



# Anomaly detection by robust statistics

Peter J. Rousseeuw\* and Mia Hubert

Real data often contain anomalous cases, also known as outliers. These may spoil the resulting analysis but they may also contain valuable information. In either case, the ability to detect such anomalies is essential. A useful tool for this purpose is robust statistics, which aims to detect the outliers by first fitting the majority of the data and then flagging data points that deviate from it. We present an overview of several robust methods and the resulting graphical outlier detection tools. We discuss robust procedures for univariate, low-dimensional, and high-dimensional data, such as estimating location and scatter, linear regression, principal component analysis, classification, clustering, and functional data analysis. Also the challenging new topic of cellwise outliers is introduced. © 2017 The Authors. *WIREs Data Mining and Knowledge Discovery* published by Wiley Periodicals, Inc.

## How to cite this article:

*WIREs Data Mining Knowl Discov* 2018, 8:e1236. doi: 10.1002/widm.1236

## INTRODUCTION

In real-world datasets it often happens that some cases behave differently from the majority of data. Such data points are called *anomalies* in machine learning, and *outliers* in statistics. Outliers may be caused by errors, but they could also have been recorded under exceptional circumstances, or belong to another population. It is very important to be able to detect anomalous cases, which may (a) have a harmful effect on the conclusions drawn from the data, or (b) contain valuable nuggets of information.

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 may 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 close to the fit we would have found without the outliers. We can then identify the outliers by their large

‘deviation’ (e.g., its distance or residual) from that robust fit.

First, we describe some robust procedures for detecting anomalies in univariate location and scale, as well as in multivariate data and in the linear regression setting. For more details on this part see Refs<sup>1–3</sup>. Next, we discuss principal component analysis (PCA) and some available robust methods for classification, clustering, and functional data analysis. Finally, we introduce the emerging research topic of detecting cellwise anomalies.

## ESTIMATING UNIVARIATE LOCATION AND SCALE

As an example of univariate data, suppose we have five measurements of a length:

$$6.27, 6.34, 6.25, 6.31, 6.28 \quad (1)$$

and we want to estimate its true value. For this, one usually computes the sample mean  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  which in this case equals  $\bar{x} = (6.27 + 6.34 + 6.25 + 6.31 + 6.28)/5 = 6.29$ . Let us now suppose that the fourth measurement has been recorded wrongly and the data become

\*Correspondence to: peter@rousseeuw.net

Department of Mathematics, KU Leuven, Leuven, Belgium

$$6.27, 6.34, 6.25, 63.1, 6.28. \quad (2)$$

In this case, we obtain  $\bar{x} = 17.65$ , which is far off. Alternatively, we could also compute the *median* of these data. For this, we sort the observations in (2) from smallest to largest:

$$6.25 \leq 6.27 \leq 6.28 \leq 6.34 \leq 63.10.$$

The median is the middle value, here yielding 6.28, which is still reasonable. We say that the median is more robust against an outlier.

More generally, 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 unknown center  $\mu$  and the unknown scale parameter  $\sigma$ .

The classical estimate of location is the mean. As we saw above, the mean is very sensitive to aberrant values among the  $n$  observations. In general, replacing even a single observation by a very large value can change the mean completely. We say that the *breakdown value*<sup>4,5</sup> of the sample mean is  $1/n$ , so it becomes 0% for large  $n$ . In general, the breakdown value is the smallest proportion of observations in the dataset that need to be replaced to carry the estimate arbitrarily far away. A breakdown value of 0% is thus the worst possible. See Ref<sup>6</sup> for precise definitions and extensions. The robustness of an estimator is also measured by its *influence function*<sup>7</sup> which measures the effect of a single outlier. The influence function of the mean is unbounded, which again illustrates that the mean is not robust.

For the general definition of the median, we denote the  $i$ th ordered observation as  $x_{(i)}$ . The median is defined as  $x_{((n+1)/2)}$  if  $n$  is odd and  $(x_{(n/2)} + x_{(n/2+1)})/2$  if  $n$  is even. Its breakdown value is about 50%, meaning that the median can resist almost 50% of outliers. This is the best possible breakdown value since the clean data need to be in the majority.

The situation for the scale parameter  $\sigma$  is similar. The classical estimator is the *standard deviation*  $s = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)}$ . Since a single outlier can already make  $s$  arbitrarily large, its breakdown value is 0%. For instance, for the clean data (1) above we have  $s = 0.035$ , whereas for the data (2) with the outlier we obtain  $s = 25.41$ !

A robust measure of scale is the median of all absolute deviations from the median (MAD), given by

$$\text{MAD} = 1.4826 \text{ median}_{i=1, \dots, n} |x_i - \text{median}_{j=1, \dots, n}(x_j)|. \quad (3)$$

The constant 1.4826 is a correction factor which makes the MAD consistent at Gaussian distributions. The MAD of (2) is the same as that of (1), namely 0.044. We can also use the  $Q_n$  estimator<sup>8</sup>, defined as

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

with  $k = \binom{h}{2} \approx \binom{n}{2}/4$  and  $h = \lfloor \frac{n}{2} \rfloor + 1$ . Here,  $\lfloor \dots \rfloor$  rounds down to the nearest integer. This scale estimator is thus the first quartile of all pairwise distances between two data points. The breakdown value of both the MAD and the  $Q_n$  estimator is 50%.

Also the (normalized) interquartile range (IQR) can be used, given by  $\text{IQR} = 0.7413(Q_3 - Q_1)$  where  $Q_1 = x_{(\lfloor n/4 \rfloor)}$  is the first quartile of the data and  $Q_3 = x_{(\lceil 3n/4 \rceil)}$  is the third quartile. The IQR has a simple expression but its breakdown value is only 25%, so it is less robust than the MAD and  $Q_n$ .

The robustness of the median comes at a price: at the Gaussian model it is less efficient than the mean. Many robust procedures have been proposed that strike a balance between robustness and efficiency, such as location M-estimators<sup>9</sup>. They are defined implicitly as the solution of the equation

$$\sum_{i=1}^n \psi \left( \frac{x_i - \hat{\mu}}{\hat{\sigma}} \right) = 0 \quad (4)$$

for a given real function  $\psi$ . The denominator  $\hat{\sigma}$  is an initial robust scale estimate such as  $Q_n$ . A solution  $\hat{\mu}$  to Eq. (4) can be found by an iterative algorithm, starting from the initial location estimate  $\hat{\mu}^{(0)} = \text{median}_i(x_i)$ . Popular choices for  $\psi$  are the Huber function  $\psi(x) = x \min(1, c/|x|)$  and Tukey's bisquare function  $\psi(x) = x(1 - (x/c)^2)^2 I(|x| < c)$ . These M-estimators contain a tuning parameter  $c$  which needs to be chosen in advance. Also M-estimators for the scale parameter  $\sigma$  exist.

People often use rules to detect outliers. The classical rule is based on the  $z$ -scores of the observations, given by

$$z_i = \frac{x_i - \bar{x}}{s} \quad (5)$$

where  $s$  is the standard deviation of the data. More precisely, the rule flags  $x_i$  as outlying if  $|z_i|$  exceeds 2.5, say. But in the above example (2) with the outlier, the  $z$ -scores are

−0.45, −0.45, −0.45, 1.79, −0.45

so none of them attains 2.5. The largest value is only 1.79, which is quite similar to the largest  $z$ -score for the clean data (1), which equals 1.41. The  $z$ -score of the outlier is small because it subtracts the nonrobust mean (which was drawn towards the outlier) and because it divides by the nonrobust standard deviation (which the outlier has made much larger than in the clean data). Plugging in robust estimators of location and scale such as the median and the MAD yields the robust scores

$$\frac{x_i - \text{median}_j(x_j)}{\text{MAD}_j(x_j)} \quad (6)$$

which yield a much more reliable outlier detection tool. Indeed, in the contaminated example (2) the robust scores are

−0.22, 1.35, −0.67, 1277.5, 0.0

where that of the outlier greatly exceeds the 2.5 cutoff.

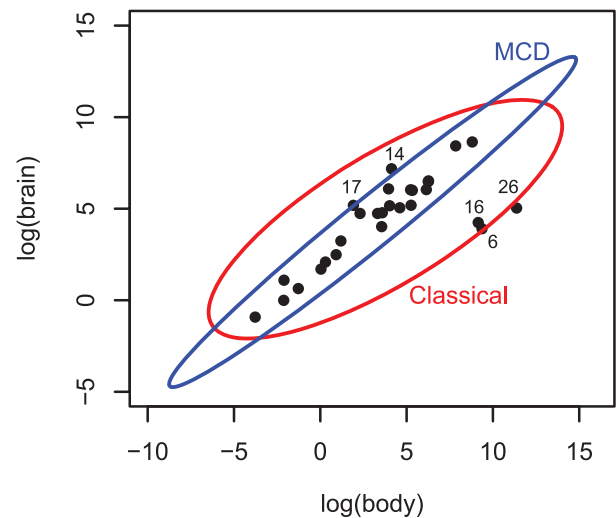
Also Tukey's boxplot is often used to pinpoint possible outliers. In this plot a box is drawn from the first quartile  $Q_1$  of the data to the third quartile  $Q_3$ . Points outside the interval  $[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}]$ , called the *fence*, are traditionally marked as outliers. Note that the boxplot assumes symmetry, since we add the same amount to  $Q_3$  as what we subtract from  $Q_1$ . At asymmetric distributions the usual boxplot typically flags many regular data points as outliers. The skewness-adjusted boxplot<sup>10</sup> corrects for this by using a robust measure of skewness<sup>11</sup> in determining the fence.

## MULTIVARIATE LOCATION AND COVARIANCE ESTIMATION

From now on, we assume that the data are  $d$ -dimensional and are stored in an  $n \times d$  data matrix  $X = (x_1, \dots, x_n)^T$  with  $x_i = (x_{i1}, \dots, x_{id})^T$  the  $i$ th data point. Classical measures of location and scatter are given by the empirical mean  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  and the empirical covariance matrix  $S_X = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T / (n-1)$ . 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.

As an illustration, we consider a bivariate dataset (Ref<sup>2</sup>, p. 59) containing the logarithms of body

## Classical and robust tolerance ellipse



**FIGURE 1** | Animal data: tolerance ellipse of the classical mean and covariance matrix (red), and that of the robust location and scatter matrix (blue).

weight and brain weight of 28 animal species, with scatterplot in Figure 1. Any point  $x$  has a so-called Mahalanobis distance (or 'generalized distance')

$$MD(x, \hat{\mu}, \hat{\Sigma}) = \sqrt{(x - \hat{\mu})^T \hat{\Sigma}^{-1} (x - \hat{\mu})} \quad (7)$$

to the mean  $\hat{\mu} = \bar{x}$ , by taking the covariance matrix  $\hat{\Sigma} = S_X$  into account. The MD is constant on ellipsoids. The so-called 97.5% tolerance ellipsoid is given by  $MD(x) \leq \sqrt{\chi_{d,0.975}^2}$  where  $\chi_{d,0.975}^2$  is the 0.975 quantile of the chi-squared distribution with  $d$  degrees of freedom. In this bivariate example  $d = 2$ , and the resulting ellipse is drawn in red. We see that it is inflated in the direction of the three outliers 6, 16, and 26 which are dinosaurs having low brain weight and high body weight. As a result these data points fall near the boundary of the tolerance ellipse, i.e., their  $MD(x_i)$  are not very high.

Alternatively, we can compute robust estimates of location and scatter (covariance), for instance, by the *Minimum Covariance Determinant* (MCD) method<sup>12,13</sup>. The MCD looks for those  $h$  observations in the dataset (where the number  $h$  is given by the user) whose classical covariance matrix has the lowest possible determinant. The MCD estimate of location  $\hat{\mu}$  is then the average of these  $h$  points, whereas the MCD estimate of scatter  $\hat{\Sigma}$  is their covariance matrix, multiplied by a consistency factor. (By default this is then followed by a

reweighting step to improve efficiency at Gaussian data.) Instead of Mahalanobis distances we can then compute *robust distances*, again given by (7) but now with the robust estimates  $\hat{\mu}$  and  $\hat{\Sigma}$ . This yields the robust tolerance ellipse shown in blue in Figure 1. This ellipse exposes the three dinosaurs, and we see two species near the upper boundary, 17 (rhesus monkey) and 14 (human).

In dimension  $d = 4$  or higher, it becomes infeasible to visualize the tolerance ellipsoid, but we still have the distances. The *distance–distance plot* (DD-plot) in Figure 2 shows the robust distance  $RD(x_i)$  of each data point versus its classical Mahalanobis distance  $MD(x_i)$ . The horizontal and vertical cutoff lines are at  $\sqrt{\chi^2_{d,0.975}}$  and the dashed line is where classical and robust distances coincide. We see that the  $RD(x_i)$  flag all the outliers in this dataset, while the  $MD(x_i)$  do not. For a dataset in which they are very similar we can trust classical statistical methods, but when they differ much (like here) the DD-plot detects the outlying data points. This does not imply we should somehow delete them, but rather that they should be investigated and understood. Outliers are not necessarily ‘errors’: they can also correspond to unusual circumstances or be members of a different population.

The MCD estimator, as well as its weighted version, has a bounded influence function and breakdown value  $(n - b + 1)/n$ , hence the number  $b$  determines the robustness of the estimator. The

MCD has its highest possible breakdown value when  $b = \lfloor (n + d + 1)/2 \rfloor$ . When a large proportion of contamination is expected,  $b$  should thus be chosen close to  $0.5n$ . Otherwise an intermediate value for  $b$ , such as  $0.75n$ , is recommended to obtain a higher finite-sample efficiency. Reference<sup>14</sup> gives a more detailed overview of the MCD estimator and its properties.

The computation of the MCD estimator is non-trivial and naively requires an exhaustive investigation of all  $b$ -subsets out of  $n$ . Fortunately, a much faster algorithm was constructed, called FastMCD<sup>15</sup>. It starts by randomly drawing many subsets of  $d + 1$  observations from the dataset. Based on these subsets,  $b$ -subsets are obtained by means of so-called C-steps (see Ref<sup>15</sup> for details). More recently an even faster algorithm called DetMCD was devised<sup>16</sup> which carries out a deterministic computation instead of random sampling.

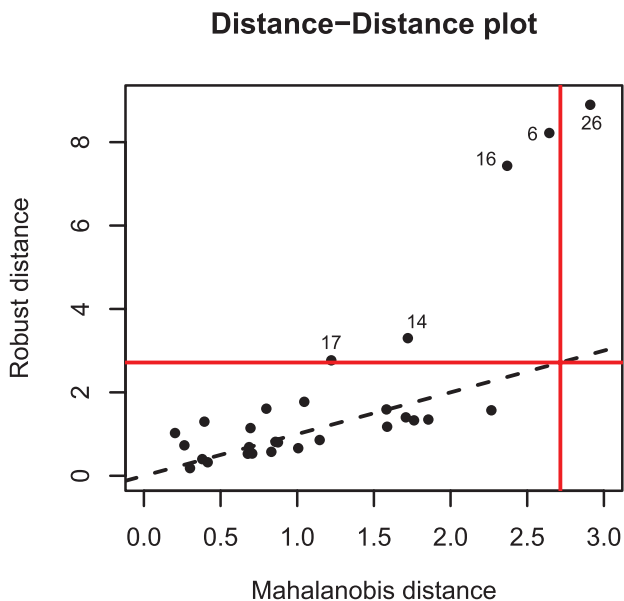
The MCD assumes that  $n > d$ , so there must be more data points than dimensions, and it works best when  $n > 5d$ . When there are more than, say, 20 dimensions and/or  $d \geq n$  other methods are needed. One is to compute robust principal components as described in a section below. Another is to use the *minimum regularized covariance determinant* (MRCD) method<sup>17</sup>. This approach minimizes  $\det\{\rho T + (1 - \rho)S_H\}$  where  $T$  is a positive definite target matrix and  $S_H$  is the covariance matrix of a subset  $H$  with  $b$  data points. The combined matrix is always positive definite, whereas  $\det\{S_H\} = 0$  when  $d \geq n$ .

Many other robust estimators of location and scatter have been presented in the literature. The first such estimator was proposed by Stahel<sup>18</sup> and Donoho<sup>19</sup> (see also Ref<sup>20</sup>). They defined the so-called Stahel–Donoho outlyingness of a data point  $x_i$  as

$$\text{outl}(x_i) = \max_{\mathbf{u}} \frac{|\mathbf{x}_i^T \mathbf{u} - \text{median}_{j=1, \dots, n}(\mathbf{x}_j^T \mathbf{u})|}{\text{MAD}_{j=1, \dots, n}(\mathbf{x}_j^T \mathbf{u})} \quad (8)$$

where the maximum is over all directions (i.e., all  $d$ -dimensional unit length vectors  $\mathbf{u}$ ), and  $\mathbf{x}_j^T \mathbf{u}$  is the projection of  $\mathbf{x}_j$  on the direction  $\mathbf{u}$ . In each direction this uses the robust  $z$ -scores Eq. (6). Recently, a version of Eq. (8) suitable for skewed distributions was proposed<sup>21</sup>.

Multivariate M-estimators<sup>22</sup> have a low breakdown value due to possible implosion of the estimated scatter matrix. More recent robust estimators of multivariate location and scatter with high breakdown value include S-estimators<sup>2,23</sup>, MM-estimators<sup>24</sup>, and the OGK estimator<sup>25</sup>.



**FIGURE 2** | Animal data: robust distance versus classical Mahalanobis distance.

## LINEAR REGRESSION

The multiple linear regression model assumes that there are  $d$  ‘explanatory’  $x$ -variables as well as a response variable  $y$  which can be approximated by a linear combination of the  $x$ -variables. More precisely, the model says that for all data points  $(x_i, y_i)$  it holds that

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_d x_{id} + \varepsilon_i \quad i = 1, \dots, n \quad (9)$$

where the errors  $\varepsilon_i$  are assumed to be independent and identically distributed with zero mean and constant variance  $\sigma^2$ . Applying a regression estimator to the data yields  $d + 1$  regression coefficients, combined as  $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_d)^T$ . The residual  $r_i$  of case  $i$  is defined as the difference between the observed response  $y_i$  and its estimated value  $\hat{y}_i$ .

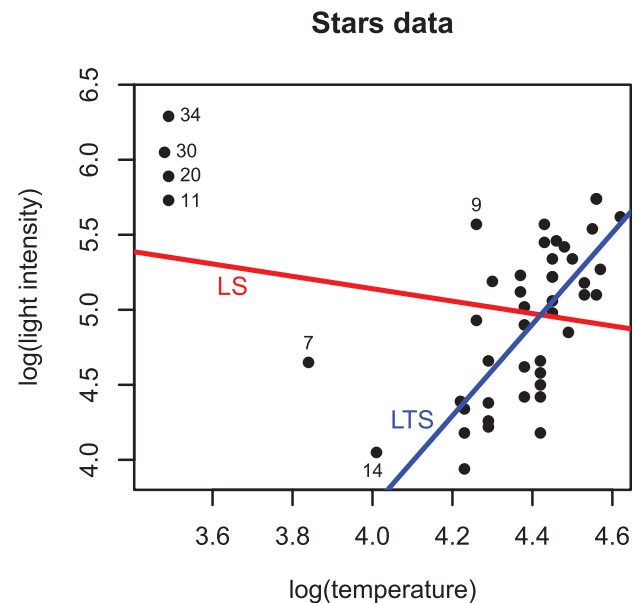
The classical least squares (LS) method to estimate  $\beta$  minimizes the sum of the squared residuals. It is popular because it allows to compute the regression estimates explicitly, and it is optimal if the errors have a Gaussian distribution. Unfortunately, LS is extremely sensitive to outliers, i.e., data points that do not obey the linear pattern formed by the majority of the data.

For instance, Figure 3 shows the Hertzsprung–Russell diagram of the star cluster CYG OB1, containing 47 stars. The  $x$ -coordinate of each star is the logarithm of its surface temperature, and the  $y$ -coordinate is the logarithm of its light intensity. Most of the stars belong to the so-called main sequence, whereas 11, 20, 30, 34 are giant stars and 7 is intermediate. The LS line is shown in red, and has a negative slope although the main sequence slopes upward. It has been pulled away by the leverage exerted by the four giant stars. As an unfortunate side effect, the giant stars do not have larger absolute residuals than some of the main sequence stars, so only looking at residuals would not allow to detect them.

The blue line on the other hand is the result of a robust regression method, the *Least Trimmed Squares* (LTS) estimator proposed by Rousseeuw<sup>12</sup>. The LTS is given by

$$\underset{\beta}{\text{minimize}} \quad \sum_{i=1}^h (r^2)_{(i)} \quad (10)$$

where  $(r^2)_{(1)} < (r^2)_{(2)} < \dots < (r^2)_{(n)}$  are the ordered squared residuals. (They are first squared, and then ordered.) By not adding *all* the squared residuals the LTS makes it possible to fit the majority of the data,



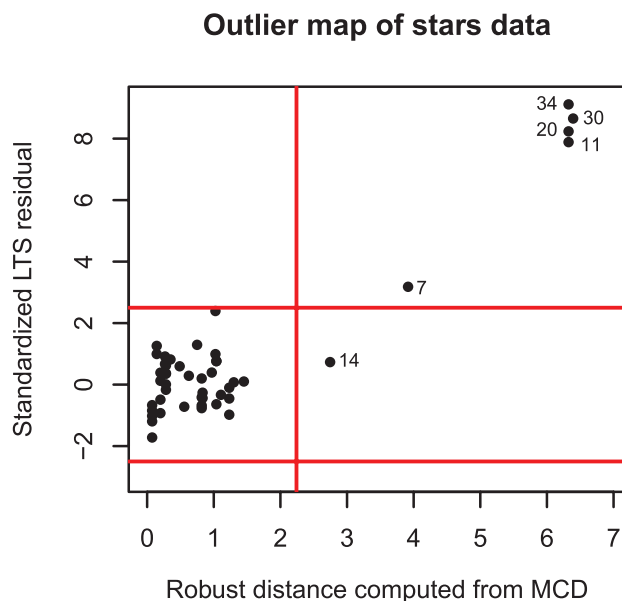
**FIGURE 3** | Stars data: classical least squares line (red) and robust line (blue).

whereas the outliers can have large residuals. In Figure 3, the blue line indeed fits the main sequence stars, and stays far from the four giant stars so the latter will have large residuals from that line. (Note that the giant stars are not ‘errors’ but correct observations of members of a different population.)

The value  $h$  in (10) plays the same role as in the MCD estimator. For  $h \approx n/2$  we find a breakdown value of 50%, whereas for larger  $h$  we obtain roughly  $(n - h)/n$ . A fast algorithm for the LTS estimator (FastLTS) has been developed<sup>26</sup>. The scale of the errors  $\sigma$  can be estimated by  $\hat{\sigma}_{\text{LTS}}^2 = c_{h,n}^2 \sum_{i=1}^h (r^2)_{(i)} / h$  where  $r_i$  are the residuals from the LTS fit, and  $c_{h,n}$  is a constant that makes  $\hat{\sigma}_{\text{LTS}}$  consistent at Gaussian error distributions, as described in<sup>27</sup>. We can then identify outliers by their large standardized LTS residuals  $r_i / \hat{\sigma}_{\text{LTS}}$ . We can also use the standardized LTS residuals to assign a weight to every observation. The weighted LS 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. Alternatively, inference for LTS can be based on the fast robust bootstrap proposed in<sup>28,29</sup>.

In most situations we have more than one explanatory variable, and for dimension  $d = 3$  and higher it is no longer possible to perceive the linear patterns by eye. It is in those cases that robust regression becomes the most useful. To flag and interpret the outliers we can use the *outlier map* of<sup>30</sup> which





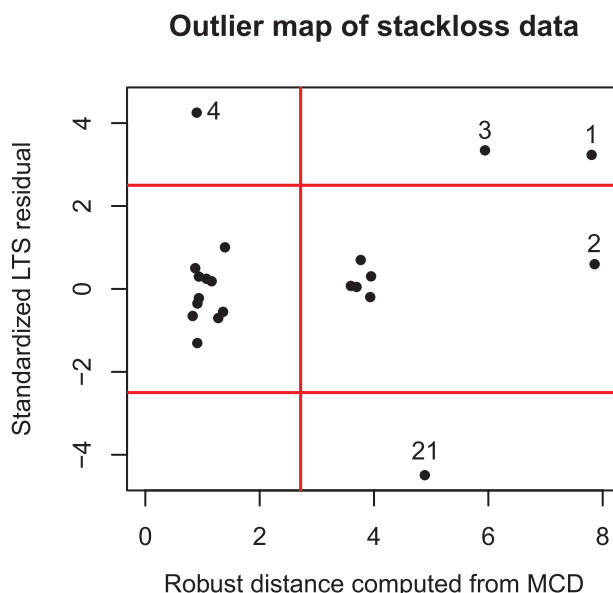
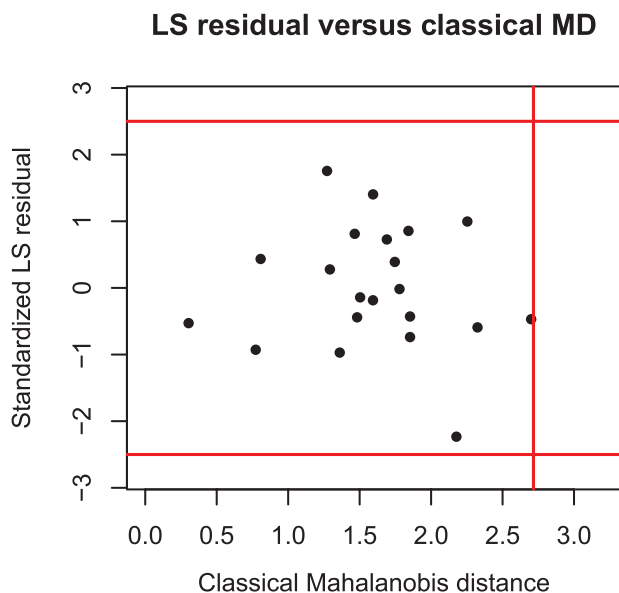
**FIGURE 4** | Stars data: standardized robust residuals of  $y$  versus robust distances of  $x$ .

plots the standardized LTS residuals versus robust distances (7) based on (for instance) the MCD estimator applied to the  $x$ -variables only. Figure 4 is the outlier map of the stars data. The tolerance band on the standardized LTS residuals is given by the horizontal lines at 2.5 and -2.5, and the vertical line corresponds to the cutoff value  $\sqrt{\chi^2_{d,0.975}}$  on the robust distances of the  $x_i$ . Data points  $(x_i, y_i)$  whose residuals fall outside the horizontal tolerance band are

called *regression outliers*. On the other hand, data points  $(x_i, y_i)$  whose robust distance  $RD(x_i)$  exceeds the cutoff are called *leverage points*, irrespective of their response  $y_i$ . So, the outlier map diagnoses four types of data points. Those with small  $|r_i|$  and small  $RD(x_i)$  are considered *regular observations*, and most points in Figure 4 fall in that rectangle. Those with large residuals  $r_i$  (positive or negative) and small  $RD(x_i)$  are called *vertical outliers* (there are none in this figure). Those with small  $|r_i|$  but large  $RD(x_i)$  (like point 14) are called *good leverage points* because they improve the accuracy of the fit. And finally, regression outliers that are also leverage points are called *bad leverage points*, like the four giant stars in this example. Note that the outlier map permits nuanced statements, for instance, point 7 is a leverage point but only slightly bad.

The main benefit of the outlier map is when the data has more dimensions. For instance, the stackloss data<sup>31</sup> is a benchmark data set with 21 points with  $d = 3$  explanatory variables, an intercept term and a response variable  $y_i$ . We cannot easily interpret such 4-dimensional data, but we can still look at the outlier map in the right panel of Figure 5. We see that 4 is a vertical outlier, 1, 3, and 21 are bad leverage points, and 2 is a good leverage point. Note that the left panel of Figure 5 does not flag any of these points because it uses the classical LS residuals and the classical distances  $MD(x_i)$ , both of which tend to mask atypical points.

It should be stressed that flagging atypical points with the outlier map (as in the right panel of



**FIGURE 5** | Stackloss data: (left) standardized nonrobust least squares (LS) residuals of  $y$  versus nonrobust distances of  $x$ ; (right) same with robust residuals and robust distances.

Figure 5) is not the end of the analysis, but rather a new start. The next step should be to try to figure out why these points are atypical and/or to improve the model by things like data transformation, model selection, higher order terms, etc. For variance selection in robust regression see Ref<sup>32</sup>. When the dimension is very high one needs to resort to sparse methods, for instance, by penalization. The first sparse methods for robust regression were developed in Refs<sup>33,34</sup>.

Historically, the earliest attempts at robust regression were least absolute deviations (LAD, also called  $L^1$ ), M-estimators<sup>35</sup>, R-estimators<sup>36</sup>, and L-estimators<sup>37</sup>. The breakdown value of all these methods is 0% because of their vulnerability to bad leverage points. Generalized M-estimators (GM-estimators)<sup>7</sup> were the first to attain a positive breakdown value, which unfortunately still went down to zero for increasing  $d$ .

The low finite-sample efficiency of LTS can be improved by replacing its objective function by a more efficient scale estimator applied to the residuals  $r_i$ . This approach has led to the introduction of high-breakdown regression S-estimators<sup>38</sup> and MM-estimators<sup>39</sup>.

## PRINCIPAL COMPONENT ANALYSIS

PCA is a popular dimension reduction method. It tries to explain the covariance structure of the data by means of a (hopefully 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. PCA is often the first step of the data analysis, followed by other multivariate techniques.

In the classical approach, the first principal component corresponds to the direction in which the projected data points have the largest variance. The second component is then taken orthogonal to the first and must again maximize the variance of the data points projected on it (subject to the orthogonality constraint). Continuing in this way produces all the principal components. It turns out that the classical principal components 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 are often attracted toward 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 weighted MCD estimator or MM-estimators<sup>40,41</sup>. Unfortunately, the use of these covariance estimators is limited to small to moderate dimensions since they are not defined when  $d > n$ .

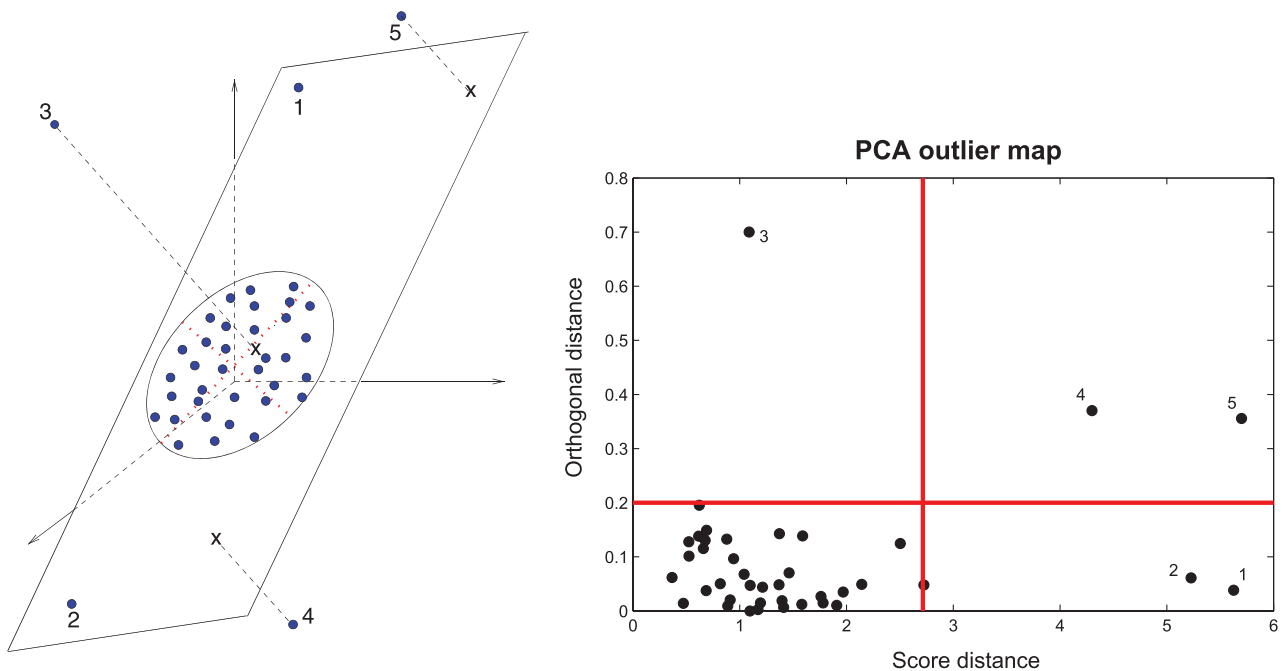
A second approach to robust PCA uses *Projection Pursuit* techniques. These methods maximize a robust measure of spread to obtain consecutive directions on which the data points are projected, see Refs<sup>42,43</sup>.

The ROBPCA<sup>44</sup> approach is a hybrid, which combines ideas of 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.

In order to diagnose outliers, we can draw an outlier map for PCA<sup>44</sup>, similar to the outlier map for regression in the previous section. A stylized example of such a PCA outlier map is shown in the right panel of Figure 6, which corresponds to the three-dimensional data in the left panel which is fitted by two principal components. On the vertical axis of the PCA outlier map, we find the *orthogonal distance* of each data point to the PCA subspace. This is just the Euclidean distance of the data point to its projection. The orthogonal distance is highest for the points 3, 4, and 5 in the example. On the horizontal axis we see the *score distance* of each data point, which is just the robust distance (7) of its projection relative to all the projected data points. The score distance is rather high for the points 1, 2, 4, and 5 in the figure.

By combining both distance measures the outlier map allows to distinguish between four types of data points. *Regular observations* have both a small orthogonal distance and a small score distance. Points with a high score distance but a small orthogonal distance, such as points 1 and 2 in Figure 6, are called *good leverage points* as they can improve the accuracy of the fitted PCA subspace. *Orthogonal outliers* have a large orthogonal distance but a small score distance, like point 3. *Bad leverage points* have both a large orthogonal distance and a large score distance, like points 4 and 5. They lie far from the space spanned by the robust principal components, and after projection on that space they lie far from most of the other projected data. They are called 'bad' because they typically they have a large influence on classical PCA, as the main eigenvectors will be tilted toward them.

As a real example, we take the glass data<sup>45</sup> consisting of spectra of 180 archeological glass vessels



**FIGURE 6** | Illustration of PCA: (left) types of outliers; (right) outlier map: plot of orthogonal distances versus score distances.

from the 16th–17th centuries. They have 750 wavelengths so  $d > n$ . The spectra are shown in Figure 7 with their outlier map based on ROBPCA, which clearly indicates a substantial number of bad leverage points and several orthogonal outliers. An analogous plot based on classical PCA (not shown) did not reveal the outliers, because they tilted the PCA subspace toward them. Also the plots of the first few principal components looked quite different.

Other proposals for robust PCA include spherical PCA<sup>46</sup> which first projects the data onto a sphere with a robust center, and then applies PCA to these projected data. To obtain sparse loadings, a robust sparse PCA method is proposed in Ref<sup>47</sup>. When linear models are not appropriate, one may use support vector machines (SVM) which are powerful tools for handling nonlinear structures<sup>48</sup>. A kernelized version of ROBPCA (KROBPCA) is introduced in Ref<sup>49</sup>. For a review of robust versions of principal component regression and partial least squares see Ref<sup>1</sup>.

## OTHER MODELS

### Classification

The goal of classification, also known as discriminant analysis or supervised learning, is to obtain rules that describe the separation between known groups  $G_j$  of  $d$ -dimensional data points, with an eye toward assigning new data points to one of the groups. We

write  $p_j$  for the membership probability, i.e., the probability for any observation to come from  $G_j$ .

For low-dimensional data, a popular classification rule results from maximizing the Bayes' posterior probability. At Gaussian distributions this yields quadratic discriminant analysis (QDA), i.e., choosing the  $j$  for which  $\mathbf{x}$  has the highest quadratic score  $d_j^Q(\mathbf{x})$  given by

$$d_j^Q(\mathbf{x}) = -\frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \Sigma_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln(p_j) \quad (11)$$

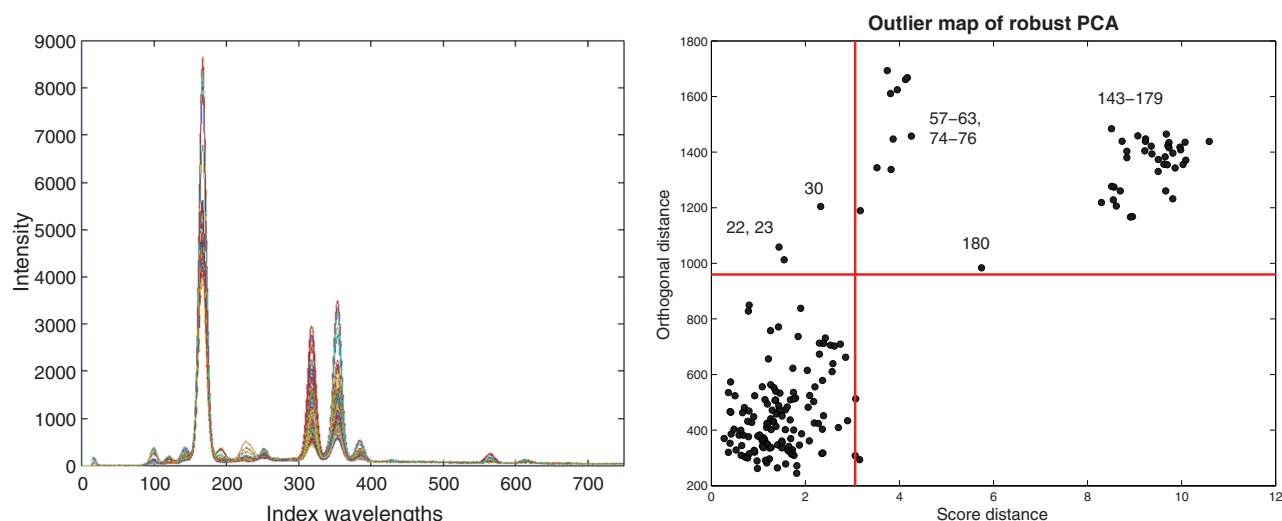
When all the covariance matrices are assumed to be equal, these scores can be simplified to

$$d_j^L(\mathbf{x}) = \boldsymbol{\mu}_j^T \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_j^T \Sigma^{-1} \boldsymbol{\mu}_j + \ln(p_j) \quad (12)$$

where  $\Sigma$  is the common covariance matrix, yielding linear discriminant analysis (LDA). Robust classification rules can be obtained by replacing the classical covariance matrices by robust alternatives such as the MCD estimator or S-estimators, as in Refs<sup>50–53</sup>.

When the data are high-dimensional, this approach cannot be applied because the robust covariance estimators are no longer computable. One approach is to first apply robust PCA to the entire dataset. Alternatively, one can also apply a PCA method to each group separately. This is the idea





**FIGURE 7** | Glass data: (left) spectra; (right) outlier map.

behind the SIMCA (Soft Independent Modeling of Class Analogy) method<sup>54</sup>. A robustification of SIMCA is obtained by first applying robust PCA to each group, and then constructing a classification rule for new observations based on their orthogonal distance to each subspace and their score distance within each subspace<sup>55</sup>.

An SVM classifier with an unbounded kernel, e.g., a linear kernel, is not robust and suffers the same problems as traditional linear classifiers. But when a bounded kernel is used, the resulting non-linear SVM classification handles outliers quite well<sup>48</sup>. As an alternative, one can apply KROBPCA combined with LDA to the scores<sup>49</sup>.

## Clustering

Cluster analysis (also known as unsupervised learning) is an important methodology when handling large datasets. It searches for homogeneous groups in the data, which afterward may be analyzed separately. Partitioning (nonhierarchical) clustering methods search for the best clustering in  $k$  groups.

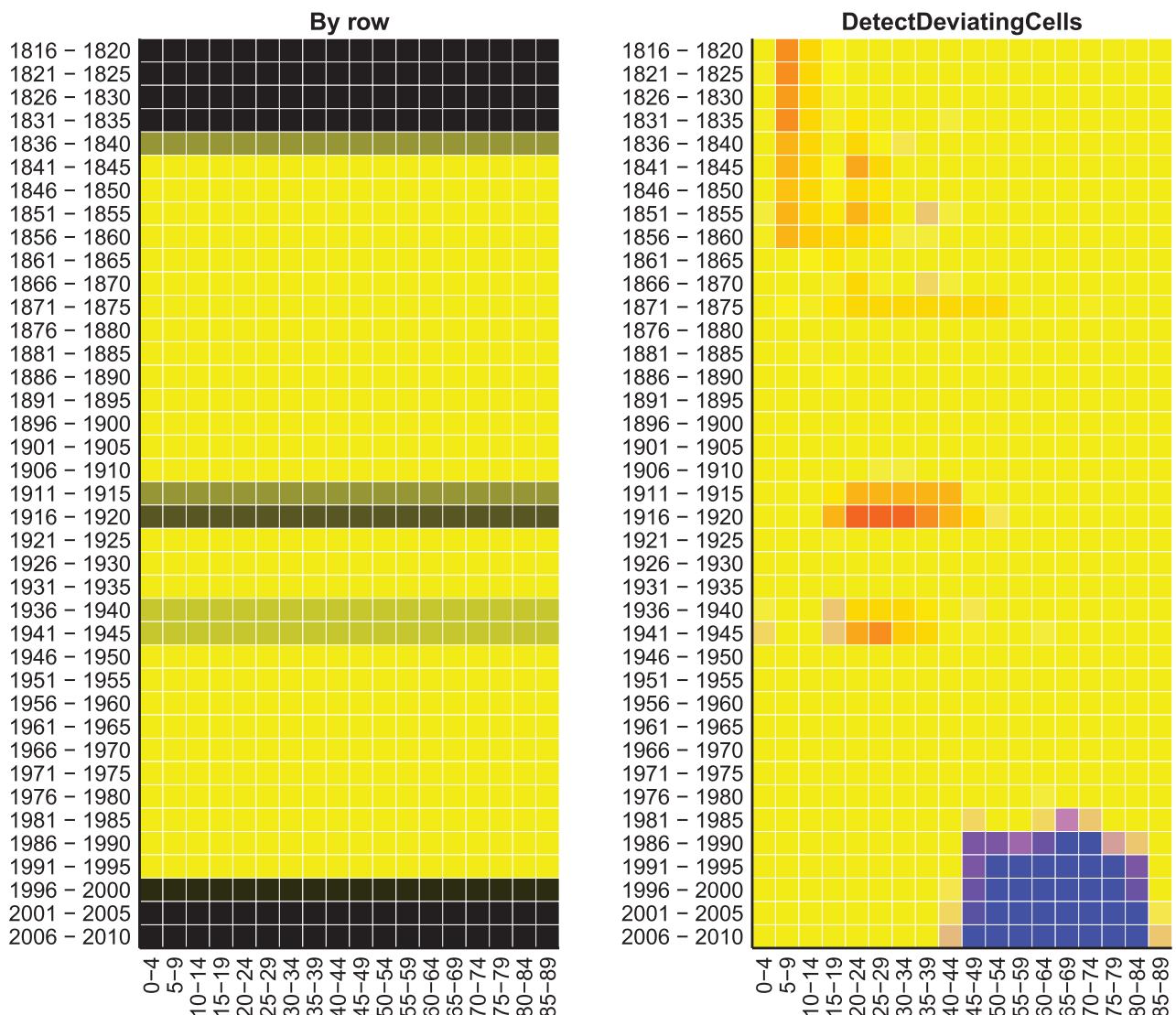
For spherical clusters, the most popular method is  $k$ -means which minimizes the sum of the squared Euclidean distances of the observations to the mean of their group<sup>56</sup>. This method is not robust as it uses averages. To overcome this problem, one of the first robust proposals was the Partitioning Around Medoids method<sup>57</sup>. It searches for  $k$  observations (called medoids) such that the sum of the unsquared distances of the observations to the medoid of their group is minimized. The CLARA algorithm<sup>57</sup> implemented this method for large datasets, and was extended to CLARANS<sup>58</sup> for spatial data mining.

Later on the more robust trimmed  $k$ -means method has been proposed<sup>59</sup>, inspired by the trimming ideas in the MCD and the LTS. It searches for the  $h$ -subset (with  $h$  as in the definition of MCD) such that the sum of the squared distances of the observations to the mean of their group is minimized. Consequently, not all observations need to be classified, as  $n - h$  cases can be left unassigned. To perform the trimmed  $k$ -means clustering an iterative algorithm<sup>60</sup> has been developed, using C-steps like those in the FastMCD algorithm. For nonspherical clusters, constrained maximum likelihood approaches<sup>61,62</sup> were developed.

## Functional Data

In functional data analysis, the cases are not data points but functions. A functional dataset typically consists of  $n$  curves observed on a set of gridpoints  $t_1, \dots, t_T$ . These curves can have smoothness properties, numerical derivatives and so on. Standard references on functional data are the books<sup>63,64</sup>. A functional dataset can be analyzed by principal components, for which robust methods are available<sup>65</sup>. To classify functional data, a recent approach is presented in Ref<sup>66</sup>.

The literature on outlier detection in functional data is rather young, and several graphical tools have been developed<sup>67-69</sup>, mainly for univariate functions. In Ref<sup>70</sup> also multivariate functions are discussed and a taxonomy of functional outliers is set up, with on the one hand functions that are outlying on most of their domain, such as shift and magnitude outliers as well as shape outliers, and on the other hand isolated outliers which are only outlying on a small part of



**FIGURE 8** | Male mortality in France in 1816–2010: (left) detecting outlying rows by a robust principal component analysis method; (right) detecting outlying cells by *DetectDeviatingCells*. After the analysis, the cells were grouped in blocks of  $5 \times 5$  for visibility.

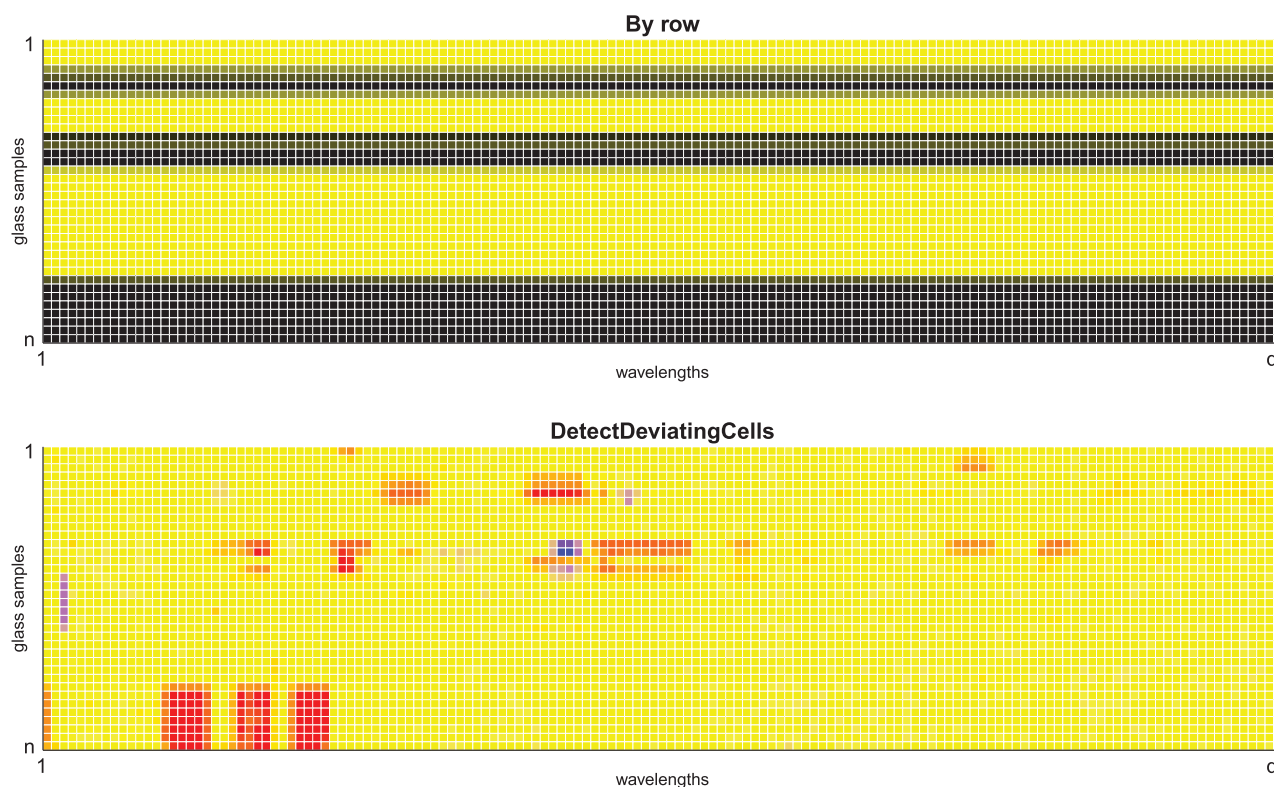
their domain. The proposed heatmap and functional outlier map are tools to flag outliers and detect their type. This work is expanded in Ref<sup>21</sup> to functional data with bivariate domains, such as images and video.

## Other Applications

Robust statistics has many other uses apart from outlier detection. For instance, robust estimation can be used in automated settings such as computer vision<sup>71,72</sup>. Another aspect is statistical inference, such as the construction of robust hypothesis tests,  $P$ -values, confidence intervals, and model selection (e.g., variable selection in regression). This aspect is studied in Refs<sup>3,7</sup> and in other works they reference.

## DETECTING OUTLYING CELLS

Until recently people have always considered outliers to be cases (data points), i.e., rows of the  $n \times d$  data matrix  $X$ . But recently, the realization has grown that this paradigm is no longer sufficient for the high-dimensional datasets we are often faced with nowadays. Typically, most data cells (entries) in a row are regular and a few cells are anomalous. The first paper to formulate the cellwise paradigm was Ref<sup>73</sup>, which showed how such outlying cells propagate in computations. In more than a few dimensions, even a small percentage of outlying cells can spoil a large percentage of rows. This is fatal for rowwise robust methods, which require at least 50% of the rows to be clean.



**FIGURE 9** | Cell map of the glass data. The positions of the deviating cells reveal the chemical contaminants.

Detecting cellwise outliers is a hard problem, since the outlyingness of a cell depends on the relation of its column to the other columns of the data, and on the values of the other cells in its row (some of which may be outlying themselves). The *DetectDeviatingCells*<sup>74</sup> algorithm addresses these issues, and apart from flagging cells it also provides a graphical output called a cellmap.

As an example, we consider the mortality by age for males in France from 1816 to 2010, obtained from <http://www.mortality.org>. Each row corresponds to the mortalities in a given calendar year. The left panel in Figure 8 shows the result of the ROBPCA method described in the section on principal components. Outlying rows are shown in black and regular rows in yellow. The analysis was carried out on the dataset with the individual years and the individual ages, but as this resolution would be too high to fit on the page we have combined the cells into  $5 \times 5$  blocks afterward. The combination of some black rows with some yellow ones has led to gray blocks. We can see that there were outlying rows in the early years, the most recent years, and during two periods in between. Note that a black row does not provide information about its cells.

By contrast, the result of *DetectDeviatingCells* in the right panel in Figure 8 identifies a lot more

information. Cells with higher values than predicted are shown in red, and those with lower values in blue, after which the colors were averaged in the  $5 \times 5$  blocks. The outlying early years saw a high infant mortality. During the Prussian war and both world wars, there was a higher mortality among young adult men. And in recent years mortality among middle-aged and older men has decreased substantially, perhaps due to medical advances.

We also return to the glass data from the section on PCA. The top panel in Figure 9 shows the rows detected by the ROBPCA method. The lower panel is the cell map obtained by *DetectDeviatingCells* on this  $180 \times 750$  dataset. After the analysis, the cells were again grouped in  $5 \times 5$  blocks. We now see clearly which parts of each spectrum are higher/lower than predicted. The wavelengths of these deviating cells reveal the chemical elements responsible.

Ideally, after running *DetectDeviatingCells* the user can look at the deviating cells and whether their values are higher or lower than predicted, and make sense of what is going on. This may lead to a better understanding of the data pattern, to changes in the way the data are collected/measured, to dropping certain rows or columns, to transforming variables, to changing the model, and so on. If the

dataset is too large for visual inspection of the results, or the analysis is automated, the deviating cells can be set to missing after which the dataset is treated by a method appropriate for incomplete data. A good rowwise robust method of this type is Ref<sup>75</sup>.

## SOFTWARE AVAILABILITY

All the examples in this paper were produced with the free software R<sup>76</sup>. The publicly available CRAN package *robustbase* contains Qn, covMcd, ltsReg, and lmrob, whereas *rrcov* has many robust covariance estimators, robust principal components, and robust LDA and QDA classification. ROBPCA and its extensions are available in *rospca* and robust SIMCA in *rrcovHD*. Robust clustering can be performed with the *cluster* and *tclust* packages. The package *mrfDepth* contains tools for functional data and *cellWise* provides cellwise outlier detection and cellmaps.

Matlab functions for many of these methods are available in the LIBRA toolbox<sup>77,78</sup>, which can be downloaded from <http://wis.kuleuven.be/stat/robust>.

The MCD and LTS methods are also built into S-PLUS as well as SAS (version 11 or higher) and SAS/IML (version 7 or higher).

## CONCLUSIONS

We have surveyed the utility of robust statistical methods and their algorithms for detecting anomalous data. These methods were illustrated on real data, in frameworks ranging from covariance matrices, the linear regression model and PCA, with references to methods for many other tasks such as supervised and unsupervised classification as well as the analysis of functional data. For high-dimensional data, sparse and regularized robust methods were developed recently.

We have described methods to detect anomalous cases (rowwise outliers) but also newer work on the detection of anomalous data cells (cellwise outliers). An important topic for future research is to further improve the efficiency of the robust methodologies, in terms of both predictive accuracy and computational cost.

## REFERENCES

- Hubert M, Rousseeuw PJ, Van Aelst S. High breakdown robust multivariate methods. *Stat Sci* 2008, 23:92–119.
- Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. New York: Wiley-Interscience; 1987.
- Maronna RA, Martin DR, Yohai VJ. *Robust Statistics: Theory and Methods*. New York: John Wiley & Sons; 2006.
- Hampel FR. A general qualitative definition of robustness. *Ann Math Stat* 1971, 42:1887–1896.
- Donoho DL, Huber PJ. The notion of breakdown point. In: Bickel P, Doksum K, Hodges JL, eds. *A Festschrift for Erich Lehmann*. Belmont, CA: Wadsworth; 1983, 157–184.
- Hubert M, Debruyne M. Breakdown value. *Wiley Interdiscip Rev Comput Stat* 2009, 1:296–302.
- Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons; 1986.
- Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J Am Stat Assoc* 1993, 88: 1273–1283.
- Huber PJ. Robust estimation of a location parameter. *Ann Math Stat* 1964, 35:73–101.
- Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. *Comput Stat Data Anal* 2008, 52:5186–5201.
- Brys G, Hubert M, Struyf A. A robust measure of skewness. *J Comput Graph Stat* 2004, 13:996–1017.
- Rousseeuw PJ. Least median of squares regression. *J Am Stat Assoc* 1984, 79:871–880.
- Rousseeuw PJ. Multivariate estimation with high breakdown point. In: Grossmann W, Pflug G, Vincze I, Wertz W, eds. *Mathematical Statistics and Applications*, vol. B. Dordrecht, The Netherlands: Reidel Publishing Company; 1985, 283–297.
- Hubert M, Debruyne M, Rousseeuw PJ. Minimum covariance determinant and extensions. *Wiley Interdiscip Rev Comput Stat* 2017, arXiv 1709.07045.
- Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999, 41:212–223.
- Hubert M, Rousseeuw PJ, Verdonck T. A deterministic algorithm for robust location and scatter. *J Comput Graph Stat* 2012, 21:618–637.
- Boudt K, Rousseeuw PJ, Vanduffel S, Verdonck T. The minimum regularized covariance determinant estimator. Technical Report, arXiv:1701.07086, 2017.

18. Stahel WA. Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. PhD Thesis, *ETH Zürich*, 1981.
19. Donoho DL. Breakdown properties of multivariate location estimators. PhD Qualifying paper, Department of Statistics, Harvard University, Boston, 1982.
20. Maronna RA, Yohai VJ. The behavior of the Stahel-Donoho robust multivariate estimator. *J Am Stat Assoc* 1995, 90:330–341.
21. Rousseeuw PJ, Raymaekers J, Hubert M. A measure of directional outlyingness with applications to image data and video. *J Comput Graph Stat* 2017, <https://doi.org/10.1080/10618600.2017.1366912>.
22. Maronna RA. Robust M-estimators of multivariate location and scatter. *Ann Stat* 1976, 4:51–67.
23. Davies L. Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *Ann Stat* 1987, 15:1269–1292.
24. Tatsuoka KS, Tyler DE. On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Ann Stat* 2000, 28:1219–1243.
25. Maronna RA, Zamar RH. Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics* 2002, 44:307–317.
26. Rousseeuw PJ, Van Driessen K. Computing LTS regression for large data sets. *Data Min Knowl Disc* 2006, 12:29–45.
27. Pison G, Van Aelst S, Willems G. Small sample corrections for LTS and MCD. *Metrika* 2002, 55:111–123.
28. Willems G, Van Aelst S. Fast and robust bootstrap for LTS. *Comput Stat Data Anal* 2005, 48:703–715.
29. Salibián-Barrera M, Van Aelst S, Willems G. Fast and robust bootstrap. *Stat. Methods Appl* 2008, 17:41–47.
30. Rousseeuw PJ, van Zomeren BC. Unmasking multivariate outliers and leverage points. *J Am Stat Assoc* 1990, 85:633–651.
31. Brownlee KA. *Statistical Theory and Methodology in Science and Engineering*. New York: John Wiley & Sons; 1965.
32. Khan J, Van Aelst S, Zamar R. Building a robust linear model with forward selection and stepwise procedures. *Comput Stat Data Anal* 2007, 52:239–248.
33. Khan J, Van Aelst S, Zamar R. Robust linear model selection based on least angle regression. *J Am Stat Assoc* 2007, 102:1289–1299.
34. Alfons A, Croux C, Gelper S. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann Appl Stat* 2013, 7:226–248.
35. Huber PJ. *Robust Statistics*. New York: John Wiley & Sons; 1981.
36. Jurecková J. Nonparametric estimate of regression coefficients. *Ann Math Stat* 1971, 42:1328–1338.
37. Koenker R, Portnoy S. L-estimation for linear models. *J Am Stat Assoc* 1987, 82:851–857.
38. Rousseeuw PJ, Yohai VJ. Robust regression by means of S-estimators. In: Franke J, Härdle W, Martin RD, eds. *Robust and Nonlinear Time Series Analysis*. New York: Springer-Verlag; 1984. Lecture Notes in Statistics No. 26, 256–272.
39. Yohai VJ. High breakdown point and high efficiency robust estimates for regression. *Ann Stat* 1987, 15:642–656.
40. Croux C, Haesbroeck G. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* 2000, 87:603–618.
41. Salibián-Barrera M, Van Aelst S, Willems G. PCA based on multivariate MM-estimators with fast and robust bootstrap. *J Am Stat Assoc* 2006, 101:1198–1211.
42. Hubert M, Rousseeuw PJ, Verboven S. A fast robust method for principal components with applications to chemometrics. *Chemom Intell Lab Syst* 2002, 60:101–111.
43. Croux C, Filzmoser P, Oliveira MR. Algorithms for projection-pursuit robust principal component analysis. *Chemom Intell Lab Syst* 2007, 87:218–225.
44. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: a new approach to robust principal components analysis. *Technometrics* 2005, 47:64–79.
45. Lemberge P, De Raedt I, Janssens KH, Wei F, Van Espen PJ. Quantitative Z-analysis of 16th–17th century archaeological glass vessels using PLS regression of EPXMA and  $\mu$ -XRF data. *J Chemom* 2000, 14:751–763.
46. Locantore N, Marron JS, Simpson DG, Tripoli N, Zhang JT, Cohen KL. Robust principal component analysis for functional data. *TEST* 1999, 8:1–73.
47. Hubert M, Reynkens T, Schmitt E, Verdonck T. Sparse PCA for high-dimensional data with outliers. *Technometrics* 2016, 58:424–434.
48. Steinwart I, Christmann A. *Support Vector Machines*. New York: Springer; 2008.
49. Debruyne M, Verdonck T. Robust kernel principal component analysis and classification. *Adv Data Anal Classif* 2010, 4:151–167.
50. Hawkins DM, McLachlan GJ. High-breakdown linear discriminant analysis. *J Am Stat Assoc* 1997, 92:136–143.
51. He X, Fung WK. High breakdown estimation for multiple populations with applications to discriminant analysis. *J Multivar Anal* 2000, 72:151–162.
52. Croux C, Dehon C. Robust linear discriminant analysis using S-estimators. *Can J Stat* 2001, 29:473–492.
53. Hubert M, Van Driessen K. Fast and robust discriminant analysis. *Comput Stat Data Anal* 2004, 45:301–320.



54. Wold S. Pattern recognition by means of disjoint principal components models. *Pattern Recogn* 1976, 8:127–139.
55. Vanden Branden K, Hubert M. Robust classification in high dimensions based on the SIMCA method. *Chemom Intell Lab Syst* 2005, 79:10–21.
56. MacQueen JB. Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Oakland: University of California Press; 1967, 281–297.
57. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons; 1990.
58. Ng RT, Han J. CLARANS: a method for clustering objects for spatial data mining. *IEEE Trans Knowl Data Eng* 2002, 14:1003–1016.
59. Cuesta-Albertos JA, Gordaliza A, Matrán C. Trimmed  $k$ -means: an attempt to robustify quantizers. *Ann Stat* 1997, 25:553–576.
60. García-Escudero LA, Gordaliza A, Matrán C. Trimming tools in exploratory data analysis. *J Comput Graph Stat* 2003, 12:434–449.
61. Gallegos MT, Ritter G. A robust method for cluster analysis. *Ann Stat* 2005, 33:347–380.
62. García-Escudero LA, Gordaliza A, Matrán C, Mayo-Isar A. A general trimming approach to robust cluster analysis. *Ann Stat* 2008, 36:1324–1345.
63. Ramsay J, Silverman B. *Functional Data Analysis*. New York: Springer; 2005.
64. Ferraty F, Vieu P. *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer; 2006.
65. Boente G, Salibián-Barrera M. S-estimators for functional principal component analysis. *J Am Stat Assoc* 2015, 110:1100–1111.
66. Hubert M, Rousseeuw PJ, Segaert P. Multivariate and functional classification using depth and distance. *Adv Data Anal Classif* 2017, 11:445–466.
67. Hyndman R, Shang H. Rainbow plots, bagplots, and boxplots for functional data. *J Comput Graph Stat* 2010, 19:29–45.
68. Sun Y, Genton M. Functional boxplots. *J Comput Graph Stat* 2011, 20:316–334.
69. Arribas-Gil A, Romo J. Shape outlier detection and visualization for functional data: the outliergram. *Bio-statistics* 2014, 15:603–619.
70. Hubert M, Rousseeuw PJ, Segaert P. Multivariate functional outlier detection (with discussion). *Stat Methods Appl* 2015, 24:177–246.
71. Meer P, Mintz D, Rosenfeld A, Kim DY. Robust regression methods in computer vision: a review. *Int J Comput Vis* 1991, 6:59–70.
72. Stewart CV. MINPRAN: a new robust estimator for computer vision. *IEEE Trans Pattern Anal Mach Intell* 1995, 17:925–938.
73. Alqallaf F, Van Aelst S, Yohai VJ, Zamar RH. Propagation of outliers in multivariate data. *Ann Stat* 2009, 37:311–331.
74. Rousseeuw PJ, Van den Bossche W. Detecting deviating data cells. *Technometrics* 2017, <https://doi.org/10.1080/00401706.2017.1340909>.
75. Agostinelli C, Leung A, Yohai VJ, Zamar RH. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST* 2015, 24:441–461.
76. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, Available at: <https://www.R-project.org/>; R Foundation for Statistical Computing; 2016.
77. Verboven S, Hubert M. LIBRA: a Matlab library for robust analysis. *Chemom Intell Lab Syst* 2005, 75:127–136.
78. Verboven S, Hubert M. MATLAB library LIBRA. *Wiley Interdiscip Rev Comput Stat* 2010, 2:509–515.