf{d} - \text{median}_{j=1,\dots,n}(\mathbf{x}_j^T \mathbf{d})|}{\text{MAD}_{j=1,\dots,n}(\mathbf{x}_j^T \mathbf{d})} \quad (11)$$

where the maximum is over all directions (i.e., all unit length vectors in  $\mathbb{R}^p$ ), and  $\mathbf{x}_j^T \mathbf{d}$  is the projection of  $\mathbf{x}_j$  on the direction  $\mathbf{d}$ . (Note that for  $p = 1$ , (11) reduces to (4).) Next they gave each observation a weight  $w_i$  based on  $\text{outl}(\mathbf{x}_i)$ , and computed robust estimates  $\hat{\boldsymbol{\mu}}_R$  and  $\hat{\boldsymbol{\Sigma}}_R$  as in (9) and (10). The Stahel-Donoho estimator was further investigated by Tyler<sup>13</sup> and Maronna and Yohai<sup>14</sup>.

Multivariate M-estimators<sup>15</sup> have a relatively low breakdown value due to possible implosion of the estimated scatter matrix. Together with the MCD estimator, Rousseeuw<sup>16</sup> introduced the Minimum Volume Ellipsoid. More recent classes of robust estimators of multivariate location and scatter include S-estimators<sup>7,17</sup>, CM-estimators<sup>18</sup>,  $\tau$ -estimators<sup>19</sup>, MM-estimators<sup>20</sup>, estimators based on multivariate ranks or signs<sup>21</sup>, and depth-based estimators<sup>22,23,24</sup>.

## 2.4 Outlier detection for skewed multivariate distributions

The Mahalanobis distance (7) and the robust distance (8) assume that the distribution underlying the majority of the data (i.e. the regular points) is elliptically symmetric, in the sense that its density is of the form  $f(\mathbf{x}) = g(\sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})})$  for some center  $\boldsymbol{\mu}$  and some positive definite scatter matrix  $\boldsymbol{\Sigma}$ . (The multivariate normal distribution is elliptically symmetric.) The density contours are then ellipses (when  $p = 2$ ) or ellipsoids (when  $p \geq 3$ ) with center  $\boldsymbol{\mu}$ . Even the Stahel-Donoho outlyingness (11) makes such an assumption [in fact, if we replace the median in (11) by the mean and the MAD by the standard deviation, (11) reduces to the Mahalanobis distance (7)].

But in many cases the underlying distribution is not elliptically symmetric, since its projections in some directions may be skewed. In such cases we can work with the (skewness-) *adjusted outlyingness*<sup>25</sup>, which is defined like (11) but where its (symmetric) denominator is replaced by:

$$\begin{aligned} (Q_3 + 1.5e^{3.5MC} \text{ IQR}) - \text{median}_{j=1,\dots,n}(\mathbf{x}_j^T \mathbf{d}) & \quad \text{if} \quad \mathbf{x}_i^T \mathbf{d} > \text{median}_{j=1,\dots,n}(\mathbf{x}_j^T \mathbf{d}) \\ \text{median}_{j=1,\dots,n}(\mathbf{x}_j^T \mathbf{d}) - (Q_1 - 1.5e^{-3.5MC} \text{ IQR}) & \quad \text{if} \quad \mathbf{x}_i^T \mathbf{d} \leq \text{median}_{j=1,\dots,n}(\mathbf{x}_j^T \mathbf{d}). \end{aligned}$$

Unlike (7) and (8), the AO does not reject too many data points at such skewed distributions.

Brys et al<sup>25</sup> applied the AO to independent component analysis (ICA). In the ICA framework it is *required* that the underlying distribution be non-elliptical (in order for the independent components to be identifiable). Classical ICA is sensitive to outliers, and downweighting observations with high AO yielded a robust version of ICA.

## 3 Linear regression in low dimensions

### 3.1 Linear regression with one response variable

The multiple linear regression model assumes that in addition to the  $p$  independent  $x$ -variables, a response variable  $y$  is measured, which can be explained as a linear combination of the  $x$ -variables. More precisely, the model says that for all observations  $(\mathbf{x}_i, y_i)$  it holds that

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i \quad i = 1, \dots, n \quad (12)$$

where the errors  $\epsilon_i$  are assumed to be independent and identically distributed with zero mean and constant variance  $\sigma^2$ . The vector  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is called the slope, and  $\beta_0$  the intercept. We denote  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$  and  $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^T)^T = (\beta_0, \beta_1, \dots, \beta_p)^T$ . Applying a regression estimator to the data yields  $p+1$  regression coefficients  $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)^T$ . The residual  $r_i$  of case  $i$  is defined as the difference between the observed response  $y_i$  and its estimated value  $\hat{y}_i$ :

$$r_i(\hat{\boldsymbol{\theta}}) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}).$$

### 3.1.1 The classical least squares estimator

The classical least squares method for multiple linear regression (MLR) to estimate  $\boldsymbol{\theta}$  minimizes the sum of the squared residuals. It is a very popular method because it allows to compute the regression estimates explicitly as  $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  (where the design matrix  $\mathbf{X}$  is enlarged with a column of ones for the intercept term) and moreover the least squares method is optimal if the errors are normally distributed.

However, MLR is extremely sensitive to regression outliers, which are observations that do not obey the linear pattern formed by the majority of the data. This is illustrated in Figure 3 for simple regression (where there is only one regressor  $x$ , or  $p = 1$ ). It contains the Hertzsprung-Russell diagram of 47 stars, of which the logarithm of their light intensity and the logarithm of their surface temperature were measured<sup>7</sup>. The four most outlying observations are giant stars, which clearly deviate from the main sequence stars. The least squares fit in this plot was attracted by the giant stars.

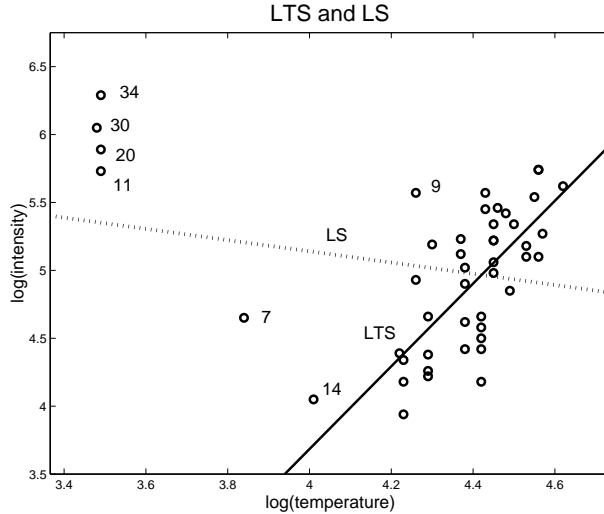


Figure 3: Stars regression data set with classical and robust fit.

An estimate of the scale of the error distribution  $\sigma$  is given by  $s^2 = \sum_{i=1}^n r_i^2 / (n - p - 1)$ . One often flags observations for which  $|r_i/s|$  exceeds a cut-off like 2.5 as regression outliers, because values generated by a Gaussian distribution are rarely larger than  $2.5\sigma$ . In Figure 4(a) this strategy fails: the standardized least squares residuals of all 47 points lie inside the tolerance band between -2.5 and 2.5. The four outliers in Figure 3 have attracted the least squares line so much that they have small residuals  $r_i$  from it.

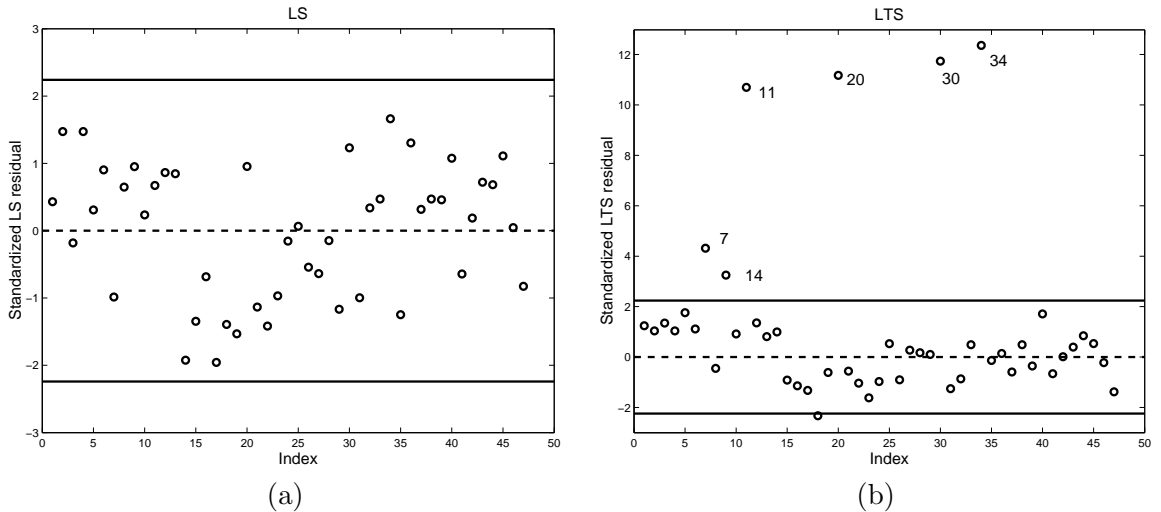


Figure 4: Standardized residuals of the stars data set, based on the (a) classical MLR; (b) robust LTS estimator.

### 3.1.2 The robust LTS estimator

On Figure 3 a robust regression fit is superimposed. The *least trimmed squares estimator* (LTS) proposed by Rousseeuw<sup>8</sup> is given by

$$\text{minimize } \sum_{i=1}^h (r^2)_{i:n} \quad (13)$$

where  $(r^2)_{1:n} \leq (r^2)_{2:n} \leq \dots \leq (r^2)_{n:n}$  are the ordered squared residuals. (They are first squared, and then ranked.) The value  $h$  plays the same role as in the definition of the MCD estimator. For  $h \approx n/2$  we find a breakdown value of 50%, whereas for larger  $h$  we obtain  $(n - h + 1)/n$ . A fast algorithm for the LTS estimator (FAST-LTS) has been developed<sup>26</sup>.

The scale of the errors  $\sigma$  can be estimated by  $\hat{\sigma}_{\text{LTS}}^2 = c_{h,n}^2 \frac{1}{h} \sum_{i=1}^h (r^2)_{i:n}$  where  $r_i$  are the residuals from the LTS fit, and  $c_{h,n}$  makes  $\hat{\sigma}$  consistent and unbiased at Gaussian error distributions<sup>27</sup>. We can then identify regression outliers by their standardized LTS residuals  $r_i/\hat{\sigma}_{\text{LTS}}$ . This yields Figure 4(b) in which we clearly see the outliers. We can also use the standardized LTS residuals to assign a weight to every observation. The reweighed MLR estimator with these LTS weights inherits the nice robustness properties of LTS, but is more efficient and yields all the usual inferential output such as t-statistics, F-statistics, an  $R^2$  statistic, and the corresponding  $p$ -values.

### 3.1.3 An outlier map

Residuals plots become even more important in multiple regression with more than one regressor, as then we can no longer rely on a scatter plot of the data. Figure 4 however only allows us to detect observations that lie far away from the regression fit. It is also interesting to detect aberrant behavior in  $\mathbf{x}$ -space. Therefore a more sophisticated outlier map can be constructed<sup>28</sup>, plotting the standardized LTS residuals versus robust distances (8) based on (for example) the MCD estimator which is applied to the  $x$ -variables only.

Figure 5 shows the outlier map of the stars data. We can distinguish three types of outliers. *Bad leverage points* lie far away from the regression fit and far away from the other observations

in  $\mathbf{x}$ -space, e.g. the four giant stars and star 7. *Vertical outliers* have an outlying residual, but no outlying robust distance, e.g. star 9. Observation 14 is a *good leverage point*: it has an outlying robust distance, but it still follows the linear trend of the main sequence, since its standardized residual does not exceed 2.5.

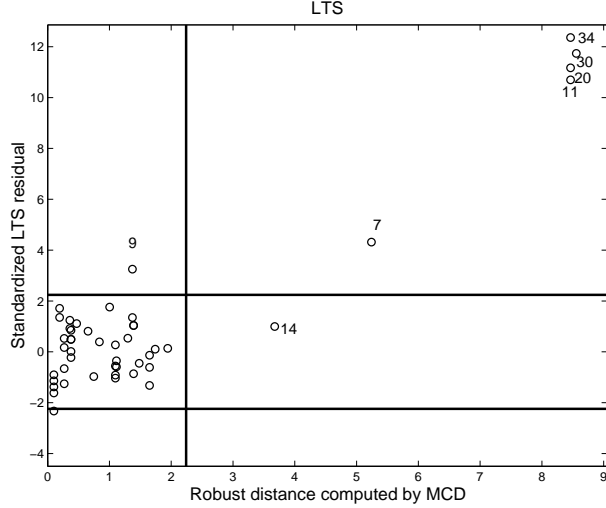


Figure 5: Outlier map for the stars data set.

### 3.1.4 Other robust regression estimators

The earliest systematic theory of robust regression was based on M-estimators<sup>29,2</sup>, followed by R-estimators<sup>30</sup> and L-estimators<sup>31</sup> of regression.

The breakdown value of all these methods is 0% because of their vulnerability to bad leverage points. Generalized M-estimators (GM-estimators)<sup>1</sup> were the first to attain a positive breakdown value, unfortunately going down to zero for increasing  $p$ .

The Least Median of Squares (LMS) of Rousseeuw<sup>8</sup> and the LTS described above were the first equivariant methods to attain a 50% breakdown value. Their low finite-sample efficiency can be improved by carrying out a one-step reweighted least squares fit afterwards, but also by replacing their objective functions by a more efficient scale estimator applied to the residuals  $r_i$ . This direction has led to the introduction of efficient positive-breakdown regression methods, such as S-estimators<sup>32</sup>, MM-estimators<sup>33</sup>, CM-estimators<sup>34</sup>, and many others.

To extend the good properties of the median to regression, the notion of regression depth<sup>35</sup> and deepest regression<sup>36,37</sup> was introduced and applied to several problems in chemistry<sup>38,39</sup>.

## 3.2 Linear regression with several response variables

The regression model can be extended to the case where we have more than one response variable. For  $p$ -variate predictors  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$  and  $q$ -variate responses  $\mathbf{y}_i = (y_{i1}, \dots, y_{iq})^T$  the *multivariate (multiple) regression model* is given by

$$\mathbf{y}_i = \boldsymbol{\beta}_0 + \mathbf{B}^T \mathbf{x}_i + \boldsymbol{\varepsilon}_i \quad (14)$$

where  $\mathbf{B}$  is the  $p \times q$  slope matrix,  $\boldsymbol{\beta}_0$  is the  $q$ -dimensional intercept vector, and the errors  $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iq})^T$  are i.i.d. with zero mean and with  $\text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}_\varepsilon$  a positive definite matrix of size  $q$ .

The least squares solution can be written as

$$\hat{\mathbf{B}} = \hat{\Sigma}_x^{-1} \hat{\Sigma}_{xy} \quad (15)$$

$$\hat{\beta}_0 = \hat{\mu}_y - \hat{\mathbf{B}}^T \hat{\mu}_x \quad (16)$$

$$\hat{\Sigma}_\varepsilon = \hat{\Sigma}_y - \hat{\mathbf{B}}^T \hat{\Sigma}_x \hat{\mathbf{B}} \quad (17)$$

where

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{pmatrix} \quad \text{and} \quad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_x & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{yx} & \hat{\Sigma}_y \end{pmatrix} \quad (18)$$

are the empirical mean and covariance matrix of the joint  $(x, y)$  variables.

In Rousseeuw et al.<sup>40</sup> it is proposed to fill in the MCD estimates for the center  $\boldsymbol{\mu}$  and the scatter matrix  $\Sigma$  of the joint  $(x, y)$  variables in (18), yielding robust estimates (15) to (17). The resulting estimates are called MCD-regression estimates. They inherit the high breakdown value of the MCD estimator. To obtain a better efficiency, the reweighed MCD estimates are used in (15)-(17) and followed by a regression reweighing step. For any fit  $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0^T, \hat{\mathbf{B}}^T)^T$ , denote the corresponding  $q$ -dimensional residuals by  $\mathbf{r}_i(\hat{\boldsymbol{\theta}}) = \mathbf{y}_i - \hat{\mathbf{B}}^T \mathbf{x}_i - \hat{\beta}_0$ . Then the *residual distance* of the  $i$ th case is defined as  $\text{ResD}_i = \sqrt{\mathbf{r}_i^T \hat{\Sigma}_\varepsilon^{-1} \mathbf{r}_i}$ . These residual distances can then be used in a reweighing step in order to improve the efficiency. Also an outlier map can be constructed. Plotting the residual distances versus the robust distances one can easily detect good leverage points, bad leverage points and vertical outliers.

## 4 Principal Component Analysis

### 4.1 PCA based on a covariance matrix

Principal component analysis is a popular statistical method which tries to explain the covariance structure of data by means of a small number of components. These components are linear combinations of the original variables, and often allow for an interpretation and a better understanding of the different sources of variation. Because PCA is concerned with data reduction, it is widely used for the analysis of high-dimensional data which are frequently encountered in chemometrics. PCA is then often the first step of the data analysis, followed by classification, cluster analysis, or other multivariate techniques<sup>41</sup>.

In the classical approach, the first principal component corresponds to the direction in which the projected observations have the largest variance. The second component is then orthogonal to the first and again maximizes the variance of the data points projected on it. Continuing in this way produces all the principal components, which correspond to the eigenvectors of the empirical covariance matrix. Unfortunately, both the classical variance (which is being maximized) and the classical covariance matrix (which is being decomposed) are very sensitive to anomalous observations. Consequently, the first components from classical PCA (CPCA) are often attracted towards outlying points, and may not capture the variation of the regular observations.

A first group of robust PCA methods is obtained by replacing the classical covariance matrix by a robust covariance estimator, such as the reweighed MCD estimator<sup>42</sup> (Section 2.3.2). Unfortunately the use of these affine equivariant covariance estimators is limited to small to moderate dimensions. When  $p$  is larger than  $n$ , the MCD estimator is not defined. A second problem is the computation of these robust estimators in high dimensions. Today's fastest algorithms can handle up to about 100 dimensions, whereas in chemometrics one often needs to analyze data with dimensions in the thousands.

## 4.2 Robust PCA based on projection pursuit

A second approach to robust PCA uses *Projection Pursuit* (PP) techniques. These methods maximize a robust measure of spread to obtain consecutive directions on which the data points are projected. In Hubert et al.<sup>43</sup> a projection pursuit (PP) algorithm is presented, based on the ideas of Li and Chen<sup>44</sup> and Croux and Ruiz-Gazen<sup>45</sup>. The algorithm is called RAPCA, which stands for *reflection algorithm for principal component analysis*.

If  $p \geq n$  the RAPCA method starts by reducing the data space to the affine subspace spanned by the  $n$  observations. This is done quickly and accurately by a singular value decomposition (SVD) of  $\mathbf{X}_{n,p}$ . (From here on the subscripts to a matrix serve to recall its size, e.g.  $\mathbf{X}_{n,p}$  is an  $n \times p$  matrix.) This step is useful as soon as  $p > r = \text{rank}(\mathbf{X})$ . When  $p \gg n$  we obtain a huge reduction. For spectral data, e.g.  $n = 50, p = 1000$ , this reduces the 1000-dimensional original data set to one in only 49 dimensions.

The main step of the RAPCA algorithm is then to search for the direction in which the projected observations have the largest robust scale. This univariate scale is measured by the  $Q_n$  estimator (2). Comparisons using other scale estimators are presented in Croux and Ruiz-Gazen<sup>46</sup> and Cui et al.<sup>47</sup>. To make the algorithm computationally feasible, the collection of directions to be investigated are restricted to all directions that pass through the  $L^1$ -median and a data point. The  $L^1$ -median is a highly robust (50% breakdown value) location estimator, also known as the spatial median. It is defined as the point  $\boldsymbol{\theta}$  which minimizes the sum of the distances to all observations.

Having found the first direction  $\mathbf{v}_1$ , the data are reflected such that the first eigenvector is mapped onto the first basis vector. Then the data are projected onto the orthogonal complement of the first eigenvector. This is simply done by omitting the first component of each (reflected) point. Doing so, the dimension of the projected data points can be reduced by one and consequently, the computations do not need to be done in the full  $r$ -dimensional space.

The method can then be applied in the orthogonal complement to search for the second eigenvector and so on. It is not required to compute all eigenvectors, which would be very time-consuming for high  $p$ , since the computations can be stopped as soon as the required number of components has been found.

Note that a PCA analysis often starts by prestandardizing the data in order to obtain variables that all have the same spread. Otherwise, the variables with a large variance compared to the others will dominate the first principal components. Standardizing by the mean and the standard deviation of each variable yields a PCA analysis based on the correlation matrix instead of the covariance matrix. We prefer to standardize each variable  $j$  in a robust way, e.g. by first subtracting its median  $\text{med}(x_{1j}, \dots, x_{nj})$  and then dividing by its robust scale estimate  $Q_n(x_{1j}, \dots, x_{nj})$ .

## 4.3 Robust PCA based on projection pursuit and the MCD

Another approach to robust PCA has been proposed in Hubert et al.<sup>48</sup> and is called ROBPCA. This method combines ideas of both projection pursuit and robust covariance estimation. The projection pursuit part is used for the initial dimension reduction. Some ideas based on the MCD estimator are then applied to this lower-dimensional data space. Simulations have shown that this combined approach yields more accurate estimates than the raw projection pursuit algorithm RAPCA. The complete description of the ROBPCA method is quite involved, so here we will only outline the main stages of the algorithm.

First, as in RAPCA, the data are preprocessed by reducing their data space to the affine subspace spanned by the  $n$  observations. In the second step of the ROBPCA algorithm, we compute the outlyingness of each data point by (11) where the median is replaced by the univariate MCD location and the MAD is replaced by the univariate MCD scale. (When dealing with skewed

multivariate distributions, one can use the adjusted outlyingness of Section 2.4.) To keep the computation time feasible,  $\mathbf{d}$  ranges over all directions determined by lines passing through two data points. Next, the covariance matrix  $\hat{\Sigma}_h$  of the  $h$  data points with smallest outlyingness is computed. The last stage of ROBPCA consists of projecting all the data points onto the  $k$ -dimensional subspace spanned by the  $k$  largest eigenvectors of  $\hat{\Sigma}_h$  and of computing their center and shape by means of the reweighed MCD estimator of (9) and (10). The eigenvectors of this scatter matrix then determine the robust principal components which can be collected in a loading matrix  $\mathbf{P}_{p,k}$  with orthogonal columns. The MCD location estimate  $\hat{\boldsymbol{\mu}}_x$  serves as a robust center. The influence functions of both  $\hat{\boldsymbol{\mu}}_x$  and  $\mathbf{P}_{p,k}$  are bounded<sup>49</sup>.

Since the loadings are orthogonal, they determine a new coordinate system in the  $k$ -dimensional subspace that they span. The  $k$ -dimensional scores of each data point  $\mathbf{t}_i$  are computed as the coordinates of the projections of the robustly centered  $\mathbf{x}_i$  onto this subspace, or equivalently  $\mathbf{t}_i = \mathbf{P}_{k,p}^T(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x)$ . The *orthogonal distance* measures the distance between an observation  $\mathbf{x}_i$  and its projection  $\hat{\mathbf{x}}_i$  in the  $k$ -dimensional PCA subspace:

$$\hat{\mathbf{x}}_i = \hat{\boldsymbol{\mu}}_x + \mathbf{P}_{p,k}\mathbf{t}_i \quad (19)$$

$$\text{OD}_i = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|. \quad (20)$$

Let  $\mathbf{L}_{k,k}$  denote the diagonal matrix which contains the  $k$  eigenvalues  $l_j$  of the MCD scatter matrix, sorted from largest to smallest. Thus  $l_1 \geq l_2 \geq \dots \geq l_k$ . The *score distance* of the  $i$ th sample measures the robust distance of its projection to the center of all the projected observations. Hence, it is measured within the PCA subspace, where due to the knowledge of the eigenvalues, we have information about the covariance structure of the scores. Consequently, the score distance is defined as in (8):

$$\text{SD}_i = \sqrt{\mathbf{t}_i^T \mathbf{L}^{-1} \mathbf{t}_i} = \sqrt{\sum_{j=1}^k (t_{ij}^2 / l_j)}. \quad (21)$$

Moreover, the  $k$  robust principal components generate a  $p \times p$  robust scatter matrix  $\hat{\Sigma}_x$  of rank  $k$  given by

$$\hat{\Sigma}_x = \mathbf{P}_{p,k} \mathbf{L}_{k,k} \mathbf{P}_{k,p}^T. \quad (22)$$

We can plot the  $\text{OD}_i$  (20) versus the  $\text{SD}_i$  (21) in an outlier map<sup>48</sup>. This way four types of observations are visualized. *Regular observations* have a small orthogonal and a small score distance. When samples have a large score distance, but a small orthogonal distance, they are called *good leverage points*. *Orthogonal outliers* have a large orthogonal distance, but a small score distance. *Bad leverage points* have a large orthogonal distance and a large score distance. They lie far outside the space spanned by the principal components, and after projection far from the regular data points. Typically they have a large influence on classical PCA, as the eigenvectors will be tilted towards them.

Other proposals for robust PCA include the robust LTS-subspace estimator and its generalizations<sup>7,50</sup>. The idea behind these approaches consists in minimizing a robust scale of the orthogonal distances, similar to the LTS estimator and S-estimators in regression. Also the Orthogonalized Gnanadesikan-Kettenring estimator<sup>51</sup> is fast and robust, but it is not orthogonally equivariant. This implies that when the data are first rotated, the principal components do not transform accordingly.

#### 4.4 Selecting the number of principal components

To choose the optimal number of loadings  $k_{\text{opt}}$  there exist many criteria. For a detailed overview, see Jolliffe<sup>52</sup>. A popular graphical technique is based on the *scree plot* which shows the eigenvalues

in decreasing order. One then selects the index of the last component before the plot flattens.

A more formal criterion considers the total variation which is explained by the first  $k$  loadings, and requires e.g. that

$$\left( \sum_{j=1}^{k_{\text{opt}}} l_j \right) / \left( \sum_{j=1}^p l_j \right) \geq 80\%. \quad (23)$$

Note that this criterion cannot be used with ROBPCA as the method does not yield all  $p$  eigenvalues. But we can apply it to the eigenvalues of the covariance matrix of  $\hat{\Sigma}_h$  that was constructed in the second stage of the algorithm.

Another criterion that is based on the predictive ability of PCA is the PREDicted Sum of Squares (PRESS) statistic. To compute the (cross-validated) PRESS value at a certain  $k$ , the  $i$ th observation is removed from the original data set (for  $i = 1, \dots, n$ ), the center and the  $k$  loadings of the reduced data set are estimated, and then the fitted value of the  $i$ th observation is computed following (19) and denoted as  $\hat{\mathbf{x}}_{-i}$ . Finally, we set

$$\text{PRESS}_k = \sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}_{-i}\|^2. \quad (24)$$

The value  $k$  for which  $\text{PRESS}_k$  is small enough is then considered as the optimal number of components  $k_{\text{opt}}$ . One could also apply formal F-type tests based on successive PRESS values<sup>53,54</sup>.

The  $\text{PRESS}_k$  statistic is however not suited at contaminated data sets because it also includes the prediction error of the outliers. Even if the fitted values are based on a robust PCA algorithm, their prediction error might increase  $\text{PRESS}_k$  because they fit the model badly. To obtain a robust PRESS value, the following procedure<sup>55</sup> can be applied. For each PCA model under investigation ( $k = 1, \dots, k_{\text{max}}$ ), the outliers are marked. These are the observations that exceed the horizontal and/or the vertical cut-off value on the outlier map. Next, all the outliers are collected (over all  $k$ ) and they are removed together from the sum in (24). Doing so, the robust  $\text{PRESS}_k$  value is based on the same set of observations for each  $k$ . Fast algorithms to compute such a robust PRESS value have been developed<sup>55</sup>.

## 4.5 An example

We illustrate ROBPCA and the outlier map on a data set which consists of spectra of 180 ancient glass pieces over  $p = 750$  wavelengths<sup>56</sup>. Three components are retained for CPCA and ROBPCA, yielding a classical explanation percentage of 99% and a robust explanation percentage (23) of 96%.

The resulting outlier maps are shown in Figure 6. In the outlier map in Figure 6(a) we see that CPCA does not find big outliers. On the other hand the ROBPCA map of Figure 6(b) clearly distinguishes two major groups in the data, a smaller group of bad leverage points, a few orthogonal outliers and the isolated case 180 in between the two major groups. A high-breakdown method such as ROBPCA treats the smaller group with cases 143–179 as one set of outliers. Later, it turned out that the window of the detector system had been cleaned before the last 38 spectra were measured. As a result of this less radiation (X-rays) is absorbed and more can be detected, resulting in higher X-ray intensities. The other bad leverage points (57–63) and (74–76) are samples with a large concentration of calcic. The orthogonal outliers (22, 23 and 30) are borderline cases, although it turned out that they have larger measurements at the channels 215–245. This might indicate a higher concentration of phosphor.

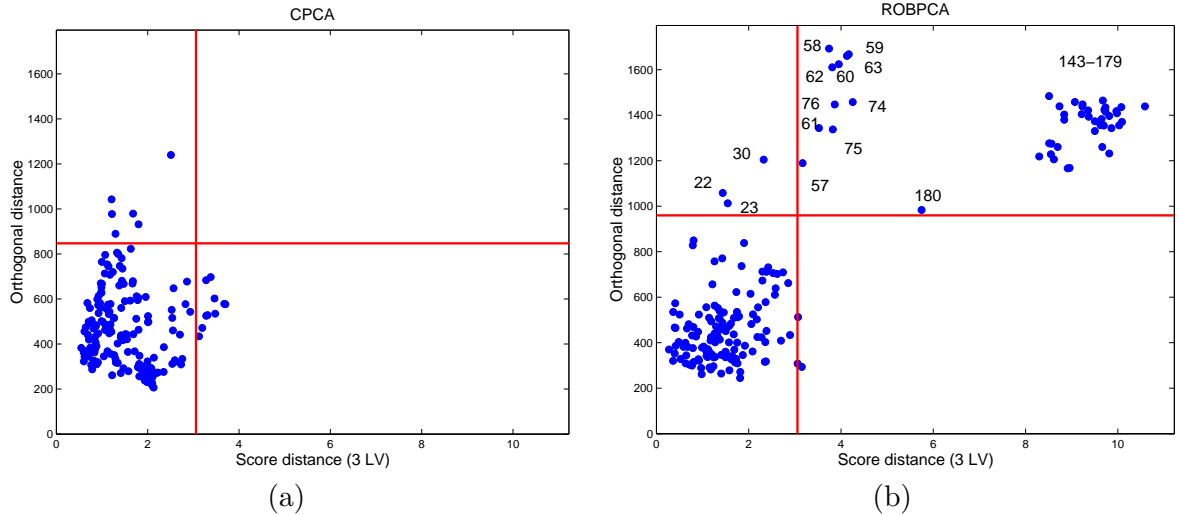


Figure 6: Outlier map of the glass data set based on three principal components computed with (a) CPCA; (b) ROBPCA.

## 5 Linear regression in high dimensions

When the number of independent variables  $p$  in a regression model is very large or when the regressors are highly correlated (this is known as *multicollinearity*), traditional regression methods such as ordinary least squares tend to fail.

An important example in chemometrics is multivariate calibration, whose goal is to predict constituent concentrations of a material based on its spectrum. Since a spectrum typically ranges over a large number of wavelengths, it is a high-dimensional vector with hundreds of components. The number of concentrations on the other hand is usually limited to at most, say, five. In the univariate approach, only one concentration at a time is modelled and analyzed. The more general problem assumes that the number of response variables  $q$  is larger than one, which means that several concentrations are to be estimated together. This model has the advantage that the covariance structure between the concentrations is also taken into account, which is appropriate when the concentrations are known to be strongly intercorrelated with each other. Here, we will write down the formulas for the general multivariate setting (14) for which  $q \geq 1$ , but they can of course be simplified when  $q = 1$ .

Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) are two methods frequently used to build regression models in very high dimensions. Both assume that the linear relation (14) between the  $x$ - and  $y$ -variables is a bilinear model which depends on  $k$  scores  $\tilde{\mathbf{t}}_i$ :

$$\mathbf{x}_i = \bar{\mathbf{x}} + \mathbf{P}_{p,k} \tilde{\mathbf{t}}_i + \mathbf{f}_i \quad (25)$$

$$\mathbf{y}_i = \bar{\mathbf{y}} + \mathbf{A}_{q,k}^T \tilde{\mathbf{t}}_i + \mathbf{g}_i. \quad (26)$$

with  $\bar{\mathbf{x}}$  and  $\bar{\mathbf{y}}$  the mean of the  $x$ - and  $y$ -variables.

Typically  $k$  is taken much smaller than  $p$ . Consequently, both PCR and PLSR first try to estimate the scores  $\tilde{\mathbf{t}}_i$ . Then a traditional regression can be used to regress the response  $y$  onto these low-dimensional scores. In order to obtain the scores, one can perform PCA on the independent variables. This is the idea behind PCR. In this case no information about the response variable is used when reducing the dimension. Therefore PLSR is sometimes more appropriate, as this method estimates the scores maximizing a covariance criterion between the independent and dependent

variable. In any case, the original versions of both PCR and PLSR strongly rely on classical PCA and regression methods, making them very sensitive to outliers.

## 5.1 Robust PCR

A robust PCR method (RPCR) was proposed by Hubert and Verboven<sup>57</sup>. In the first stage of the algorithm, robust scores  $\tilde{\mathbf{t}}_i$  are obtained by applying ROBPCA to the  $x$ -variables and retaining  $k$  components. In the second stage of RPCR, the original response variables  $\mathbf{y}_i$  are regressed on the  $\tilde{\mathbf{t}}_i$  using a robust regression method. If there is only one response variable ( $q = 1$ ), the reweighted LTS estimator (Section 3.1.2) is applied. If  $q > 1$  the MCD-regression is performed (Section 3.2). Note that the robustness of the RPCR algorithm depends on the value of  $h$  which is chosen in the ROBPCA algorithm and in the LTS and MCD-regression. Although it is not really necessary, it is recommended to use the same value in both steps. Using all robust distances in play, one can again construct outlier maps to visualize outliers and classify observations as regular observations, PCA outliers or regression outliers.

## 5.2 Robust PLSR

In PLSR the estimation of the scores (25) is a little bit more involved as it also includes information about the response variable. Let  $\tilde{\mathbf{X}}_{n,p}$  and  $\tilde{\mathbf{Y}}_{n,q}$  denote the mean-centered data matrices. The normalized PLS weight vectors  $\mathbf{r}_a$  and  $\mathbf{q}_a$  (with  $\|\mathbf{r}_a\| = \|\mathbf{q}_a\| = 1$ ) are then defined as the vectors that maximize

$$\text{cov}(\tilde{\mathbf{Y}}\mathbf{q}_a, \tilde{\mathbf{X}}\mathbf{r}_a) = \mathbf{q}_a^T \frac{\tilde{\mathbf{Y}}^T \tilde{\mathbf{X}}}{n-1} \mathbf{r}_a = \mathbf{q}_a^T \mathbf{S}_{yx} \mathbf{r}_a \quad (27)$$

for each  $a = 1, \dots, k$ , where  $\mathbf{S}_{yx}^T = \mathbf{S}_{xy} = \frac{\tilde{\mathbf{X}}^T \tilde{\mathbf{Y}}}{n-1}$  is the empirical cross-covariance matrix between the  $x$ - and the  $y$ -variables. The elements of the scores  $\tilde{\mathbf{t}}_i$  are then defined as linear combinations of the mean-centered data:  $\tilde{t}_{ia} = \tilde{\mathbf{x}}_i^T \mathbf{r}_a$ , or equivalently  $\tilde{\mathbf{T}}_{n,k} = \tilde{\mathbf{X}}_{n,p} \mathbf{R}_{p,k}$  with  $\mathbf{R}_{p,k} = (\mathbf{r}_1, \dots, \mathbf{r}_k)$ .

The computation of the PLS weight vectors can be performed using the SIMPLS algorithm<sup>58</sup>. The solution of the maximization problem (27) is found by taking  $\mathbf{r}_1$  and  $\mathbf{q}_1$  as the first left and right singular eigenvectors of  $\mathbf{S}_{xy}$ . The other PLSR weight vectors  $\mathbf{r}_a$  and  $\mathbf{q}_a$  for  $a = 2, \dots, k$  are obtained by imposing an orthogonality constraint to the elements of the scores. If we require that  $\sum_{i=1}^n t_{ia} t_{ib} = 0$  for  $a \neq b$ , a deflation of the cross-covariance matrix  $\mathbf{S}_{xy}$  provides the solutions for the other PLSR weight vectors. This deflation is carried out by first calculating the  $x$ -loading

$$\mathbf{p}_a = \mathbf{S}_x \mathbf{r}_a / (\mathbf{r}_a^T \mathbf{S}_x \mathbf{r}_a) \quad (28)$$

with  $\mathbf{S}_x$  the empirical covariance matrix of the  $x$ -variables. Next an orthonormal base  $\{\mathbf{v}_1, \dots, \mathbf{v}_a\}$  of  $\{\mathbf{p}_1, \dots, \mathbf{p}_a\}$  is constructed and  $\mathbf{S}_{xy}$  is deflated as

$$\mathbf{S}_{xy}^a = \mathbf{S}_{xy}^{a-1} - \mathbf{v}_a (\mathbf{v}_a^T \mathbf{S}_{xy}^{a-1})$$

with  $\mathbf{S}_{xy}^1 = \mathbf{S}_{xy}$ . In general the PLSR weight vectors  $\mathbf{r}_a$  and  $\mathbf{q}_a$  are obtained as the left and right singular vector of  $\mathbf{S}_{xy}^a$ .

A robust method RSIMPLS has been developed in<sup>59</sup>. It starts by applying ROBPCA to the joint  $x$ - and  $y$ -variables in order to replace  $\mathbf{S}_{xy}$  and  $\mathbf{S}_x$  by robust estimates, and then proceeds analogously to the SIMPLS algorithm.

More precisely, to obtain robust scores, ROBPCA is first applied to the joint  $x$ - and  $y$ -variables  $\mathbf{Z}_{n,m} = (\mathbf{X}_{n,p}, \mathbf{Y}_{n,q})$  with  $m = p + q$ . Assume we select  $k_0$  components. This yields a robust estimate  $\hat{\boldsymbol{\mu}}_z$  of the center of  $\mathbf{Z}$ , and following (22) an estimate  $\hat{\boldsymbol{\Sigma}}_z$  of its shape. These estimates can

then be split into blocks, just like (18). The cross-covariance matrix  $\Sigma_{xy}$  is then estimated by  $\hat{\Sigma}_{xy}$  and the PLS weight vectors  $\mathbf{r}_a$  are computed as in the SIMPLS algorithm, but now starting with  $\hat{\Sigma}_{xy}$  instead of  $\mathbf{S}_{xy}$ . In analogy with (28) the  $x$ -loadings  $\mathbf{p}_j$  are defined as  $\mathbf{p}_j = \hat{\Sigma}_x \mathbf{r}_j / (\mathbf{r}_j^T \hat{\Sigma}_x \mathbf{r}_j)$ . Then the deflation of the scatter matrix  $\hat{\Sigma}_{xy}^a$  is performed as in SIMPLS. In each step, the robust scores are calculated as  $t_{ia} = \check{\mathbf{x}}_i^T \mathbf{r}_a = (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x)^T \mathbf{r}_a$  where  $\check{\mathbf{x}}_i = \mathbf{x}_i - \hat{\boldsymbol{\mu}}_x$  are the robustly centered observations.

Next, we need a robust regression of  $\mathbf{y}_i$  on  $t_i$ . This could again be done using the MCD-regression method of Section 3.2. A faster approach is also possible<sup>59</sup>, by explicitly making use of the prior information given by ROBPCA in the first step of the algorithm.

This RSIMPLS approach yields bounded influence functions for the weight vectors  $\mathbf{r}_a$  and  $\mathbf{q}_a$  and for the regression estimates<sup>60,49</sup>. Also the breakdown value is inherited from the MCD estimator.

Another robustification of PLSR has been proposed in<sup>61</sup>. A reweighing scheme is introduced based on ordinary PLSR, leading to a fast and robust procedure. The algorithm can however only deal with the univariate case ( $q = 1$ ).

### 5.3 Model calibration and validation

An important issue in PCR and PLSR is the selection of the optimal number of scores  $k_{\text{opt}}$ . A popular approach consists of minimizing the root mean squared error of cross-validation criterion  $\text{RMSECV}_k$ , defined as

$$\text{RMSECV}_k = \sqrt{\frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i - \hat{\mathbf{y}}_{-i,k}\|^2} \quad (29)$$

with  $\hat{\mathbf{y}}_{-i,k}$  the cross-validated prediction of  $\mathbf{y}_i$  based on  $k$  scores. The goal of the  $\text{RMSECV}_k$  statistic is twofold. It yields an estimate of the root mean squared prediction error  $E(y - \hat{y})^2$  when  $k$  components are used in the model, whereas the curve of  $\text{RMSECV}_k$  for  $k = 1, \dots, k_{\text{max}}$  is a graphical tool to choose the optimal number of scores.

As argued for the PRESS statistic (24) in PCA, also this  $\text{RMSECV}_k$  statistic is not suited for contaminated data sets because it includes the prediction error of the outliers. A robust  $\text{RMSECV}$  (R- $\text{RMSECV}$ ) measure was constructed in<sup>62</sup> by omitting the outliers from the sum in (29). In a naive algorithm this approach would of course be extremely time consuming, since we have to run the entire RPCR (or RSIMPLS) algorithm  $n$  times (each deleting another observation in the cross-validation) for every possible choice of  $k$ . In<sup>62</sup> a faster algorithm was proposed, which efficiently reuses results and information from previous runs. The R- $\text{RMSECV}_k$  criterion is a robust measure of how well the model predicts the response for *new* observations. If we want to see how well the model fits the *given* observations, we can define a very similar goodness-of-fit criterion. The root mean squared error ( $\text{RMSE}_k$ ) is calculated by replacing  $\hat{\mathbf{y}}_{-i,k}$  in (29) by the fitted value  $\hat{\mathbf{y}}_{i,k}$  obtained using all observations including the  $i$ th one. As for R- $\text{RMSECV}_k$ , a robust R- $\text{RMSE}_k$  does not include the outliers to compute the average squared error. Finally, a Robust Component Selection (RCS) statistic is defined<sup>62</sup> by

$$\text{RCS}_k = \sqrt{\gamma \text{R-RMSECV}_k^2 + (1 - \gamma) \text{R-RMSE}_k^2}$$

with a tuning parameter  $\gamma \in [0, 1]$ . If the user selects a small  $\gamma$ , then the goodness-of-fit becomes the most important term. Choosing  $\gamma$  close to one on the other hand emphasizes the importance of the quality of predictions. If the user has no a priori preference,  $\gamma = 0.5$  can be selected in order to give equal weight to both terms. Finally, a plot of  $\text{RCS}_k$  versus  $k$  offers an easy way of visually selecting the most appropriate  $k$ .

## 5.4 An example

The robustness of RSIMPLS is illustrated on the octane data set<sup>63</sup> consisting of NIR absorbance spectra over  $p = 226$  wavelengths ranging from 1102nm to 1552nm with measurements every two nm. For each of the  $n = 39$  production gasoline samples the octane number  $y$  was measured, so  $q = 1$ . It is known that the octane data set contains six outliers (25, 26, 36–39) to which alcohol was added.

Figure 7 shows the RCS curves for  $\gamma$  equal to 0, 0.5 and 1. Choosing  $k = 1$  is clearly a bad idea, since the prediction error is very high in that case. From two components on, the RCS curve becomes rather stable. For  $\gamma \geq 0.5$ , the minimal error is attained at  $k = 6$ , which would be a good choice. The difference with  $k = 2$  is not very big however, so for the sake of simplicity we decided to retain  $k = 2$  components.

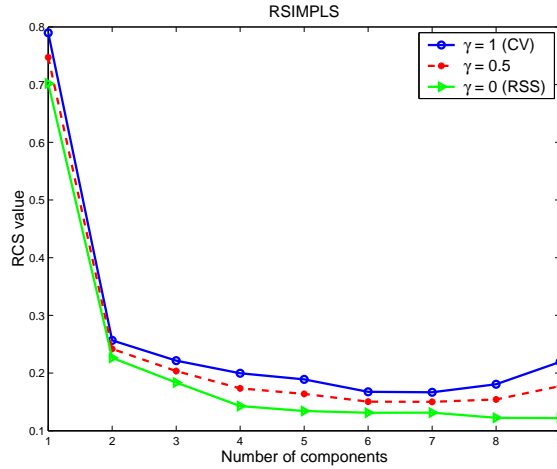


Figure 7: RCS curve for the octane data set,  $\gamma = 0$  (line with triangles),  $\gamma = 0.5$  (dashed line) and  $\gamma = 1$  (line with circles).

The resulting outlier maps are shown in Figure 8. The robust PCA outlier map is displayed in Figure 8(a). According to model (25), it displays the score distance  $SD_i = \sqrt{(\mathbf{t}_i - \hat{\boldsymbol{\mu}}_t)^T \hat{\boldsymbol{\Sigma}}_t^{-1} (\mathbf{t}_i - \hat{\boldsymbol{\mu}}_t)}$  on the horizontal axis, where  $\hat{\boldsymbol{\mu}}_t$  and  $\hat{\boldsymbol{\Sigma}}_t$  are derived in the regression step of the RSIMPLS algorithm. On the vertical axis it shows the orthogonal distance of the observation to  $t$ -space, so  $OD_i = \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x - \mathbf{P}_{p,k} \mathbf{t}_i\|$ . We immediately spot the six samples with added alcohol. The SIMPLS outlier map is shown in Figure 8(b). We see that this analysis only detects the outlying spectrum 26, which does not even stick out much above the border line. The robust regression outlier map in Figure 8(c) shows that the outliers are good leverage points, whereas SIMPLS again only reveals case 26.

## 6 Classification

### 6.1 Classification in low dimensions

The goal of classification, also known as discriminant analysis or supervised learning, is to obtain rules that describe the separation between known groups of observations. Moreover, it allows to classify new observations into one of the groups. We denote the number of groups by  $l$  and assume that we can describe our experiment in each population  $\pi_j$  by a  $p$ -dimensional random variable

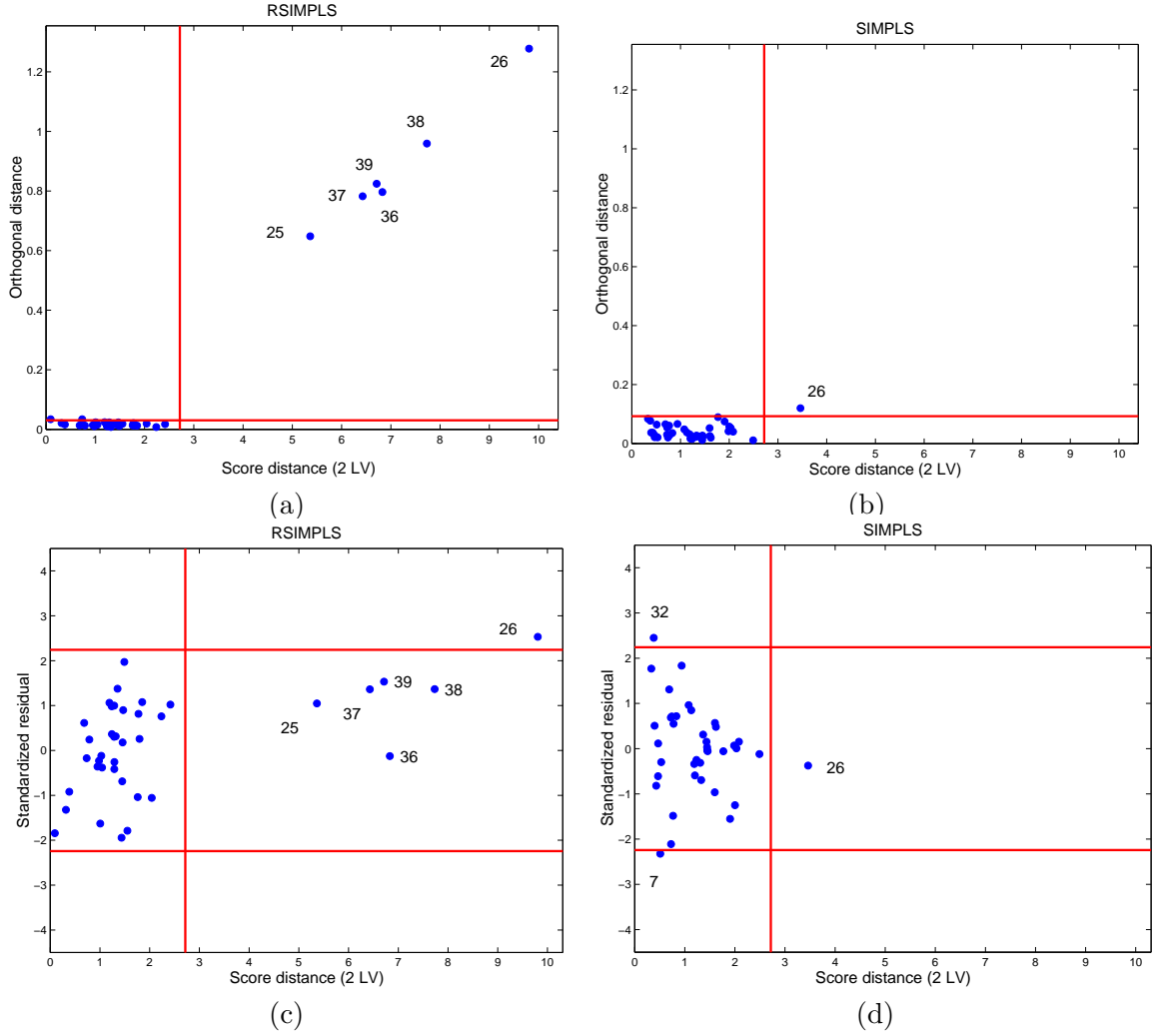


Figure 8: (a) PCA outlier map of the octane data set obtained with RSIMPLS; (b) with SIMPLS; (c) Regression outlier map obtained with RSIMPLS; (d) with SIMPLS.

$X_j$  with distribution function (density)  $f_j$ . We write  $p_j$  for the membership probability, i.e. the probability for an observation to come from  $\pi_j$ .

The maximum likelihood rule classifies an observation  $\mathbf{x}$  in  $\mathbb{R}^p$  into  $\pi_m$  if  $\ln(p_m f_m(\mathbf{x}))$  is the maximum of the set  $\{\ln(p_j f_j(\mathbf{x})); j = 1, \dots, l\}$ . If we assume that the density  $f_j$  for each group is gaussian with mean  $\boldsymbol{\mu}_j$  and covariance matrix  $\boldsymbol{\Sigma}_j$ , then it can be seen that the maximum likelihood rule is equivalent to maximizing the discriminant scores  $d_j^Q(\mathbf{x})$  with

$$d_j^Q(\mathbf{x}) = -\frac{1}{2} \ln |\boldsymbol{\Sigma}_j| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln(p_j). \quad (30)$$

That is,  $\mathbf{x}$  is allocated to  $\pi_m$  if  $d_m^Q(\mathbf{x}) \geq d_j^Q(\mathbf{x})$  for all  $j = 1, \dots, l$  (see e.g. Johnson and Wichern<sup>64</sup>).

In practice  $\boldsymbol{\mu}_j$ ,  $\boldsymbol{\Sigma}_j$  and  $p_j$  have to be estimated. Classical Quadratic Discriminant Analysis (CQDA) uses the group's mean and empirical covariance matrix to estimate  $\boldsymbol{\mu}_j$  and  $\boldsymbol{\Sigma}_j$ . The membership probabilities are usually estimated by the relative frequencies of the observations in each group, hence  $\hat{p}_j^C = n_j/n$  with  $n_j$  the number of observations in group  $j$ .

A Robust Quadratic Discriminant Analysis<sup>65</sup> (RQDA) is derived by using robust estimators of  $\boldsymbol{\mu}_j$ ,  $\boldsymbol{\Sigma}_j$  and  $p_j$ . In particular, if the number of observations is sufficiently large with respect to the dimension  $p$ , we can apply the reweighted MCD estimator of location and scatter in each group (Section 2.3.2). As a byproduct of this robust procedure, outliers (within each group) can be distinguished from the regular observations. Finally, the membership probabilities can be robustly estimated as the relative frequency of the *regular* observations in each group, yielding  $\hat{p}_j^R$ .

When all the covariance matrices are assumed to be equal, the quadratic scores (30) can be simplified to

$$d_j^L(\boldsymbol{x}) = \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \ln(p_j) \quad (31)$$

where  $\boldsymbol{\Sigma}$  is the common covariance matrix. The resulting scores (31) are linear in  $\boldsymbol{x}$ , hence the maximum likelihood rule belongs to the class of *linear discriminant analysis*. It is well known that if we have only two populations ( $l = 2$ ) with a common covariance structure and if both groups have equal membership probabilities, this rule coincides with Fisher's linear discriminant rule. Again the common covariance matrix can be estimated by means of the MCD estimator, e.g. by pooling the MCD estimates in each group. Robust linear discriminant analysis based on the MCD estimator (or S-estimators) has been studied by several authors<sup>66,67,68,65</sup>.

## 6.2 Classification in high dimensions

### 6.2.1 Classical SIMCA

When the data are high-dimensional, the approach of the previous section cannot be applied anymore because the MCD becomes uncomputable. This problem can be solved by applying a dimension reduction procedure (PCA) on the *whole* set of observations. Instead, one can also apply a PCA method on each group separately. This is the idea behind the SIMCA method (Soft Independent Modelling of Class Analogy)<sup>69</sup>.

Suppose again that we have  $l$  groups with  $p$ -dimensional data matrices  $\boldsymbol{X}^j$ ,  $j = 1, \dots, l$ . Denote  $n_j$  the number of observations in group  $j$ . The SIMCA method starts by performing PCA on each group  $\boldsymbol{X}^j$  separately. Let  $k_j$  denote the number of retained principal components in group  $j$ . New observations are then classified by means of their distances to the different PCA models. The choice of an appropriate distance however is a difficult task. A first idea is to use the orthogonal distances obtained from the PCA analysis, cfr. (20). Denote  $OD^{(j)}$  the orthogonal distance from a new observation  $\boldsymbol{x}$  to the PCA hyperplane for the  $j$ th group. Denote  $OD_i^j$  the orthogonal distance from the  $i$ th observation in group  $j$  to the PCA hyperplane for the  $j$ th group. Then for  $j$  ranging from 1 to  $l$ , an  $F$ -test is performed with test statistic  $(s^{(j)}/s_j)^2$  where

$$(s^{(j)})^2 = \frac{(OD^{(j)})^2}{p - k_j} \quad \text{and} \quad s_j^2 = \frac{\sum_{i=1}^{n_j} (OD_i^j)^2}{(p - k_j)(n_j - k_j - 1)}.$$

If the observed  $F$ -value is smaller than the critical value,  $\boldsymbol{x}$  is said to belong to group  $j$ . Note that an observation can be classified in many different groups, hence the term *Soft* in SIMCA.

This approach based on the orthogonal distances only, turned out not to be completely satisfactory. To fully exploit applying PCA in each group separately, it was suggested to include the score distances (21) as well. They can be used to construct a multidimensional box around the  $j$ th PCA model. In Figure 9(a) these boxes are plotted for a three-dimensional example with three groups. A boundary distance  $BD^{(j)}$  is then defined as the distance of a new observation  $\boldsymbol{x}$  to the box for the  $j$ th group. Assigning  $\boldsymbol{x}$  to any of the  $l$  classes is then done by means of an  $F$ -test based on a linear combination of  $(BD^{(j)})^2$  and  $(OD^{(j)})^2$ .

### 6.2.2 Robust SIMCA

A first step in robustifying SIMCA can be obtained by applying a robust PCA method, such as ROBPCA (Section 4.3), to each group<sup>70</sup>. The number of components in each group can e.g. be selected by robust cross-validation, as explained in Section 4.4. Also the classification rule needs to be changed, since the SIMCA-boxes are defined in such a way that they contain all the observations. When outliers are present these boxes can thus be highly inflated, as demonstrated in Figure 9(a).

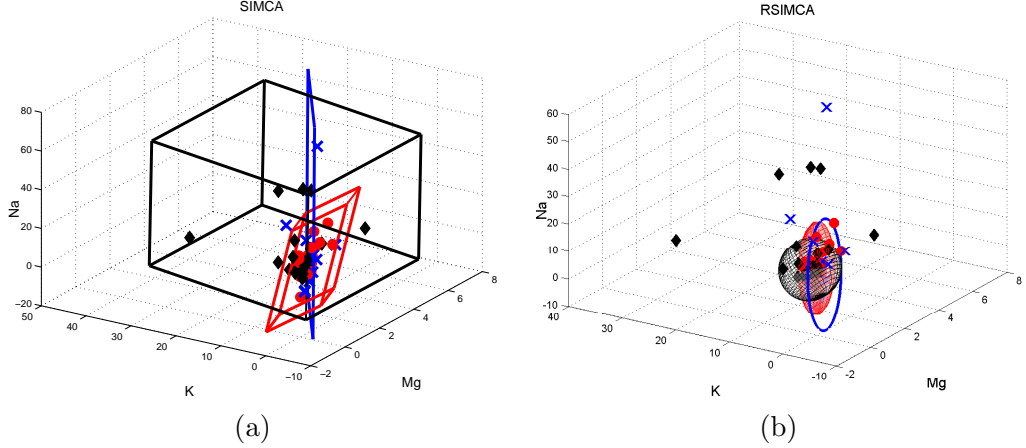


Figure 9: (a) Boxes based on classical PCA, which are blown up by some outliers; (b) ellipsoids based on robust PCA.

Figure 9(b) shows robust ellipsoids<sup>70</sup> of the groups. A new observation  $\mathbf{x}$  is classified in group  $m$  if

$$\gamma \left( \frac{\text{OD}_j(\mathbf{x})}{c_j^v} \right)^2 + (1 - \gamma) \left( \frac{\text{SD}_j(\mathbf{x})}{c_j^h} \right)^2$$

is smallest for  $j = m$ , where OD (resp. SD) now denotes the orthogonal (resp. score) distance to a robust PCA model. The numbers  $c_j^v$  and  $c_j^h$  are carefully chosen normalizers. The tuning parameter  $0 \leq \gamma \leq 1$  is added for two reasons. If the user a priori judges that the OD (resp. the SD) is the most important criterion to build the classifier, the parameter  $\gamma$  can be chosen close to one (resp. zero). Otherwise,  $\gamma$  can be selected such that the misclassification probability is minimized. This probability can be estimated by means of a validation set or cross-validation.

Also in this setting, outlier maps are helpful graphical tools to gain more insight in the data. We illustrate RSIMCA on the fruit data set<sup>65</sup>. It contains the spectra of three different cultivars of a melon. The cultivars (named D, M and HA) have sizes 490, 106 and 500, and all spectra are measured in 256 wavelengths. The RSIMCA classification rules were derived on a training set, consisting of a random subset of 60% of the samples. Figure 10(a) shows the outlier map of cultivar HA. It plots the  $(\text{SD}_i, \text{OD}_i)$  for all 500 observations, using circles for the training data and crosses for the validation data. We immediately detect a large group of outliers. As it turns out, these outliers were caused by a change in the illumination system. Figure 10(b) depicts the corresponding outlier map using the classical SIMCA results. In this case the outliers remain undetected, another example of the masking effect that non-robust methods can suffer from.

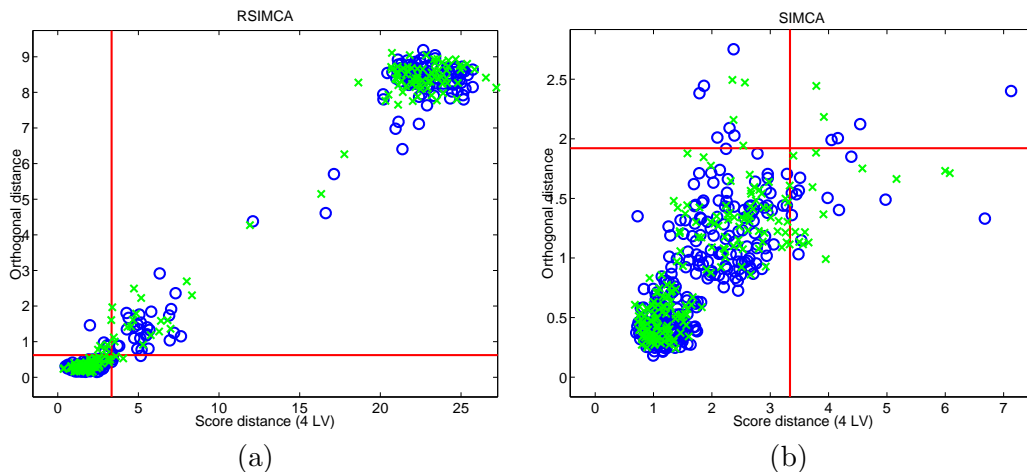


Figure 10: Fruit data, cultivar HA. (a) Outlier map for RSIMCA; (b) outlier map for classical SIMCA.

## 7 Support Vector Machines

In all methods considered so far, a linear model was explicitly assumed. In real applications this can sometimes be too restrictive, as more complicated non-linear structures might underlie the data. In this section we briefly describe Support Vector Machines, a powerful tool for non-linear modelling. Excellent introductory material on this subject can be found in<sup>71,72,73</sup>.

### 7.1 Linear classification

Assuming two groups of data, the classification data can be formulated as  $n$  observations  $(\mathbf{x}_i, y_i)$  with input data  $\mathbf{x}_i$  in  $\mathbb{R}^p$  and group labels  $y_i$  in  $\{-1, 1\}$ . Consider a linear classifier

$$\hat{y} = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$$

defined by the slope parameter  $\mathbf{w}$  in  $\mathbb{R}^p$  and the offset  $b$  in  $\mathbb{R}$ . We say that the two groups of data are *separable* if there exists a linear hyperplane that provides a perfect classification. Many such separating hyperplanes may exist. To select an optimal classifier, Vapnik<sup>74</sup> considered a rescaling of the problem such that the points closest to the hyperplane satisfy  $|\mathbf{w}^T \mathbf{x}_i + b| = 1$ . When the data of the two classes are separable, we have that  $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$  for all  $i = 1, \dots, n$ . After rescaling the problem in this way, one can prove that the *margin* equals  $2/\|\mathbf{w}\|_2$ . The margin is twice the distance between the hyperplane and the nearest point (see Figure 11(a) for an illustration). The optimal classifier is then defined as the one that maximizes this margin, or equivalently minimizes  $\|\mathbf{w}\|_2/2$ . Thus the primal optimization problem for linear Support Vector Machines (SVM) in case of separable data is formulated as follows:

$$\min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w} \quad \text{such that} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, n. \quad (32)$$

One of the key ingredients of SVM's is the possibility to translate this primal formulation into a so-called dual formulation. Using Lagrange multipliers, one can prove that the decision boundary

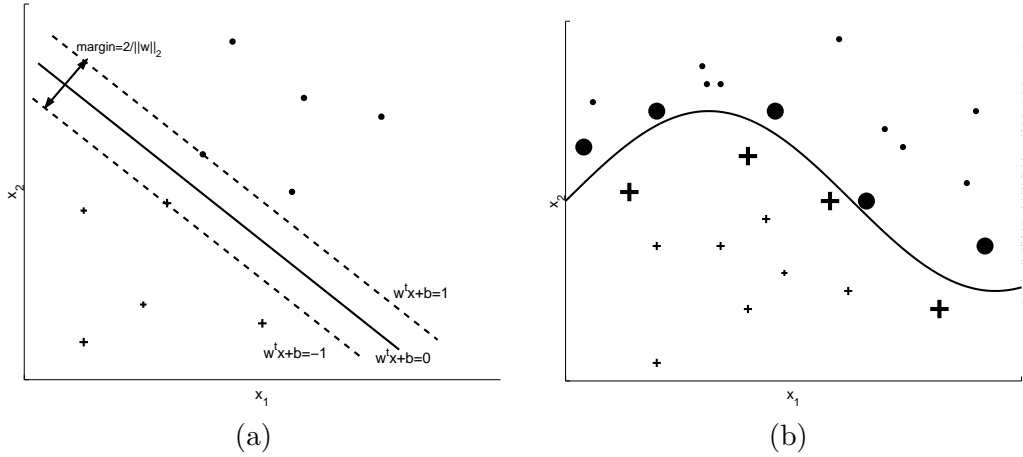


Figure 11: (a) Maximizing the margin; (b) Non-linear SVM classification. The enlarged observations are the support vectors determining the decision boundary.

corresponding to the solution of (32) can be written in the form

$$\hat{y}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b\right). \quad (33)$$

The numbers  $\alpha_i$ ,  $i = 1, \dots, n$  can be obtained from a quadratic programming problem:

$$\max_{\alpha} \left( -\frac{1}{2} \sum_{i,l=1}^n y_i y_l \mathbf{x}_i^T \mathbf{x}_l \alpha_i \alpha_l + \sum_{i=1}^n \alpha_i \right) \quad \text{such that} \quad \sum_{i=1}^n \alpha_i y_i = 0. \quad (34)$$

This dual formulation has some interesting properties:

- It can be solved using common techniques from optimization theory. Moreover, the solution is unique under mild conditions.
- Many of the resulting  $\alpha_i$  values are zero. Hence the obtained solution vector is sparse. This means that the classifier in (33) can be constructed using only the non-zero  $\alpha_i$ . Therefore, fast evaluation of the classification rule is possible even for very large data sets. Moreover the non-zero  $\alpha_i$  have geometrical meaning since they are the observations close to the decision boundary (see Figure 11(b)). For these reasons the non-zero  $\alpha_i$  received the name *support vectors*.
- The complexity of the dual problem grows with the number of observations  $n$  and not with the dimension  $p$ . Therefore Support Vector Machines can easily handle very high-dimensional data sets.

## 7.2 General SVM classification

The assumption of separable data is of course very unrealistic. It is however possible to extend the SVM framework to non-separable and even non-linear cases<sup>75</sup>. The primal optimization problem becomes

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \mathbf{w}^T \mathbf{w} + c \sum_{i=1}^n \xi_i \quad \text{such that} \quad y_i (\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i = 1, \dots, n. \quad (35)$$

Comparing with (32) there are two main differences. In order to cover the case of non-separable data, slack variables  $\xi_i$  were introduced to allow for misclassification. The penalty for misclassifications is determined by the constant  $c$ . A good choice of  $c$  is very important. If  $c$  is too small, misclassifications are not penalized enough, resulting in a poor classification. If  $c$  is too large, the method will overfit the given observations and suffer from a lack of generalization, leading to poor performance when classifying new observations. In practice  $c$  is usually determined through cross-validation. A second difference between (32) and (35) is the appearance of the function  $\phi$ , which is called the *feature map*, allowing non-linear decision boundaries  $\hat{y}(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$ .

This primal problem is impossible to solve due to the unknown feature map  $\phi$ . However, in the dual formulation this  $\phi$  miraculously disappears in favor of a pre-defined kernel function  $K : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R} : (\mathbf{x}, \mathbf{t}) \rightarrow K(\mathbf{x}, \mathbf{t})$ . This is the *kernel trick*. Many possible kernels are proposed in the literature. The most common choices are the linear kernel  $\mathbf{x}^T \mathbf{t} + 1$  for linear classification, the polynomial kernel for polynomial decision boundaries (e.g. a parabola), and the Gaussian kernel (with bandwidth  $\sigma$ )  $K(\mathbf{x}, \mathbf{t}) = e^{-\|\mathbf{x}-\mathbf{t}\|^2/(2\sigma^2)}$  for general nonlinear, semiparametric decision boundaries.

The dual problem of (35) then becomes

$$\max_{\alpha} \left( -\frac{1}{2} \sum_{i,l=1}^n y_i y_l K(\mathbf{x}_i, \mathbf{x}_l) \alpha_i \alpha_l + \sum_{i=1}^n \alpha_i \right) \quad \text{such that} \quad \sum_{i=1}^n \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq c.$$

This is again a straightforward quadratic programming problem, yielding a sparse vector of solutions  $\alpha_i$  defining the decision boundary

$$\hat{y}(\mathbf{x}) = \text{sign} \left( \sum_{i=1}^n \alpha_i y_i K(\mathbf{x}_i^T, \mathbf{x}) + b \right)$$

that one can use to classify a new observation  $\mathbf{x}$  in  $\mathbb{R}^p$ .

Several related methods have been proposed in recent years. An important extension consists of replacing the slack variables  $\xi_i$  in (35) by some loss function of  $\xi_i$ . One could for example use  $\xi_i^2$ , resulting in Least Squares SVM<sup>76</sup>. This way the dual formulation becomes a linear programming problem, which can generally be solved faster than a quadratic one. On the other hand sparseness is lost, since all solutions  $\alpha_i$  are non-zero for Least Squares SVM. Another possibility is to consider a logistic loss function<sup>77</sup>. The resulting Kernel Logistic Regression method can be appealing if one wants to estimate the probability that an observation belongs to a certain class. Such probabilities can only be estimated via this method, not via the standard SVM formulation. On the other hand, kernel logistic regression is often very involved computationally.

### 7.3 Robustness

Support Vector Machines and related methods are gaining popularity because they offer good predictions in complex and high dimensional data structures. In recent years some interesting results were obtained with respect to their robustness. In<sup>78</sup> the influence function of a broad class of kernel classifiers was investigated. It was proven that the robustness properties of SVM classification strongly depend on the choice of the kernel. For an unbounded kernel, e.g. a linear kernel, the resulting SVM methods are not robust and suffer the same problems as traditional linear classifiers (such as the classical SIMCA method from Section 6.2.1). However, when a bounded kernel is used, such as the Gaussian kernel, the resulting non-linear SVM classification handles outliers quite well. Thus non-linear modelling is not only appealing when one suspects a complex data structure for which a linear decision boundary is too simple: it can also be an interesting option when the data might contain outliers.

## 7.4 Regression

Originally Support Vector Machines were introduced for classification, as outlined in the previous paragraphs. Nowadays the framework is much broader. Similar ideas have been successfully proposed in many other areas of statistical modelling<sup>79,76</sup>. For regression, the main ideas are very similar to the classification case including

1. Choice of a loss function. The most common choice is Vapnik's  $\epsilon$ -insensitive loss function, a modification of the traditional  $L_1$  loss function, leading to Support Vector Regression (SVR). Also an  $L_2$  loss function (Least Squares SVR) can be appealing from a computationally.
2. Choice of a kernel. The main choices are the linear, polynomial, and Gaussian kernel, leading to linear, polynomial, and nonlinear semiparametric regression. Within the chemometrical literature, the Pearson Universal Kernel was very recently introduced<sup>80</sup> and good results were reported.
3. A primal formulation that can be translated into a dual formulation with all the same benefits as in the classification case: the solution vector is sparse; obtaining the solution is computationally very feasible (quadratic programming problem for Vapnik's  $\epsilon$ -insensitive loss, linear programming for least squares loss); the complexity of the dual problem to solve is independent of the dimension of the problem at hand.

The robustness of these kernel-based regression methods was recently investigated<sup>81</sup>. As in classification, the kernel plays an important role. A linear kernel leads to non-robust methods. Hence linear support vector regression can suffer the same problems with outliers as traditional methods, such as ordinary least squares regression in low dimensions or PCR and PLS in high dimensions. On the other hand, a bounded kernel (e.g. the Gaussian kernel) leads to quite robust methods with respect to outliers in  $\mathbf{x}$ -space. In order to reduce the effect of vertical outliers, one should choose a loss function with a bounded derivative. Vapnik's  $\epsilon$ -insensitive loss function is thus a good choice, in contrast with the nonrobust least squares loss function. However, the latter can be improved drastically by some reweighting steps<sup>82</sup>.

## 7.5 An example

We reconsider the octane data set from Section 5.4. There we found six outlying observations using RSIMPLS and its outlier map. Let us now look at the predictive power of RSIMPLS and kernel based regression. For this, we compute the RMSECV criterion (29) but take the sum only over all regular observations. For RSIMPLS this yields a prediction error of 0.32 when  $k = 2$  scores are considered, and a prediction error of 0.24 at  $k = 6$ . When we use a support vector regressor with Vapnik's  $\epsilon$ -insensitive loss function ( $\epsilon=0.05$ ) and a Gaussian kernel, we obtain RMSECV = 0.22.

SVR thus yields the smallest prediction error. It is however very important to note that RSIMPLS and SVR deal with the outliers in a very different way. RSIMPLS explicitly tries to detect the outliers and downweight their effects. The robustness of SVR is due to its non-linearity and inherent flexibility, making it possible to build one model giving good prediction results for both the good data and the outliers. However, as the outliers are taken into the model, it is impossible to detect them in a straightforward way. In this respect the methods are quite complementary. The raw predictive power of non-linear SVM's and the exploratory, insightful tools of linear robust methods can provide a powerful combination for high-dimensional data analysis.

## 8 Multi-way analysis

So far only data *matrices* were considered, but many experiments give rise to more complex data structures, where different sets of variables are measured at the same time. It is shown in e.g.<sup>83</sup> that preserving the nature of the data set by arranging it in a higher-dimensional tensor, instead of forcing it into a matrix, leads to a better understanding and more precise models. Such complex data structures are called *multi-way* data sets. Different techniques exist to analyze these multi-way arrays, among which PARAFAC and Tucker3 are the most popular ones (see e.g.<sup>84,85,86</sup>). They aim at constructing scores and loadings to express the data in a more comprehensive way. From this point of view, PARAFAC and Tucker3 can be seen as generalizations of principal components analysis (PCA) to higher-order tensors<sup>87</sup>.

The methods discussed below are only described for three-way data, but generalizations to higher-order tensors are possible. Three-way data  $\underline{X}$  contain  $I$  observations that are measured for  $J$  and  $K$  variables. PARAFAC decomposes the data into trilinear components. The structural model can be described as

$$x_{ijk} = \sum_{f=1}^F a_{if} b_{jf} c_{kf} + e_{ijk} \quad (36)$$

where  $a_{if}$ ,  $b_{jf}$  and  $c_{kf}$  are parameters describing the importance of the samples/variables to each component. The residual  $e_{ijk}$  contains the variation not captured by the PARAFAC model. In terms of the unfolded matrix  $X^{I \times JK}$ , this model can also be written as

$$X^{I \times JK} = A(C \odot B)^T + E \quad (37)$$

with  $A$  an  $(I \times F)$ -matrix of scores,  $B$  a  $(J \times F)$ -matrix of loadings, and  $C$  a  $(K \times F)$ -matrix of loadings. The number  $F$  stands for the number of factors to include in the model,  $E$  is the error term and  $\odot$  is the Kathri-Rao product, which is defined by  $C \odot B = [\text{vec}(\mathbf{b}_1 \mathbf{c}_1^T), \dots, \text{vec}(\mathbf{b}_F \mathbf{c}_F^T)]$ . The *vec* operator yields the vector obtained by unfolding a matrix column-wise to one column. The PARAFAC model is often used to decompose fluorescence data into tri-linear components according to the number of fluorophores ( $F$ ) present in the samples. The observed value  $x_{ijk}$  then corresponds to the intensity of sample  $i$  at emission wavelength  $j$  and excitation wavelength  $k$ .

The scores and loadings of the PARAFAC model are estimated by minimizing the objective function

$$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (x_{ijk} - \hat{x}_{ijk})^2 = \|X - \hat{X}\|_F^2 = \|X - \hat{A}(\hat{C} \odot \hat{B})^T\|_F^2. \quad (38)$$

An algorithm based on alternating Least Squares (ALS) is used for this purpose (see e.g.<sup>87</sup>). This means that given initial estimates for  $B$  and  $C$ ,  $A$  is estimated conditionally on  $B$  and  $C$  by minimizing (38). If we define  $Z = C \odot B$  the optimization problem can be reduced to minimizing  $\|X - AZ^T\|_F^2$ , which gives rise to the classical least squares regression problem. A least squares estimate for  $A$  is therefore given by  $\hat{A} = XZ(Z^T Z)^+$  with  $(Z^T Z)^+$  the Moore-Penrose inverse of  $(Z^T Z)$ . Estimates for  $B$  and  $C$  are found analogously.

It is obvious that the model will be highly influenced by outliers, as a least squares minimization procedure is used. It is important to distinguish two types of outliers. The first type, called outlying samples, are observations that have a deviating profile compared to the majority of observations. A second type of outliers, called elementwise outliers, are individual data elements that are unexpected, whereas other data elements of the same sample or the same data element for other samples may be fine. A typical example is scattering in fluorescence data.

Dealing with both types of outliers requires different approaches. For the elementwise corruption an automated method for identifying scattering in fluorescence data was proposed<sup>88</sup>, which uses the ROBPCA method of Section 4.3.

To cope with entire outlying samples, a robust counterpart for PARAFAC has been constructed<sup>89</sup>. The procedure starts by looking for  $\frac{I}{2} < h < I$  points that minimize the objective function (38). The value of  $h$  plays the same role as in the MCD estimator (Section 2.3.2) and the LTS estimator (Section 3.1.2). To find this optimal  $h$ -subset, ROBPCA is applied to the unfolded data  $X^{I \times JK}$  and the  $h$  points with the smallest residuals from the robust subspace are taken as initial  $h$ -subset. After performing the classical PARAFAC algorithm on these  $h$  points, the  $h$ -subset is updated by taking the  $h$  observations with smallest residuals. This whole procedure is iterated until the relative change in fit becomes small. Finally, a reweighting step is included to increase the accuracy of the method.

We illustrate the robust methods on the three-way Dorrit data<sup>90,91</sup>, which is a fluorescence data set containing both scattering and outlying samples. Four fluorophores are mixed together for different sets of concentrations, so  $F = 4$ . The data set contains 27 excitation-emission (EEM) landscapes with emission wavelengths ranging from 200nm to 450nm every 5nm for excitation at wavelengths from 200nm to 350nm at 5nm intervals. Some noisy parts situated at the excitation wavelengths from 200 - 230nm and at emission wavelengths below 260nm are excluded before the PARAFAC models are built. This means that we end up with a data array of size  $27 \times 116 \times 18$ .

First we detect the scatter regions by the automated method<sup>88</sup> and replace the identified elementwise outliers by interpolated values. To find the outlying EEM-landscapes, we apply the robust PARAFAC method and construct outlier maps (defined analogously to PCA). The classical and robust outlier maps are depicted in Figure 12. In the classical outlier map, observation 10 is marked as a residual outlier whereas samples 2, 3, 4 and 5 are flagged as good leverage points. The robust analysis yields a very different conclusion. It shows that that samples 2, 3 and 5 are bad leverage points. In order to find out which of these results we should trust, the emission and

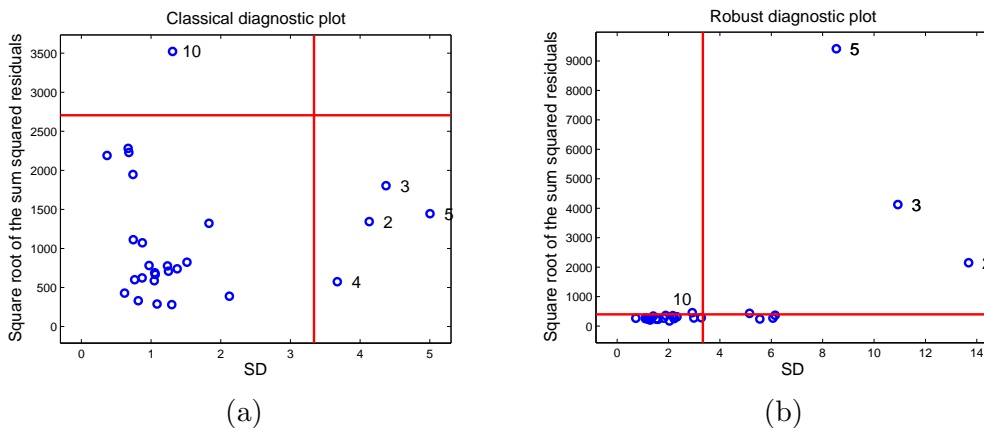


Figure 12: Outlier maps of the Dorrit data set based on (a) classical PARAFAC and (b) robust PARAFAC.

excitation loadings are plotted in Figure 13. The classical results in Figure 13(a–b) are corrupted by the outliers, as neither the emission loadings nor the excitation loadings correspond to the expected profiles of the known chemical compounds (see e.g.<sup>90</sup>). Moreover, we can compute the angle between the estimated and the reference subspaces for the  $B$ - and  $C$ -loadings. This yields 1.34 and 0.44, which confirms that the classical estimates are far off. On the other hand, based on visual

inspection of Figure 13(c-d) and the angles 0.06 for the  $B$ -loadings and 0.18 for the  $C$ -loadings, we can conclude that the robust algorithm has succeeded in estimating the underlying structure of the data much more accurately.

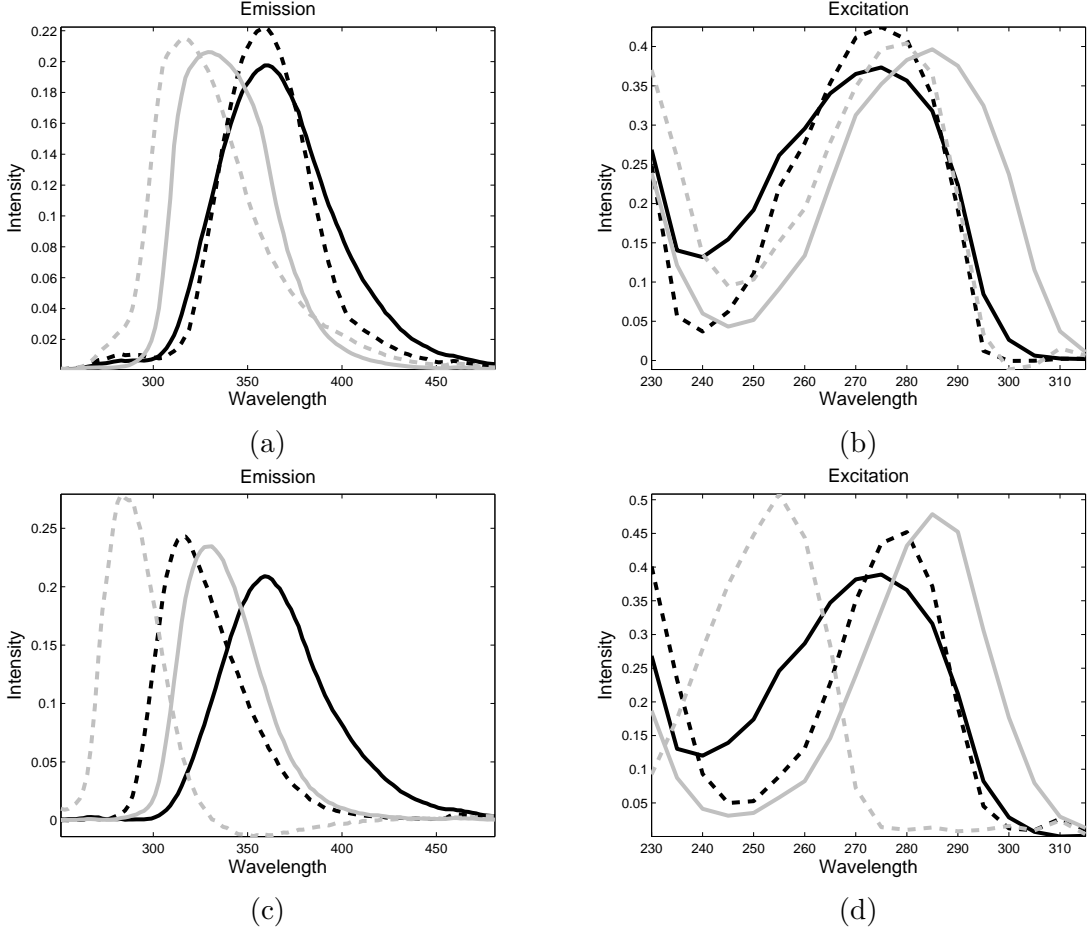


Figure 13: Emission (left) and excitation (right) loadings for the Dorrit data set, using the classical (top) and robust (bottom) PARAFAC algorithms.

The PARAFAC model can be generalized to the Tucker3 model<sup>92</sup> which adds an  $L \times M \times N$  core matrix  $\underline{Z}$  in (36):

$$x_{ijk} = \sum_{l=1}^L \sum_{m=1}^M \sum_{n=1}^N a_{il} b_{jm} c_{kn} z_{lmn} + e_{ijk}.$$

In<sup>93</sup> a robust algorithm is proposed, based on the MCD estimator and iterated steps as in the robust PARAFAC method. From our experience this method works well in practice, although it can probably be improved by using a more robust initial  $h$ -subset and by computing more precise outlier cutoff values.

## 9 Software availability

Matlab functions for most of the procedures mentioned in this paper are part of the LIBRA Toolbox<sup>94</sup>, which can be downloaded from

<http://www.wis.kuleuven.ac.be/stat/robust/LIBRA.html>.

Stand-alone programs carrying out FAST-MCD and FAST-LTS can be downloaded from the website <http://www.agoras.ua.ac.be/>, as well as Matlab versions. The MCD is already available in the packages S-PLUS and R as the built-in function *cov.mcd* and has been included in SAS Version 11 and SAS/IML Version 7. These packages all provide the one-step reweighted MCD estimates. The LTS is available in S-PLUS and R as the built-in function *ltsreg* and has also been incorporated in SAS Version 11 and SAS/IML Version 7.

Support Vector Machines are implemented in numerous software packages. An extensive list can be found at <http://www.kernel-machines.org>. For Least Squares SVM applications, a Matlab toolbox is available at <http://www.esat.kuleuven.be/sista/lssvmlab>.

PARAFAC and related multiway methods are incorporated in the PLS-toolbox at <http://software.eigenvector.com>. Stand-alone tools are freely available from <http://www.models.kvl.dk/source>.

## References

1. F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions* (New York: Wiley, 1986).
2. P.J. Huber, *Robust Statistics* (New York: Wiley, 1981).
3. P.J. Rousseeuw and C. Croux, Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88 (1993):1273–1283.
4. C. Reimann and P. Filzmoser, Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology*, 39 (2000):1001–1014.
5. E. Vandervieren and M. Hubert, An adjusted boxplot for skewed distributions, in *Proceedings in Computational Statistics*, ed. J. Antoch (Heidelberg: Springer, Physica-Verlag, 2004), 1933–1940
6. G. Brys, M. Hubert, and A. Struyf, A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13 (2004):996–1017.
7. P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection* (New York: Wiley, 1987).
8. P.J. Rousseeuw, Least median of squares regression. *Journal of the American Statistical Association*, 79 (1984):871–880.
9. C. Croux and G. Haesbroeck, Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71 (1999):161–190.
10. P.J. Rousseeuw and K. Van Driessen, A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics*, 41 (1999):212–223.
11. W.A. Stahel, *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen* (ETH Zürich: PhD thesis, 1981).
12. D.L. Donoho, *Breakdown properties of multivariate location estimators* (Harvard University, Boston: Qualifying paper, 1982).

13. D.E. Tyler, Finite-sample breakdown points of projection-based multivariate location and scatter statistics. *The Annals of Statistics*, 22 (1994):1024–1044.
14. R.A. Maronna and V.J. Yohai, The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90 (1995):330–341.
15. R.A. Maronna, Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4 (1976):51–67.
16. P.J. Rousseeuw, Multivariate estimation with high breakdown point, in *Mathematical Statistics and Applications, Vol. B*, ed. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (Dordrecht: Reidel Publishing Company, 1985), 283–297.
17. L. Davies, Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15 (1987):1269–1292.
18. J.T. Kent and D.E. Tyler, Constrained M-estimation for multivariate location and scatter. *The Annals of Statistics*, 24 (1996):1346–1370.
19. H.P. Lopuhaä, Multivariate  $\tau$ -estimators for location and scatter. *The Canadian Journal of Statistics*, 19 (1991):307–321.
20. K.S. Tatsuoka and D.E. Tyler, On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *The Annals of Statistics*, 28 (2000):1219–1243.
21. S. Visuri, V. Koivunen, and H. Oja, Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91 (2000):557–575.
22. D.L. Donoho and M. Gasko, Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20 (1992):1803–1827.
23. R.Y. Liu, J.M. Parelius, and K. Singh, Multivariate analysis by data depth: descriptive statistics, graphics and inference. *The Annals of Statistics*, 27 (1999):783–840.
24. P.J. Rousseeuw and A. Struyf, Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8 (1998):193–203.
25. G. Brys, M. Hubert, and P.J. Rousseeuw, A robustification of Independent Component Analysis. *Journal of Chemometrics*, 19 (2005):364–375.
26. P.J. Rousseeuw and K. Van Driessen, An algorithm for positive-breakdown methods based on concentration steps, in *Data Analysis: Scientific Modeling and Practical Application*, ed. W. Gaul, O. Opitz, and M. Schader (New York: Springer-Verlag, 2000), 335–346.
27. P.J. Rousseeuw and M. Hubert, Recent developments in PROGRESS, in *L1-Statistical Procedures and Related Topics* (Hayward, California: Institute of Mathematical Statistics Lecture Notes-Monograph Series Vol 31, 1997), 201–214.
28. P.J. Rousseeuw and B.C. van Zomeren, Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85 (1990):633–651.
29. P.J. Huber, Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1 (1973):799–821.

30. J. Jurecková, Nonparametric estimate of regression coefficients. *The Annals of Mathematical Statistics*, 42 (1971):1328–1338.
31. R. Koenker and S. Portnoy, L-estimation for linear models. *Journal of the American Statistical Association*, 82 (1987):851–857.
32. P.J. Rousseeuw and V.J. Yohai, Robust regression by means of S-estimators, in *Robust and Nonlinear Time Series Analysis*, ed. J. Franke, W. Härdle, and R.D. Martin (New York: Lecture Notes in Statistics No. 26, Springer-Verlag, 1984), 256–272.
33. V.J. Yohai, High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15 (1987):642–656.
34. B. Mendes and D.E. Tyler, Constrained M-estimates for regression, in *Robust Statistics; Data Analysis and Computer Intensive Methods*, ed. H. Rieder (New York: Lecture Notes in Statistics No. 109, Springer-Verlag, 1996), 299–320.
35. P.J. Rousseeuw and M. Hubert, Regression depth. *Journal of the American Statistical Association*, 94 (1999):388–402.
36. S. Van Aelst and P.J. Rousseeuw, Robustness of deepest regression. *Journal of Multivariate Analysis*, 73 (2000):82–106.
37. S. Van Aelst, P.J. Rousseeuw, M. Hubert, and A. Struyf, The deepest regression method. *Journal of Multivariate Analysis*, 81 (2002):138–166.
38. P.J. Rousseeuw, S. Van Aelst, B. Rambali, and J. Smeyers-Verbeke, Deepest regression in analytical chemistry. *Analytica Chimica Acta*, 446 (2001):243–254.
39. B. Rambali, S. Van Aelst, L. Baert, and D.L. Massart, Using deepest regression method for optimization of fluidized bed granulation on semi-full scale. *International Journal of Pharmaceutics*, 258 (2003):85–94.
40. P.J. Rousseeuw, S. Van Aelst, K. Van Driessen, and J. Agulló, Robust multivariate regression. *Technometrics*, 46 (2004):293–305.
41. M. Hubert and S. Engelen, Robust PCA and classification in biosciences. *Bioinformatics*, 20 (2004):1728–1736.
42. C. Croux and G. Haesbroeck, Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87 (2000):603–618.
43. M. Hubert, P.J. Rousseeuw, and S. Verboven, A fast robust method for principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60 (2002):101–111.
44. G. Li and Z. Chen, Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association*, 80 (1985):759–766.
45. C. Croux and A. Ruiz-Gazen, A fast algorithm for robust principal components based on projection pursuit. In *Proceedings in Computational Statistics*, ed. A. Prat (Heidelberg: Physica-Verlag, 1996), 211–217.

46. C. Croux and A. Ruiz-Gazen, High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95 (2005):206–226.
47. H. Cui, X. He, and K.W. Ng, Asymptotic distributions of principal components based on robust dispersions. *Biometrika*, 90 (2003):953–966, 2003.
48. M. Hubert, P.J. Rousseeuw, and K. Vanden Branden, ROBPCA: a new approach to robust principal components analysis. *Technometrics*, 47 (2005):64–79.
49. M. Debruyne and M. Hubert, The influence function of Stahel-Donoho type methods for robust covariance estimation and PCA (2006), Submitted.
50. R.A. Maronna, Principal components and orthogonal regression based on robust scales. *Technometrics*, 47 (2005):264–273.
51. R.A. Maronna and R.H. Zamar, Robust multivariate estimates for high dimensional data sets. *Technometrics*, 44 (2002):307–317.
52. I.T. Jolliffe, *Principal Component Analysis* (New York: Springer-Verlag, 1986).
53. S. Wold, Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20 (1978):397–405.
54. H.T. Eastment and W.J. Krzanowski, Cross-validatory choice of the number of components from a principal components analysis. *Technometrics*, 24 (1982):73–77.
55. M. Hubert and S. Engelen, Fast cross-validation for high-breakdown resampling algorithms for PCA (2006), Submitted.
56. P. Lemberge, I. De Raedt, K.H. Janssens, F. Wei, and P.J. Van Espen, Quantitative Z-analysis of 16th–17th century archaeological glass vessels using PLS regression of EPXMA and  $\mu$ -XRF data. *Journal of Chemometrics*, 14 (2000):751–763.
57. M. Hubert and S. Verboven, A robust PCR method for high-dimensional regressors. *Journal of Chemometrics*, 17 (2003):438–452.
58. S. de Jong, SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18 (1993): 251–263.
59. M. Hubert and K. Vanden Branden, Robust methods for Partial Least Squares Regression. *Journal of Chemometrics*, 17 (2003):537–549.
60. K. Vanden Branden and M. Hubert, Robustness properties of a robust PLS regression method. *Analytica Chimica Acta*, 515 (2004):229–241.
61. S. Serneels, C. Croux, P. Filzmoser, and P.J. Van Espen, Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, 76 (2005):197–204.
62. S. Engelen and M. Hubert, Fast model selection for robust calibration. *Analytica Chimica Acta*, 544 (2005):219–228.
63. K.H. Esbensen, S. Schönkopf, and T. Midtgaard, *Multivariate Analysis in Practice* (Trondheim: Camo, 1994).

64. R.A. Johnson and D.W. Wichern, *Applied multivariate statistical analysis* (Englewood Cliffs, NJ: Prentice Hall Inc., 1998).
65. M. Hubert and K. Van Driessen, Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45 (2004):301–320.
66. D.M. Hawkins and G.J. McLachlan, High-breakdown linear discriminant analysis. *Journal of the American Statistical Association*, 92 (1997):136–143.
67. X. He and W.K. Fung, High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, 72 (2) (2000):151–162.
68. C. Croux and C. Dehon, Robust linear discriminant analysis using S-estimators. *The Canadian Journal of Statistics*, 29 (2001):473–492.
69. S. Wold, Pattern recognition by means of disjoint principal components models. *Pattern Recognition*, 8 (1976):127–139.
70. K. Vanden Branden and M. Hubert, Robust classification in high dimensions based on the SIMCA method. *Chemometrics and Intelligent Laboratory Systems*, 79 (2005):10–21.
71. C.J.C. Burges, A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2 (1998):121–167.
72. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines* (Cambridge: Cambridge University Press, 2000).
73. B. Schölkopf and A. Smola, *Learning with Kernels* (Cambridge, MA: MIT Press, 2002).
74. V. Vapnik, *The Nature of Statistical Learning Theory* (New York: Springer-Verlag, 1995).
75. V. Vapnik, *Statistical Learning Theory* (New York: Wiley, 1998).
76. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines* (Singapore: World Scientific, 2002).
77. G. Wahba, Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, in *Advances in Kernel Methods - Support Vector Learning*, ed. B. Schölkopf, C. Burges, and A. Smola (Cambridge, MA: MIT Press, 1999), 69–88.
78. A. Christmann and I. Steinwart, On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5 (2004):1007–1034.
79. B. Schölkopf, A. Smola, and K-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10 (1998):1299–1319.
80. B. Üstün, W.J. Melsen, and L.M.C. Buydens, Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems*, 81 (2006):29–40.
81. A. Christmann and I. Steinwart, Consistency and robustness of kernel based regression (2005), University of Dortmund, SFB-475, Technical Report.
82. J.A.K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, Weighted least squares support vector machines: Robustness and sparse approximation. *Neurocomputing*, 48 (2002):85–105.

83. R. Bro, *Multi-way Analysis in the Food Industry* (Royal Veterinary and Agricultural university, Denmark: PhD thesis, 1998).
84. C.M. Andersen and R. Bro, Practical aspects of PARAFAC modelling of fluorescence excitation-emission data. *Journal of Chemometrics*, 17 (2003):200–215.
85. R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics and Intelligent Laboratory Systems*, 46 (1999):133–147.
86. R.D. Jiji, G.G. Andersson, and K.S. Booksh, Application of PARAFAC for calibration with excitation-emission matrix fluorescence spectra of three classes of environmental pollutants. *Journal of Chemometrics*, 14 (2000):171–185.
87. A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis with Applications in the Chemical Sciences* (Chichester: Wiley, 2004).
88. S. Engelen, S. Frosch Möller, and M. Hubert, Automatically identifying scatter in fluorescence data using robust techniques (2006), Submitted.
89. S. Engelen and M. Hubert, A robust version of PARAFAC (2006), In preparation.
90. D. Baunsgaard, *Factors affecting 3-way modelling (PARAFAC) of fluorescence landscapes* (Royal Veterinary and Agricultural University, Department of Dairy and Food technology, Frederiksberg, Denmark: PhD thesis, 1999).
91. J. Riu and R. Bro, Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics and Intelligent Laboratory Systems*, 65 (2003):35–49.
92. L.R. Tucker, Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31 (1966):279–311.
93. V. Pravdova, B. Walczak, and D.L. Massart, A robust version of the Tucker3 model. *Chemometrics and Intelligent Laboratory Systems*, 59 (2001):75–88.
94. S. Verboven and M. Hubert, LIBRA: a Matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75 (2005):127–136.