Serhii W. ZABOLOTNII [1], **Zygmunt Lech WARSZA** [2]
[1] CHERKASY STATE UNIVERSITY OF TECHNOLOGY, Cherkasy, Ukraine
[2] RESEARCH INSTITUTE OF AUTOMATION AND MEASUREMENTS PIAP, 202 Jerozolimskie Ave., 02-486 Warsaw, Poland

Semi-parametric modification of cumulative sum algorithms for the change-point detection of non-Gaussian sequences

Abstract

The expansion of logarithm likelihood ratio in the stochastic series to find the sequential change-point detection of non-Gaussian sequences is used. The moment criteria of the minimum of upper limit error probabilities sum to find the expansion coefficients is applied. The proposed method is a semi-parametric type of cumulative sum (CUSUM) algorithm which needs of higher-order statistics. Results show that polynomial algorithms are more effective in comparison with similar non-parametric procedures.

Keywords: change point, CUSUM algorithm, non-Gaussian sequence, stochastic polynomial, high order statistics.

1. Introduction

In modern systems of measurement and technical diagnostics of random processes the important are problems of sequential analysis, in which it is necessary to detect abrupt changes (disorder) in properties of these processes. For analyzing a fixed volume of sample containing all the information about the tested object the retrospective (a posteriori) statistical methods are used [1]. But the measurement and diagnosis in real time is the more typical problem in the automatic measurement and control of industrial processes, fault diagnosis of technical systems, monitoring and control of the environmental, geophysical and many others processes [2]. In such cases the statistical sequential analysis is applied. This range of applications requires the development of a variety of models and methods of statistical processing. However, the most theoretical works devoted to the problem of "disorder" are focused on the Gaussian processes. The considerable part of the real process differs from this model.

Parametric methods are based on the probability density function (PDF). The main of their problem is the requirement of a priori information about the form of the distribution laws, as well as the high complexity of their implementation. Therefore, a lot of the research in this area is associated with the construction of non-parametric methods of the change-point detecting, not tied to specific types of PDF [3]. The price of such simplification is the worsening of qualitative characteristics in comparison with the optimal parametric methods. Thus, there is an actual problem of creating the new approach, which on the one hand would allow to take into account the properties of the real non-Gaussian random processes and would be potentially adaptive. On the other hand, this method should be characterized by the simplicity, both in terms of training, tuning and the algorithmic implementation.

As one of perspective directions to apply is the mathematical statistics of higher then second order moments or cumulants. An example is the application of this method to construct the probabilistic models in various areas of the change point detection. Such works are devoted to the detection of the moment of arrival the acoustic emission signals [4], to the segmentation of video streams [5], and to the detection of hackers in telecommunication networks [6].

In this work, a new approach to the problem solving associated with the processing of non-Gaussian signals is given. These signals are described by their moments and cumulants and the apparatus of stochastic polynomials is used [7]. The aim is to modify a known cumulative sum (CUSUM) procedure by expanding the log-likelihood ratio (LLR) in the stochastic series with coefficients which are optimized on the basis of the moment quality criteria for testing the statistical hypothesis [8].

2. Formulation of the problem

Suppose there is a sequence of independent random variables x_1, $x_2 \cdots x_n, \cdots$ obtained by the regular sampling of the tested process. The probabilistic model can be described by the mean value θ, variance σ^2 and cumulant coefficients γ_l up to a given order $l = \overrightarrow{3, 2s}$. Up to some (a priori unknown) point of the discrete time τ–1 values of these parameters are constant (hypothesis H_0). Then, at the time τ, one or more parameters abruptly changes its value (hypothesis H_1). The challenge is to detect as quickly as possible through continuous analysis of sample values $x\nu$ the disorder while ensuring a fixed probability (the average time of occurrence) of a false alarm.

3. Polynomial modification of CUSUM algorithm

Classic version of CUSUM algorithm, used for the sequential change-point detection, based on the statistics generated on the basis of the logarithm of the LLR [2] is:

$$\Lambda_v = \sum_{j=1}^{v} \ln \frac{f_1(x_j)}{f_0(x_j)}, \; v = 1,...,n \, , \tag{1}$$

where $f_0(\cdot), f_1(\cdot)$,- density distribution before and after change.

Page [9] proposed the rule, which discovers the moment of disorder. It has the form:

$$\hat{\tau} = \inf\left\{ n \geq 1 : \Lambda_n - \min_{0 \leq j \leq n} \Lambda_j \geq h \right\}, \tag{2}$$

where; h>0 is the detection threshold.

In practice, it is more convenient to use a modified algorithm called "holding barrier" [2], which uses a recursive form of statistics:

$$g_n = \left(g_{n-1} + \ln \frac{f_1(x_n)}{f_0(x_n)} \right)^+ \tag{3}$$

where g_0=0, $(A)^+ = \max\{0, A\}$.

The nonparametric modification of sequential change-point detection, known from literature, is based on the statistics of cumulative sums [2].

In this work, the semi-parametric version of CUSUM algorithm is applied. The non-Gaussian statistics in the form of high-order cumulant coefficients obtained by decomposition of the LLR into stochastic power series, proposed by Kunchenko [7], is used

$$\Lambda_v = \sum_{j=1}^{v} \ln \frac{f_1(x_j)}{f_0(x_j)} = k_0 + \sum_{j=1}^{v}\sum_{i=1}^{\infty} k_i x_v^i \, , \; v = 1,...,n \, , \tag{4}$$

where: $f_0(\cdot), f_0(\cdot)$ - density distributions before and after the change point.

To find the coefficients k_i of the expansion (4), so-called the criteria for the formation of decision rules for statistical hypothesis testing are used. One of them is the criterion of the minimum upper limit of the sum of the wrong decision probabilities (criterion Ku). It is used in [7, 8] and defined by:

$$Ku = \frac{D_0 + D_1}{[E_1 - E_0]^2}, \qquad (5)$$

where: E_r, and D_r, (r=0, 1) are mathematical expectations and variances respectively, of a decision rule significance test H_0 against the alternative test H_1.

If as decision rule, the comparison of LLR with the threshold $0.5(E_1+E_0)$ is used, then this probability criterion has a minimum value, which can be written as:

$$Ku_{\min} = \frac{\sum_{v=1}^{n}\left(D_{0v} + D_{1v}\right)}{\left[\sum_{v=1}^{n}\left[I_v(1:0) + I_v(0:1)\right]\right]^2} = \frac{D_{0v} + D_{1v}}{nI_v(1:0)} = J_n^{-1}, \qquad (6)$$

where: $I_v(1.\ 0)$ - Kullback-Leibler average information contained in a value of v-th sample for making decision in favor of the hypothesis H_0 against the alternative H_1; D_{rv} - is equal to the variance LLR of v-th sample value at an appropriate hypothesis H_r, r=0, 1; J_n - is called the maximum amount of information contained in the sample of n elements.
Amount of information J_n is used for testing the difference between hypotheses H_0 and H_1 by the selected quality criteria.

For the limited number of terms of series (6), the approximation of LLR by polynomial of s degree is used, i.e.

$$\Lambda_v^{(s)} = k_0 + \sum_{j=1}^{v}\sum_{i=1}^{s} k_i x_v^i, \ v = 1,...,n. \qquad (7)$$

Coefficients k_i, $i = \overrightarrow{1,s}$, which are optimized by the criterion Ku, should be found by solving the system of linear algebraic equations [8]:

$$\sum_{j=1}^{s} k_j F_{i,j} = m_i - u_i, \ i = \overrightarrow{1,s}, \qquad (8)$$

where: $F_{i,j} = m_{i+j} - m_i m_j + u_{i+j} - u_i u_j$, $u_i = E\{x_v^i / H_0\}$ and $m_i = E\{x_v^i / H_1\}$ are the initial moments of the random variable under the appropriate hypotheses.

Optimal value of coefficient k_0 (with the appropriate degree s) has the form:

$$k_0 = -\frac{n}{2}\sum_{i=1}^{s} k_i\left(m_i + u_i\right). \qquad (9)$$

The limitation of polynomial degree leads to errors of LLR value. The replacement of it in the final rule by the polynomial approximation reduces the amount of information for testing the difference between hypotheses. At the optimal values of the coefficients k_i, $J_n^{(s)}$ is given by:

$$J_n^{(s)} = n\sum_{i=1}^{s}\sum_{j=1}^{s} k_i k_j F_{i,j} = n\sum_{i=1}^{s} k_i\left(m_i - u_i\right). \qquad (10)$$

With increasing of the polynomial degree s $J_n^{(s)}$ has the limit:

$$\lim_{s\to\infty} J_n^{(s)} = J_n. \qquad (11)$$

Thus, the value $J_n^{(s)}$ can be interpreted as information convergence criterion of stochastic polynomial (7) to the LLR value in terms of its use in the construction of decision rules that are optimal for the selected criterion.

Using the approximate representation of LLR as a stochastic polynomial (7) and taking into account the optimal (according to the criterion Ku) coefficient k_0 of the form (9), (by analogy with (3)), a recursive polynomial form statistics can be written as:

$$g_n^{(s)} = \left(g_{n-1}^{(s)} + \sum_{i=1}^{s} k_i\left[x_n^i - \frac{m_i + u_i}{2}\right]\right)^+. \qquad (12)$$

Obviously, statistics (12) represents a semi-parametric polynomial modification of the CUSUM algorithm. It is so, because in its construction the information about the laws of probability distribution of a random sequence is not used. Formula (12) is based on incomplete probabilistic description as a sequence of moments up to 2s-th order.

4. Modelling of polynomial CUSUM algorithms

Based on the above results the software package for statistical modeling of the proposed semi-parametric CUSUM procedures was developed. It is dedicated for detecting a disorder mean and/or variance of non-Gaussian random sequences. This package allows to make both: the single experiments to detect a disorder, and multiple tests (according to the Monte Carlo method) for experimentally comparing the accuracy of the proposed polynomial algorithms.

The main criterion for efficiency of the sequential detection algorithms is the average value of time needed for the detection of disorder while providing the same probability (the average time of occurrence) of a false alarm.

Fig. 1 shows as example the results of the simulated change-point detection procedures based on polynomial CUSUM algorithm with "holding barrier" form (12). It is obtained for the relative values of the mean changes $q=(\theta_1-\theta_2)/\sigma_0=1$ and variance changes $d=\sigma_1/\sigma_2=2$.

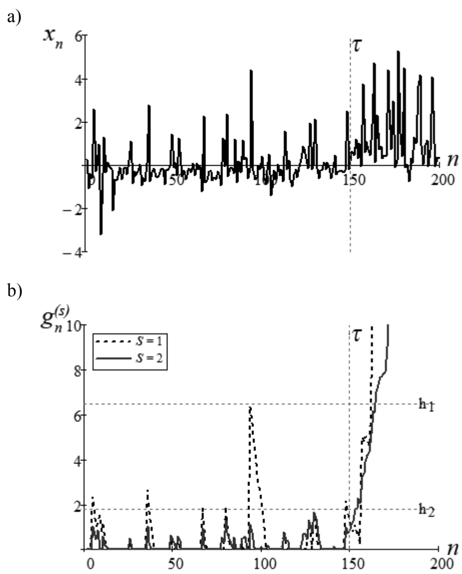

Fig. 1. Example of the sequential change-point detection: a) non-Gaussian sequence with a disorder; b) polynomial CUSUM statistics with a "holding barrier".

The cumulant coefficients of skewness γ_3=2 and of kurtosis γ_4=5, characterize the degree of *non-Gaussianity* of the random sequence. The presented results show a decreasing of time to make a decision on the change-point detection for the polynomial CUSUM algorithm, synthesized at a power s=2, if compared with the CUSUM algorithm, obtained at s=1. It should be also noted that the linear CUSUM algorithm can be used as a nonparametric procedure, which is optimal for detection the disorder of the mean value in case of a Gaussian distribution of the elements of a random sequence.

Increasing the efficiency factor explains a significant reduction in the threshold values $h_2<h_1$ of decision-making of the occurrence of a disorder (Fig. 1b). It is provided with a fixed probability (mean time to emergence) of false alarms for both statistics. Results of single experiments do not allow to compare adequately the accuracy of decision-making algorithms. Therefore, the empirical estimate of the range of winning is the average delay time $T=\tau-\hat{\tau}$ for the detection of a disorder, which can be obtained by series of repeated experiments with the same initial values of the model parameters.

Fig. 2 shows the mean values (for W=2000 trials) of relations delays $T^{(2)}/T^{(1)}$ of the polynomial (at s=2) and non-parametric CUSUM algorithms.

a)

b)

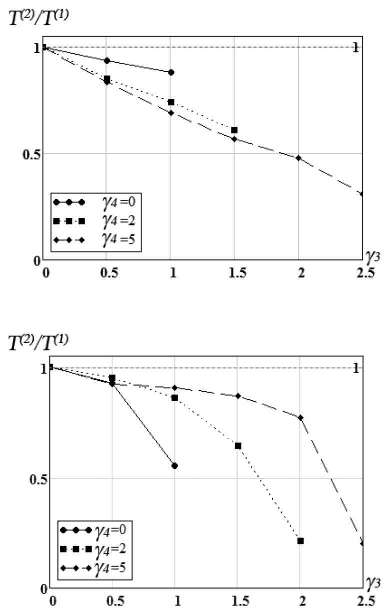

Fig. 2. The experimental values reduce the average time to detect disorder, depending on the characteristics of random sequence:
a) disorder of the mean value (q=1); b) a disorder of the variance (d=2)

These curves were obtained at discrete values of the skewness coefficient (of increments $\Delta\gamma_3$=0.5) and for the given choice of the decision threshold to ensure probability of false alarm $\alpha=10^{-3}$, the same for both algorithms. It should be noted that limits of the ranges of plots shown in Figure 2b are due to inequality $\gamma_3^2 \le \gamma_4+2$.

Presented in Fig. 2 curves are obtained on the basis of the Monte Carlo method. They indicate the substantial increase of the effectiveness of the proposed polynomial algorithms. Degree of that depends primarily on the value of skewness (at γ_3=0 increase of the amount of information is not observed). Another important result is that the synthesized algorithms manifest themselves more effectively for small values of the relative disorder. That makes them potentially more accurate in the detection of weak structural changes.

Thus, the results of statistical modeling confirm the theoretical assumption about the growth of the effectiveness of the proposed approach for constructing polynomial algorithms such as CUSUM. This growth is achieved by incorporation of the more information about the probabilistic nature of random sequences, such as values of cumulant coefficients of skewness and kurtosis. Natural price to pay for this effect is a certain complexity of the algorithm (non-linear processing), as well as an additional requirement for a priori information to its configuration.

5. Conclusions

Results of the theoretical and experimental research point to the high efficiency of the log-likelihood ratio (LLR) decomposition for the synthesis of sequential change-point detection algorithms of mean value disorder with simple implementation when random sequences have the asymmetric probabilistic nature.

Scientific novelty of the obtained results is the new approach to construction of semi-parametric decision-making procedures for analysis of non-Gaussian random sequences based on Kunchenko stochastic polynomials. Among the possible directions for further research of constructing sequential change-point detection procedures by means of stochastic polynomials are following:

- increase of the degree of stochastic polynomial for more effective solutions of symmetrical non-Gaussian sequences;
- use of other moment's criterion for create decision rules (e.g. Neyman-Pearson torque criterion [8]);
- modification of classical procedures based on the log-likelihood ratio, e.g. GRSh (Girshik-Rubin-Shiryaev) algorithm [3].

6. References

[1] Zabolotnii S. W., and Warsza Z.L.: Semi-parametric estimation of the change-point of mean value of non-Gaussian random sequences by polynomial maximization method. Proceedings of 13th IMEKO TC10 Workshop on Technical Diagnostics. Warsaw UT, pp. 189-194, 2014.

[2] Basseville M., and Nikiforov I. V.: Detection of abrupt changes: theory and application. Prentice-Hall, New Jork USA, 1993.

[3] Brodsky B., and Darkhovsky B.: Nonparametric Methods in change-point problems. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.

[4] Lokajicek T., and Klima K.: A first arrival identification system of acoustic emission (AE) signals by means of a higher-order statistics approach. Measurement Science and Technology, vol. 17, pp. 2461-66, 2006.

[5] Yih-Ru Wang: The signal change-point detection using the high-order statistics of log-likelihood difference functions. Proccedings of Acoustics, Speech and Signal Processing, ICASSP IEEE International Conf. pp. 4381-4384, 2008.

[6] Hilas C. S., Rekanos I. T., and Mastorocostas P. A.: Change point detection in time series using higher-order statistics: a heuristic approach. Math. Problems in Engineering, ID 317613, 2013.

[7] Kunchenko Y.: Polynomial Parameter Estimates of Close to Gaussian Random variables. Aachen, Shaker Verlag, 2002.

[8] Kunchenko Y. P.: A moment performance criterion of a decision making for testing simple statistical hypothesis. IEEE, International Symposium on Information Theory, Ulm, Germany, p. 40, 1997.

[9] Page E. S.: A test for a change in a parameter occuring at an unknown point. Biometrika, vol. 42, pp. 523-527, 1955.

Received: 17.09.2015 *Paper reviewed* *Accepted: 03.11.2015*

Assoc. Prof. Serhii V. ZABOLOTNII, PhD, eng

Born in 1973, Cherkasy, Ukraine. Graduated: Faculty of Information Technologies and Systems of Cherkasy Technological Institute: diploma of radio-eng. (1995). PhD in information-measuring systems (2000), DSc of Technical Sciences, 2015. Work: Department of Radio Engineering, Cherkasy State Technological University: Assistant 1998 – 2003, docent from 2003, vice-Rector 2015. Author or co-author above 100 scientific publications, 1 monograph and 7 patents.

e-mail: zabolotni@ukr.net

Assoc. Prof. Zygmunt L. WARSZA, PhD, eng

Born in Wilno. Graduated at Electrical Faculty of Warsaw Technical University: MSc. El. Eng. 1959. PhD in El. Measurements 1967. Assoc. Prof. (docent) from 1970. Worked in 3 Technical Universities and 3 research institutes 1959 -2002. Presently with Industrial Research Institute of Automation and Measurements (PIAP) Warszawa. President of Polish Society of Metrology. Author or co-author above 220 papers, 4 monographs, many reports and 11 patents on new instruments.

e-mail: zlw@op.pl