Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 5186–5201
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

An adjusted boxplot for skewed distributions

M. Hubert[a,*], E. Vandervieren[b]

[a] Department of Mathematics - Leuven Statistics Research Center, K.U.Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium
[b] Department of Mathematics & Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium

Received 11 June 2007; received in revised form 30 October 2007; accepted 10 November 2007
Available online 21 November 2007

Abstract

The boxplot is a very popular graphical tool for visualizing the distribution of continuous unimodal data. It shows information about the location, spread, skewness as well as the tails of the data. However, when the data are skewed, usually many points exceed the whiskers and are often erroneously declared as outliers. An adjustment of the boxplot is presented that includes a robust measure of skewness in the determination of the whiskers. This results in a more accurate representation of the data and of possible outliers. Consequently, this adjusted boxplot can also be used as a fast and automatic outlier detection tool without making any parametric assumption about the distribution of the bulk of the data. Several examples and simulation results show the advantages of this new procedure.

1. Introduction

One of the most frequently used graphical techniques for analyzing a univariate data set is the *boxplot*, proposed by Tukey (1977). If $X_n = \{x_1, x_2, \ldots, x_n\}$ is a univariate data set, the boxplot is constructed by:

- putting a line at the height of the sample median Q_2;
- drawing a box from the first quartile Q_1 to the third quartile Q_3. The length of this box equals the interquartile range $\mathrm{IQR} = Q_3 - Q_1$, which is a robust measure of the scale;
- classifying all points outside the interval (the fence)

$$[Q_1 - 1.5\ \mathrm{IQR};\ Q_3 + 1.5\ \mathrm{IQR}] \tag{1}$$

  as *potential* outliers and marking them on the plot;
- drawing the whiskers as the lines that go from the ends of the box to the most remote points within the fence.

Note that quartiles can be defined in different ways. In the original proposal, so-called 'fourths' are used. As the finite-sample differences between these different definitions are not important in the rest of the paper, we will not specify a particular definition for the quartiles.

It is well recognized in the original work (e.g. Hoaglin et al. (1983) p. 39, 59–65) that observations outside the fence are not necessarily 'real' outliers that behave differently from the majority of the data. At thick tailed symmetric

* Corresponding author. Tel.: +32 (0)16 322023; fax: +32 (0)16 327998.
E-mail address: mia.hubert@wis.kuleuven.be (M. Hubert).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2007.11.008

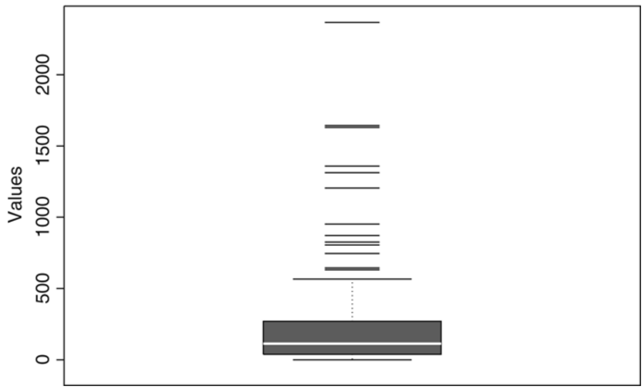

Fig. 1. Standard boxplot of the time intervals between coal mining disasters.

distributions, many regular observations will exceed the outlier cutoff values defined in (1), whereas data from thin tailed distributions will hardly exceed the fence. The same phenomenon applies to skewed distributions. If the data come for example from a χ^2_1-distribution, the probability to exceed the lower fence is zero, whereas it can be expected that 7.56% of the (regular) data exceed the upper fence. Similarly, we can expect a 7.76% upper exceedance probability at the lognormal distribution (with $\mu = 0$ and $\sigma = 1$). Note that for the normal distribution, the expected exceedance percentage is only 0.7%, i.e. 0.35% on both sides of the distribution.

As an example of real data, we consider the time intervals between coal mining disasters (Jarret, 1979). This data set contains 190 time intervals, measured in days, between explosions in coal mines from 15th March 1851 to 22nd March 1962 inclusive. From the boxplot of these data, shown in Fig. 1, it can be seen that the underlying distribution of the data set is skewed to the right. The median does not lie in the middle of the box and the lower whisker is much smaller than the upper whisker (because it is only drawn up to the smallest observation in the fence). Here, 6.84% of the observations exceed the upper whisker. Clearly, it would not be correct to classify them all as outliers.

It can be argued that this phenomenon is not a major problem in exploratory data analysis. On the contrary, the percentage of observations outside the fence give an additional graphical indication of the shape of the distribution (thick or thin tailed, degree of skewness). Unfortunately, many textbooks including manuals and help files of statistical software, do not distinguish between 'potential' outliers and 'real' outliers. They simply classify all points outside the fence as outliers, and thus implicitly assume that the regular data points are normally distributed. Consequently, many practitioners nowadays incorrectly use the boxplot as a tool for outlier detection.

In order to construct a graphical method that distinguishes better between regular observations and outliers, several modifications to the boxplot have been proposed in the literature.

Some authors (Kimber, 1990; Aucremanne et al., 2004) have adjusted the fence towards skewed data by use of the lower and upper semi-interquartile range $\text{SIQR}_L = Q_2 - Q_1$ and $\text{SIQR}_U = Q_3 - Q_2$, i.e. they define the fence as

$$[Q_1 - 3\text{SIQR}_L;\, Q_3 + 3\text{SIQR}_U].$$

Unfortunately, this SIQR boxplot does not sufficiently adjust itself for skewness. To illustrate this, Fig. 2 shows the SIQR boxplot of the coal mine data. We see that the upper whisker has only slightly been enlarged, and consequently many observations are still flagged as outliers. In Section 3 we present more examples which demonstrate the poor behavior of this SIQR boxplot.

In Carling (2000), the quartiles Q_1 and Q_3 in (1) are replaced with the median Q_2, and the constant 1.5 is changed by a size-dependent formula, in order to adjust the boxplot for sample size. Similar ideas have been proposed in Schwertman et al. (2004) and Schwertman and de Silva (2007). Their idea is to attain a pre-specified outside rate, i.e. the probability that an observation from the non-contaminated distribution exceeds the fence. These approaches are interesting, but they have the serious drawback that the resulting constants are not only size-dependent; they also depend on some characteristics of the uncontaminated distribution which might be difficult to estimate. In Carling (2000), rules are derived for the class of generalized lambda distributions, which are characterized by a location, a scale and two shape parameters. It is not clear how the method performs when these shape parameters first need to be

Fig. 2. SIQR boxplot of the time intervals between coal mining disasters.

estimated, nor how the outlier rule applies to other distributions. The methods presented in Schwertman et al. (2004) and Schwertman and de Silva (2007) have the limitation that the resulting constants are based on the expected value of the quartiles, and on some theoretical quantiles, which are in general not known in advance. Only for normal or almost normal data, the authors provide fixed values.

We propose an adjustment to the boxplot that can be applied to all distributions, even without finite moments. Moreover, we estimate the underlying skewness with a robust measure, to avoid masking the real outliers. The structure of this paper is as follows. In Section 2, we present our generalization of the boxplot that includes a robust measure of skewness in the determination of the whiskers. To construct this adjusted boxplot we will derive new outlier rules at the *population* level. To draw the boxplot at a particular data set, we then just need to plug in the finite-sample estimates. In Section 3, we apply the new boxplot to several real data sets, whereas in Section 4 a simulation study is performed at uncontaminated as well as contaminated data sets. Section 5 concludes and gives directions for future research.

2. Skewness adjustment to the boxplot

2.1. A robust measure of skewness

To measure the skewness of a univariate sample $\{x_1, \ldots, x_n\}$ from a continuous unimodal distribution F, we use the *medcouple* (MC), introduced in Brys et al. (2004). It is defined as

$$\mathrm{MC} = \underset{x_i \leqslant Q_2 \leqslant x_j}{\mathrm{med}} h(x_i, x_j)$$

with Q_2 the sample median and where for all $x_i \neq x_j$ the kernel function h is given by

$$h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}.$$

For the special case $x_i = Q_2 = x_j$ (which occurs with zero probability) a specific definition applies, see Brys et al. (2004) for the details. The medcouple thus equals the median of all $h(x_i, x_j)$ values for which $x_i \leqslant Q_2 \leqslant x_j$. Naively this takes $\mathrm{O}(n^2)$ time, but a fast $\mathrm{O}(n \log n)$ algorithm is available.

Note that this definition is inspired by the quartile skewness (QS), introduced in Bowley (1920) and Moors et al. (1996), defined as

$$\mathrm{QS} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}.$$

It clearly follows from this definition that the medcouple always lies between −1 and 1. A distribution that is skewed to the right has a positive value for the medcouple, whereas the MC becomes negative at a left skewed distribution. Finally, a symmetric distribution has a zero medcouple. As shown in Brys et al. (2004), this robust measure of skewness has a bounded influence function and a breakdown value of 25%, which means that at least 25% outliers are needed to obtain a medcouple of $+1$ (or -1). Besides, the MC turned out to be the overall winner when compared with two other robust skewness measures which are solely based on quantiles, namely the QS and the octile skewness (OS), given by

$$\mathrm{OS} = \frac{(Q_{0.875} - Q_2) - (Q_2 - Q_{0.125})}{Q_{0.875} - Q_{0.125}}$$

with $Q_{0.875}$ and $Q_{0.125}$ the 0.875 and 0.125 sample quantiles. The MC combines the strengths of OS and QS: it has the sensitivity of OS to detect skewness and the robustness of QS towards outliers.

2.2. *Incorporating skewness into the boxplot*

In order to make the standard boxplot skewness adjusted, we incorporate the medcouple into the definition of the whiskers. This can be done by introducing some functions h_l (MC) and h_u (MC) into the outlier cutoff values. Instead of using the fence

$$[Q_1 - 1.5\ \mathrm{IQR};\ Q_3 + 1.5\ \mathrm{IQR}],$$

we propose the boundaries of the interval to be defined as

$$[Q_1 - h_l(\mathrm{MC})\ \mathrm{IQR};\ Q_3 + h_u(\mathrm{MC})\ \mathrm{IQR}].$$

Additionally, we require that $h_l(0) = h_u(0) = 1.5$ in order to obtain the standard boxplot at symmetric distributions. As the medcouple is location and scale invariant, this interval is location and scale equivariant. Note that by using different functions h_l and h_u, we allow the fence to be asymmetric around the box, so that adjustment for skewness is indeed possible. Also note that at the population level, no distinction should be made between the whiskers and the fence. At finite samples on the other hand, the whiskers are drawn up to the most remote points before the fence.

Three different models have been studied, namely a

(1) *linear model*:

$$\begin{aligned} h_l(\mathrm{MC}) &= 1.5 + a\ \mathrm{MC} \\ h_u(\mathrm{MC}) &= 1.5 + b\ \mathrm{MC} \end{aligned} \tag{2}$$

(2) *quadratic model*:

$$\begin{aligned} h_l(\mathrm{MC}) &= 1.5 + a_1\ \mathrm{MC} + a_2\ \mathrm{MC}^2 \\ h_u(\mathrm{MC}) &= 1.5 + b_1\ \mathrm{MC} + b_2\ \mathrm{MC}^2 \end{aligned} \tag{3}$$

(3) *exponential model*:

$$\begin{aligned} h_l(\mathrm{MC}) &= 1.5\mathrm{e}^{a\ \mathrm{MC}} \\ h_u(\mathrm{MC}) &= 1.5\mathrm{e}^{b\ \mathrm{MC}} \end{aligned} \tag{4}$$

with $a, a_1, a_2, b, b_1, b_2 \in \mathbb{R}$. Note that each of these models is simple and contains only a few parameters. This is very important for exploratory data analysis.

2.3. *Determination of the constants*

In order to find good values for a, a_1, a_2, b, b_1 and b_2, we fit a whole range of distributions and try to define the fence such that the expected percentage of marked outliers is close to 0.7%, which coincides with the outlier rule of the standard boxplot at the normal distribution. At the linear model (2), this implies that the constants a and b should satisfy

$$\begin{cases} Q_1 - (1.5 + a\ \mathrm{MC})\ \mathrm{IQR} \approx Q_\alpha \\ Q_3 + (1.5 + b\ \mathrm{MC})\ \mathrm{IQR} \approx Q_\beta \end{cases}$$

where in general Q_p denotes the pth quantile of the distribution, $\alpha = 0.0035$ and $\beta = 0.9965$. The previous system can be rewritten as

$$\begin{cases} \dfrac{Q_1 - Q_\alpha}{\text{IQR}} - 1.5 \approx a \text{ MC} \\ \dfrac{Q_\beta - Q_3}{\text{IQR}} - 1.5 \approx b \text{ MC}. \end{cases}$$

Linear regression without intercept can then be used to obtain estimates of the parameters a and b. The parameter determination at the quadratic and at the exponential model is analogous to that of the linear case. At the exponential model we obtain the linear system

$$\begin{cases} \ln\left(\dfrac{2}{3}\dfrac{Q_1 - Q_\alpha}{\text{IQR}}\right) \approx a \text{ MC} \\ \ln\left(\dfrac{2}{3}\dfrac{Q_\beta - Q_3}{\text{IQR}}\right) \approx b \text{ MC} \end{cases}$$

so that again linear regression without intercept can be applied.

To derive the constants we used 12,605 distributions from the family of Γ, χ^2, F, Pareto and G_g-distributions (Hoaglin et al., 1985). More precisely, we used $\Gamma(\beta, \gamma)$ distributions with scale parameter $\beta = 0.1$ and shape parameter $\gamma \in [0.1; 10]$, χ^2_{df} distributions with $df \in [1; 30]$, F_{m_1,m_2} distributions with $(m_1, m_2) \in [1; 100] \times [1; 100]$, Pareto distributions $\text{Par}(\alpha, c)$ with $c = 1$ and $\alpha \in [0.1; 20]$, and G_g-distributions with $g \in [0; 1]$.

The parameters of the distributions were selected such that the medcouple did not exceed 0.6. Doing so, we retained a large collection of distributions that are not extremely skewed. It appeared that constructing one good and easy model that also includes the cases with MC > 0.6 is hard, hence we only concentrated on the more common distributions with moderate skewness. Note that we only considered symmetric and right skewed distributions, as the boundaries just need to be switched for left skewed distributions.

To obtain the population values of the medcouple and the quartiles at all these distributions, we generated 10,000 observations from each of them, and used their finite-sample estimates as the true values.

In Fig. 3(a) we show the fitted regression curves for the lower whisker, after applying LS regression for the linear, quadratic and exponential models, and based on the whole set of distributions we considered. For reasons of clarity, we have set on the vertical axis the response value of the exponential model, which is $\ln(\frac{2}{3}\frac{Q_1 - Q_\alpha}{IQR})$. Hence, only the exponential fit is presented by a straight line. Fig. 3(b) only displays the G_g-distributions (with the same fits superimposed). Fig. 4(a) and (b) show analogous results for the upper whisker.

2.4. The adjusted boxplot

From Figs. 3 and 4 it can be seen that the linear model completely fails to determine accurate lower whiskers, whereas the exponential and quadratic models perform much better. For the upper whiskers, the quadratic model gives less accurate estimates than the exponential model. Consequently, as the exponential model is appropriate for both the lower and the upper tails, we will use the *exponential model* in the definition of our adjusted boxplot, rather than the quadratic model. We also remark that the exponential model only includes one parameter (on each side), which makes it simpler than the quadratic model.

Although the exponential fit will produce an underestimate of Q_α (respectively Q_β) for some distributions, the same quantile will be overestimated for others. Consequently it gives a good compromise for the whole set of distributions we considered. If we would have a priori information of the distribution, for example we would know that it belongs to the class of G_g-distributions, it is clear from Fig. 3(b) and Fig. 4(b) that a more appropriate model could be constructed. But as the boxplot is often used as a first exploratory tool for further data analysis, we do not want to include assumptions about the data distribution.

To ease the model and for robustness reasons, we rounded off the estimated values of the exponential model $a = -3.79$ and $b = 3.87$ to $a = -4$ and $b = 3$. Note that rounding off the constants to the nearest *smaller* integer, yields a smaller fence and consequently a more robust model. The current values of a and b are constructed on outlier-free populations. But as robust estimators also show some bias at contaminated samples, the estimate of the medcouple might increase (with outliers on the upper side of the distribution) or decrease (with outliers on the lower

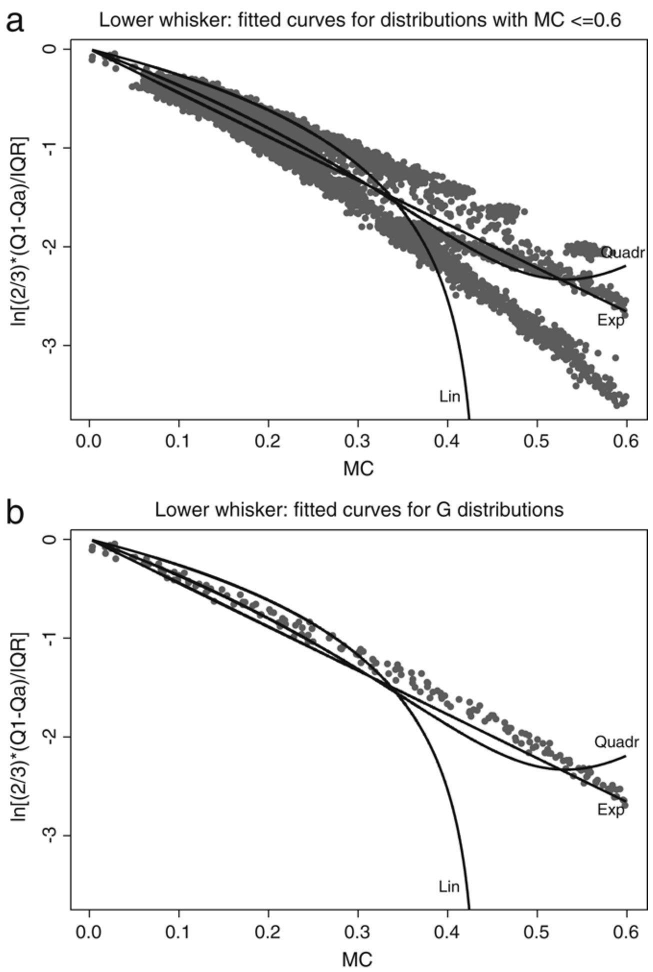

Fig. 3. Lower whisker: Regression curves for the linear, quadratic and exponential models.

side). To decrease the potential influence of outliers in our exponential model, we therefore prefer lower values of a and b. We also considered the fence given by $a = -3.5$ and $b = 3.5$, which has the advantage of being even simpler. As the resulting boxplot is then less robust to a higher percentage of contamination, we only recommend the use of that model if a small number of outliers is presumed (say, at most 5%).

To summarize, we thus can say that when $\mathrm{MC} \geqslant 0$, all observations outside the interval

$$[Q_1 - 1.5\mathrm{e}^{-4\,\mathrm{MC}}\,\mathrm{IQR};\ Q_3 + 1.5\mathrm{e}^{3\,\mathrm{MC}}\,\mathrm{IQR}] \tag{5}$$

will be marked as potential outliers. For $\mathrm{MC} < 0$, the interval becomes

$$[Q_1 - 1.5\mathrm{e}^{-3\,\mathrm{MC}}\,\mathrm{IQR};\ Q_3 + 1.5\mathrm{e}^{4\,\mathrm{MC}}\,\mathrm{IQR}].$$

Note that while this adjusted boxplot accounts for skewness, it does not yet account for tail heaviness. This could be done by including tail information of the distribution as well. We could for example try to construct a model which includes robust measures of left and right tails, such as those proposed in Brys et al. (2006), or by including a robust estimator of the tail index (e.g. Vandewalle et al. (2007)). We see however several disadvantages for such a procedure. First of all, the model would become more complex with more estimators and parameters. The robustness would decrease as the tail measures have a lower breakdown value, and the variability of the whisker's length would increase, due to the variability of the tail measures.

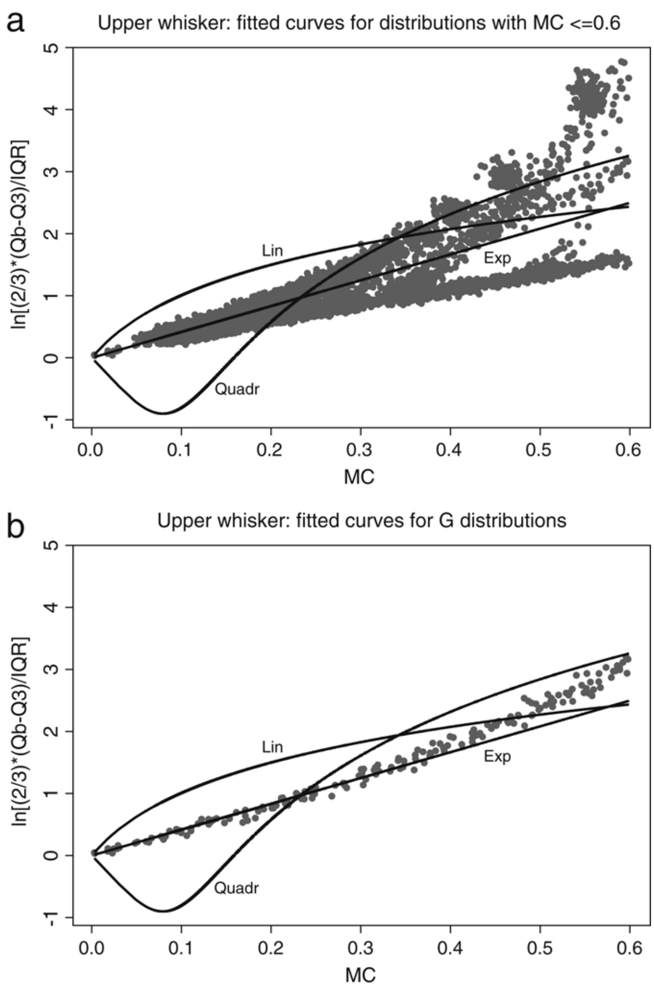

Fig. 4. Upper whisker: Regression curves for the linear, quadratic and exponential models.

3. Examples

3.1. Coal mine data

We recall the coal mine data from the introduction where we have illustrated that the standard and the SIQR boxplots flag many observations as potential outliers. Fig. 5 shows in addition the adjusted boxplot, obtained using our S-PLUS code. It can be seen that the adjusted boxplot yields a more accurate representation of the data. The upper whisker has become much larger and now reflects better the skewness of the underlying distribution. Besides, it causes fewer observations to be marked as upper outliers.

3.2. Condroz data

The Condroz data (Goegebeur et al., 2005) contain the pH-value and the Calcium (Ca) content in soil samples, collected from different communities of the Condroz region in Belgium. As in Vandewalle et al. (2007), we focus on the subset of 428 samples with a pH-value between 7.0 and 7.5.

From the normal quantile plot in Fig. 6 it can be seen that the distribution of Calcium is right skewed. This also follows from the MC value, which equals 0.16. Besides, we notice 6 upper outliers and 3 lower outliers in Fig. 6. In Vandewalle et al. (2007), they were also identified as such, based on a robust estimator of the tail index. The

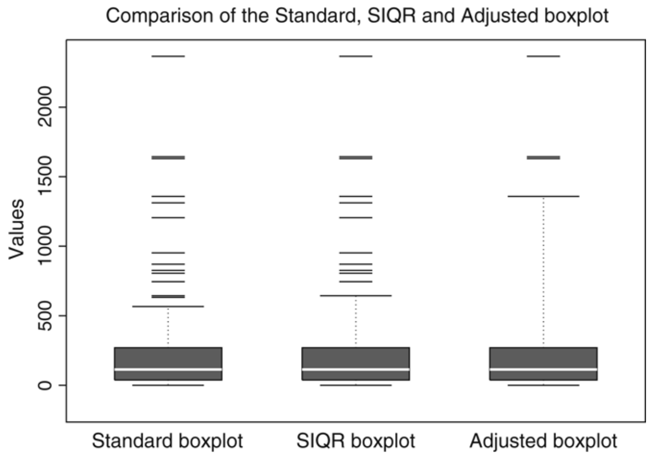

Fig. 5. Coal mine data: A comparison of the standard, the SIQR and the adjusted boxplots.

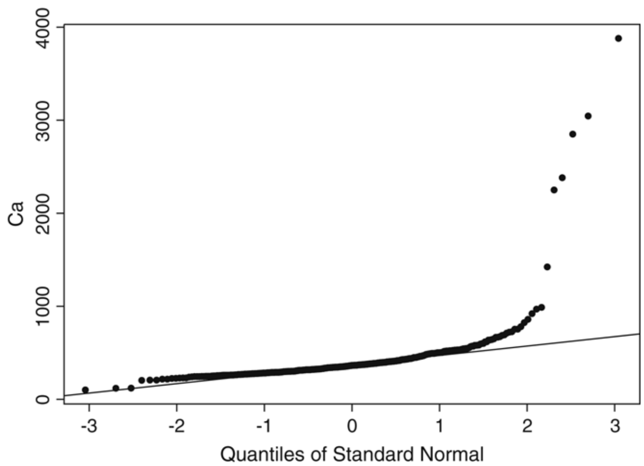

Fig. 6. Normal quantile plot of the Condroz data with pH between 7.0 and 7.5.

outliers appeared to be measurements from communities at the boundary of the Condroz region and hence, can be considered to be sampled from another distribution.

We see from the standard boxplot in Fig. 7(a) that a substantial number of observations exceed the upper whisker, leading to a black box in which the 'outlying' observations can no longer be recognized. Moreover, none of the cases is indicated as a lower outlier. Another visualization of the data is given in the index plot in Fig. 7(b). Full lines were drawn at the median, the first and third quartiles of the data. The dotted lines refer to the whiskers of the standard boxplot. We see that 20 of the regular data points are marked upper outliers. The SIQR boxplot is hardly different. The upper whisker is almost identical, whereas the lower whisker only flags the smallest observation.

The adjusted boxplot, also shown in Fig. 7(a), has a longer upper whisker and a shorter lower whisker. The skewness of the underlying distribution is more pronounced and the shorter left tail is better reflected. Now, less observations exceed the upper cutoff, whereas the three smallest cases are marked lower outliers. Note that only two marks are visible, as two of the three smallest observations almost coincide as can be seen on the index plot in Fig. 7(b). Here, the dashed lines refer to the whiskers of the adjusted boxplot.

As the data are highly skewed, it is common practice to apply a log transformation to the data to make them more symmetric. The resulting boxplots are shown in Fig. 8(a).

We see that all boxplots do not differ very much as the MC of the log-transformed Calcium values equals only 0.044. We also notice that the standard and the SIQR boxplots applied to the transformed data now show the same

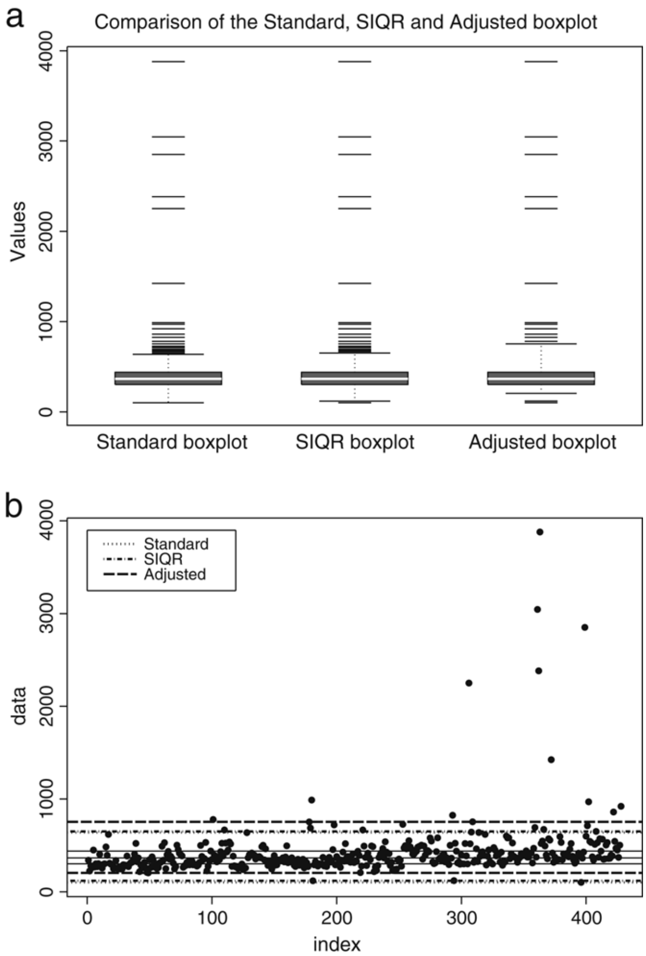

Fig. 7. (a) Standard, SIQR and adjusted boxplots of the Condroz data. (b) Plot of the Ca measurements versus their index. Full lines were drawn at the median, the first and third quartiles. Dashed lines were used to indicate the boundaries of the adjusted boxplot. The dotted lines refer to the boundaries of the standard boxplot. The dashed–dotted lines correspond to the SIQR boxplot.

outliers as does the adjusted boxplot of the raw data. From this example, one could conclude that the adjusted boxplot is not needed at all and that alternatively, first a transformation could be applied, after which the whiskers could be retransformed to the original unit scale. This approach would certainly work out well in many examples, but it has the drawback that first an appropriate symmetrizing transformation has to be found. This additional data analysis might be a problem for less experienced users, or for applications that require an automatic outlier detection procedure.

For illustrative purposes, Fig. 8(b) shows boxplot-type figures, obtained by taking the log transformation of the boxplots of Fig. 7(a). As the whiskers are not equivariant to monotone transformations, the resulting figures are different from the boxplots in Fig. 8(a).

3.3. Air data

In Fig. 8(a) we have already noticed that the SIQR and the adjusted boxplots are very similar to the standard boxplot when data are nearly symmetric. Here, we consider another example, the wind speed variable from the air data (Chambers and Hastie, 1992), measuring the wind speed (in miles per hour) for 111 consecutive days. The resulting boxplots are depicted in Fig. 9.

As MC = 0.012, we see that all three boxplots are equal. Note that the small value of MC slightly effects the fence, as defined in (5), but does not effect the whiskers here as they are drawn to the largest (smallest) non-outlier. This

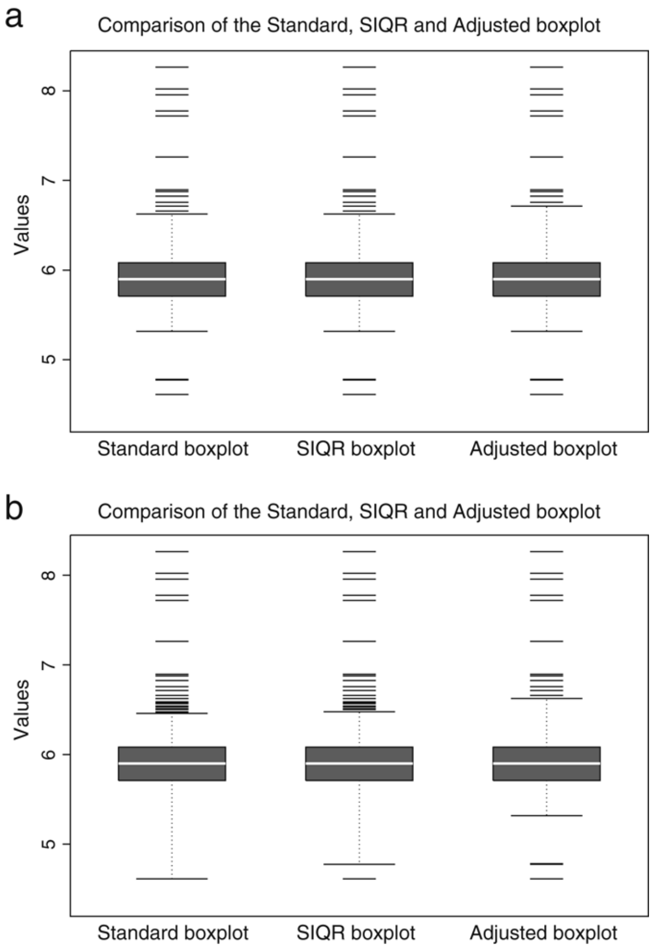

Fig. 8. (a) Standard, SIQR and adjusted boxplots of the log-transformed Condroz data; (b) Log transformations of the boxplots of Fig. 7(a).

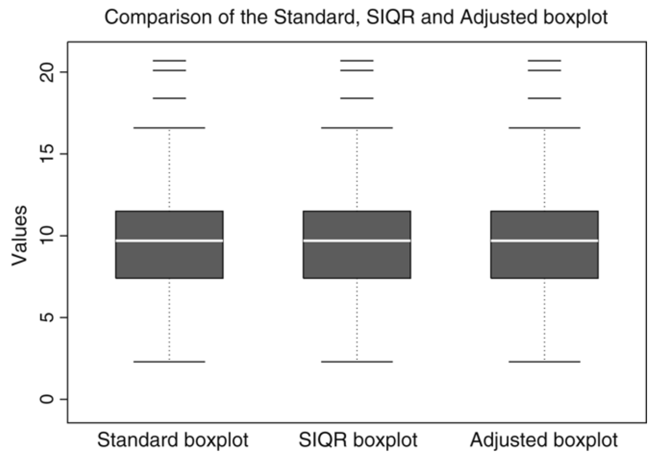

Fig. 9. Air data: A comparison of the standard, the SIQR and the adjusted boxplots.

example thus again illustrates that we can consider the adjusted boxplot as a generalization of the standard boxplot towards skewed distributions.

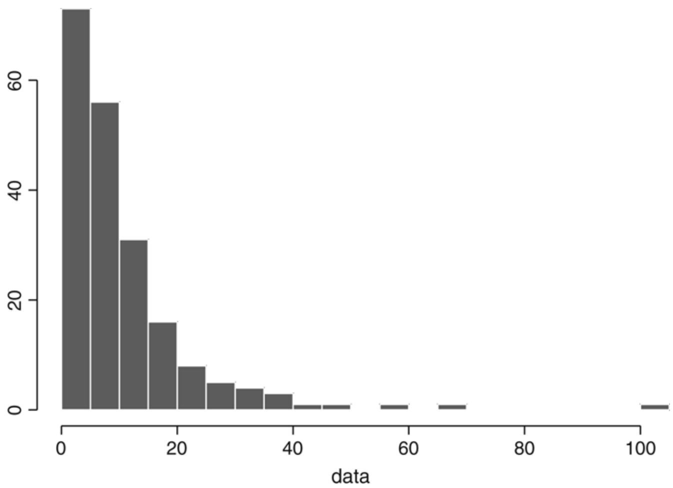

Fig. 10. Histogram of the 'Length of Stay' data.

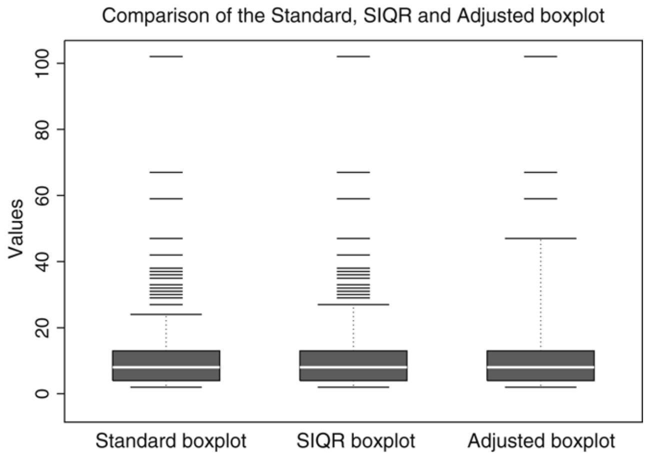

Fig. 11. The standard, the SIQR and the adjusted boxplots of the 'Length of Stay' data.

3.4. *Length of stay data*

Our next example is concerned with the data of 201 patients, who stayed in the University Hospital of Lausanne in the year 2000. The data are kindly provided by A. Marazzi (Institute of Social and Preventive Medicine, Lausanne). One of the main objectives is to estimate and predict the total resource consumption of this group of patients (Ruffieux et al., 2000). For this purpose, one can focus on the variable 'length of stay' (LOS) in days, which is an easily available indicator of hospital activity and is used for various purposes, such as management of hospital care, quality control, appropriateness of hospital use and hospital planning. The most natural way to compute an estimate of the expected LOS, is to use the arithmetic mean. However, the underlying distribution of the LOS data has two features, which make the use of this simple statistic questionable. First of all, the distribution of the LOS data is skew distributed to the right, as can be seen on the histogram in Fig. 10. Besides, three observations are clearly isolated from the majority of the data and may therefore be regarded as outlying values for the length of stay.

The adjusted boxplot of the LOS data is shown in Fig. 11. Due to the upward shift of the upper whisker, only the three largest observations are flagged as outliers. At the standard and SIQR boxplots on the other hand, many (17 and 14) observations are detected as potential upper outliers. This illustrates again that only the adjusted boxplot accounts sufficiently for skewness. Consequently, one could use the adjusted boxplot as a trimming rule which accounts for skewness. Taking the mean of all observations within the lower and upper fence of the adjusted boxplot then gives a more realistic estimate of the expected LOS.

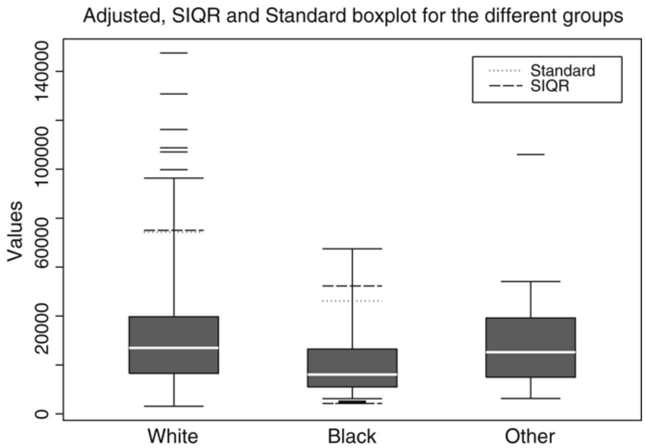

Fig. 12. Boxplots of the different groups in the CES data. The adjusted boxplots are plotted, with dotted lines at the height of the standard whiskers and dashed lines at the height of the SIQR whiskers superimposed.

3.5. Consumer expenditure survey data

Boxplots are often used to compare the distribution of a variable within several groups. Our adjusted boxplot can be used for this purpose as well. To illustrate, we consider data derived from the Consumer Expenditure Survey (CES) of 1995, collected by the Bureau of Labor Statistics, U.S. Department of Labor, and available at http://econ.lse.ac.uk/courses/ec220/G/iedata/ces/.

In this paper we focus on the variables ‘exp’ and ‘refrace’, which represent the total household expenditure and the ethnicity of the reference person respectively, which can be either ‘white’, ‘black’ or ‘other’ (e.g. American Indian, Aleut, Eskimo, Asian, Pacific Islander etc.).

Applying the adjusted boxplot to each of the ethnicity subgroups, yields Fig. 12. On the adjusted boxplots, we have superimposed dotted and dashed lines which refer to the whiskers of the standard and the SIQR boxplot. At the ‘white’ and ‘black’ groups, the upper whisker has been shifted upwards, yielding less upper outliers and emphasizing the skewness of the underlying distribution. Furthermore, at the ‘black’ group the lower whisker has shifted too, now better reflecting the shorter left tail. Finally, all three boxplots give the same result at the ‘other’ group, from which we can conclude that the majority of the observations in this subgroup come from a symmetric distribution.

4. Simulation study

To compare more thoroughly our adjusted boxplot with the standard one, a simulation study has been done. We focussed on several right skewed distributions, such as the normal, G_g, χ^2, Γ, Pareto and F-distributions. More detailed information can be found in Table 1.

4.1. Performance at uncontaminated data sets

For each of the considered distributions, we generated 100 samples of size 1000 and computed the percentage of lower and upper outliers (observations that fall outside the boundaries defined by (1) and (5) resp.). Note that we consider large samples as we want to have an idea of the *expected* proportions of exceedance when data are contaminated. The average percentages of lower and upper outliers for the standard boxplot (crosses) and for the adjusted boxplot (circles) are reported in Fig. 13. Fig. 13(a) gives the result for the upper tail, whereas Fig. 13(b) concentrates on the lower tail.

At the normal distribution (distribution 1), we notice that, slightly remarkable, the adjusted boxplot classifies more observations as outliers than before. This is because the finite-sample medcouple is not exactly zero, hence the adjusted whiskers are slightly different from the original ones. Though, the discrepancy is rather small (the total percentage of outliers, classified by the adjusted boxplot is about 0.96% as opposed to about 0.7% at the standard boxplot).

Table 1
The 20 different distributions that are used in the simulation study

No.	Distribution	No.	Distribution
1	N(0, 1)	11	$\Gamma(0.1, 0.75)$
2	$G_{0.1}$	12	$\Gamma(0.1, 1.25)$
3	$G_{0.25}$	13	$\Gamma(0.1, 5)$
4	$G_{0.5}$	14	Pareto(1, 1)
5	G_1	15	Pareto(3, 1)
6	G_3	16	Pareto(6, 1)
7	χ_1^2	17	F(90, 10)
8	χ_5^2	18	F(10, 10)
9	χ_{20}^2	19	F(10, 90)
10	$\Gamma(0.1, 0.5)$	20	F(80, 80)

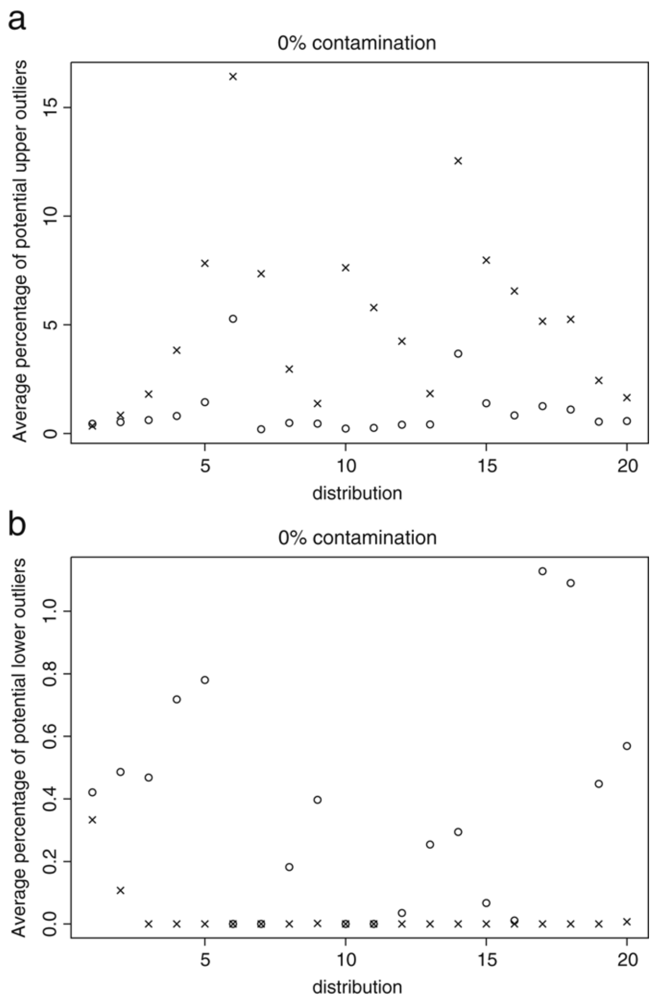

Fig. 13. For different distributions, the average percentage of data points exceeding (a) the upper whisker and (b) the lower whisker. Results for the standard boxplot (crosses) and the adjusted boxplot (circles).

Much more pronounced differences can be seen at the skewed distributions. At the χ_5^2 distribution (distribution 8) for example, the average number of flagged outliers is less than 0.6% at the adjusted boxplot as opposed to more

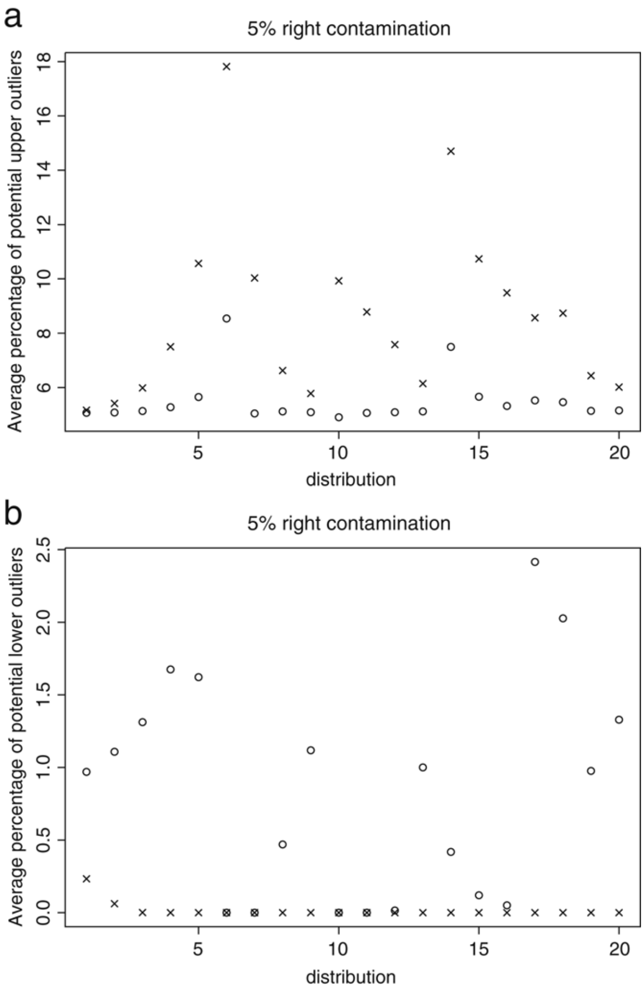

Fig. 14. For different distributions, the average percentage of data points exceeding (a) the upper whisker and (b) the lower whisker, with 5% contamination in the *upper* tail. Results for the standard boxplot (crosses) and the adjusted boxplot (circles).

than 2.7% at the standard boxplot. The adjusted boxplot of the Pareto(3, 1) distribution (distribution 15) now yields on average at most 1.48% outliers, whereas on average more than 8% of the observations are marked as outliers at the standard boxplot. Note that the G_3-distribution (distribution 6) was not used in the calibration of the exponential model, but also here we see that our model highlights much fewer outliers than before.

As we see, the improvements differ somewhat over the distributions. The overall improvement is mainly due to a substantial increase of the upper whiskers. This causes the adjusted boxplot to flag less observations as upper outliers than before. Besides, due to the exponential factor we added, the lower whiskers are shorter, which allows one to detect better possible lower outliers.

4.2. Performance at contaminated data sets

To get an idea of the robustness of the skewness-adjusted boxplot, we looked at its performance when applied to contaminated data sets. We generated again 100 samples of size 1000 for each distribution, but now replaced 5% of the data by upper (respectively lower) outliers, coming from a normal distribution. The results with 5% of upper contamination are depicted in Fig. 14.

Fig. 14(a) clearly shows that the adjusted boxplot detects the expected 5% of upper outliers, without marking too many observations as potential upper outliers. This is not always the case when we apply the standard boxplot. For example at the G_1-distribution (distribution 5, which is in fact a lognormal distribution), on average more than 10% of

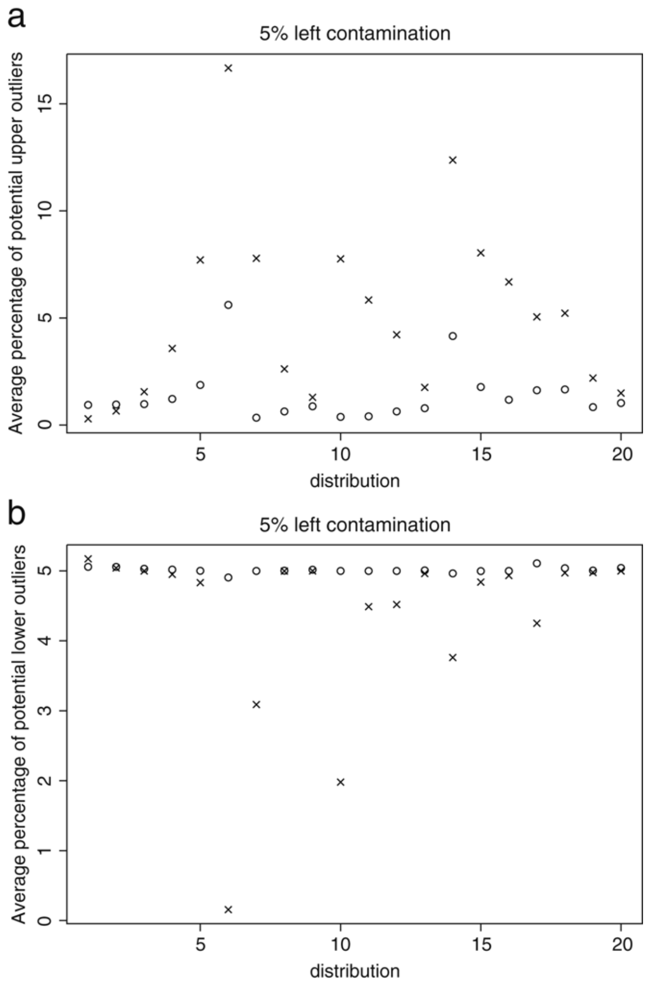

Fig. 15. For different distributions, the average percentage of data points exceeding (a) the upper whisker and (b) the lower whisker, with 5% contamination in the *lower* tail. Results for the standard boxplot (crosses) and the adjusted boxplot (circles).

the observations were marked as upper outliers. Furthermore, the adjusted boxplot gives again a more balanced view of potential lower outliers, which can be seen from Fig. 14(b).

The results of the simulation study with 5% lower contamination are reported in Fig. 15. Fig. 15(b) clearly shows that the adjusted boxplot succeeds in marking the 5% lower outliers that were added. This is not always the case at the standard boxplot. For example at $\Gamma(0.1; 0.5)$ (distribution 10), on average less than 2% of the lower outliers were detected. Also in the upper tail, the adjusted boxplot performs well, as can be seen from Fig. 15(a). For most distributions the percentage of observations that have been marked as potential upper outliers are rather small. Only at the G_3-distribution (distribution 6) more than 5% of upper outliers were detected. This effect is due to the extreme skewness of the underlying distribution and was already visible at the uncontaminated data shown in Fig. 13(a).

5. Discussion and conclusion

The boxplot is a frequently used graphical tool for analyzing a univariate data set. Unfortunately, when drawing the boxplot of a skewed distribution, many regular observations are typically flagged as potential outliers. The SIQR boxplot does not adequately solve this problem. Therefore, we have presented an adjustment of the boxplot, by modifying the whiskers such that the skewness is sufficiently taken into account. To measure skewness of the data, the medcouple has been used and different models for generalizing the standard boxplot have been studied. The overall winner seems to be an exponential model.

The results on real and simulated data indicate that by using the adjusted boxplot at skewed distributions, a better distinction is made between regular observations and outliers. This makes the adjusted boxplot an interesting and fast tool for automatic outlier detection, without making any assumption about the distribution of the data.

We have implemented the adjusted boxplot in S-PLUS and Matlab, the latter being part of the LIBRA toolbox (Verboven and Hubert, 2005). Our functions are available from http://wis.kuleuven.be/stat/robust. Also an R implementation will become available as part of the *robustbase* package.

While in this paper we focussed on the detection of univariate outliers, the idea of the adjusted boxplot has been applied to detect multivariate outliers in the context of independent component analysis (Brys et al., 2005). This methodology can be extended to find outlying observations at multivariate skewed distributions (Hubert and Van der Veeken, 2007). This leads to an extension of the adjusted boxplot for bivariate data, such as the bagplot (Rousseeuw et al., 1999). Also skewness-adjusted modifications of the robust PCA method ROBPCA (Hubert et al., 2005) have been studied (Hubert et al., 2007).

References

Aucremanne, L., Brys, G., Hubert, M., Rousseeuw, P.J., Struyf, A., 2004. A study of belgian inflation, relative prices and nominal rigidities using new robust measures of skewness and tail weight. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology. Birkhauser, Basel, pp. 13–25.

Bowley, A.L., 1920. Elements of Statistics. Charles Scribner's Sons, New York.

Brys, G., Hubert, M., Rousseeuw, P.J., 2005. A robustification of independent component analysis. Journal of Chemometrics 19, 364–375.

Brys, G., Hubert, M., Struyf, A., 2004. A robust measure of skewness. Journal of Computational and Graphical Statistics 13, 996–1017.

Brys, G., Hubert, M., Struyf, A., 2006. Robust measures of tail weight. Computational Statistics and Data Analysis 50, 733–759.

Carling, K., 2000. Resistant outlier rules and the non-Gaussian case. Computational Statistics and Data Analysis 33, 249–258.

Chambers, J.M., Hastie, T.J., 1992. Statistical Models in S. Wadsworth and Brooks, Pacific Grove, pp. 348 – 351.

Goegebeur, Y., Planchon, V., Beirlant, J., Oger, R., 2005. Quality assessment of pedochemical data using extreme value methodology. Journal of Applied Science 5, 1092–1102.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1983. Understanding Robust and Exploratory Data Analysis. Wiley, New York, pp. 58–77.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1985. Exploring Data Tables, Trends and Shapes. Wiley, New York, pp. 463–478.

Hubert, M., Rousseeuw, P.J., Vanden Branden, K., 2005. ROBPCA: A new approach to robust principal components analysis. Technometrics 47, 64–79.

Hubert, M., Van der Veeken, S., 2007. Outlier detection for skewed data. Journal of Chemometrics (in press).

Hubert, M., Rousseeuw, P.J., Verdonck, T., 2007. Robust PCA for skewed data (submitted for publication).

Jarret, R.G., 1979. A note on the intervals between coal mining disasters. Biometrika 66, 191–193.

Kimber, A.C., 1990. Exploratory data analysis for possibly censored data from skewed distributions. Applied Statistics 39, 21–30.

Moors, J.J.A., Wagemakers, R.Th.A., Coenen, V.M.J., Heuts, R.M.J., Janssens, M.J.B.T., 1996. Characterizing systems of distributions by quantile measures. Statistica Neerlandica 50, 417–430.

Rousseeuw, P.J., Ruts, I., Tukey, J.W., 1999. The Bagplot: A bivariate boxplot. The American Statistician 53, 382–387.

Ruffieux, C., Paccaud, F., Marazzi, A., 2000. Comparing rules for truncating hospital length of stay. Casemix Quaterly 2 (1).

Schwertman, N.C., Owens, M.A., Adnan, R., 2004. A simple more general boxplot method for identifying outliers. Computational Statistics and Data Analysis 47, 165–174.

Schwertman, N.C., de Silva, R., 2007. Identifying outliers with sequential fences. Computational Statistics and Data Analysis 51, 3800–3810.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts, pp. 39–49.

Vandewalle, B., Beirlant, J., Christmann, A., Hubert, M., 2007. A robust estimator for the tail index of pareto-type distributions. Computational Statistics and Data Analysis 51, 6252–6268.

Verboven, S., Hubert, M., 2005. LIBRA: A MATLAB library for robust analysis. Chemometrics and Intelligent Laboratory Systems 75, 127–136.