DETECTION OF OUTLIERS WITH BOXPLOTS

K. Andrea, G.L. Shevlyakov, P.O. Smirnov
St. Petersburg State Polytechnic University
St. Petersburg, RUSSIA
E-mail: Kliton.Andrea@gmail.com

Abstract

Low-complexity robust modifications to the Tukey boxplot based on fast highly efficient robust estimates of scale are proposed. The performance of the Tukey boxplot and its modified robust versions is measured relative to identification of outliers in Monte Carlo experiments at contaminated normal distributions. The obtained results show that the proposed methods outperform the conventional Tukey boxplot and the classical Grubbs test.

1 Introduction

Robust statistics provides 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.

In parallel with robust statistics, practical methods for analyzing data evolved known as *Exploratory Data Analysis* (EDA). 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.

Our work represents new results in robust data analysis technologies, providing alternatives to the boxplot technique. The univariate Tukey method 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, two new robust versions of the Tukey boxplot based on the highly efficient robust estimates of scale are proposed. In Section 3, two new rules for detection of outliers based on the proposed robust boxplots are introduced and examined on the contaminated Gaussian data. In Section 4, some conclusions are drawn.

2 Robust Modifications of the Tukey Boxplot

The Tukey univariate boxplot [5] 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

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

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

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 [2].

Since the interquartile range is less resistant to outliers than the median absolute deviation $MAD_n\, x = med_i\, |x_i - med\, x|$ [2], a more robust rule for constructing the boxplot extremes can be given by

$$x_L = \max\{x_{(1)},\, LQ - k_{MAD}\, MAD_n\}, \quad x_U = \min\{x_{(n)},\, UQ + k_{MAD}\, MAD_n\}, \quad (2)$$

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

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 [3], 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 time complexity of Q_n is of a greater order than of MAD_n.

In [4], an M-estimate of scale denoted by FQ_n whose influence function is approximately equal to the influence function of the estimate Q_n is proposed

$$FQ_n = 1.483\, MAD_n \left(1 - \frac{Z_0 - n/\sqrt{2}}{Z_2}\right), \quad (3)$$

where

$$Z_k = \sum_{i=1}^{n} u_i^k e^{-u_i^2/2}, \quad u_i = \frac{x_i - med\, x}{1.483\, MAD_n},\ k = 0, 2; \quad i = 1, \ldots, n.$$

The efficiency and breakdown point of FQ_n are equal to 0.81 and to 0.5, respectively.

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

$$x_L = \max\{x_{(1)},\, LQ - k_{FQ}\, FQ_n\}, \quad x_U = \min\{x_{(n)},\, UQ + k_{FQ}\, FQ_n\}. \quad (4)$$

3 Performance Evaluation

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 is an observation that is numerically distant from the rest of the data. A frequent cause of outliers is a mixture of two distributions, namely, a combination of "good data" and "bad data".

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 at the normal distribution. This rule is the classical Grubbs test [1].

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 $\alpha = 0.1$.

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

$$f(x) = (1 - \varepsilon)N(x; 0, 1) + \varepsilon N(x; \mu, s), \tag{5}$$

where $0 \leq \varepsilon < 1$ is the probability of outliers 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 tests with different values of the false alarm probability can be effectively compared. In our study, the false alarm rates for the Tukey and modified boxplots are $\alpha = 0.06$ and $\alpha = 0.1$, respectively.

The results of Monte Carlo experiment are given in Tables 1-2 with the best performing statistics represented in boldface.

Table 1: H-means 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 Table 1 it follows that under scale contamination, the performances of boxplot tests, generally, 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.

Table 2: H-means under shift contamination: $\mu = 3$, $s = 1$, $n = 100$.

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

Further, the robust MAD and FQ versions are slightly but systematically better than the Tukey boxplot test. Similar results are also obtained for the gross error models with shift contamination.

In Table 2, 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 [2].

4 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: 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.

References

[1] Grubbs F.E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*. Vol. **11**, pp. 121.

[2] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* Wiley, New York.

[3] Rousseeuw P.J., Croux C. (1993). Alternatives to the Median Absolute Deviation. *J. Amer. Stat. Assoc.*, Vol. **88**, pp. 1273-1283.

[4] Smirnov P.O., Shevlyakov G.L. (2010). On Approximation of the Q_n-Estimate of Scale by Fast M-Estimates. In: *Book of Abstracts of the International Conference on Robust Statistics (ICORS 2010)*. Prague, Czech Republic, pp. 94-95.

[5] Tukey J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, MA.