A New Measure of Outlier Detection Performance

Kliton Andrea[1], Georgy L. Shevlyakov[1], Natalia Vassilieva[2], Alexander Ulanov[2]

[1] Department of Applied Mathematics, St.Petersburg Polytechnic University
[2] St. Petersburg HP Research Labs

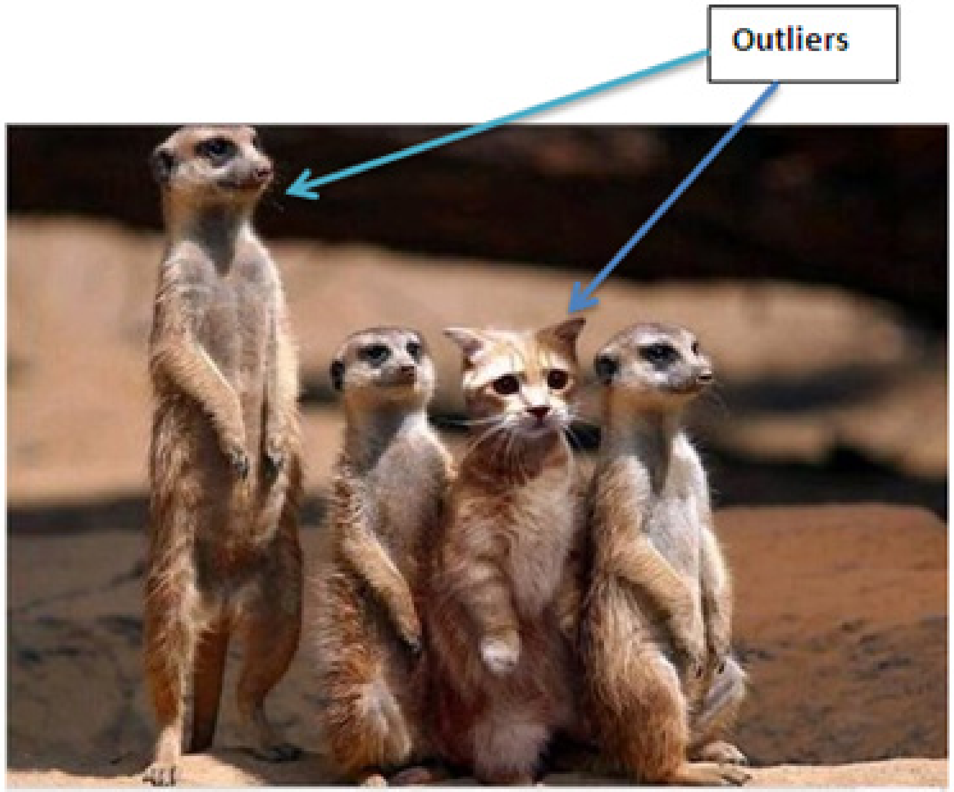
MLDM 2014, St. Petersburg, July 21, 2014

Outline

- Introduction
- Motivation and Principal Idea
- H-measure
- Performance Evaluation
- Conclusions
- References

Introduction

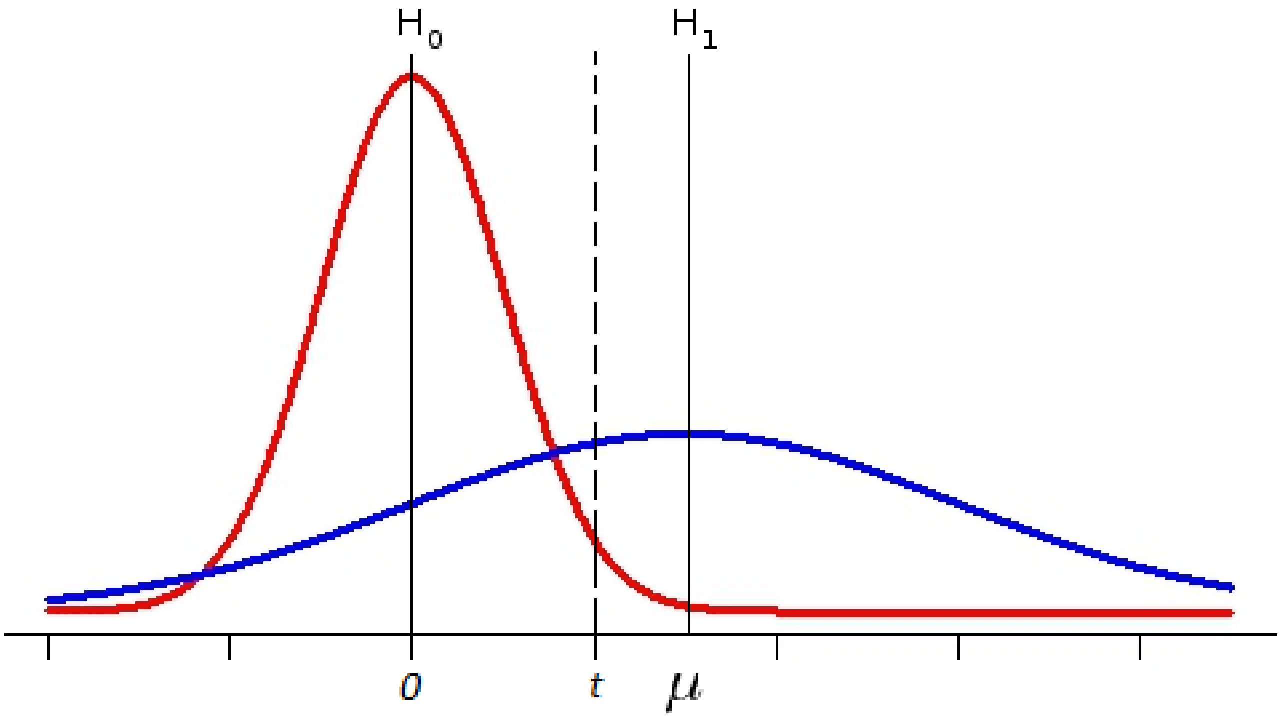
Grubbs' definition: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. (Barnett 1978, Hawkins 1980)



Introduction

H_0: an observation X is regular – $X \sim N(0, 1)$,

H_1: X is an outlier – $X \sim N(\mu, \sigma)$, $\mu > 0, \sigma > 1$.



Introduction

Example of detection rule:

H$_0$: $X < t$

H$_1$: $X \geq t$

Detection power $\qquad P_D = P(X \geq t | H_1)$

False alarm rate $\qquad P_F = P(X \geq t | H_0)$

Introduction

Neyman-Pearson Approach:

$P_D \to \max_t$,

$P_F \leq \alpha$,

where the false alarm bound rates or significance values α are usually chosen as follows

$\alpha = 0.001, 0.005, 0.01, 0.05, 0.1.$

Introduction

Generally, within the Neyman-Pearson approach to hypotheses testing, a decision rule has the following form

$$t(x_1, ..., x_n) \geq \lambda_\alpha,$$

$$P_F = \alpha,$$

where $t(x_1, ..., x_n)$ is a test statistic, λ_α is a threshold defined by the last equation.

In particular, the classical Grubbs' test (1969) for outliers has the form

$$\max_{i=1,...,n} \frac{|x_i - \bar{x}|}{s} \geq \lambda_\alpha,$$

where $x_1, ..., x_n$ is the observed sample, $\bar{x}$ and s are the sample mean and standard deviation.

Motivation and Principal Idea

There exist numerous methods for detecting outliers (Chauvenet 1863, Dixon 1950, Grubbs 1950, 1969; Tukey 1977, Barnett and Lewis 1978, 1994; Hawkins 1980, Shevlyakov and Vilchevski 2002, 2011; Manoj and Senthamarai Kennan 2013).

Motivation point 1: **The existing comparison studies** based on the Neyman-Person measures (power and false alarm rate) for outlier detection tests **are fragmentary**.

In theoretical (most asymptotic) **studies, a comparison**, generally, **can be well-defined**: given a false alarm rate, the performance of detection is measured in its power, **but the cases when it is possible are relatively rare.**

In real-life applications, when one deals **with small samples and Monte Carlo techniques,** this does not work — **it is difficult to provide** stability of small false alarm probability rates and thus **stability of statistical inference**.

Motivation and Principal Idea

Motivation point 2: **If a researcher is interested in optimization of** the parameters of **an outlier detection procedure, then it is more convenient to use a single scalar measure of performance**, that is, rather the unconditional optimization tools than the conditional or multi-criteria ones.

Principal Idea: **In IR classification studies, the F-measure of classification performance** (the harmonic mean between Precision P and Recall R – $H(P,R)$) **is widely and successfully used, so we propose to apply it to outlier detection as a measure of comparison and optimization** (the harmonic mean between the detection power P_D and unit minus the false alarm rate $1-P_F$:

$$H(P_D, 1-P_F) = 2\,\frac{P_D(1-P_F)}{P_D+1-P_F}\,.$$

H-measure

Recall **the properties of the harmonic mean** $H = H(x_1, x_2)$ with $0 < x_1 \leq 1, 0 < x_2 \leq 1$, **the reciprocal of the average of reciprocals** $H = 1/(1/2(1/x_1 + 1/x_2)) = 2x_1x_2/(x_1 + x_2)$:

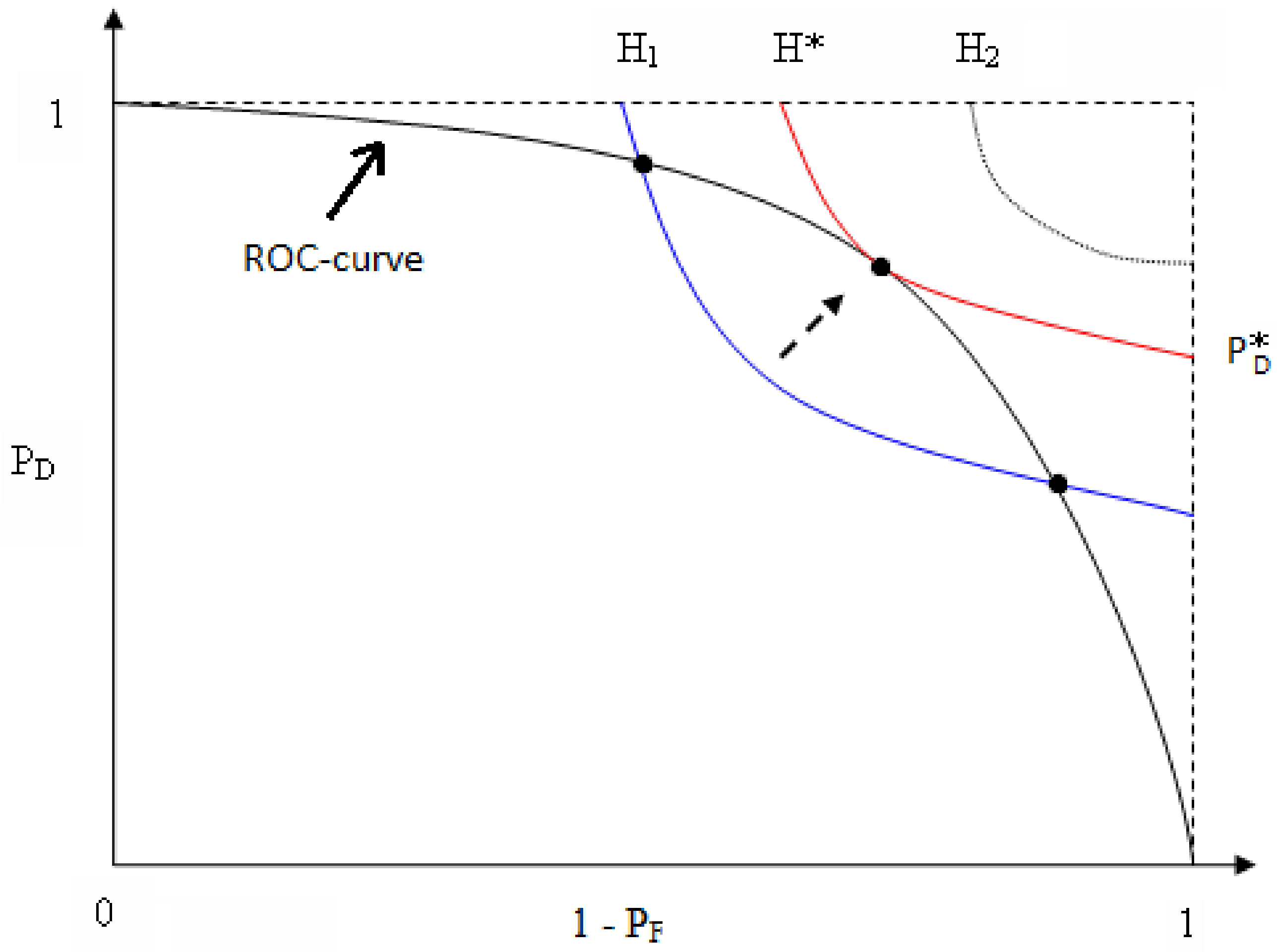
- H is **one of the three Pythagorean means**, the average $A = (x_1 + x_2)/2$, the geometric mean $G = \sqrt{x_1x_2}$ and H with the following relation between them:

$$H \leq G \leq A.$$

- **The harmonic mean mitigates the impact of a larger number and aggravates the impact of a smaller one**.

- **The harmonic mean aims at the processing of rates and ratios** (numerous examples in physics, genetics, finance).

H-measure



H-measure

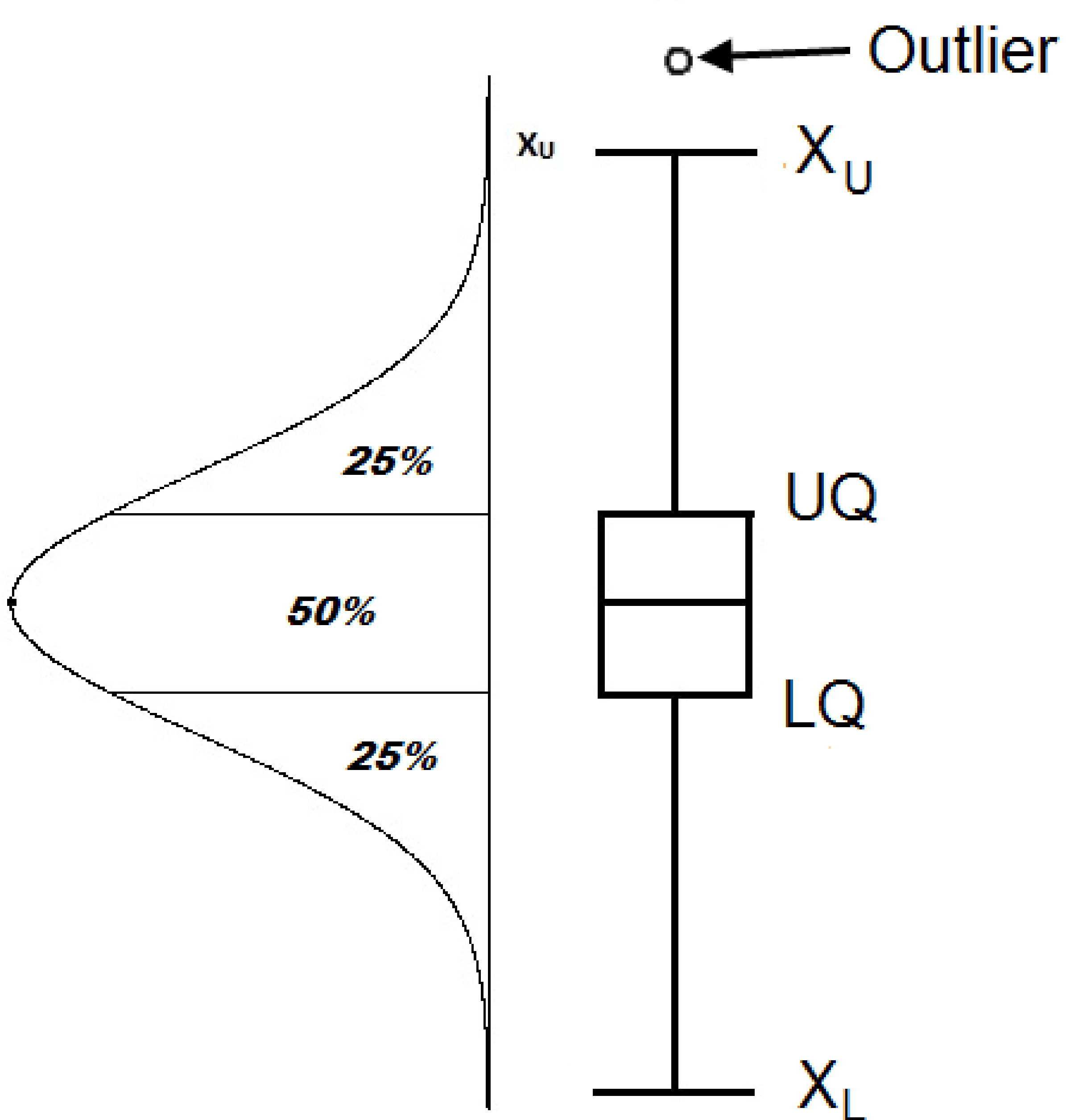
Theorem Given an H-measure value, the following inequalities hold:

$$P_D > P_D^{min} = \frac{H}{2-H}, \quad P_F < P_F^{max} = 1 - P_D^{min},$$

where P_D is the detection power and P_F is the false alarm rate.

H	0.8	0.9	0.92	0.94	0.95	0.96	0.97	0.98	0.99	0.999
P_D^{min}	0.67	0.82	0.85	0.87	0.90	0.92	0.94	0.96	0.98	0.998
P_F^{max}	0.33	0.18	0.15	0.13	0.10	0.08	0.06	0.04	0.02	0.002

Performance Evaluation



Performance Evaluation

Tukey's boxplot (Tukey 1977)

The Tukey's boxplot extremes x_L and x_U are of the form:

$$\begin{cases} x_L = \max\{x_{(1)}, LQ - \frac{3}{2} IQR\} \\ \\ x_U = \min\{x_{(n)}, UQ + \frac{3}{2} IQR\}, \end{cases} \tag{1}$$

where $x_{(1)}$ and $x_{(n)}$ are the extremal sample order statistics, $IQR = UQ - LQ$ is the sample interquartile range - a robust estimate of scale.

Performance Evaluation

Robust modifications of Tukey's boxplot (Andrea and Shevlyakov 2011)

The boxplot extremes x_L and x_U have the following general form:

$$\begin{cases} x_L = \max\{x_{(1)}, LQ - k_S\, S\} \\ \\ x_U = \min\{x_{(n)}, UQ + k_S\, S\} \end{cases} \tag{2}$$

where S is a robust estimate of scale and k_S is a threshold coefficient.

The following robust modifications of Tukey's boxplot will be considered: MAD-boxplot and FQ-boxplot.

Performance Evaluation

The median absolute deviation MAD_n (Hampel 1974):

$$MAD_n = 1.4826\, med\, \{|x - med\, x|\}.$$

Its efficiency at the Gaussian is 37% ($eff = 0.37$) and the breakdown point is 50% ($\varepsilon^* = 0.5$).

The Q_n-estimate (Rousseeuw and Croux 1993):

$$Q_n = c_n \, \text{1st quartile}\, \{|x_i - x_j|, \quad i < j\},$$

with $eff = 0.82$ and $\varepsilon^* = 0.5$.

The fast low-complexity FQ_n-estimate (Smirnov and Shevlyakov 2014): $eff = 0.81$ and $\varepsilon^* = 0.5$.

Performance Evaluation

Optimization problem

Consider the following optimization problem

$$k_S^* = \arg\max_{k_S} H(k_S) \tag{3}$$

in Tukey's gross error models:

- Scale contamination $X \sim F(x) = (1-\varepsilon)\Phi(x) + \varepsilon\,\Phi(x/\sigma)$,

- Shift contamination $X \sim F(x) = (1-\varepsilon)\Phi(x) + \varepsilon\,\Phi(x-\mu)$,

where $\varepsilon < 0.5$, $\mu > 0$ and $\sigma > 1$.

Performance Evaluation

The optimal threshold values for the scale and shift contamination in terms of the H-measure.

	Scale Contamination				Shift Contamination			
$\varepsilon = 0.1$	k^*	H	P_D	P_F	k^*	H	P_D	P_F
Tukey's boxplot	0.36	0.73	0.67	0.20	0.51	0.78	0.74	0.13
MAD-boxplot	0.47	0.73	0.68	0.21	0.77	0.78	0.73	0.11
FQ-boxplot	0.5	0.73	0.66	0.19	0.61	0.78	0.75	0.14

Performance Evaluation

The optimal threshold values for the scale and shift contamination in terms of the F-measure.

	Scale Contamination				Shift Contamination			
$\varepsilon = 0.1$	k^*	F	P_D	P_F	k^*	F	P_D	P_F
Tukey boxplot	1.4	0.96	0.35	0.007	1.5	0.97	0.51	0.004
MAD-boxplot	1.9	0.96	0.35	0.007	2	0.97	0.52	0.005
FQ-boxplot	1.8	0.96	0.36	0.007	1.9	0.97	0.51	0.003

Conclusions

A new H-measure of outlier detection performance is proposed.

The relationships between the H-measure and the Neyman-Pearson approach characteristics (detection power P_D, false alarm rate P_F and ROC-curve) are established.

The H-measure aims both at comparison of different detection tests and at their optimization.

The H-measure is recommended for outlier detection when it is more important to identify a real outlier than to assign a regular observation to outliers; in other words, the cost of the errors of the *2*nd kind is greater than of the *1*st kind.

The classical Tukey's boxplot threshold $k = 3/2$ is optimal with respect to the F-measure.

References

Barnett V. and Lewis, T. (1994). *Outliers in Statistical Data.* Wiley, New York.

Grubbs F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics,* 11, 1-21.

Shevlyakov G. and Vilchevski N. (2011). *Robustness in Data Analysis: criteria and methods.* De Gruyter, Boston.

Shevlyakov G., Andrea K., Choudur L., Smirnov P., Ulanov A. and Vassilieva N. (2013). Robust Versions of the Tukey Boxplot with Their Application to Detection of Outliers. *ICASSP 2013 Proceedings,* Vancouver.

Smirnov P. and Shevlyakov G. (2014). Fast highly efficient and robust one-step M-estimators of scale based on Q_n. *Computational Statistics and Data Analysis.* (in press).

References

THANK YOU !