Comparison of Outliers Elimination Algorithms

P. Mostarac, R. Malarić, H. Hegedušić

Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
Email: {petar.mostarac; roman.malaric; hrvoje.hegedusic}@fer.hr

Abstract. *In this paper one general approach to elimination the single outlier is described, and shown that Chauvenet's criterion is a one particular solution. Peirce's criterion is also described and compared for the rejection of suspicious data with Chauvenet's criterion, their sensitivity to outliers and their characteristics. All expressions needed for calculation the borders for rejection suspicious observations and their solutions are given.*

Keywords: Outliers Elimination, Comparison of Peirce's Criterion vs. Chauvenet's Criterion

1. Introduction

When modeling error of measurement, which is defined as difference between real and measured value, insignificant part of an error is made by impulse error. Impulse errors are common fluctuation of significant deviation and can increase result of measurement interpreted by mean and standard deviation statistics.

If probability of having suspicious measurement, which can be described as impulse error, with assumption of Gauss distribution and estimation of parameters (mean and variance), given by the all measurements, less than the real number of suspicious data, it can be discuss on their rejection. That decision is explained with probability that rejected data are a result of impulse error. By setting boundaries for outlier elimination, strictness of criterion is defined and arbitration over suspicious data is done.

The rest of this paper the focus will be on setting the criterion for one suspicious data, and Peirce's criterion which is able for multiple outlier elimination.

2. Criterions

Criterion for one suspicious data
Criterion for eliminating outliers can be defined by the amount of allowed deviation comparing to standard deviation σ. For N measurements, with standard deviation σ, mean μ and suspicious data *x* defined as:

$$n = \frac{\max\{|x - \mu|\}}{\sigma} \tag{1}$$

It can be defined [1] that probability in N measurements must be greater or equal to *p* so it can be kept:

$$p \geq N \times (1 - P(n\sigma)), \tag{2}$$

P is a function defined like integral of probability density function (*pdf*) on interval ±nσ. Function *P* can be substituted with Cumulative distribution function (*cdf*) *Φ*,

$$P(n\sigma)=(1-2\Phi(n\sigma)), \tag{3}$$

and introducing (3) in (2), with defining *cdf* function through Error function the solution is:

$$n \leq \sqrt{2} erf^{-1}(\frac{p}{N} - 1)\,. \tag{4}$$

For Chauvenet's criterion p=0,5. And p can be chosen such that our criterion be less or more rigorous.

Peirce's criterion

Peirce's criterion is a more rigorous than Chauvenet's criterion. Peirce's criterion is also able to remove several suspicious data. It is an iterative method based on theory of probability. In [2] it is described and also required conditions for rejection are presented. We will explain the crucial parts of Peirce's criterion. The principle is that the k suspicious data *should be rejected when the probability* P, with all data including the suspicious data, *of the system of the errors is less then* a probability without suspicious data multiplied by *the probability of making* those suspicious observations P_1.

$$P<P_1 \tag{5}$$

Whence [3],

$$\lambda^{N-k} R^k < Q^N . \tag{6}$$

$$\lambda^2 = \frac{N-m-kn^2}{N-m-k} \tag{7}$$

$$R = e^{\frac{(n^2-1)}{2}} erf\left(\frac{n}{\sqrt{2}}\right) \tag{8}$$

$$Q^N = \frac{k^k (N-k)^{N-k}}{N^N}, \tag{9}$$

where

k number of suspicious data

m the number of unknown quantities contained in the observations.

Equation (6) doesn't have explicit solution. We can solve it on several ways. With some numerical method like Newton-Raphson method or we can use Gould's proposal that he gave in [3]. First to calculate the value for Q^N, it is a constant, then R with arbitrary n_a, and then recalculate new n_n from (7). For large N use logarithm. After that we can repeats all process with updated $n_a = n_n$, until the error $|n_n - n_a|$ is enough small, $n_a \cong n_n = n$. Final n is the number of σ. If suspicious data excide ±nσ, they can be removed. The number of suspicious data $k \in [1, \lfloor N/2 \rfloor]$, there is no point to calculate n for grater k because, if $k > \lfloor N/2 \rfloor$ then we can consider that all data are suspicious and it will be smart to repeat measurements.

3. Results

Results of equation (4) for $p \in (0,1]$ and $N \in [3,50]$ are shown in Fig. 1. Above the function is area of outliers and under the function is area of valid data.

Results for calculations of n with equation (6) for $N \in [3,50]$ and $k \in [1, \lfloor N/2 \rfloor]$ are shown in Fig. 1. The value of n for k greater then $\lfloor N/2 \rfloor$ is zero only for easier representation, but the real value isn't defined. For small number of data and only one suspicious observation the Peirce's criterion is more rigorous as it is shown in Fig. 3.

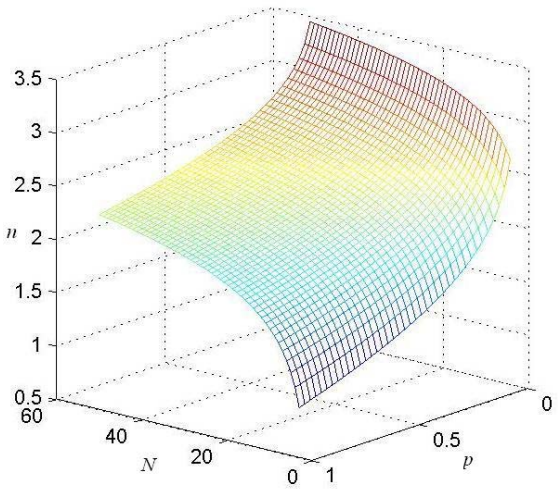

Fig. 1. Parameter *n*, from function (4), for $p \in (0,1]$.

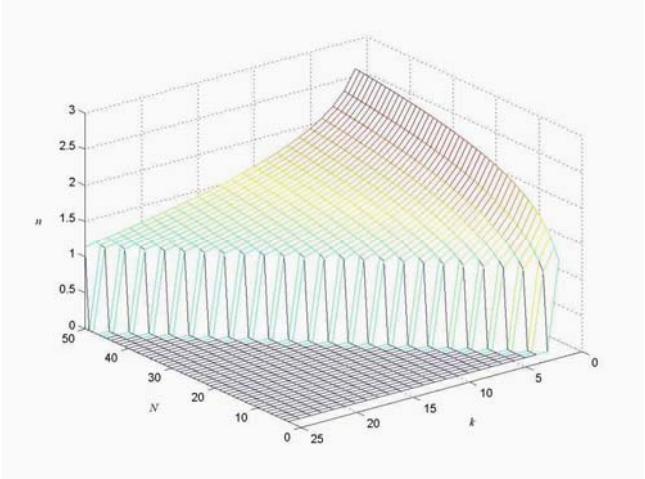

Fig. 2. Parameter *n* calculated from function (6) for $k \in [1, \lfloor N/2 \rfloor]$.

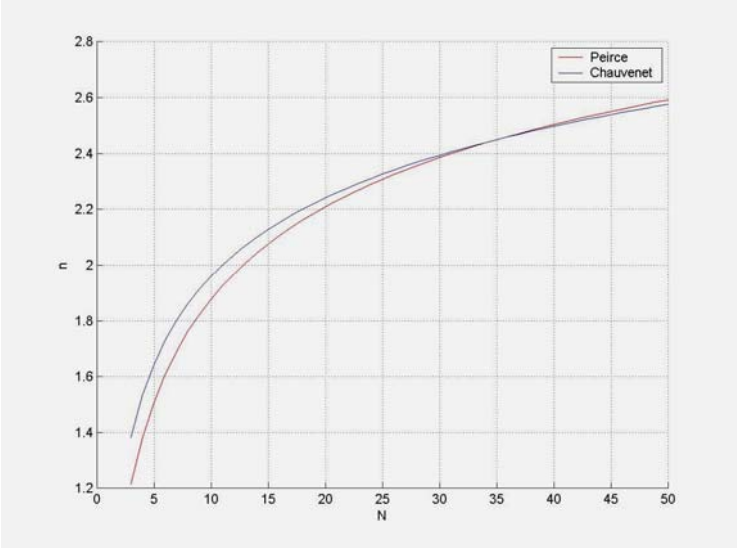

Fig. 3. Parameter *n* calculated from function (4) with *p*=0,5 and from function (6).

4. Conclusions

Chauvenet's criterion is frequently used for removing suspicious data, and also for removing several suspicious data without exact justification. For those data, where are possible more then one suspicious data we must use Peirce's criterion, or some other iterative algorithm. Chauvenet's criterion create fix borders independent from number of suspicious data, so second and other removed data aren't removed because their probability to be outliers, but due to the probability of one data to be outlier.

References

[1] Taylor J. R.: An introduction to Error Analysis, University Science Books, Sausalito, 1997.

[2] Benjamin P., Criterion for the rejection of doubtful observations, *Astronomical Journal*, 45, 161-163, 1852.

[3] Gould B. A., On Peirce's criterion for the rejection of doubtful observations, with tables for facilitating its application, *Astronomical Journal*, 83, 81-83, 1855.