Doc. CCT/01-31

AN ALGORITHM FOR ON-LINE OUTLIER REJECTION BY SEQUENCE-ANALYSIS IN DATA ACQUISITION

F.PAVESE AND D.ICHIM
CNR, Istituto di Metrologia "G.Colonnetti",
Torino, Italy
E-mail: f.pavese@imgc.to.cnr.it, d.ichim@imgc.to.cnr.it

P.CIARLINI
CNR, Istituto per le Applicazioni del Calcolo "M.Picone",
Roma, Italy
E-mail: ciarlini@iac.rm.cnr.it

In automatic acquisition of thermal data, a valid sample is generally made up of several instrumental readings. These readings are generally reduced to a single value by simple methods, such as averaging. To avoid ambiguous results or a computation time too long in comparison with the experimental purposes, the data acquisition has to be performed automatically to reproduce the physical property possibly independently on outliers values. The paper introduces an algorithm, named a sequence-analysis outliers rejection (SAOR) that takes into account the most usual problems affecting the measurand during the acquisition, i.e. a non linear drift , sequences of outliers due to noise peaks. The algorithm uses the ordering of the sequence and of the "distances" between successive readings. The case of equispaced data is discussed. Results on tests performed for this case are reported using simulated thermal data affected by sequences of outliers.

1. Introduction

To describe a thermo-physical property y by a mathematical model f(T), several samples of the property are to be acquired in a certain range of T. The number M of the model parameters represents the minimum number of values T_i where the property must be measured. In the following, we call "i-th sample " of the property the value y_i in the experimental data (y_i, T_i). The number S of samples is always larger than the number M when the uncertainty associated to them must be evaluated. However, the increase of the sample number should kept to a minimum value, since the cost in resources of each sample is generally high and, also, the necessary experimental preparation may take a long time.

The experiment for acquiring each sample y_i, at the set-point value T_i consists of measurements of signals that are the output of a set of instruments. Each measurement may consist of several instrumental readings y_{ij} in sequence, made at different times t_j, (y_{ij}, T_{ij}, t_j)$_{j=1,...,P}$, where $T_{i1} = \ldots = T_{iP}$ for each i, (stationary case) or there is a small change of T in time (quasi-stationary case, showing the so-called baseline drift). In automatic acquisition the cost of each reading is obviously low.

In data acquisition two sampling strategies are usually followed:

A) acquiring single-reading property samples (y_i, T_i), $i=1,\ldots M$;

B) acquiring property samples $(t_j, y_{ij}, T_{ij},)_{j=1,..,P;\ i=1,...,M}$, each one made of a sequence of P readings that are *on-line* processed, often by the instrument firmware (usually by simply averaging them), to obtain $(\overline{y}_i , \overline{T}_i)$, $i=1,\ldots,M$.

Both strategies are not entirely satisfactory when outliers rejection is a "must".

(A) strategy minimises the acquisition time of a single sample, but requires the acquisition of a bigger number of samples, since any *off-line* outliers rejection procedure applied "*a posteriori*" to the set $\{(y_i, T_i)\}$ might reduce the number of valid data below the minimum value M. Hence the experimental cost can become unsatisfactorily high, as common in thermal metrology.

(B) strategy produces samples less influenced by the presence of outliers, since statistical analysis is performed *on-line* to the readings; however a simple average can still be highly affected by the presence of outlying readings. As a drawback, a too large number of readings can also too much increase the overall measurement time, for example in the case of calibration of a huge number of thermometers. General methods of managing outliers or unusual values in a dataset can be found in the literature in book addressing to diagnostics or robust statistics [1]. The latter uses inferential methods that are rather insensitive to certain types of failures in the parametric model of the data. Diagnostics has traditionally a different approach, since it attempts instead to identify the incorrect assumptions, allowing the analyst to refit under a new set of assumptions. Both approaches mainly deal with single outliers and less frequently with multiple occurrences of outliers.

For the outlier identification, the simplest statistical methods are based on the Normality assumption for the zero-mean random errors affecting the data. Some methods label the suspected outliers, for example on the basis of the z-score test, while others accommodate the outliers by means of the trimmed or windsorized mean/median [2].

When also a baseline drift affects the dataset, as in the quasi-stationary case of thermal data, the zero-mean assumption fails and the previous methods cannot be directly applied. A current strategy to estimate the baseline is to apply regression followed by a testing of the studentised residuals to identify the outlying values. However, it has to be embedded in an iterative procedure, which is time consuming.

To *on-line* manage outliers, a different approach may be used, a filtering process [3] performed in firmware. However it can be successfully used only when the noise frequency spectrum is sufficiently well known. This is not a common case in thermal measurements.

In automatic acquisition of a sequence of readings, a suitable strategy is necessary when a non-linear baseline drift and outlying readings simultaneously may affect the measurand.

A two-steps procedure has been designed, to improve strategy B): for each sample, i) a pre-processing step rejects outlying readings, ii) the baseline drifts is suppressed by means of the regression on “cleaned" readings, to obtain the valid

output sample value. The first step consists of a pre-processing algorithm, named "sequence-analysis outliers rejection" (SAOR), that aimes at being easily fitted in the instrument firmware. It is based on the *distances between successive readings* for the analysis of the outlying sequences. The paper introduces the algorithm and reports on the results obtained on simulated test sequences of uniformly spaced readings.

2. Assumptions on signal characteristics

The instrumental system generally has a typical and stationary noise level. However, during the acquisition of the readings, noise spikes may occur due to power switching in the mains or because of activities in the nearby laboratories or for many other reasons. The occurrence of noise spikes can be quite different from case to case: in a sequence of readings this spot noise may affect only a single reading or several consecutive readings, and may occur more than once over a certain sequence of readings. The number of readings of the sequence necessary to obtain a valid sample should be optimised for each case. On the other hand, the necessity to acquire more of one reading to compute a valid sample makes, in general, the acquisition time longer: this increase in time must be minimised. In addition, the signal cannot, in general, be assumed to be stationary during the sequence of readings, i.e., the reading sequence is affected by a non-random low-frequency variation (baseline drift).

The characteristics of the signal shown in Fig.1 can be summarised as follows:

- data acquisition consists in an ordered sequence of readings from an instrument (e.g., a digital voltmeter, bridge, …);
- the signal is assumed to be quasi-stationary during a sequence of readings;
- three are the reasons of the signal value changes, of different and generally independent origin:
  1. **signal drift**, due to the signal being quasi-stationary: it is assumed to have a frequency spectrum limited to sufficiently low frequencies to allow modelling with a low-order polynomial;
  2. **signal noise**, due to electrical instrumental noise: random variations of the readings with a broad frequency band (typically white noise), assumed to be *stationary* within each sequence of readings
  3. **pulse noise**, due to spot events in the environment, of magnetic, electric (switching), mechanical (shocks), etc. nature. It occurs as random spikes assumed not affecting more than a maximum number K of consecutive readings (generally a few), but with no limits as to the number of occurrences, the position within the sequence and the magnitude and sign.

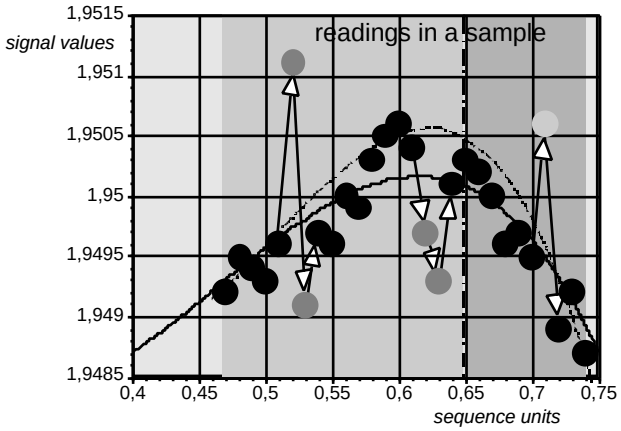

Fig.1 – A typical acquired sample of thermal data, made of a sequence of instrumental readings. They show a trend, due to a low-frequency signal drift, and a system noise. Due to external reasons, noise peaks can randomly occur (grey dots in the figure), affecting one or more consecutive readings. Drift is shown, fitted either with (solid curve) and without (broken curve) taking into account the pulse noise.

Two problems in computing from the readings the "sample value" with its associated uncertainty estimate can produce misleading results: (a) the number, position and size of the outliers are unknown; (b) the low-frequency drift of the signal can be of the same order of magnitude as the outlying readings. Therefore, methods for outlier rejection based only on the values of the readings cannot discriminate between the components 1) and 3). Moreover, any simple statistical estimate of the data –eg, the standard deviation– is affected by the presence of outliers and signal drift and, therefore, cannot be simply used as a robust threshold to discriminate the outliers.

3. The pre-processing algorithm SAOR

According to the assumed signal characteristics, a pre-processing algorithm has been designed, able to clear off outliers due to type 3 noise, to be robust against type 2 noise and to allow for subsequent compensation for type 1 drift.

The SAOR algorithm does not compute the baseline, but analyses the sequences of distances between consecutive instrumental readings. For readings non-uniformly spaced in time, the divided differences or the Euclidean distances should be adopted to take into account different scales in the two variables. However, since equispaced data are the most common case in automatic acquisition, SAOR was presently developed for this case, where the vectorial distance between two consecutive readings simply is the projection on the *y*-axis.

The algorithm input are: the number N of output data $\{(\tau_1, \psi_1), \ldots, (\tau_N, \psi_N)\}$, the number $M > N$ of input initial instrumental readings (t_i, y_i), the assumed maximum number K of consecutive outlying *reading* values, and a threshold value

to discriminate possible outliers according to the knowledge of the signal characteristics. The algorithm steps are:

1. acquire a number $M > N$ of instrumental readings in sequence: $\{(t_i , y_i)\}$, $i = 1,\ldots, M$;
2. compute the projection on the y axis of the (vectorial) distance between two consecutive readings. Associate to each $d_i = |y_{i+1} - y_i|$, $i = 1,\ldots, M$-1 and $s_i = \pm 1$ for $(y_{i+1} - y_i) \gtrless 0$. In the following d_i is called "distance" and the sign s_i is called "direction";
3. given a threshold d_0 for these *distances*, define as *candidate outliers* those distances for which $d_j > d_0$, and compute the number C of their occurrences;
4. if $M–C–1 < N$, acquire supplementary instrumental readings as necessary and return to step 2;
5. starting from *each candidate outlier*, say the j-th, analyse the sub-sequence of length $L = K+1$ for consecutive distances, $d_j \ldots, d_{j+K}$;
6. define as **not** an *outlier* a candidate distance, whenever the relative sub-sequence does **not** satisfy one of the following conditions:
   a) more than one candidate exists, except for the first and last distances;
   b) at least one change of the direction occurs.
7. for each one of the remaining sub-sequences, that do contain outlier readings, use a "truth table" of decisions to identify the outlying *readings* (from one to K). Clear off the outliers readings to output the cleaned sequence of data (τ_κ, ψ_k), $k = 1,\ldots,N$.

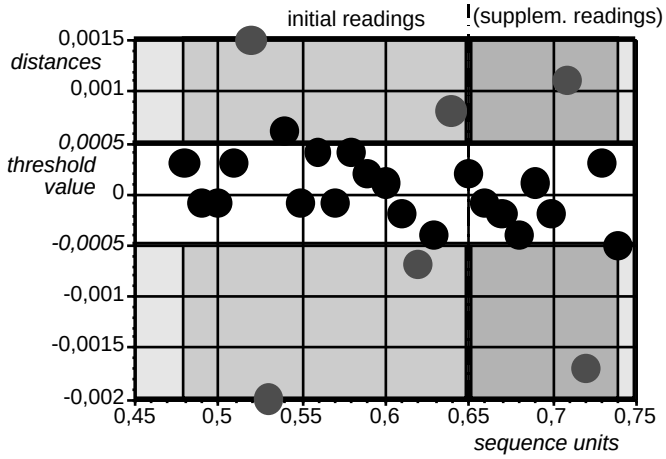

Fig.2 – For the values in Fig.1, a threshold value is defined to identify the "candidate outliers".

To obtain the sample value and its uncertainty, regression can then be applied to the "cleaned" sequence of N output data (gray readings in Fig.1) using a suitable

functional model of the baseline. In step 3 the value of the threshold depends on the scale of the measures, as discussed in the next chapter.

4. A first implementation for thermal data

A main choice in the algorithm is about the suitable location statistics used to select the candidate outliers, together with the value of the needed threshold (step 3), since it represents the discrimination level for the candidate outliers. We used the *augmented median.* The median, it is well known, effects a robust discrimination of the "regular" distances (ordered by size) from the potential outliers. However, rejecting all the distances above the simple median value would be too crude and "expensive", because half of the readings are rejected. Consequently, a higher threshold was defined, augmenting the median by a factor depending on the standard deviation of the distances below the median, multiplied by an adjustable constant.

In this first implementation, it is assumed that the noise spikes affect two consecutive readings ($K = 2$) at last, and that the distance between two consecutive readings ($t_{i+1} - t_i$) is constant. Hence, two vectors are associated to represent the values: $d_i = |y_{i+1} - y_i|$ on the y-axis and the *sign* of this difference, to indicate a direction (in step 2).

From the point of view of the cost of the data acquisition, the size of the increase $A = M - N$ of the number of readings, with respect to the target number N, is another decision critical for the time constrains. In the present work, this number was kept to the minimum required by condition b) in step 6: $A = 2$.

Table I: Sub-sequence analysis ("truth table") for SAOR algorithm: $K = 2$; $L = 3$, four readings.

Sub-sequence					Outlier: reading #	Index	Notes
Outlying distance °			Sign change				
1°	2°	3°	#1	#2			
true	false	false	n.a.	n.a.	none	0*	
true	false	false	n.a.	n.a.	1 (4)	0	First (or last) distance
true	true	false	false	true	2	9	
true	true	false	true	false	2	10	
true	false	true	false	true	2&3	5	
true	false	true	true	false	2&3	6	
true	true	true	true	true	2	15	
true	true	true	true	false	2&3	14	
true	true	true	false	true	2&3	13	
true	true	true	false	false	none	12*	Simple drift

° a sub-sequence always starts with an outlying distance. * same for indexes 1,2,3,4,7,8 and 11

Speed is important, since the algorithm should be adopted also at the very basic level of the instrumental reading, i.e. in firmware. Hence the present implementation was optimised to allows stopping the procedure as soon as the presence of further outliers can be excluded: e.g., if none or only one candidate is found, the procedure obviously ends. The analysis of the sub-sequences is the core of the algorithm. A sub-sequence of K+1 distances starting with a candidate outlier gives rise to a finite number of combinations, 2^{K+1}. Since they can be listed in full, the action that must be taken for each one can be associated, according to a "table of truth". As shown in Table I, for $K = 2$ there are only 8, out of 16 possible sub-sequences containing one or more outliers defined according to the above definition. Of course, for longer sequences the number of the sub-sequences that contains outliers rapidly increases, and so the analysing time.

5. Tests and results

A routine in FORTRAN 77 has been written to implement the SAOR algorithm, not including the post-processing regression. In order to check the robustness of the procedure against the presence of multiple outliers, a simulation study has been performed to generate data with multiple outliers.

The basic simulated sequence includes a random noise component and a non-monotonic baseline drift (Type 1) affecting the reading values by a factor of about 2, but no candidate outlier. Then, this basic sequence is automatically altered by including outlier elements, random in number, up to a maximum value O_{max}, in position in the sequence, in relative size, $R = y_{max}/y$, and in "sign".

By varying all these simulation parameters and the discrimination threshold, it is also possible to check the ability of the algorithm to identify outlying values closer and closer to the random noise level, in order to check the sharpness of the discrimination threshold. The chosen values of the other simulation parameters are: $N = 18$; $M = N+2$, $O_{max} = 4$ (maximum number allowed for the *candidate outliers* in a sub-sequence: $O_{max} < (M/2 - 1)/2)$, with $O_{max} > K$.

The routine takes a few microseconds to run on a modern PC (>300 MHz clock speed, RISK processor iMac). Tests of groups of 10 000 trials gave essentially the same results, therefore no extension of the tests above 60 000 random sequences was considered. The number of algorithm failures, i.e., of mismatches between the imposed outliers and the ones recognised by the algorithm, was tested as a function of the maximum outlier size relative to the signal value, R. Figure 3 shows the effect of chosing O_{max} from 1 to 4.

When O_{max} matches the discrimination threshold $K = 2$, i.e., when $O_{max} = 1$, the efficiency of the algorithm is 100%.

When $O_{max} > 1$, there is a non-zero statistical probability for the outliers within the sequence to form sub-sequences affecting more than two consecutive readings, violating the present assumptions of the algorithm. This is the main reason for the

resulting ≈ 0.5 % inefficiency at O_{max} = 4 for outliers of size much larger than the signal (R = 15). Inefficiency (< 0.1 %, as shown in Fig.3 for O_{max} = 2) can also be due to the sequences involving special cases, such as outliers sequence starting from the first reading: they could be taken into account by a more complicated decision tables, but the cost/benefit ratio may be too high in most experimental cases.

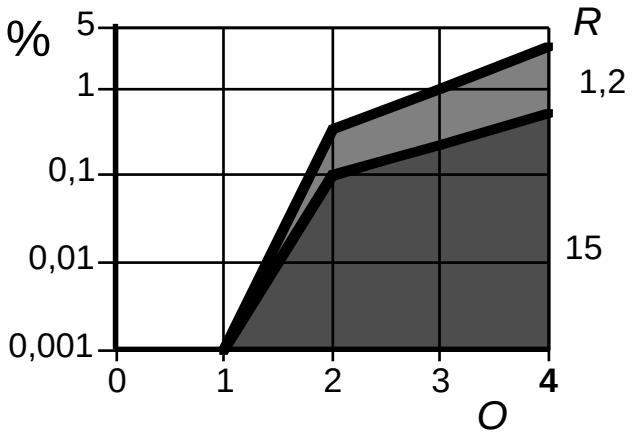

Fig.3 – SAOR efficiency in outlier detection as a function of the maximum number O of outliers injected, for two extreme values of the outlier relative maximum size R.

When the R value is lowered to approach the signal value, the inefficiency grows gradually for R = 3-2 to a value of about 3 % (O_{max} = 4). When $R \rightarrow 1$ the algorithm obviously breaks, since most of the readings are declared candidate outliers.

6. Conclusions

A pre-processing algorithm, named SAOR, for outliers rejection during automatic data acquisition has been introduced. It is based on the distances between consecutive readings and on the analysis of unusual sub-sequences of these distances. The computation of the baseline drift is postponed to an off-line step.

SAOR performs the analysis according to a truth-table in order to discriminate artefacts due to the measurement instruments/environment from those due to the characteristics of the property under study. A first implementation for thermal data affected by sequences of outliers has been discussed. The robustness of the algorithm against the presence of sequence of outliers was tested via data simulation. It allows to discriminate outliers readings defined as a broad class of occurrences in a sequence of uniformly spaced data. With a few assumptions on the

signal, the robustness of the algorithm was 99.9-99.5 % when 4 outliers occur in a sequence.

Extensions of the algorithm to non-uniformly spaced data are possible, where suitable definition of "distance" and of the relative threshold are to be adopted to take into account different scales in the two variables.

In conclusions, since the SAOR algorithm is simple and fast, and it is independent on the presence of a non-linear baseline drift, it can be fitted in instrumentation firmware, to obtain sample values and their uncertainty not affected by the presence of outlying signal values.

Acknowledgements

We wish to thank G.Regoliosi of IAC-CNR for his helpful contribution to discussions.

References

4. W. Stahel and S. Weisberg (Eds) *Directions in Robust Statistics and Diagnostics – Part I- II*, Springer-Verlag (1991)
5. J. S. Simonoff *Comm. Statist.- Simula. Computa.*, **13**(2),275 (1984)
6. G.E.P. Box and G.M. Genkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco (1976)