Special Issue

Journal of CHEMOMETRICS

Received: 10 September 2007; Revised: 7 December 2007; Accepted: 17 December 2007 Published online in Wiley Interscience: 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1123

Outlier detection for skewed data

Mia Hubert[a]* and Stephan Van der Veeken

Most outlier detection rules for multivariate data are based on the assumption of elliptical symmetry of the underlying distribution. We propose an outlier detection method which does not need the assumption of symmetry and does not rely on visual inspection. Our method is a generalization of the Stahel–Donoho outlyingness. The latter approach assigns to each observation a measure of outlyingness, which is obtained by projection pursuit techniques that only use univariate robust measures of location and scale. To allow skewness in the data, we adjust this measure of outlyingness by using a robust measure of skewness as well. The observations corresponding to an outlying value of the adjusted outlyingness (AO) are then considered as outliers. For bivariate data, our approach leads to two graphical representations. The first one is a contour plot of the AO values. We also construct an extension of the boxplot for bivariate data, in the spirit of the bagplot [1] which is based on the concept of half space depth. We illustrate our outlier detection method on several simulated and real data. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: outlier detection; boxplot; bagplot; skewness; outlyingness

1. INTRODUCTION

To detect outliers in multivariate data, it is common practice to estimate the location and scatter of the data by means of robust estimators. Well-known high-breakdown and affine equivariant estimators of location and scatter are, for example, the MCD-estimator [2], the Stahel–Donoho estimator [3,4], S-estimators [5,6] and MM-estimators [7]. Their high-breakdown property implies that the estimators can resist up to 50% of outliers, whereas their affine equivariance allows for any affine transformation of the data (such as rotations, rescaling, translations).

To classify the observations into regular points and outliers, one can then compute robust Mahalanobis-type distances, and use a cutoff value based on the distribution of these distances, see for example Reference [8–10]. All these estimators assume that the data are generated from an elliptical distribution, among which the multivariate gaussian is the most popular one.

Consequently these outlier detection methods will not work appropriately when data are *skewed*. A typical way to circumvent this problem is then to apply a symmetrizing transformation on some (or all) of the individual variables. Common examples are the logarithmic transformation or, more general, a Box-Cox transformation, see for example Reference [11]. This is certainly often a very useful approach, especially when the transformed variables also have a physical meaning. However, this procedure needs more preprocessing, is not affine invariant and leads to new variables which are not always well interpretable. Moreover, the standard Box-Cox transformation is based on maximum likelihood estimation and consequently not robust to outliers.

In this paper, we propose an automatic outlier detection method for skewed multivariate data, which is applied on the raw data. Our method is inspired by the Stahel–Donoho estimator [12]. This estimator is based on the outlyingness of the data points, which are essentially obtained by projecting the observations on many univariate directions and computing a robust center and scale in each projection. The observations are then weighted according to their outlyingness and the robust Stahel–Donoho estimates are obtained as a weighted mean and covariance matrix (see Section 2.4 for the details).

In the first step of our procedure we adjust the Stahel–Donoho outlyingness to allow for asymmetry, which leads to the so-called *adjusted outlyingness* (AO). The method is based on the adjusted boxplot for skewed data [13] and essentially defines for univariate data a different scale on each side of the median. This scale is obtained by means of a robust measure of skewness [14].

In the second step of our outlier detection method, we declare an observation as outlying when its AO is 'too' large. As the distribution of the AO's is in general not known, we apply again the adjusted boxplot outlier rule. All details are provided in Section 2.

In Section 3, we show how our approach can be used to easily obtain two graphical representations of bivariate data that well reflect their center and shape. Section 4 is devoted to a simulation study. Finally, we show in the Appendix that the AO of univariate data has a bounded influence function, which reflects its robustness towards outliers.

It is well known that skewness is only an issue in small dimensions. As the dimensionality increases, the data are more and more concentrated in an outer shell of the distribution, see for example Reference [15]. Hence, in this paper we only consider low-dimensional data sets with, say, at most 10 variables. Of course, it is possible that data are represented in a high-dimensional space, but in fact lie close to a low-dimensional space. Dimension reduction methods are then very helpful preprocessing techniques. One could, for example, first apply a robust PCA method (e.g. [16]), and then apply our new outlier detection method on the principal components scores. A somewhat more refined approach is recently proposed in Reference [17], based on the work presented here.

* *Department of Mathematics—LSTAT, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001, Leuven, Belgium.*
E-mail: mia.hubert@wis.kuleuven.be

a *M. Hubert, S. Van der Veeken*
Department of Mathematics—LSTAT, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001, Leuven, Belgium

2. OUTLIER DETECTION FOR SKEWED DATA

2.1. Outlier detection for skewed univariate data

Since our proposal is based on looking for outliers in one-dimensional projections, we first describe how we detect outliers in skewed *univariate* data. This problem has been addressed in Reference [13], where a skewness-adjusted boxplot is proposed. If $X_n = \{x_1, x_2, \ldots, x_n\}$ is a univariate (continuous, unimodal) data set, the standard boxplot [18] is constructed by drawing a line at the sample median med_n, a box from the first Q_1 to the third Q_3 quartile and whiskers w_1 and w_2 from the box to the furthest non-outlying observations. These observations are defined as all cases inside the interval

$$[Q_1 - 1.5\,\text{IQR}\,,\, Q_3 + 1.5\,\text{IQR}] \tag{1}$$

with the interquartile range $\text{IQR} = Q_3 - Q_1$.

For data coming from a normal distribution, the probability to lie beyond the whiskers is approximately 0.7%. However, if the data are skewed, this percentage can be much higher. For example, in the case of the lognormal distribution (with $\mu = 0$ and $\sigma = 1$), this probability is almost 7%. In Reference [13] the whiskers w_1 and w_2 are adjusted such that for skewed data, much less regular data points fall outside the whiskers. This is obtained by replacing the interval (1) into

$$\left[Q_1 - 1.5e^{-4\text{MC}}\,\text{IQR}\,,\, Q_3 + 1.5e^{3\text{MC}}\,\text{IQR}\right] \tag{2}$$

if $\text{MC} > 0$ and

$$\left[Q_1 - 1.5e^{-3\text{MC}}\,\text{IQR}\,,\, Q_3 + 1.5e^{4\text{MC}}\,\text{IQR}\right]$$

for $\text{MC} < 0$. Here, MC stands for the medcouple which is a robust measure of skewness [14]. It is defined as

$$\text{MC}\,(X_n) = \underset{x_i < \text{med}_n < x_j}{\text{med}} h(x_i, x_j)$$

with med_n the sample median, and

$$h(x_i, x_j) = \frac{(x_j - \text{med}_n) - (\text{med}_n - x_i)}{x_j - x_i}$$

Remark that at symmetric distributions, $\text{MC} = 0$ and hence Equation (2) reduces to Equation (1) from the standard boxplot. It has been shown in Reference [14] that the MC on one hand has a good ability to detect skewness, and on the other hand attains a high resistance to outliers. It has a 25% breakdown value, and a bounded influence function. This means that up to 25% of the regular data can be replaced by contamination before the estimator breaks down, whereas adding a small probability mass at a certain point has a bounded influence on the estimate. Moreover, the medcouple can be computed fast by an $O(n \log n)$ algorithm.

To illustrate the difference between the standard and the adjusted boxplot, we consider an example from geochemistry. The data set comes from a geological survey on the composition in agricultural soils from 10 countries surrounding the Baltic Sea [19]. Top soil (0–25 cm) and bottom soil (50–75 cm) samples from 768 sites were analysed. As an example, we consider the MgO-concentration which was apparently quite skew (MC = 0.39). The original and the adjusted boxplot are shown in Figure 1. We see that the standard boxplot marks many observations as possible outliers, whereas the adjusted boxplot finds no cases with abnormal high concentration of magnesium oxide. There are 15 observations that lie under the lower whisker, but they are clearly boundary cases.

2.2. From the adjusted boxplot to the adjusted outlyingness

The adjusted boxplot introduced in the previous section now allows us to define a skewness-AO for univariate data. According to Stahel [3] and Donoho [4], the outlyingness of a univariate point x_i is defined as

$$\text{SDO}_i = \text{SDO}^{(1)}(x_i, X_n) = \frac{|x_i - \text{med}(X_n)|}{\text{mad}(X_n)}$$

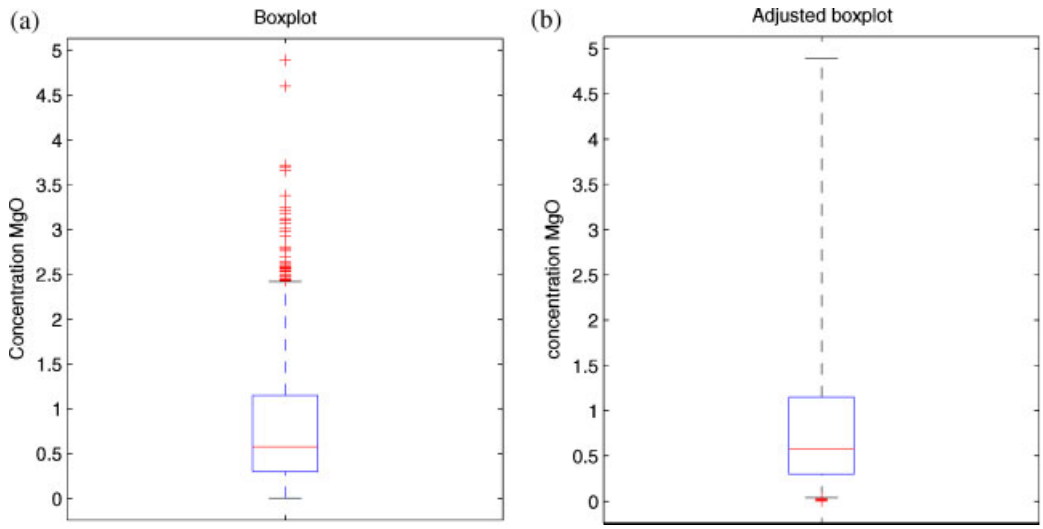

Figure 1. Geological data: (a) standard boxplot; (b) adjusted boxplot. This figure is available in colour online at www.interscience.wiley.com/journal/cem

where $\text{med}(X_n) = \text{med}_n$ is the sample median, and $\text{mad}(X_n) = b\,\text{med}_i|x_i - \text{med}_n|$, the median absolution deviation. The constant $b = 1.483$ is a correction factor which makes the MAD unbiased at the normal distribution. Note that instead of the median and the MAD also other robust estimators of location and scale can be used [16,20]. The outlyingness of a data point tells us how far the observation lies from the centre of the data, standardized by means of a robust scale. In this definition, it does not matter whether the data point is smaller or larger than the median. However, when the distribution is skewed, we propose to apply a different scale on each side of the median. More precisely the AO is defined as

$$\text{AO}_i = \text{AO}^{(1)}(x_i, X_n) = \begin{cases} \frac{x_i - \text{med}(X_n)}{w_2 - \text{med}(X_n)} & \text{if } x_i > \text{med}(X_n) \\ \frac{\text{med}(X_n) - x_i}{\text{med}(X_n) - w_1} & \text{if } x_i < \text{med}(X_n) \end{cases} \tag{3}$$

with w_1 and w_2 the lower and upper whisker of the adjusted boxplot applied to the data set X_n. Again note that AO_i reduces to SDO_i at symmetric distributions.

This AO is illustrated in Figure 2. Observation x_1 has $\text{AO}_1 = d_1/s_1 = (\text{med}(X_n) - x_1)/(\text{med}(X_n) - w_1)$, whereas for x_2 we have $\text{AO}_2 = d_2/s_2 = (x_i - \text{med}(X_n))/(w_2 - \text{med}(X_n))$. So although x_1 and x_2 are located at the same distance from the median, x_1 has a higher value of outlyingness, because the scale on the lower side of the median is smaller than the scale on the upper side. Note that $\text{SDO}^{(1)}$ and $\text{AO}^{(1)}$ are location and scale invariant, hence they are not affected by changing the center and/or the scale of the data.

As the AO is based on robust measures of location, scale and skewness, it is resistant to outliers. In theory, a resistance up to 25% of outliers can be achieved, although we noticed in practice that the medcouple often has a substantial bias when the contamination is more than 10%. Moreover, it can be shown that the influence function [21] of the AO is bounded. We refer to the Appendix for a formal proof.

2.3. Outlier detection for multivariate data

Consider now a p-dimensional sample $\mathbf{X}_n = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n)^T$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$. The Stahel–Donoho outlyingness of $\boldsymbol{x}_i$ is then defined as

$$\text{SDO}_i = \text{SDO}(\boldsymbol{x}_i, \mathbf{X}_n) = \sup_{\boldsymbol{a} \in \mathbb{R}^p} \text{SDO}^{(1)}(\boldsymbol{a}^T\boldsymbol{x}_i, \mathbf{X}_n\boldsymbol{a}) \tag{4}$$

Definition (4) can be interpreted as follows: for every univariate direction $\boldsymbol{a} \in \mathbb{R}^p$ we consider the standardized distance of the projection $\boldsymbol{a}^T\boldsymbol{x}_i$ of observation $\boldsymbol{x}_i$ to the robust center of all the projected data points. Suppose now that $\text{SDO}(\boldsymbol{x}_i, \mathbf{X})$ is large, then there exists a direction in which the projection of $\boldsymbol{x}_i$ lies far away from the bulk of the other projections. As such, one might suspect $\boldsymbol{x}_i$ being an outlier.

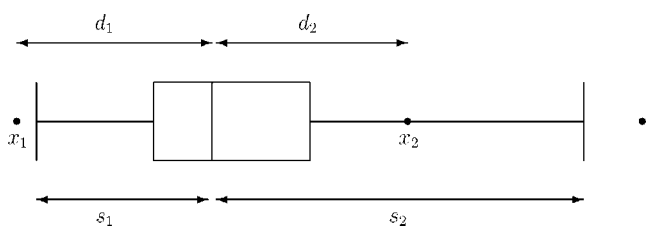

Figure 2. Illustration of the adjusted outlyingness.

It is clear from its definition that the SD outlyingness does again not account for any skewness, and hence it is only suited for elliptical symmetric data. To allow skewness, we analogously define the AO of a multivariate observation $\boldsymbol{x}_i$ as

$$\text{AO}_i = \text{AO}(\boldsymbol{x}_i, \mathbf{X}_n) = \sup_{\boldsymbol{a} \in \mathbb{R}^p} \text{AO}^{(1)}(\boldsymbol{a}^T\boldsymbol{x}_i, \mathbf{X}_n\boldsymbol{a}) \tag{5}$$

Note that in practice the AO can not be computed by projecting the observations on *all* univariate vectors $\boldsymbol{a}$. Hence, we should restrict ourselves to a finite set of random directions. Many simulations have shown that considering $m = 250p$ directions yields a good balance between 'efficiency' and computation time. Random directions are generated as the direction perpendicular to the subspace spanned by p observations, randomly drawn from the data set (as in Reference [12]). As such, the AO is invariant to affine transformations of the data. Moreover, in our implementation we always take $\|\boldsymbol{a}\| = 1$, although this is not required as $\text{AO}^{(1)}$ is scale invariant.

Once the AO is computed for every observation, we can use this information to decide whether an observation is outlying or not. Unless for normal distributions for which the AOs (or SDOs) are asymptotically χ^2_p distributed, the distribution of the AO is in general unknown (but typically right-skewed as they are bounded by zero). Hence, we compute the adjusted boxplot of the AO-values and declare a multivariate observation outlying if its AO_i exceeds the upper whisker of the adjusted boxplot. More precisely, our outlier cutoff value equals

$$\text{cutoff} = Q_3 + 1.5e^{3\text{MC}}\,\text{IQR} \tag{6}$$

where Q_3 is the third quartile of the AO_i and similarly for IQR and MC.

Remark 1. Note that the construction of the adjusted boxplot and the AO does not assume any particular underlying skewed distribution (only unimodality), hence it is a distribution-free approach. For univariate skewed data, several more refined robust estimators and outlier detection methods are available, see for example Reference [22–24], but then one needs to assume that the data are sampled from a specific class of skewed distributions (such as the gamma distribution). Our approach is in particular very useful when no information about the data distribution is available and/or when an automatic and fast outlier detection method is required.

Remark 2. A similar outlier detection method has also been proposed in Reference [25] to robustify independent component analysis (ICA). However, in Reference [25] a different definition of AO was used, by replacing the constants 3 and 4 in Equation (2) by 4 and 3.5, yielding

$$\left[Q_1 - 1.5e^{-3.5\text{MC}}\,\text{IQR}\,,\, Q_3 + 1.5e^{4\text{MC}}\,\text{IQR}\right] \tag{7}$$

for right-skewed distributions (and similarly for left-skewed data).

Definition (7) yields a larger fence than when we apply our current definition (2). This affects both the scale estimates in Equation (3) as well as the cutoff value (6) which separates the regular points from the outliers. When the proportion of contamination is small, which is the typical problem in the context of ICA, such a rule will work very well. Compared to our current approach, it will even often misclassify less regular observations

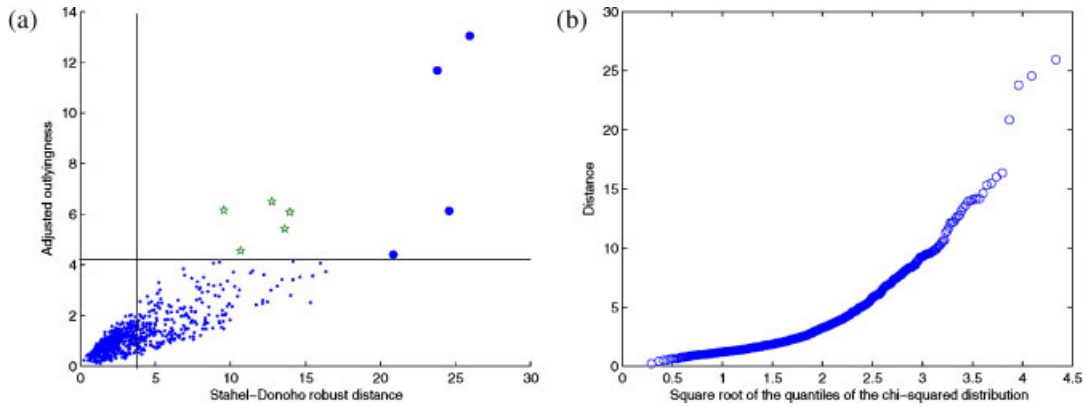

Figure 3. (a) Adjusted outlyingness versus Stahel–Donoho robust distances; (b) χ^2_4-quantile plot of the SD distances. This figure is available in colour online at www.interscience.wiley.com/journal/cem

as outliers. However, when the contamination percentage is larger, say 5–10%, the medcouple will show more bias and the factor e^{4MC} might become too large, resulting in whiskers that might mask some or all of the outliers. Therefore, in the general setting considered here, we prefer to work with the new rules.

Remark 3. Note that the concept of 'robustness towards outliers' can become ambiguous in the context of skewed distributions. Assume that a large majority of observations is sampled from a symmetric distribution, and that some smaller group (at most 25%) is outlying. When the outliers are located far from the regular points, a robust estimator of skewness should be able to detect the symmetry of the main group. An outlyingness-approach based on such a robust estimator of skewness, combined with robust estimators of location and scale, can then be able to flag the outlying measurements. When the same methodology would be used with non-robust estimators of location, scale and skewness, the outlyingness-values would be affected by the outliers (e.g. yielding a high value of skewness, and an inflated scale) such that the outlying group could be masked. This difference between a robust and non-robust approach also applies when the majority group has an asymmetric distribution. In such a situation, outliers could for example give the impression that the whole distribution is highly asymmetric, whereas this might not hold for the large majority. If on the other hand there are no outliers and the whole distribution is indeed skewed, a robust estimator of skewness should also be able to detect the asymmetry. This is why we prefer to work with the medcouple. In Reference [26], it is shown that the MC is not too conservative (such that asymmetry of the main group can be found) but robust enough (asymmetry due to outliers is detected when the outliers are far enough in the tails of the distribution).

However, when the outliers are located not very far in the tails of the main distribution, the distinction between the regular and outlying points might become very small. From our point of view, no estimator (robust or not) can then be able to make the correct distinction. If one then presumes that the asymmetry is caused by the outliers, and that the main group has a symmetric distribution, we advise to compare the AO-values with the SDO-values (or any other outlier detection method for symmetric data). If the conclusions are very different, it is then up to the analyst to decide whether he/she believes in the symmetry of the main group or not.

2.4. Example

We reconsider the geological data set of Section 2.1, and now consider the variables that measure the concentration of MgO, MnO, Fe_2O_3 and TiO_2. Hence $n = 768$ and $p = 4$. The medcouple of the individual variables is 0.39, 0.2, 0.26 and 0.14, respectively which clearly indicates the presence of skewness in this data set. Moreover the adjusted boxplots of the four variables marked several observations as (univariate) outliers.

When we apply our outlier detection method based on the AO, we find nine observations that exceed the outlier cutoff. Figure 3(a) plots the AO-values on the vertical axis, together with the adjusted boxplot cutoff (6). We see that two cases are really far outlying, whereas five observations have a somewhat larger AO, and the other two are merely boundary cases.

For this data set we also computed the robust distances

$$\mathrm{RD}_i = \sqrt{(\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})} \tag{8}$$

with $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ the Stahel–Donoho estimates of location and scatter. The SD estimator is defined by assigning a weight to every data point, inversely proportional to its outlyingness, and computing the weighted mean and covariance matrix. According to Reference [20], we applied the gaussian weights

$$w_i = \frac{\phi(\mathrm{SDO}_i^2/c)}{\phi(1)}$$

with ϕ the gaussian density and $c = \chi^2_{p,0.9}$ the 90% quantile of the χ^2 distribution with $p = 4$ degrees of freedom. This weight function decreases exponentially for $\mathrm{SDO}_i^2 > c$ and accords relatively high weights for (squared) SDO values smaller than c.

Figure3(a) shows the robust SD distances on the horizontal axis, together with the common cutoff value $\sqrt{\chi^2_{4,0.99}}$ (since the robust distances are approximately χ^2_p distributed at normal data). We see that the SD estimator detects four clear outliers (indicated with a large dot), but also yields a huge number of observations outside the outlier cutoff value. From the χ^2_4 quantile plot of the robust SD distances in Figure 3(b), we can deduce that the robust distances are not χ^2_4 distributed (as the data are skewed) and hence the cutoff value is not appropriate.

In Figure 4 we show several pairwise scatterplots indicating the observations with outlying AO value. The four outliers with highest robust SD distance are marked with a large dot. The remaining five observations with outlying AO are marked with

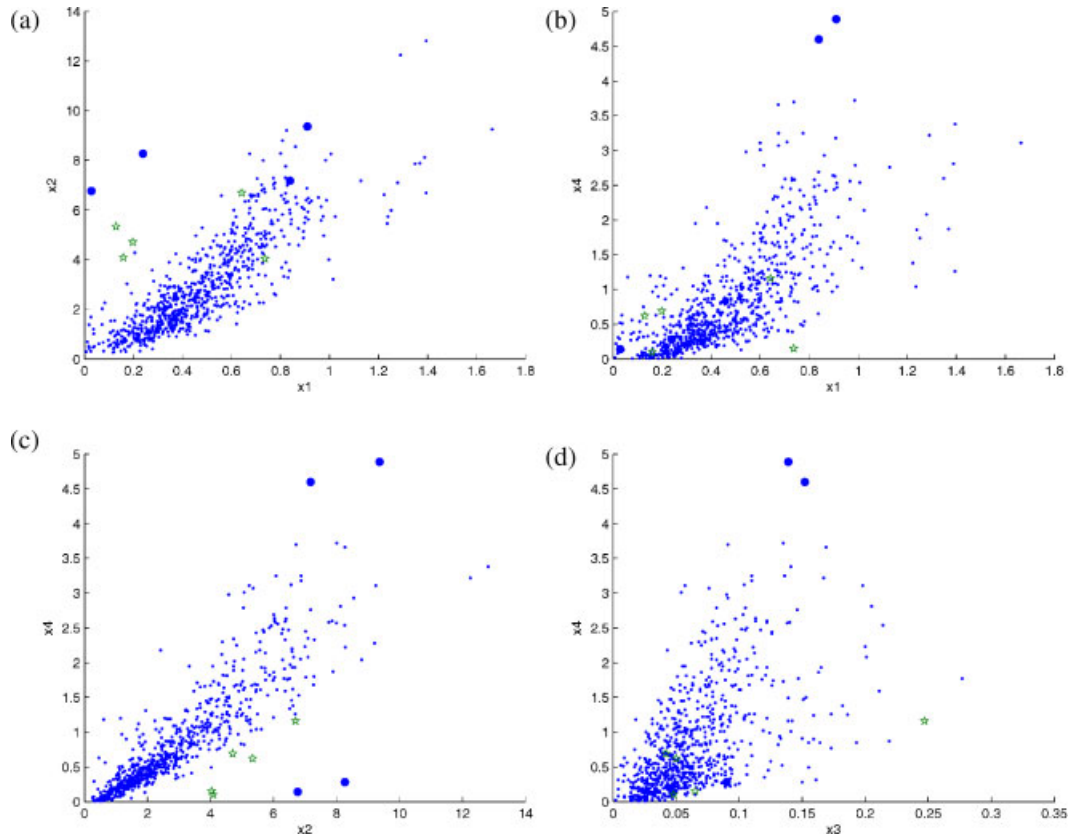

Figure 4. Several scatterplots of the geological data with outliers marked. (a) MnO versus MgO; (b) TiO_2 versus MgO; (c) TiO_2 versus MnO; (d) TiO_2 versus Fe_2O_3. This figure is available in colour online at www.interscience.wiley.com/journal/cem

a star. These scatterplots show the multivariate skewness in the data, and illustrate why these nine cases are different from the others. Figures 4(a) and (c) are the most informative ones, and demonstrate that the outliers merely have outlying (x_1, x_2) and/or (x_2, x_4) measurements.

3. GRAPHICAL REPRESENTATIONS FOR BIVARIATE DATA

For bivariate data, the AO-values can be used to easily obtain two graphical representations of the data that well reflect their center and shape.

3.1. Contour plot

The first representation consist of a contour plot of the AO values. To illustrate such a contour plot, we consider the bloodfat data from Reference [27]. For 371 male patients, data were collected on the concentration of plasma cholesterol and plasma triglycerides. The units are mg/dl. For 51 patients, there was no evidence of heart disease; for the remaining 320 patients there was evidence of narrowing of the arteries. Only those last 320 data points are used in the analysis. Both the SD and the AO of the data are computed. Using cubic interpolation (by means of the Matlab function *interp2*), contour plots of the two outlyingness measures are constructed. These plots are shown in Figure 5. We see that the contours of the AO show the underlying skewed distribution

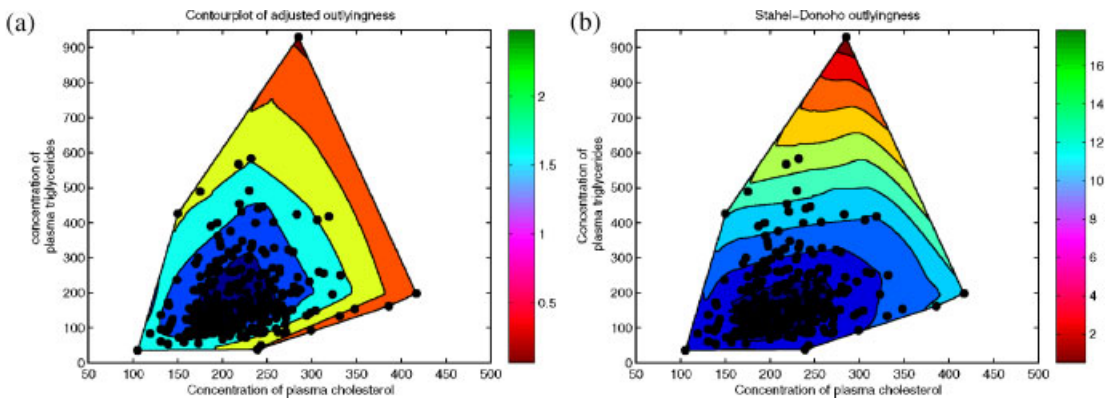

Figure 5. Contourplots of the (a) adjusted outlyingness and (b) Stahel–Donoho outlyingness of the bloodfat data. This figure is available in colour online at www.interscience.wiley.com/journal/cem

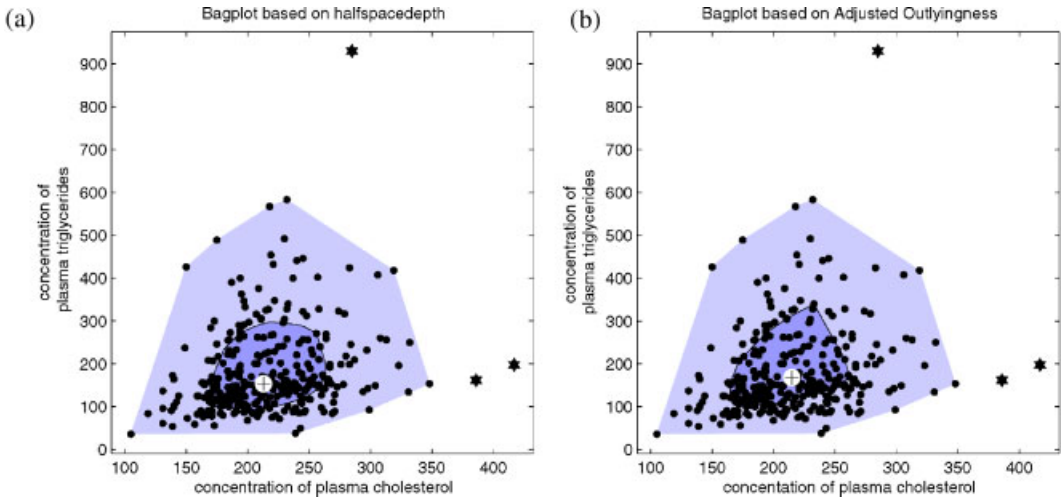

Figure 6. Bagplots of the bloodfat data based on (a) halfspacedepth and (b) adjusted outlyingness. This figure is available in colour online at www.interscience.wiley.com/journal/cem

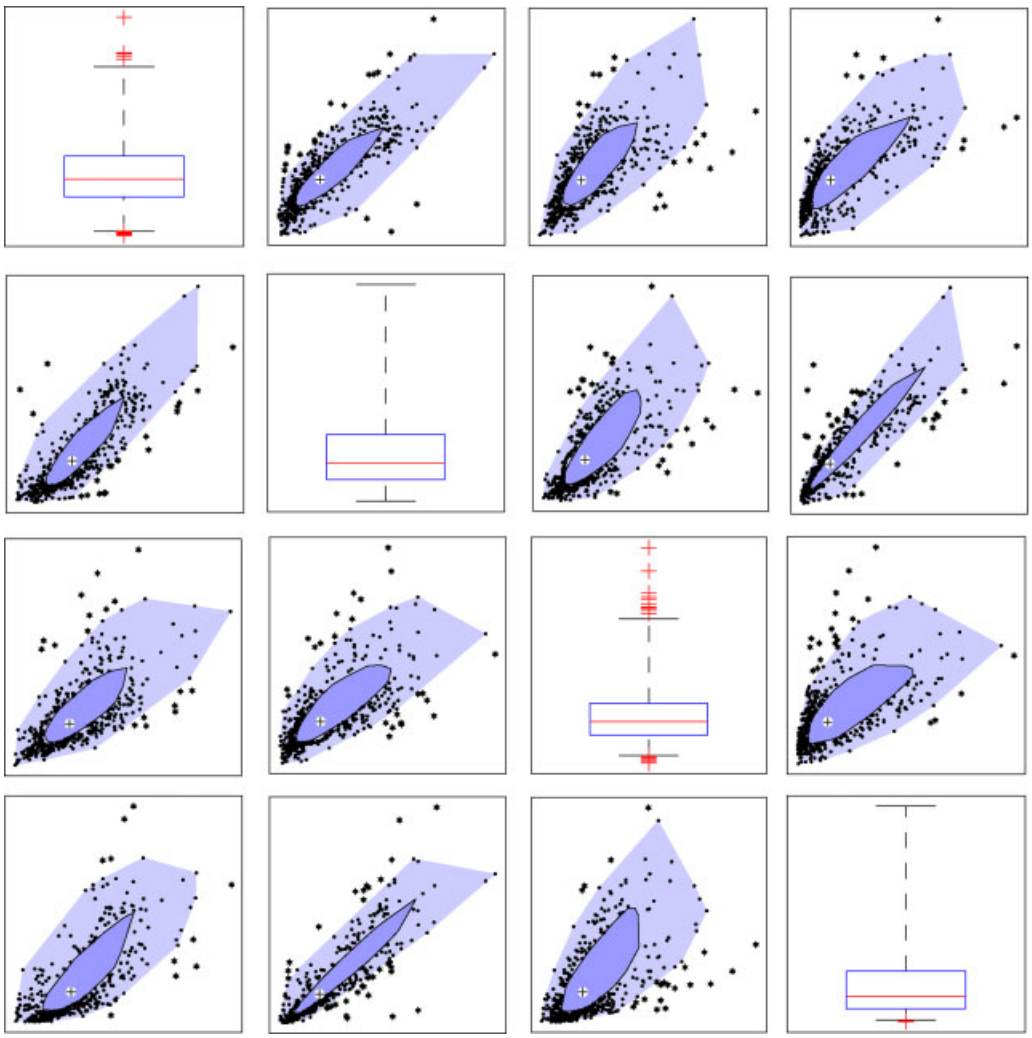

Figure 7. Bagplot matrix of the geological data. This figure is available in colour online at www.interscience.wiley.com/journal/cem

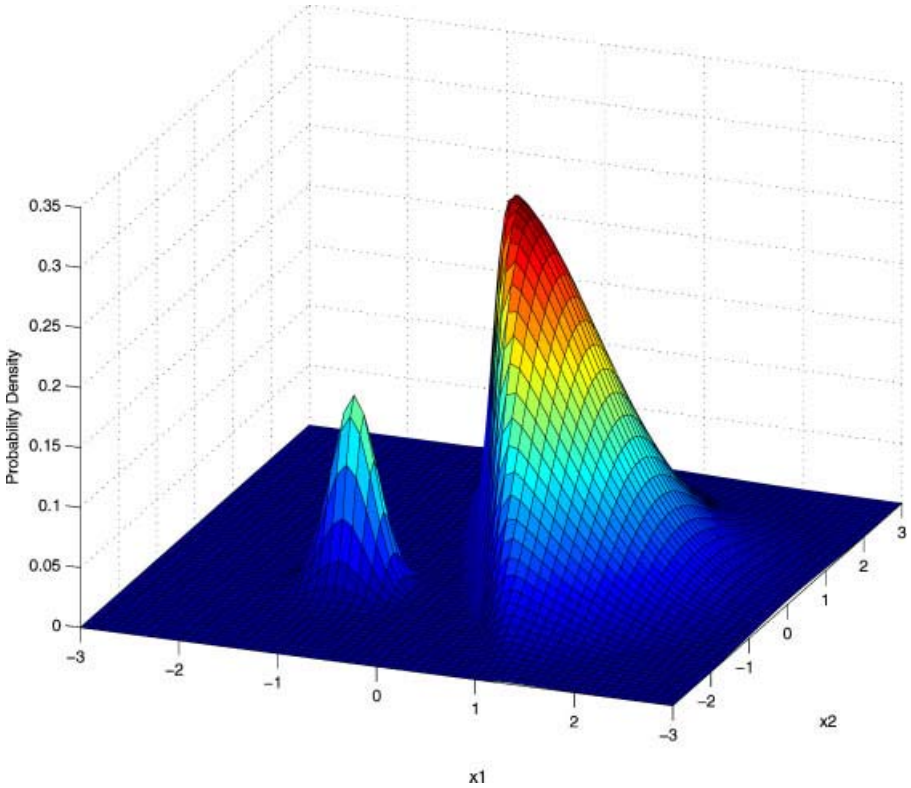

Figure 8. Density plot of simulated data from a skew-normal distribution and 10% outliers. This figure is available in colour online at www.interscience.wiley.com/journal/cem

very well. On the other hand, the inner contours of the SDO values are closer to elliptical.

3.2. Bagplot

The bagplot is introduced in Reference [1] as an extension of the boxplot for bivariate data. Just as the boxplot, the construction of the bagplot relies on a ranking of the data points. This ranking is based on the concept of halfspace depth, which was introduced in Reference [28]. The halfspace depth of a bivariate point $\boldsymbol{x}$ is defined as the smallest number of data points, lying in a closed halfplane bounded by a line through $\boldsymbol{x}$. Using this halfspace depth, a bivariate equivalent of the median can be defined as the point (not necessarily an observation) with the highest depth, called the Tukey median. If this point is not unique, the center of gravity of the deepest depth region is taken (see Reference [1] for more details). The bagplot consists of the Tukey median, the bag and the fence. The bag contains the 50% data with highest depth. The fence is defined by inflating the bag (relative to the Tukey median) by a factor 3. All observations outside the fence are considered to be outliers. The outer loop consists of the convex hull of the non-outlying observations. In Figure 6(a) the bagplot of the bloodfat data is shown. We clearly see the skewness in the data, as the Tukey median (indicated with the + symbol) does not lie in the centre of the (dark-coloured) bag, which itself is not elliptically shaped. Also the light-coloured loop is skewed and separates the three outliers (with star symbol) from the other observations. As illustrated in this example, the bagplot is very useful to show the shape of bivariate data as the halfspace depth does not make any distributional assumptions. Moreover the bagplot is equivariant to affine transformations. Its only drawback is its computational complexity, which is $O(n^2(\log n)^2)$. For larger datasets, the computation time can be reduced by drawing a random subset from the data and performing the computations on this smaller data set. This approach has been proposed and applied in Reference [1]. This explains why the bagplot of the bloodfat data in Reference [1], based on a random subset of size 150, is slightly different from Figure 6(a) which uses all observations.

The concept of AO allows us to make a similar bagplot in much lower computation time. Instead of the Tukey median we mark the observation with lowest AO, and we define the bag as the convex hull of the half sample with lowest outlyingness. If we look at the bagplot based on AO in Figure 6(b) we see that it is very similar to the depth-based bagplot and the same observations are classified as outliers. As the AO-values can be computed in $O(mnp \log n)$ time with m the number of directions considered, and as we usually set $m = 250p$, this approach thus yields a fast alternative to the depth-based bagplot.

Note that there exist alternative graphical representations of bivariate data, such as those based on kernel density estimation. As kernel methods concentrate on local properties, they are in particular suitable to detect multimodality. However, the notion of outlier is different from what we have used in this paper. Kernel methods will consider isolated points as outliers, whereas we try to detect observations which are far away from the bulk of the data. We refer to Reference [1] for an overview of alternative graphs and more discussion.

The AO-based bagplot can easily be extended to higher dimensions, as long as the software accurately supports high-dimensional graphs. To visualize multivariate data, we can alternatively also construct a bagplot matrix (as in Reference [1]). This is illustrated in Figure 7 for the geological data of Section 2.4.

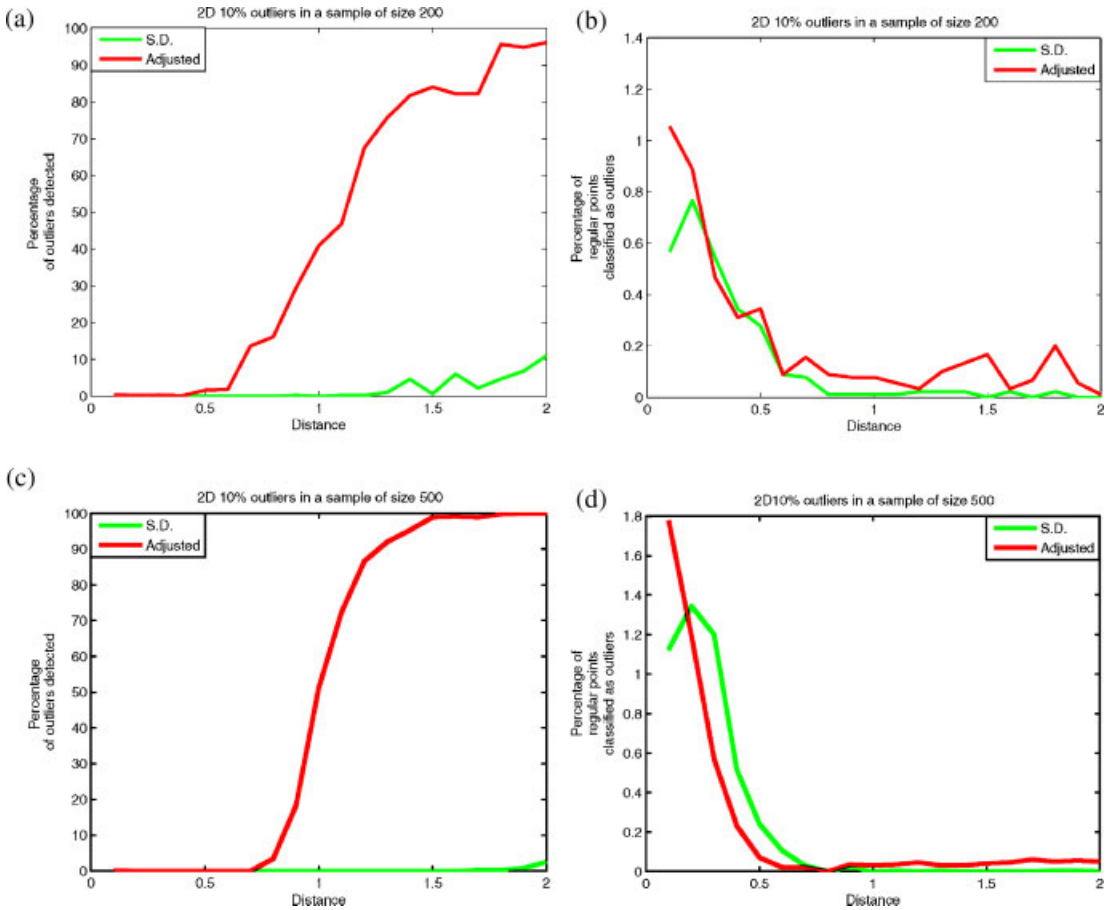

Figure 9. Simulation results for two-dimensional data of size $n = 200$ and $n = 500$. Figures (a,c) show the percentage of outliers correctly identified, Figures (b,d) show the percentage of regular observations incorrectly classified as outliers. This figure is available in colour online at www.interscience.wiley.com/journal/cem

On the diagonal we have plotted the adjusted boxplot of each variable, whereas the other cells of the matrix contain the AO-bagplot of each pair of variables. Note that as the number of observations in the bag is quite large, we have not drawn all these observations.

4. SIMULATION STUDY

In this section, we study the outlier detection ability of our approach by means of a simulation study. To this end we have generated data from a multivariate skew-normal distribution [29]. A p-dimensional random variable X is said to be multivariate skew-normal distributed if its density function is of the form

$$f(\boldsymbol{x}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^T\boldsymbol{x}) \tag{9}$$

where $\phi_p(\boldsymbol{x}; \boldsymbol{\Omega})$ is the p-dimensional normal density with zero mean and correlation matrix $\boldsymbol{\Omega}$, Φ is the standard normal distribution and $\boldsymbol{\alpha}$ is a p-dimensional vector that regulates the shape. Note that if $\boldsymbol{\alpha} = 0$, the skew-normal density reduces to the standard normal density. In our simulations we set $\boldsymbol{\Omega} = \mathbf{I}_p$ the identity matrix, and $\boldsymbol{\alpha}$ a vector with elements equal to 10 or 4. For $p = 2$ we used $\boldsymbol{\alpha} = (10, 4)^T$, for $p = 5$ we set $\boldsymbol{\alpha} = (10, 10, 4, 4, 4)^T$, whereas for $p = 10$ we took $\boldsymbol{\alpha} = (10, 10, 10, 10, 10, 4, 4, 4, 4, 4)^T$. Outliers are randomly generated from a normal distribution with $\mathbf{I}_p/20$ as covariance matrix and a centre located along the $-\mathbf{1}_p$ direction (all components equal to -1). This is on purpose not the direction of maximal directional skewness [30], but just a direction in which there is a considerable amount of skewness. The contamination was chosen to be clustered as from the simulation study in Reference [25] this setting appeared to be the most difficult to handle. We considered situations with 1% or 10% contamination in data sets of size $n = 200$, 500 and 1000. An example of such a simulation data set with 10% contamination is illustrated in Figure 8.

We compare two methods for outlier detection. The first is our approach based on the AO-values, as introduced in Section 2.3. For comparison, the second approach is based on the SD outlyingness. It would have been possible to use the robust distances from the SD estimator. However, as we have noticed in the previous sections and in our simulations, this method always yields a huge number of observations that are (erroneously) indicated as outliers. This stems from the fact that the SD method assumes symmetry in the definition of the outlyingness, as well as in the use of the χ^2_p cutoff value. To eliminate the effect of the cutoff value, we therefore consider another outlier detection approach, obtained by applying our adjusted boxplot rule to the SD outlyingness. So the two methods used in this simulation study only differ in the definition of the outlyingness, and not in how they define outliers. This makes it more easy to quantify the improvements arising from the skewness adjustment in the outlyingness.

In Figures 9–11 we report some results of our simulation study. The left figures present the percentage of outliers that were

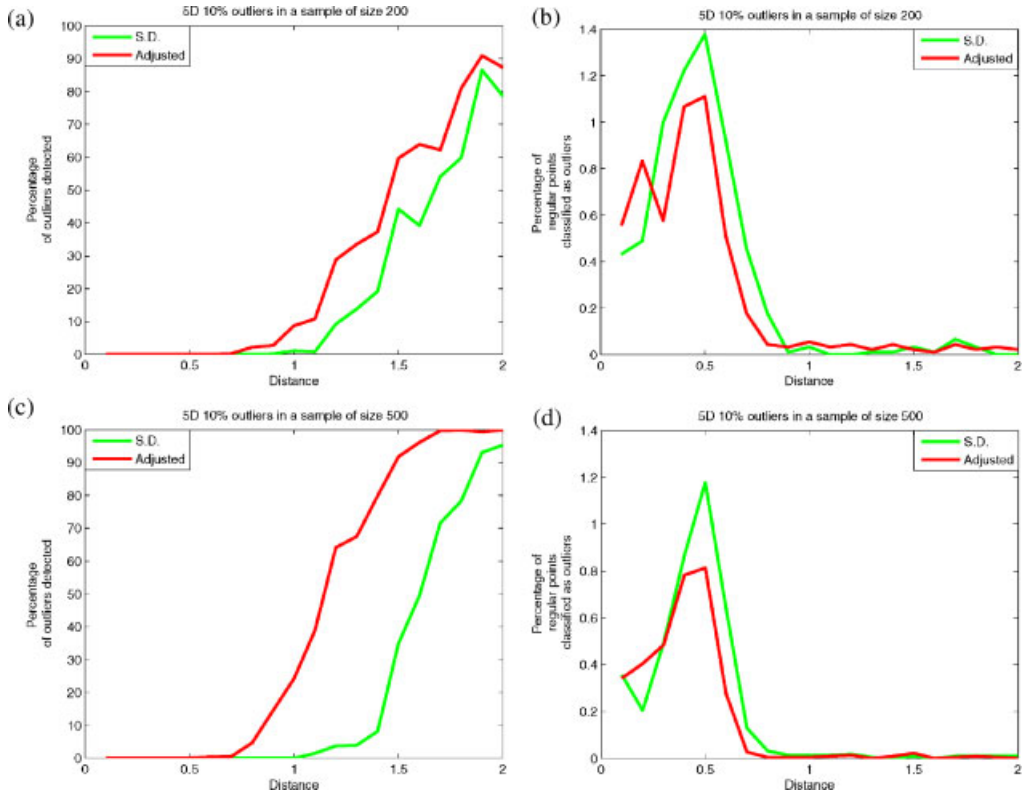

Figure 10. Simulation results for five-dimensional data of size n =200 and n =500. Figures (a,c) show the percentage of outliers correctly identified, Figures (b,d) show the percentage of regular observations incorrectly classified as outliers. This figure is available in colour online at www.interscience.wiley.com/journal/cem

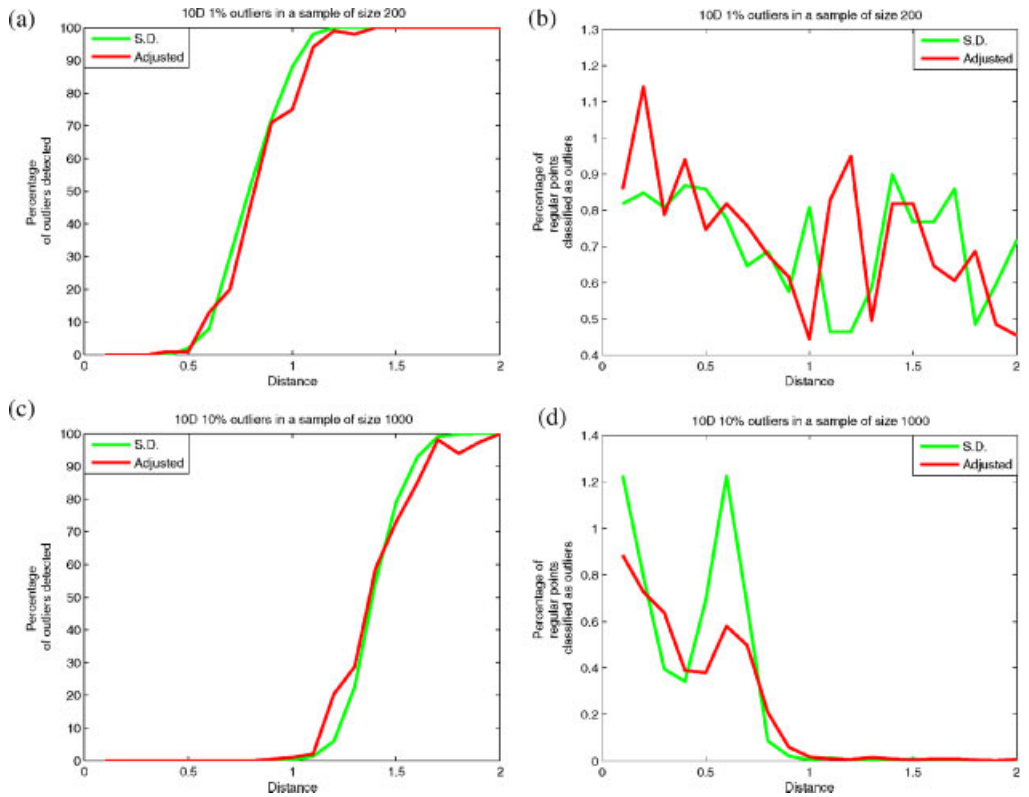

Figure 11. Simulation results for 10-dimensional data of size n = 200 and n = 1000. Figures (a,c) show the percentage of outliers correctly identified, Figures (b,d) show the percentage of regular observations incorrectly classified as outliers. This figure is available in colour online at www.interscience.wiley.com/journal/cem

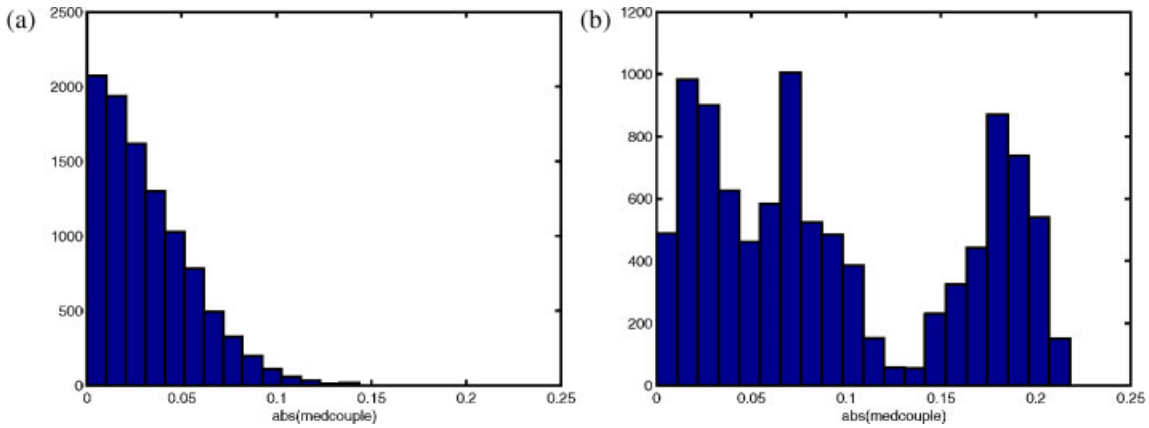

Figure 12. Histogram of the absolute MC values on all projections for a simulated data set of dimension (a) 10 and (b) 2. This figure is available in colour online at www.interscience.wiley.com/journal/cem

detected by the two methods, as a function of the distance of the center of the outliers from the origin. The figures to the right show the percentage of regular observations that were erroneously classified as outliers.

In two dimensions (Figure 9), it is clear that the AO method outperforms the SD approach considerably with respect to the detection of the correct outliers. The improvement becomes even more apparent as the sample size increases. Both methods are comparable in misclassifying regular observations.

In five dimensions (Figure 10), the gain of the skewness adjustment is still present and again more pronounced when n increases. In 10 dimensions (Figure 11) on the other hand, both methods are comparable. This is again because the data considered here do not expose a lot of skewness in 10 dimensions. To illustrate, Figure 12(a) shows for one of our simulated data sets ($n = 1000$) in 10 dimensions a histogram of the (absolute) MC values on 10 000 projections. For a two-dimensional data set, we obtain Figure 12(b). We see that the skewness on average is much smaller when $p = 10$. Consequently the AO-values will be very similar to the SDO-values.

5. CONCLUSION

In this paper, we have proposed an outlier detection method for multivariate skewed data. The procedure is based on the skewness-AO, is distribution-free and easy to compute. Moreover, we have presented contourplots and a bagplot based on the AO to visualize the distribution of bivariate data. Simulations and examples on real data have illustrated that our method outperforms robust methods that are designed for elliptical data. Software to compute these AO-values and to drawn the bagplot (based on the AO or on the halfspace depth) are available at `wis.kuleuven.be/stat/robust` as part of LIBRA: Matlab Library for Robust Analysis [31].

APPENDIX: INFLUENCE FUNCTION

In this section, we derive the influence function of the AO of a univariate continuous distribution F with density f. This influence function describes the effect on the AO of a point $x \in \mathbb{R}$ when we put an infinitesimally small contamination in a point $z \in \mathbb{R}$ [21].

More precisely, consider the contaminated distribution

$$F_{\epsilon,z} = (1 - \epsilon)F + \epsilon\Delta_z$$

for small ϵ. The distribution Δ_z is the Dirac distribution which puts all probability mass at the point z. Then the influence function of an estimator T at the distribution F is defined as

$$\mathrm{IF}(z; T, F) = \lim_{\epsilon \downarrow 0} \frac{T(F_{\epsilon,z}) - T(F)}{\epsilon} \tag{10}$$

Here, T is the univariate AO in some x. Therefore, the influence function depends both on the position of the contamination as well as on the position of the observation in which the AO is computed. We compute the influence function at a skew-normal distribution with, according to Reference [29], density function:

$$f_\alpha(z) = 2\phi(z)\Phi(\alpha z)$$

Its distribution function is then given by $\Phi_\alpha(z) = \Phi(z) - 2T_\alpha(z)$ with the T-function defined as

$$T_\alpha(z) = \frac{1}{2\pi}\int_0^\alpha \frac{\exp(-1/(2z^2(1+x^2)))}{1+x^2}\mathrm{d}x$$

We derive the IF at the skew-normal distribution $F = F_1$ with the skewness parameter α equal to 1. This distribution has $\mathrm{Med}(F) = \Phi^{-1}(1/\sqrt{2})$. Another choice of α could have been considered as well, but then the median can only be obtained by numerical integration. The theoretical value of the medcouple can be found as the solution of

$$\mathrm{MC}_F = H_F^{-1}(0.5)$$

with

$$H_F(u) = 4\int_{\mathrm{Med}_F}^{\infty} F\left(\frac{x_2(u-1) + 2\mathrm{Med}_F}{u+1}\right)\mathrm{d}F(x_2)$$

Solving this equation gives as result that the population medcouple equals 0.021.

To compute the influence function, two different cases now have to be considered: points located on the lower side of the

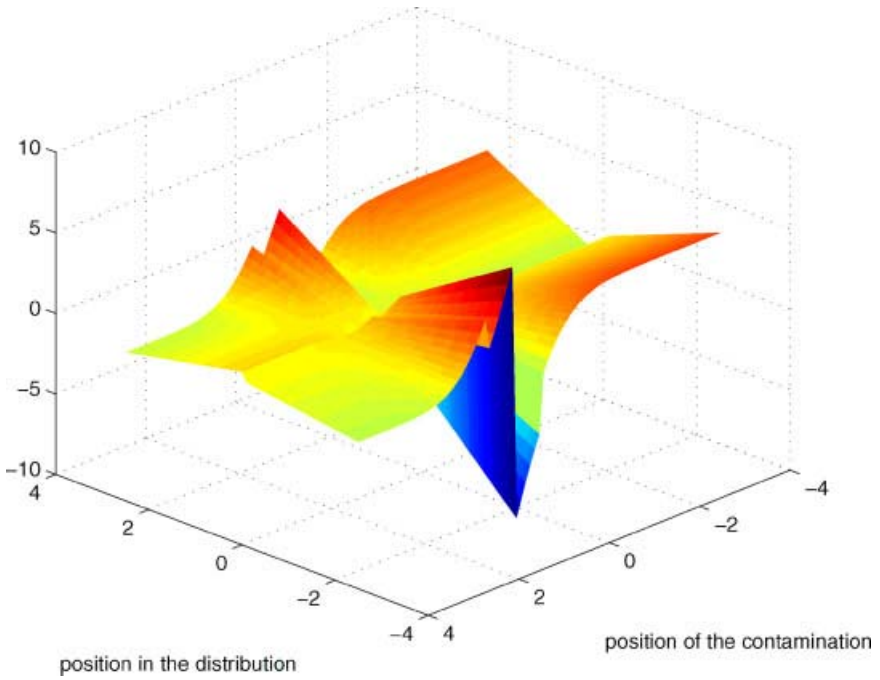

Figure 13. Influence function of the univariate adjusted outlyingness at a skew-normal distribution. This figure is available in colour online at www.interscience.wiley.com/journal/cem

median and points on the upper side. Consider first $x < \text{Med}(F)$. The AO is then defined as

$$\text{AO}^{(1)}(x, F) = \frac{\text{Med}_F - x}{\text{Med}_F - Q_1 + 1.5e^{-4\,\text{MC}}\,\text{IQR}}$$

(since the skew-normal has MC > 0). When we contaminate F, we may assume that ϵ is sufficiently small such that $x < \text{Med}(F_\epsilon)$ and $\text{MC}(F_\epsilon) > 0$. Since

$$\text{IF}\,(z, \text{AO}^{(1)}(x, F), F) = \frac{\partial}{\partial \epsilon}\text{AO}^{(1)}(x, F_\epsilon)|_{\epsilon=0}$$

we can easily derive that

$$\begin{aligned}&\text{IF}\,(z, \text{AO}^{(1)}(x, F), F)\\ &= \frac{1}{4.43}\big(2.105\,\text{IF}\,(z, \text{Med}, F) + (\text{Med}(F) - x)[\text{IF}\,(z, \text{Med}, F)\\ &\quad - \text{IF}\,(z, Q_1, F) + 1.41\,\text{IF}\,(z, \text{IQR}, F) - 4.67\,\text{IF}\,(z, \text{MC}, F)]\big) \quad (11)\end{aligned}$$

Expressions for the influence function of quantiles can, for example, be found in Reference [32], whereas the influence function of the medcouple is given in Reference [14]. The influence function for points located at the upper side of the median is calculated in a similar way. The resulting function is plotted in Figure 13. Since all the influence functions that appear in expression (11) are bounded, the influence function of the AO is bounded (in z) as well, showing its robustness. Note that the adjusted outlyingness $\text{AO}^{(1)}(x, F)$ is not bounded in x, but when x is fixed, the effect of contamination in any point (even in $z = x$) is bounded. Mathematically, the derivative with respect to z tends to a constant.

Acknowledgement

This research was supported by GOA/07/04-project of the Research Fund KULeuven and by the IAP research network nr. P6/03 of the Federal Science Policy, Belgium'.

REFERENCES

1. Rousseeuw PJ, Ruts I, Tukey JW. The bagplot: a bivariate boxplot. *Am. Statistician* 1999; **53**: 382–387.
2. Rousseeuw PJ. Least median of squares regression. *J. Am. Statistical Assoc.* 1984; **79**: 871–880.
3. Stahel WA. Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. *PhD Thesis*, ETH Zürich, 1981.
4. Donoho DL. Breakdown properties of multivariate location estimators. *Qualifying paper*, Harvard University, Boston, 1982.
5. Rousseeuw PJ, Yohai VJ. Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*, Franke J, Härdle W, Martin RD (eds). Springer-Verlag: New York, 1984; 256–272, Lecture Notes in Statistics No. 26.
6. Davies L. Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *Ann. Stat.* 1987; **15**: 1269–1292.
7. Tatsuoka KS, Tyler DE. On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Ann. Stat.* 2000; **28**: 1219–1243.
8. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. Wiley-Interscience: New York, 1987.
9. Maronna RA, Martin DR, Yohai VJ. *Robust Statistics: Theory and Methods*. Wiley: New York, 2006.
10. Rousseeuw PJ, Debruyne M, Engelen S, Hubert M. Robustness and outlier detection in chemometrics. *Crit. Rev. Anal. Chem.* 2006; **36**: 221–242.
11. Rayens WS, Srinivasan C. Box-Cox transformations in the analysis of compositional data. *J. Chemometrics* 1991; **5**: 227–239.
12. Maronna RA, Yohai VJ. The behavior of the Stahel–Donoho robust multivariate estimator. *J. Am. Statistical Assoc.* 1995; **90**: 330–341.
13. Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.* 2008. In press. Doi: 10.1016/j.csda.2007.11.008
14. Brys G, Hubert M, Struyf A. A robust measure of skewness. *J. Comput. Graph. Stat.* 2004; **13**: 996–1017.

15. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer: New York, 2001.
16. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: a new approach to robust principal components analysis. *Technometrics* 2005; **47**: 64–79.
17. Hubert M, Rousseeuw PJ, Verdonck T. Robust PCA for skewed data. 2007. Submitted.
18. Tukey JW. *Exploratory Data Analysis*. Reading (Addison-Wesley): MA, 1977.
19. Reimann C, Siewers U, Tarvainen T, Bityukova L, Eriksson J, Gilucis A, Gregorauskiene V, Lukashev V, Matinian NN, Pasieczna A. Baltic soil survey: total concentrations of major and selected trace elements in arable soils from 10 countries around the Baltic Sea. *Sci. Total Environ.* 2000; **257**: 155–170.
20. Gervini D. The influence function of the Stahel–Donoho estimator of multivariate location and scatter. *Stat. Probab. Lett.* 2002; **60**: 425–435.
21. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics: The Approach Based on Influence Functions*. Wiley: New York, 1986.
22. Marazzi A, Ruffieux C. The truncated mean of an asymmetric distribution. *Comput. Stat. Data Anal.* 1999; **32**: 79–100.
23. Markatou M, Basu A, Lindsay BG. Weighted likelihood equations with bootstrap root search. *J. Am. Stat. Assoc.* 1998; **93**: 740–750.
24. Victoria-Feser M-P, Ronchetti E. Robust methods for personal-income distribution models. *Can. J. Stat.* 1994; **22**: 247–258.
25. Brys G, Hubert M, Rousseeuw PJ. A robustification of independent component analysis. *J. Chemometrics* 2005; **19**: 364–375.
26. Brys G, Hubert M, Struyf A. A comparison of some new measures of skewness. In *Developments in Robust Statistics: International Conference on Robust Statistics 2001* (Vol. 114), Dutter R, Filzmoser P, Gather U, Rousseeuw PJ (eds). Physika Verlag: Heidelberg, 2003; 98–113.
27. Hand DJ, Lunn AJ, McConway AD, Ostrowski E. *A Handbook of Small Data Sets*. Chapman and Hall: London, 1994.
28. Tukey JW. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (Vol. 2). Vancouver, Canada, 1975; 523–531.
29. Azzalini A, Dalla Valle A. The multivariate skew-normal distribution. *Biometrika* 1996; **83**: 715–726.
30. Ferreira JT, Steel MF. On describing multivariate skewed distributions: a directional approach. *Can. J. Stat.* 2006; **34**: 411–429.
31. Verboven S, Hubert M. LIBRA: a Matlab library for robust analysis. *Chemometrics Intell. Lab. Sys.* 2005; **75**: 127–136.
32. Huber PJ. *Robust Statistics*. Wiley: New York, 1981.

