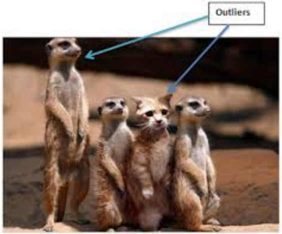
A Simple Approach to Deriving Outlier Labeling Rules for Skewed Distributions

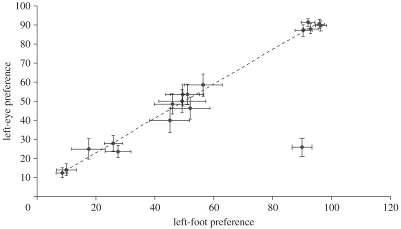
Kelly Cristina Ramos da Silva
Helder Luiz Costa de Oliveira
Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

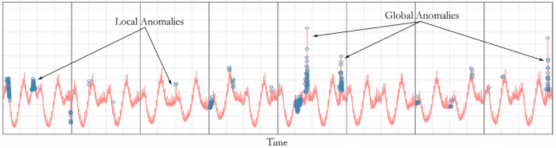
Universidade de São Paulo - ICMC

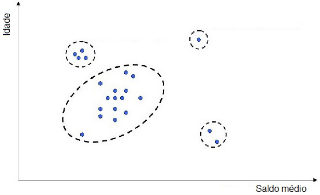
27 de agosto de 2018

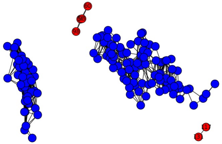
Presence of Abnormality



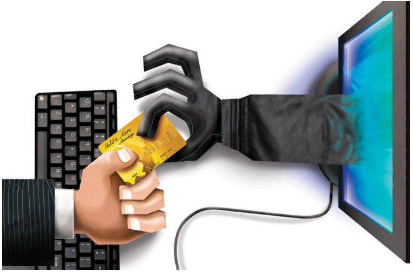








Presence of Frauds [1]



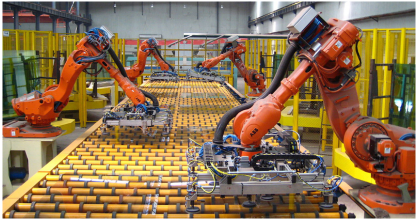
1. Intrusion Detection Systems;
2. Credit Card Fraud;
3. Law Enforcement;
4. Interesting Sensor Events.

[1] www.zoomtech.com.br/fraudes-na-internet-o-que-fazer-se-voce-for-vitima/

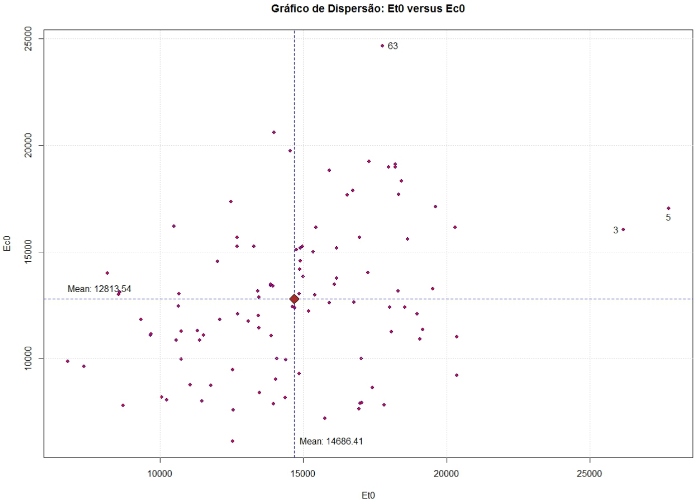
Disease Detection [2]

[2] https://www.heraldsun.com.au/news/victoria/melbourne-scientists-develop-blood-test-for-early-alzheimers-disease-detection/news-story/b61c32f26cf9286e9124f0ea86b21ebb

Human Errors or Mechanical Errors [3]



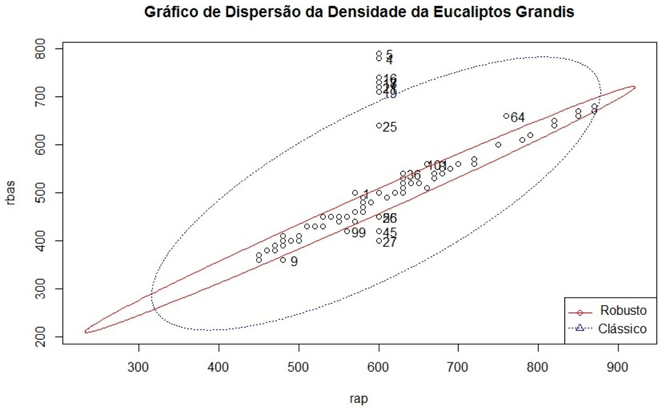
[3] http://www.hojeaprendi.com.br/2014/12/21/encontre-erros-de-digitacao-com-mais-facilidade/



Non-Robust Methods

1. Support Vector Machines - SVM;
2. Maximum Likelihood Estimators - MLE;
3. Least Square Method - LS.

Traditional vs. Robust Estimators



1. Independence;
2. Distribution Gaussian;
3. Absence of Outliers.

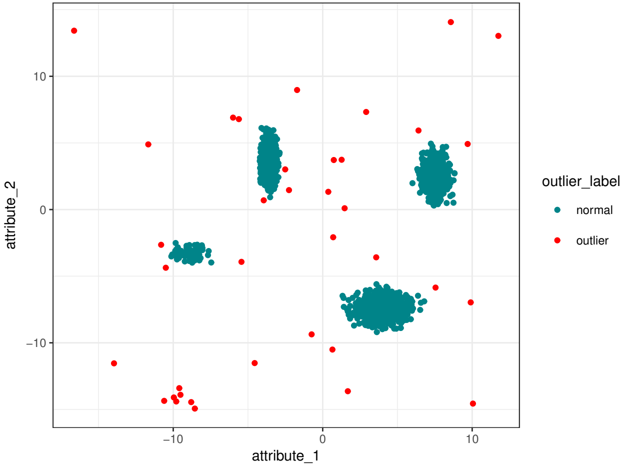
Measure Robustness [4]

Breakdown Point - BP

1. (max breakp. of 0%) $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
2. (max breakp. of 25%) $\frac{IQR}{a} = \frac{(Q_3 - Q_1)}{a}$
3. (max breakp. of 50%) $MAD = b * med|x_i - med(x_i)|$
4. (max breakp. of 50%) $S_n = c * med_i(med_j|x_i - x_j|)$

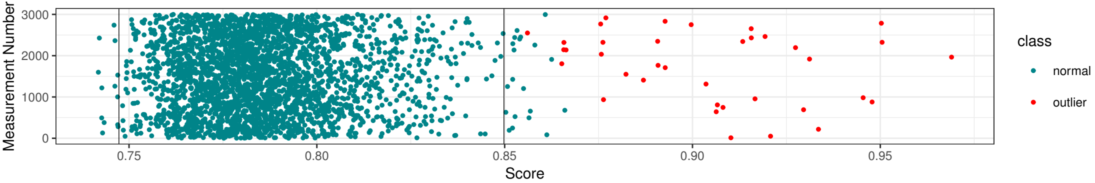
[4] Hampel, F. R. (1974). The influence curve and its role in robust estimation.

Rules for Outliers Detection [5]



[5] http://dx.doi.org/10.7910/DVN/OPQMVF

Rules for Outliers Detection



Model of Location and Scale [6]

Let the cumulative distribution function $F_{\mu,\sigma}$, where F is a continuous function with parameters of location μ and scale σ.

$$\mu \pm (factor)\sigma \tag{1}$$

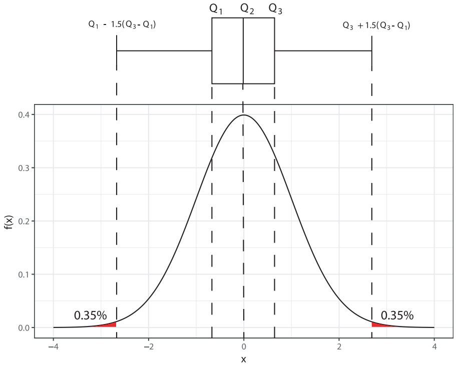
factor = g*k.

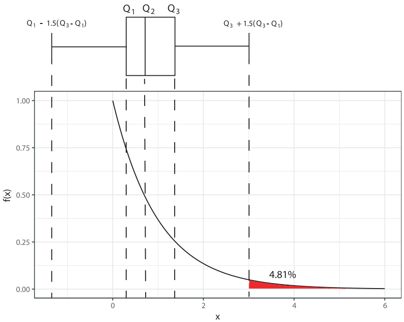
$$k = \frac{F^{-1}(1-\alpha/2) - F^{-1}(0.75)}{F^{-1}(0.75) - F^{-1}(0.25)}. \tag{2}$$

g parameter that consider the size of the sample.
$\alpha = P(X_i \in out(\alpha_n, \mu, \sigma^2) | X_i \notin out(\alpha_n, \mu, \sigma^2))$

[6] Dovoedo, Y. H., and Chakraborti, S. (2015). Boxplot-based outlier detection for the location-scale family.

Classic Boxplot [7]





$$[Q_1 - kIQR; Q_3 + kIQR] \quad (3)$$

$$k = \frac{\phi^{-1}(1-\alpha/2) - \phi^{-1}(0.75)}{\phi^{-1}(0.75) - \phi^{-1}(0.25)} \quad (4)$$

[7] Tukey, J. W. (1977). Exploratory data analysis (Vol. 2).

Adjusted Boxplot [8]

Let $S = \{x_1, x_2, \cdots, x_n\}$ be a sample from a continuous unimodal distribution. The medcouple can be determined as follows:

$$[MC = Med\frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i},] \quad (5)$$

where Q_2 is the median of S, $x_i \leq Q_2 \leq x_j$ and $x_i \neq x_j$.
If $MC \geq 0$

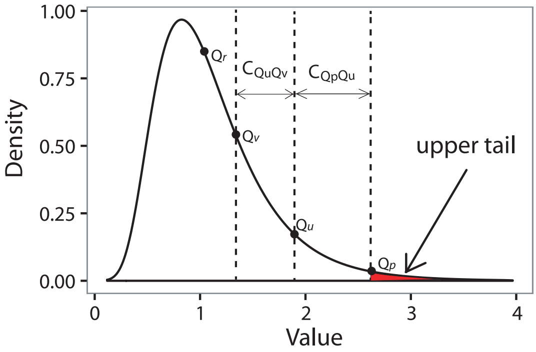
$$[Q_1 - 1.5e^{-4MC}IQR; \quad Q_3 + 1.5e^{3MC}IQR]. \quad (6)$$

If $MC < 0$

$$[Q_1 - 1.5e^{-3MC}IQR; \quad Q_3 + 1.5e^{4MC}IQR]. \quad (7)$$

[8] Hubert, M., and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions.

New Method: Contrast between Quantiles

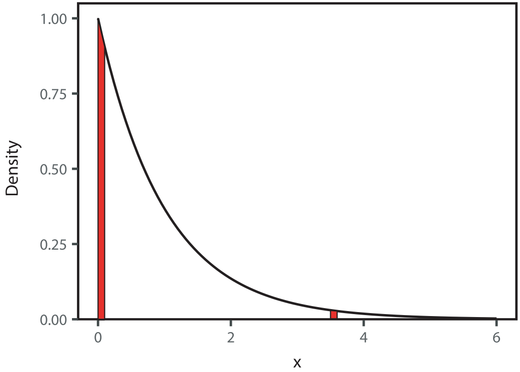


1. Non Parametric;
2. Form Simetrics and Skews;
3. Nominal Outside Rate 0.7%;
4. Scale;

$$[SIQR_L = (Q_2 - Q_1); \quad (8)$$

$$SIQR_U = (Q_3 - Q_2)] \quad (9)$$

5. Contrast between quantiles;
6. Sensitivity constant.



Upper tail

Scenario without contamination

Lower Tail

Scenario with contamination

Tabela: Summary of the outlier rules

Rule	Lower Threshold	Upper Threshold
Boxplot	$Q_1 - 1.5IQR$	$Q_3 + 1.5IQR$
Adj. Boxplot	$\begin{cases} Q_1 - 1.5\,e^{-4MC}IQR & : \quad MC \geq 0 \\ Q_1 - 1.5\,e^{-3MC}IQR & : \quad MC < 0 \end{cases}$	$\begin{cases} Q_3 + 1.5\,e^{3MC}IQR & : \quad MC \geq 0 \\ Q_3 + 1.5\,e^{4MC}IQR & : \quad MC < 0 \end{cases}$
QCR(0.0035)	$\begin{cases} Q_1 - 3\,e^{\left(1.64\left(1/C_{Q_3Q_2}-1\right)\right)}SIQR_L & : \quad C_{Q_3Q_2} \geq 1 \\ Q_1 - 3\,\log_{1.66}\left(1.66/C_{Q_3Q_2}\right)SIQR_L & : \quad C_{Q_3Q_2} < 1 \end{cases}$	$\begin{cases} Q_3 + 3\,\log_{1.66}\left(1.66C_{Q_3Q_2}\right)SIQR_U & : \quad C_{Q_3Q_2} \geq 1 \\ Q_3 + 3\,e^{\left(1.64\left(C_{Q_3Q_2}-1\right)\right)}SIQR_U & : \quad C_{Q_3Q_2} < 1 \end{cases}$
QCR(0.007)	$\begin{cases} Q_1 - 3\,e^{\left(1.64\left(1/C_{Q_3Q_2}-1\right)\right)}SIQR_L & : \quad C_{Q_3Q_2} \geq 1 \\ Q_1 - 3\,\log_{2.1}\left(2.1/C_{Q_3Q_2}\right)SIQR_L & : \quad C_{Q_3Q_2} < 1 \end{cases}$	$\begin{cases} Q_3 + 3\,\log_{2.1}\left(2.1C_{Q_3Q_2}\right)SIQR_U & : \quad C_{Q_3Q_2} \geq 1 \\ Q_3 + 3\,e^{\left(1.64\left(C_{Q_3Q_2}-1\right)\right)}SIQR_U & : \quad C_{Q_3Q_2} < 1 \end{cases}$

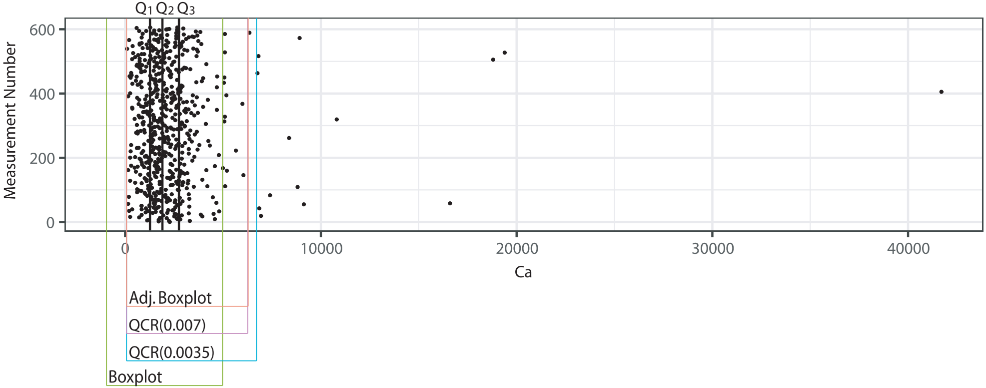
Tabela: Average outside rates determined by both upper and lower thresholds of each considered rule for uncontaminated samples of size 10,000.

Distribution	MC	Boxplot	Adj. Boxplot	QCR(0.0035)	QCR(0.007)
$exp(1)$	0.333 ± 0.001	0.0481 ± 2e-04	0.00288 ± 6e-05	0.00353 ± 9e-05	0.0070 ± 1e-04
$\Gamma(0.1, 0.5)$	0.505 ± 0.001	0.0756 ± 2e-04	0.00191 ± 6e-05	0.00349 ± 8e-05	0.0078 ± 1e-04
$\Gamma(0.1, 0.75)$	0.396 ± 0.001	0.0579 ± 2e-04	0.00243 ± 6e-05	0.00347 ± 8e-05	0.0073 ± 1e-04
$\Gamma(0.1, 1.25)$	0.292 ± 0.001	0.0419 ± 2e-04	0.00320 ± 9e-05	0.0036 ± 1e-04	0.0068 ± 1e-04
$\Gamma(0.1, 5)$	0.136 ± 0.001	0.0193 ± 2e-04	0.0062 ± 1e-04	0.00398 ± 9e-05	0.0060 ± 1e-04
$N(0, 1)$	0.000 ± 0.001	0.0070 ± 1e-04	0.0073 ± 1e-04	0.0072 ± 1e-04	0.0074 ± 1e-04
χ^2_1	0.505 ± 0.001	0.0756 ± 3e-04	0.00192 ± 5e-05	0.00350 ± 8e-05	0.0078 ± 1e-04
χ^2_5	0.197 ± 0.001	0.0279 ± 2e-04	0.0041 ± 1e-04	0.0037 ± 1e-04	0.0062 ± 1e-04
χ^2_{20}	0.095 ± 0.001	0.0139 ± 2e-04	0.0074 ± 1e-04	0.0051 ± 1e-04	0.0066 ± 1e-04
$F(90, 10)$	0.259 ± 0.001	0.0516 ± 2e-04	0.0222 ± 3e-04	0.0128 ± 1e-04	0.0177 ± 2e-04
$F(10, 10)$	0.274 ± 0.001	0.0535 ± 2e-04	0.0194 ± 3e-04	0.0119 ± 1e-04	0.0170 ± 2e-04
$F(10, 90)$	0.155 ± 0.001	0.0238 ± 2e-04	0.0076 ± 1e-04	0.0048 ± 1e-04	0.0072 ± 1e-04
$F(80, 80)$	0.099 ± 0.001	0.0163 ± 1e-04	0.0100 ± 1e-04	0.0071 ± 1e-04	0.0089 ± 1e-04

Note: MC is the medcouple - robust measure of skewness

C-horizon of the Kola Data

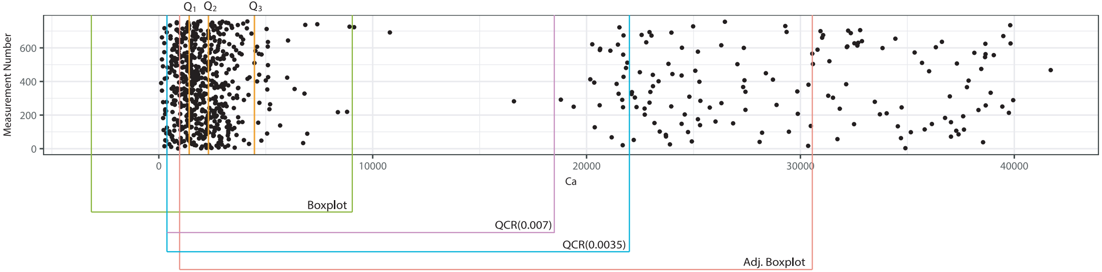
C-horizon data of 606 observations collected in the Kola Project (1993-1998, Geological Surveys of Finland and Norway and Central Kola Expedition, Russia [9])



[9] Reimann, C., and Garrett, R. G. (2005). Geochemical background—concept and reality.

C-horizon of the Kola Data

20% contamination.



Conclusions

1. High Performance for more than 15% outliers;
2. It is describe well skewed;
4. its computational complexity is $O(n)$.
5. It does not need of simulation to chance "outside rate".

References

1 Aggarwal, C. C. (2015). Outlier analysis. In Data mining (pp. 237-263). Springer, Cham.

2 Brys, G., Hubert, M., & Struyf, A. (2004). A robust measure of skewness. Journal of Computational and Graphical Statistics, 13(4), 996-1017.

3 Dovoedo, Y. H., & Chakraborti, S. (2015). Boxplot-based outlier detection for the location-scale family. Communications in Statistics-Simulation and Computation, 44(6), 1492-1513.

4 Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the american statistical association, 69(346), 383-393.

5 Hubert, M., and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. Computational statistics and data analysis, 52(12), 5186-5201.

6 Kimber, A. C. (1990). Exploratory data analysis for possibly censored data from skewed distributions. Applied Statistics, 21-30.

7 Reimann, C., and Garrett, R. G. (2005). Geochemical background—concept and reality. Science of the total environment, 350(1-3), 12-27.

8 Tukey, J. W. (1977). Exploratory data analysis (Vol. 2).