

Outliers Elimination by Reverse Approximation

Lojacono R., Mencattini A., Rabottino G., Salmeri M.

University of Rome Tor Vergata – Rome – Italy

Email: lojacono@uniroma2.it

Abstract:

1. Introduction

A process control is normally driven by some measured data evaluated by a suitable algorithm. However, the set of the measured data can contain outliers.

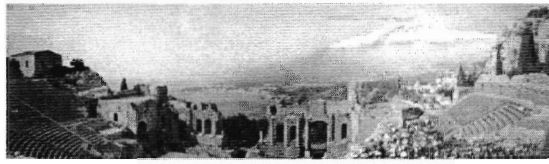
An outlier is intended as an abnormal or unexpected data or observation. This means that the characterization of outliers depends on what we consider the normal variation of data. Often, the variation of the observed data is supposed to follow a Gaussian distribution characterized by the expected value μ and the standard deviation σ . In a brute application, μ is simply the mean of the given data and σ^2 is determined as the mean square deviation. In an automated procedure, when a previous screening of the data is not present, both μ and σ are estimated taking into account also the outliers. This fact represents a possible corruption of the process control, so the outliers must be previously eliminated.

In this paper, we propose a deterministic elimination of the outliers based on the here called *reverse approximation (RA)*. As explained in the following section, we propose the use of approximation algorithms in the outlier elimination. This proposal starts firstly from *minmax* approach implemented by Remez algorithm. Then, in order to make simpler the procedure, we use the least squares method. From a theoretical point of view, we operate in the L_∞ norm instead of the L_2 norm.

2. The Remez algorithm

The purpose of the Remez algorithm, in terms of our present interest, is to find a particular subset of the given data, having a cardinality $n+2$ if n is the degree of the approximant polynomial, called the *optimal subset (OS)*. This optimal subset possess two properties related to the polynomial determined as solution of the linear system

$$P_n(\xi_i) = v_i + (-1)^i E; \quad v_i \in OS; i = 1, \dots, n+2$$



Session: 6

Presentation: 2

- 1) The polynomial deviates from the remaining data less than E ;
- 2) E obtained over OS is the greatest error that can be obtained solving the system over all the remaining subset of the given data having cardinality $n+2$.

An illustration of the progression of Remez algorithm is given in Figures 1 and 2.

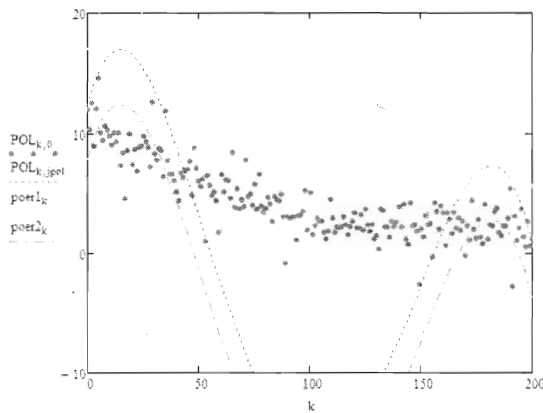


Figure 1 – Initial step of the Remez procedure

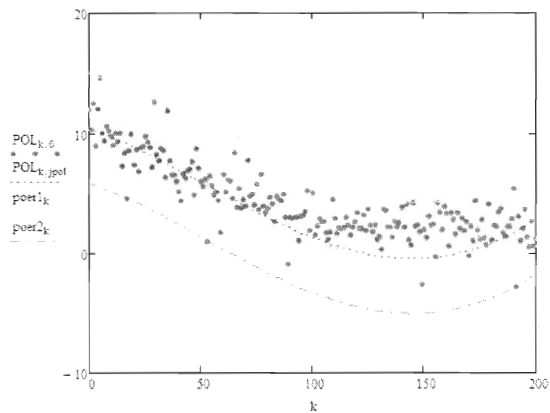
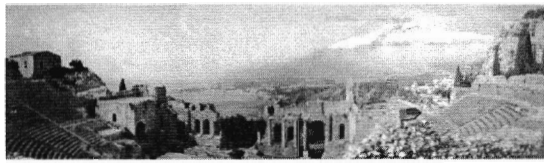


Figure 2 – Final step of the Remez procedure.

3. The Remez procedure used as current step of an outlier elimination algorithm

To our purposes, the second property of the Remez algorithm above mentioned is interpreted as follows: since operating over this subset OS led to the greatest error E , all the values of OS can be suspected to be outliers but since the polynomial has $n+1$ parameters, namely the coefficients, so it can interpolate exactly $n+1$ values of OS , only one of the values of OS , which has cardinality $n+2$, is surely an outlier. To individuate this outlier, we repeat the Remez procedure to all the subset of $n+1$ values obtained by eliminating one of the values of the OS : the outlier is the value which after its elimination produces the smallest E in the corresponding Remez procedure. The determination of this particular member of OS is clearly not simple in the general case but becomes easy if the polynomial has low degree n . If $n=0$, i.e. if we approximate a constant, the procedure becomes trivial. Let us indicate this particular case as *monodimensional case*.



Session: 6

Presentation: 2

4. Monodimensional case

Let us consider the simple case in which we want to eliminate the outliers in a set of measures of the same constant. This represents a monodimensional problem and we use for the Remez procedure a polynomial of degree 0, i.e. $P_0 = a_0$. So, the OS is composed by:

$$OS = \{ \min \{v_i\}, \max \{v_i\} \}, \quad i = 1, 2, \dots, N$$

and

$$a_0 = \frac{\min \{v_i\} + \max \{v_i\}}{2}, \quad E = \frac{\min \{v_i\} - \max \{v_i\}}{2}$$

Now the suspected outliers are the values of max and min, so we firstly eliminate the max and repeat the calculation of a_0 and E over the set of the remaining $N-1$ values; then eliminate the min and include the previous max and repeat the procedure. If the elimination of the max has determined the lower value of E , this max is definitively eliminated. Otherwise the previous min is definitively eliminated.

An example of the described procedure is given in Fig. 3. We consider a sequence of normal distributed values and superimpose over this sequence two set of outliers generated by a random function. The simulation is performed using MATHCAD tool. The example shows the usefulness of the proposed algorithm.

5. Monodimensional case

We consider here the case in which the approximant polynomial has a degree n . A great simplification of the algorithm is obtained if we eliminate in the current step all the values of the OS set. This is a useful strategy if the number of the values is large enough. However, it can become very hazardous when the number of the remaining values becomes small. A good tradeoff is obtained if we, instead of eliminate, substitute the corresponding values of the current polynomial for the values belonging to OS.

However, this procedure, which employs in the current step the Remez algorithm, appears to be excessively time consuming. The Remez algorithm can be advantageously substituted by the least squares method, though maintaining the philosophy of the proposed procedure.

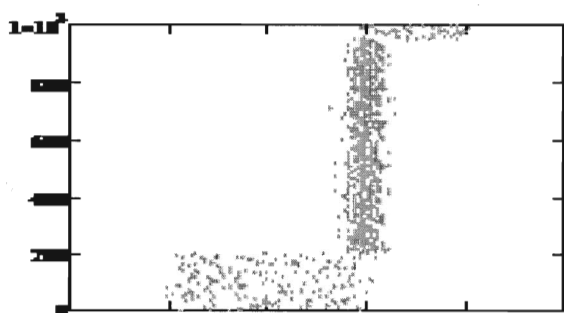


Session: 6

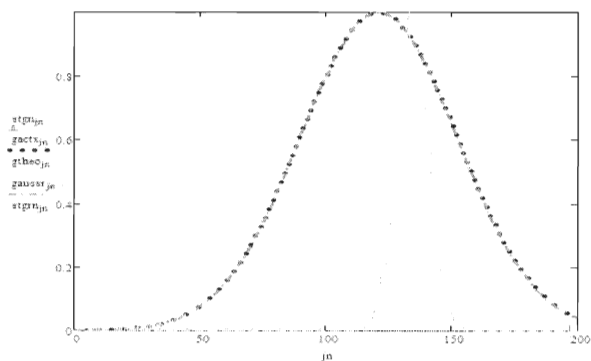
Presentation: 2

Indeed, the least square algorithm gives the most likelihood values starting from the given data. A progressive elimination of the more distant values from the founded ones leads to a situation in which in two successive steps the same values are obtained, so the procedure can be stopped.

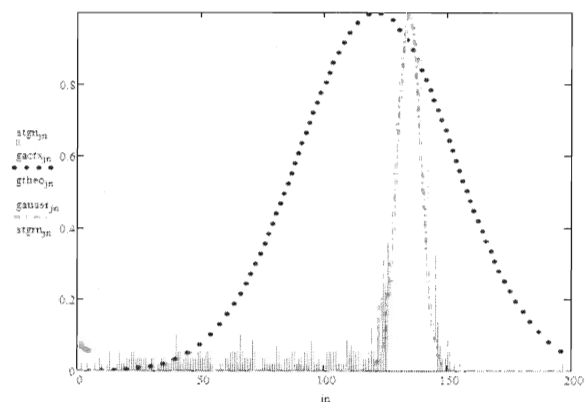
Figure 3 – An example of the proposed algorithm applied to the simple monodimensional case.



a) The sequence of normal distributed values (middle) with some superimposed random values (up and down).



b) Histogram of the above values (cyan), the Gaussian function of the normal distributed values (green dashed line) and the Gaussian calculated from the mean of the whole quadratic average deviations (violet dashed line and blue dots).



c) Histogram of the eliminated values (red), histogram of the remaining data (cyan) after the application of the proposed algorithm, Gaussian function of the overall data (blue dots) and of the remaining data (violet dashed line).