ROBUST VERSIONS OF THE TUKEY BOXPLOT WITH THEIR APPLICATION TO DETECTION OF OUTLIERS

Georgy Shevlyakov[⋆] *Kliton Andrea*[⋆] *Lakshminarayan Choudur*[†]
Pavel Smirnov[⋆] *Alexander Ulanov*[‡] *Natalia Vassilieva*[‡]

[⋆] Department of Applied Mathematics, St. Petersburg State Polytechnic University, 29 Polytechnicheskaya St. Petersburg, 195251, RUSSIA
[†] Hewlett-Packard Labs, 1501 Page Mill Rd. Palo Alto, CA 94304, USA
[‡] Hewlett-Packard Labs, 1 Artillerijskaya St. Petersburg, 191104, RUSSIA

ABSTRACT

The need for fast on-line algorithms to analyze high data-rate measurements is a vital element in production settings. Given the ever-increasing number of data sources coupled with increasing complexity of applications, and workload patterns, anomaly detection methods should be light-weight and must operate in real-time. In many modern applications, data arrive in a streaming fashion. Therefore, the underlying assumption of classical methods that the data is a sample from a stable distribution is not valid, and Gaussian and non-parametric based methods such as the control chart and boxplot are inadequate. Streaming data is an ever-changing superposition of distributions. Detection of such changes in real-time is one of the fundamental challenges. We propose low-complexity robust modifications to the conventional Tukey boxplot based on fast highly efficient robust estimates of scale. Results using synthetic as well as real-world data show that our methods outperform the Tukey boxplot and methods based on Gaussian limits.

Keywords: robustness, boxplot, outlier

1. INTRODUCTION

Robust statistics originates from the pioneering contributions of Tukey [1], Huber [2], and Hampel [3]. Roughly speaking, robustness means stability of statistical inferences under departures from the accepted distribution models. Although robust statistical procedures involve highly refined asymptotic tools, they exhibit satisfactory behavior within small samples and therefore are quite useful in real-world applications.

Concurrently, in parallel with robust statistics, practical methods for analyzing data evolved known as *Exploratory Data Analysis* (EDA). These days it is more popularly alluded to as *data mining*. A significant feature of EDA is that it does not assume an underlying probability distribution for the data which is typical in classical statistical methods and therefore is flexible in practical settings.

This paper presents new results in robust data analysis technologies, providing alternatives to the boxplot technique. The univariate Tukey boxplot summarizes the characteristics of a data distribution allowing for a quick visual inspection of streams of data over windows. Despite being a simple data analysis tool, it concisely summarizes information about the location, scale, asymmetry, tails, and outliers in the data distribution. In our study, we concentrate on visualization of distribution tails and on detection of outliers in the data.

The remainder of the paper is organized as follows. In Section 2, the state of the art in boxplot techniques is presented. In Section 3, two new robust versions of the Tukey boxplot based on the highly efficient robust estimates of scale are proposed. In Section 4, several new rules for detection of outliers based on the proposed robust boxplots are introduced and examined on the contaminated Gaussian data as well as on real-life data. In Section 5, conclusions are drawn.

2. STATE OF THE ART

A univariate boxplot [4] is specified by five parameters: the two extremes, the upper UQ (75th percentile) and lower LQ (25th percentile) quartiles and the median (50th percentile). The lower and upper extremes of a boxplot are defined as

$$x_L = \max\left\{x_{(1)},\, LQ - \frac{3}{2}\,IQR\right\},$$

$$x_U = \min\left\{x_{(n)},\, UQ + \frac{3}{2}\,IQR\right\}. \qquad (1)$$

Different streams of data are compared via their respective boxplots in a quick and convenient way. It is a common practice to identify points which are located beyond the extremes (maximum and minimum) as outliers, and mark them in the corresponding boxplots.

Many modifications have been proposed as improvement to the standard boxplot. Notable among them are [5] which displays confidence intervals around the median. This is especially useful to distinguish difference between medians of

different data windows. Another variant is due to [6] which incorporates the density information in addition to the five descriptive parameters in the the standard boxplot. Other solutions include a histplot, in which the underlying probability density function is estimated at the median and the two quartiles with its modification vaseplot [7]. And lastly, a box-percentile plot, where information regarding the empirical cumulative data distribution is used in conjunction with the boxplot [8] and a violin plot, which is a combination of a boxplot with a box-percentile [9].

The alternatives to the classical Tukey boxplot seek to exploit additional information about the underlying data distribution. However, it comes at the cost of computational complexity. In this paper, we propose a modification by incorporating a highly efficient robust estimate of scale, while maintaining low complexity structure of the Tukey boxplot.

3. NOVEL BOXPLOTS BASED ON FAST HIGHLY EFFICIENT ROBUST ESTIMATES OF SCALE

3.1. The *MAD*-Modification of the Tukey Boxplot

Although the Tukey boxplot is a widely used tool for anomaly detection, it can be modified for better performance. For estimating the width of the central part of a data distribution, (the box part of the boxplot), the sample interquartile range (IQR) can hardly be improved, since it is a natural choice for representation of the half of the data distribution mass. The remaining possibilities of improving most refer to the choice of robust estimates of scale used for visualization of tail areas and anomalies in the data (the boxplot lower and upper extremes). In this case, the sample interquartile range IQR as a robust estimate of scale is not the best choice as its efficiency and robustness can be considerably improved [10]. Efficiency is the ratio of variances of a baseline and proposed estimates. The baseline typically is the variance under normal assumptions as it is still successfully and ubiquitously used in practice of data analysis [11]. Robustness of an estimate is measured by the gross error breakdown point $0 \leq \varepsilon^* \leq 0.5$, which is the largest fraction of gross errors (anomalies) in the data that still keeps the bias of an estimator bounded [10]. For instance, the breakdown point of the sample standard deviation is $\varepsilon^* = 0$ – it means that this estimate is not robust at all, whereas the interquartile range has the moderate value of the breakdown point $\varepsilon^* = 0.25$ and the median absolute deviation $MAD_n\, x = med_i\, |x_i - med\, x|$ has the maximal value $\varepsilon^* = 0.5$.

Since the interquartile range is less resistant to outliers than the median absolute deviation $MAD_n\, x$, a more robust rule for constructing the boxplot extremes can be given by

$$x_L = \max\{x_{(1)},\, LQ - k_{MAD}\, MAD_n\},$$

$$x_U = \min\{x_{(n)},\, UQ + k_{MAD}\, MAD_n\}, \qquad (2)$$

where k_{MAD} is a threshold coefficient chosen from additional considerations.

3.2. Fast Highly Efficient Robust Estimates of Scale

Although the median absolute deviation MAD_n is a highly robust estimate of scale with the maximal value of the breakdown point $\varepsilon^* = 0.5$, its efficiency is only 0.37 at the normal distribution. In [12], a highly efficient robust estimate of scale Q_n has been proposed: it is close to the lower quartile of the absolute pairwise differences $|x_i - x_j|$, and it has the maximal breakdown point 0.5 as for MAD_n but much higher efficiency 0.82. The drawback of this estimate is its low computation speed. The computation of Q_n requires an order of greater time than of MAD_n.

In [13], an M-estimate of scale denoted by S_n^* whose influence function is approximately equal to the influence function of the estimate Q_n is proposed

$$\sum_{i=1}^{n} \chi\left(x_i / S_n^*\right) = 0, \qquad (3)$$

where the score $\chi(x) = 1/\sqrt{\pi}\left(1 - \sqrt{2}\exp(-x^2/2)\right)$.

The breakdown point of S_n^* is 0.293 with the corresponding efficiency equal to 0.808. The breakdown point of S_n^* is further improved by finding a one-step ahead M-estimate of scale. The one-step ahead estimates are obtained by solving (3) iteratively. The iterative process proceeds by using the highly robust MAD estimate as an initial estimate of scale

$$FQ_n = 1.483\, MAD_n \left(1 - \frac{Z_0 - n/\sqrt{2}}{Z_2}\right), \qquad (4)$$

where

$$Z_k = \sum_{i=1}^{n} u_i^k e^{-u_i^2/2},\ u_i = \frac{x_i - med\, x}{1.483\, MAD},$$

$$k = 0, 2; \quad i = 1, \ldots, n.$$

The FQ_n statistic is consistent under the normality assumption when it is multiplied by the constant 1.483.

The general property of a one-step M-estimate is that it has both the same asymptotic efficiency as the estimate defined by the implicit estimating equation (3) and the same breakdown point as the initial estimate [10]: in our case the efficiency and breakdown point of FQ_n are equal to 0.81 and to 0.5, respectively.

In Table 1, we present numerical results related to the performance of the proposed robust estimates under the standard normal distribution on large samples $n = 1000$; the results on small samples $n = 20$ are similar to those on large samples. The Monte Carlo experiment is based on $50,000$ trials.

The columns in Table 1 are: the average denoted by Ave, the standardized variance denoted by Var, the efficiency denoted by Eff, breakdown points denoted by ε^*, and computation times on Intel Core i7 at approximately 2.8 GHz denoted by Time (the best values are boldfaced).

Table 1. Performance of Robust Estimates at the Standard Normal, $n = 1000$

	Ave	Var	Eff	ε^*	Time (ms)
MAD_n	**0.999**	1.364	0.37	**0.50**	**0.17**
Q_n	1.004	**0.605**	**0.82**	**0.50**	1.02
S_n^*	**0.999**	0.624	0.81	0.29	0.23
FQ_n	1.005	0.630	0.81	**0.50**	0.20

The values of estimate efficiencies and breakdown points in Table 1 are obtained analytically, using standard asymptotic techniques [14]. The computation times on small samples ($n = 20$) are approximately the same for all the competitors. From Table 1 it follows that for large samples, S_n^* and FQ_n dominate over Q_n in computation time. Finally, we recommend the estimate FQ_n with the high efficiency 0.81, the maximal breakdown point 0.50 and the much faster computation time $0.20\,ms$ as compared to the Q_n computation time $1.02\,ms$.

3.3. The FQ-Modification of the Tukey Boxplot

Based on the highly efficient robust estimate FQ_n of scale, we propose a new rule for the boxplot extremes defined as

$$x_L = \max\{x_{(1)},\, LQ - k_{FQ}\, FQ_n\},$$
$$x_U = \min\{x_{(n)},\, UQ + k_{FQ}\, FQ_n\}. \qquad (5)$$

4. PERFORMANCE EVALUATION

4.1. Detection of Outliers

The proposed robust boxplots as alternatives to the Tukey boxplot, differ in estimating tail areas and consequently in detecting outliers. Therefore, we undertake a comparison study involving the robust and Tukey versions relative to detection of outliers.

In statistics, an outlier [15] is an observation that is numerically distant from the rest of the data. Outliers can occur by chance in any distribution, but they are often indicative either of a measurement error or that the underlying population has a heavy-tailed distribution. In the former case, these anomalous observations can have occurred due to transcription errors or measurement system malfunctions. In the latter case, they indicate that the underlying distribution may be set by large kurtosis. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, namely, a combination of "good data" and "bad data". This scenario can be modeled by the Tukey gross error model [1].

Within the classical approach to detection of outliers, an observation x is taken as an outlier if $|x - \overline{x}|/S > k_\alpha$, where $\overline{x}$ is the sample mean, S is the standard deviation, and the threshold k_α is determined from the given false alarm rate (the probability of Type I error) at the normal distribution. This rule is the classical Grubbs test [16].

In this paper, we most consider the boxplot (BP) detection tests of the form: an observation x is regarded as an outlier if $x < x_L$ or $x > x_U$, where x_L and x_U are the lower and upper extremes, respectively. In this setting, these thresholds also depend on a free parameter k, which is chosen from the false alarm rate $\alpha = 0.1$.

4.2. Data Generation and Performance Measure

The Monte Carlo experiments are conducted by generating 300 samples of observations from the mixture of normal distributions (Tukey's gross error model) [1]

$$f(x) = (1 - \varepsilon)N(x; 0, 1) + \varepsilon N(x; \mu, s), \qquad (6)$$

where $0 \le \varepsilon < 1$ is the probability of outliers (the fraction of contamination) in the data and $s > 1$ is their scale.

For evaluating the performance of different tests, the sensitivity (SE) and specificity (SP) measures are used in the comparative study. Note that the sensitivity is nothing but the test power, and the specificity is just unit minus the false alarm probability. These two metrics are combined into a single measure, namely, the harmonic mean between SE and SP: H-mean$= 2\,SE\,SP/(SE+SP)$. The introduced H-mean is an analog to the widely used in IR studies F-measure, which is the harmonic mean between the recall (R) and the precision (P): $F = 2\,R\,P/(R+P)$. The H-mean can be naturally used for performance evaluation in detection of outliers, since in this case, tests with the different values of the false alarm probability can be effectively compared. In our study, we just have this case: the false alarm rates for the Tukey and modified boxplots are $\alpha = 0.06$ and $\alpha = 0.1$, respectively.

4.3. Scale and Shift Contamination

The results of Monte Carlo experiment are given in Tables 2-3 with the best performing statistics represented in boldface.

Table 2. H-means for detection tests under scale contamination: $\mu = 0$, $s = 3$.

$\varepsilon = 0.1$	20	50	100	1000	10000
Tukey BP	0.64	**0.72**	0.72	0.72	0.72
MAD-BP	**0.67**	**0.72**	**0.73**	**0.73**	**0.73**
FQ-BP	0.66	**0.72**	0.72	0.72	**0.73**
Grubbs test	0.17	0.29	0.30	0.30	0.30

From Tables 2-3 it follows that both under scale and shift contamination, the performances of boxplot tests, generally,

Table 3. H-means for detection tests under shift contamination: $\mu = 3$, $k = 1$.

$\varepsilon = 0.1$	20	50	100	1000	10000
Tukey BP	**0.75**	0.79	0.80	0.80	0.80
MAD-BP	0.73	**0.80**	0.80	0.80	0.80
FQ-BP	0.73	0.79	**0.81**	**0.81**	**0.81**
Grubbs test	0.32	0.39	0.40	0.39	0.39

are close to each other, and all of them outperform the classical Grubbs test, which is catastrophically bad. This effect can be explained by non-robustness of the Grubbs test forming statistics, the sample mean and standard deviation, under contamination.

Further, the robust MAD and FQ versions are slightly but systematically better than the Tukey boxplot test. Similar results are obtained for the gross error models with $\varepsilon = 0.2$.

Table 4. H-means for detection tests under shift contamination with the different values of ε: $\mu = 3$, $s = 1$, $n = 100$.

ε	0.05	0.10	0.20	0.30	0.40	0.50
Tukey BP	0.63	0.62	0.59	0.55	0.51	0.43
MAD-BP	0.65	0.65	0.60	**0.56**	**0.52**	**0.44**
FQ-BP	**0.67**	**0.67**	**0.61**	**0.56**	0.50	0.40
Grubbs test	0.65	0.56	0.41	0.31	0.25	0.21

In Table 4, the dependence of detection performance w.r.t. the contamination parameter ε is studied. It is observed that with small and moderate levels of shift contamination, the FQ-boxplot is marginally better than its competitors. For larger fractions of contamination ($\varepsilon \geq 0.3$), the MAD-boxplot outperforms its competitors. It can be explained by the fact that the MAD is a minimax bias estimate of scale under the Tukey gross error model [14].

4.4. Real-Life Data Results

We tested our algorithms on a real-world dataset obtained from an experimental set up with representative cloud applications. It is a data intensive application implemented on Hadoop based on a distributed set of auctioning services. The analyzed data consist of 10 hours worth of service requests collected at 30 second intervals, into which 50 anomalies are injected over the duration of the experimental time-period. The anomalies are major failures or performance issues. We consider the metrics such as the server idle time (*% idle*), the traffic per second (*tpc*), the speed of reading and writing of data blocks (*bread/s+bwrtn/s*), and the speed of receiving and transmitting of data blocks (*rxpck/s+txpck/s*).

From Table 5, it follows that the *MAD* and FQ-boxplots considerably outperform the Tukey boxplot in terms of H-mean. Similar results are also observed if to use the false

Table 5. H-means for boxplot tests applied to server data

	Tukey BP	MAD-BP	FQ-BP
% idle	0.51	0.55	**0.58**
tpc	0.47	**0.57**	**0.57**
bread/s+bwrtn/s	0.47	**0.56**	**0.56**
rxpck/s+txpck/s	0.20	**0.33**	0.31

alarm rate α and the power of detection P_D. For instance, in case of the traffic per second (*tpc*), we have $\alpha = 0.06$, $P_D = 0.31$ for the Tukey boxplot and $\alpha = 0.1$, $P_D = 0.42$ for the MAD- and FQ-boxplots. Generally, all those results exhibit a rather low level of detection power; to raise it, we should increase the false alarm rate.

5. CONCLUSIONS

The two robust versions of the Tukey boxplot are proposed. Both versions aim at the symmetric distribution as their classical counterpart, the first MAD-BP being preferable under heavy contamination, while the second FQ-BP – under moderate contamination. The thresholds k can be adjusted to the adopted level of the false alarm probability α when detecting outliers. We recommend the values $k_{MAD} = 1.44$ and $k_{FQ} = 0.97$ corresponding to the rate $\alpha = 0.1$ under normality. All the boxplot tests considerably outperform the classical Grubbs test, which is catastrophically bad under contamination.

6. REFERENCES

[1] J.W. Tukey, "A survey of sampling from contaminated distributions," *Contributions to Probability and Statistics*, pp. 448–485, 1960.

[2] P.J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, pp. 73–101, 1964.

[3] F.R. Hampel, *Contributions to the Theory of Robust Estimation*, Ph.D. thesis, University of California, Berkeley, 1968.

[4] J.W. Tukey, *Exploratoy Data Analysis*, Addison-Wesley Series in Behavioral Science. Addison-Wesley, 1977.

[5] Robert McGill, John W. Tukey, and Wayne A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. pp. 12–16, 1978.

[6] H. Hagen A. Kerren Potter, K. and P. Dannenmann, "Methods for presenting statistical information: The box plot," *Visualization of Large and Unstructured Data Sets*, vol. S-4, pp. 97–106, 2006.

[7] Yoav Benjamini, "Opening the box of a boxplot," *The American Statistician*, vol. 42, no. 4, pp. 257–262, 1988.

[8] Warren W. Esty and Jeff Banfield, "The box-percentile plot," *Journal of Statistical Software*, vol. 8, no. 17, pp. 1–14, 10 2003.

[9] Jerry L. Hintze and Ray D. Nelson, "Violin plots: A box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.

[10] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, *Robust statistics: the approach based on influence functions*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 2005.

[11] Shevlyakov G.L. Kim, K., "Why gaussianity?," *The IEEE Signal Processing Magazine*, vol. 25, pp. 100–113, 2008.

[12] Peter J. Rousseeuw and Christophe Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 12731283, 1993.

[13] Shevlyakov G.L. Smirnov, P.O., "On approximation of the qn-estimate of scale by fast m-estimates," in *Int. Conf. on Robust Statistics*, Parma, Italy, 2010.

[14] P.J. Huber, *Robust Statistics*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics. John Wiley & Sons, 1981.

[15] Vic Barnett and T. Lewis, *Outliers in statistical data*, Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley & Sons, 1994.

[16] Frank E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, February 1969.