# Semi-parametric estimation of the change-point of mean value of non-gaussian random sequences by polynomial maximization method

Serhii W. Zabolotnii[1], Zygmunt Lech Warsza[2]

[1] *Cherkasy State University of Technology, Ukraine (zabolotni@ukr.net)*
[2] *Research Institute of Automation and Measurements PIAP*
*(Al. Jerozolimskie 202, 02-486 Warszawa, Poland, e-mail: zlw@op.pl)*

***Abstract-*** An application of the maximization technique in the synthesis of polynomial adaptive algorithms for a posterior (retrospective) estimation of the change-point of the mean value of random sequences is presented. Statistical simulation shows a significant increase in the accuracy of polynomial estimates, which is achieved by taking into account the non-Gaussian character of statistical data.

## I. Introduction

One of the important tasks of the diagnosis of random processes is the measurement of the point at which the properties of the observed process are subject to a change (disorder). This point is called "the change-point". Statistical methods of detecting the change-point can be used in real time or a posteriori. The latter ones are also called the retrospective methods. A posterior statistical estimation is based on the analysis of a fixed volume sample of all information received from the diagnosed object. This approach requires a longer time of reaction, but it provides a more reliable and accurate estimation of the change-point [1].

A posterior estimation of the time of changes the parameters of stochastic processes is needed in many practical applications, such as the diagnosis of some industrial processes, the detection of climate change [2], analysis of the genetic time series [3], identification of intrusion in computer networks [4], segmentation of speech signals and messages of social networks [5]. For such a wide range of tasks the development of a large variety of mathematical models and statistical processing tools is required. It should be noted that most of the theoretical studies connected with the estimation of  the change point is focused on the class of random processes described by the Gaussian probability distribution. However, the real statistical data are often different from the Gaussian model. The classical methods, which are based on the probability density, are called the parametric methods. The main problems in parametric approaches (Bayesian and maximum-likelihood) are connected with the requirement of an a priori information about the form of distribution, as well as with the potentially high complexity of their implementation and the analysis of the properties. Thus, a significant amount of the contemporary research concerns the construction of applied statistical methods which would allow to remove or minimize the required amount of the a priori information. Such methods are based on robust statistical processing procedures that are insensitive to "non-exactness" of  probabilistic models, or on nonparametric criteria, independent of specific types of distributions. The price for ”omision” of probabilistic properties in handled statistical data is the deterioration of quality characteristics in comparison with optimal parametric methods [6].

The use of higher-order statistics (described by moments or cumulants) is one of the alternative approaches in solving problems related to processing of non-Gaussian signals and data. This mathematical tool can find applications  in various areas, where the estimation of the change-points is important, e.g. in: defining the moment of arrival the acoustic emission signals [7], the detection changes of video signals [8], detection of signal changes in telecommunication networks [9].

In this paper there is considered the application a new unconventional statistical method, in solving problems of a posterior type estimation of change points. The method is called the polynomial maximization method (MMP) and it was proposed by Kunchenko [10]. It uses stochastic polynomials as the mathematical tool. The method used in conjunction with the moment-cumulant description allows to simplify substantially the process of synthesis of adaptive statistical algorithms.  Moreover – since the information about probabilistic properties of data is used -  an improvement  of the accuracy is obtained, i.e. reduction of  value of estimated variance and decrease of the probability of erroneous decisions are achieved.

Since the moment-cumulant description of the probability distribution is only approximate,  statistical method. Based on that description allow to obtain only asymptotically optimal results (i.e., its accuracy increases with the order of the used statistics. Thus, the proposed methods can be classified as semi-parametric.

The aims of this paper are the following:

- using the method of polynomial optimization to synthesis o algorithms of a posteriori estimation of the of change-point of the mean-value of the non-Gaussian random sequences,
- investigation of effectiveness of those algorithms, using statistical modelling.

## II. Mathematical formulation of the problem

Suppose there is a random sample $\vec{x} = \{x_1, x_2, ... x_n\}$. Elements of this sample can be interpreted as a set of $n$ independent random variables. The probabilistic nature of this sample can be described by the mean value $\theta$, variance $\sigma^2$ and cumulant coefficients $\gamma_l$ up to a given order $l = \overrightarrow{3, 2s}$. Up too some (a priori unknown) point of the discrete time $\tau$, the mean value is equal to $\theta_0$, and then, at the time $\tau + 1$ its value jumps to $\theta_1$. Our purpose is to estimate the value $\tau$ of the change-point of the mean value of the sequence, on the basis of the analysis of the sample. Various variants of the change point estimation may differ, depending on the availability of an a priory information about values of a variable parameter (before and/or after the change), as well as on the knowledge of the probabilistic character of the other parameters of the random sequence model.

## III. A posteriori estimation of the change-point of mean value by maximum likelihood method

One of the basic directions in investigations of a posteriori problems of the change point study is based on the idea of the maximization of the likelihood. It was elaborated in details by Hinckley [11]. He proposed a general asymptotic approach to obtain distributions of a priori change-point estimates by method of maximum likelihood (MML). Application of this approach requires an a priori information about the distribution law of statistical data, before and after the change.

For a Gaussian distribution it is known that estimation of the mean value by MML method is the same as a linear estimation by method of moments (MM), i.e.

$$\hat{\theta} = \frac{1}{n} \sum_{v=1}^{n} x_v \tag{1}$$

Estimate of the form (1) is consistent and not shifted. So non-parametric MM estimator can be used for estimation of the mean value of random variables of any arbitrary distribution. However, this assessment is effective only for the Gaussian model. For this probabilistic model the logarithm of the maximum likelihood function (MML) with known variance $\sigma^2$ is transformed [1] into statistics of the form:

$$T_r(\theta_0, \theta_1) = r \sum_{v=1}^{r} (x_v - \theta_0)^2 + (n - r) \sum_{v=r+1}^{n} (x_v - \theta_1)^2 \ . \tag{2}$$

$T_r(\theta_0, \theta_1)$ has a maximum in a neighbourhood of the true value of the change-point $\tau$. Thus, the desired change-point estimate can be finding by algorithm:

$$\hat{\tau} = \arg \max_{1 \le r \le n-1} T_r(\theta_0, \theta_1) \tag{2a}$$

Hinckley considered also the case when parameters $\theta_1$ and $\theta_2$ of the Gauss distribution are unknown. In this case, the MML estimation for the change-point of the mean value takes on the form:

$$\hat{\tau} = \arg \max_{1 \le r \le n-1} \left[ r \left( \hat{\theta}_{0,r} - \hat{\theta} \right)^2 + (n - r) \left( \hat{\theta}_{1,r} - \hat{\theta} \right)^2 \right], \tag{3}$$

where $\hat{\theta}_{0,r} = \frac{1}{r} \sum_{v=1}^{r} x_v$ , $\hat{\theta}_{1,r} = \frac{1}{n-r} \sum_{v=r+1}^{n} x_v$ . (4)

Since statistics (2) and (3) do not depend on any other probabilistic parameters, they can be used for nonparametric estimation of the change-point of the average of random sequences with an arbitrary distribution. However, in such situations (similarly, in the case where the average is evaluating according (1), in general the nonparametric algorithms lose their optimality. To overcome this difficulty, new nonlinear estimation algorithms based on the of the minimization of the polynomial are described below. They allow to take into account, in a simple way, the degree of non-Gaussian character of the statistical data

## IV. Two-parameter measurement of strain and temperature change

Let $\vec{x}$ be equally distributed sampled elements. Consider the algorithm presented in [10] and denoted by MMP. It is shown in that paper that the estimate of an arbitrary parameter $\vartheta$ can be found by solving the following stochastic equations with respect to $\vartheta$:

$$\sum_{i=1}^{s} h_i(\vartheta)\left[\frac{1}{n}\sum_{v=1}^{n} x_v^i - \alpha_i(\vartheta)\right]\Bigg|_{\vartheta_r=\hat{\vartheta}_r} = 0,$$

where: $s$ – is the order of the polynomial for parameter estimation, $\alpha_i(\vartheta)$ - is the theoretical initial moment of the *i-th* order.

Coefficients $h_i(\vartheta)$ (for $i=\overline{1,s}$ ) can be found by solving the system of linear algebraic equations, given by conditions of minimization of variance (with the appropriate order $s$) of the estimate of the parameter $\vartheta$, namely:

$$\sum_{i=1}^{s} h_i(\vartheta)\, F_{i,j}(\vartheta) = \frac{d}{d\vartheta}\alpha_j(\vartheta),\ \ j=\overline{1,s}\,, \tag{5}$$

where $F_{i,j}(\vartheta)=\alpha_{i+j}(\vartheta)-\alpha_i(\vartheta)\alpha_j(\vartheta)$ - centered correlants of $(i,j)$ dimensions.

Equations (5) can be solved analytically using the Kramer method, i.e.

$$h_i(\vartheta)=\frac{\Delta_{is}}{\Delta_s},\ \ i=\overline{1,s}\,,$$

where $\Delta_s=\det\left\|F_{i,j}\right\|$; $i,j=\overline{1,s}$ - volume of the stochastic polynomial body of dimension $s$, $\Delta_{is}$ - is the determinant obtained from $\Delta_s$ by replacing the *i-th* column by the column of free terms of eq. (5).

A new approach for finding the posteriori estimates of change-point, proposed in this paper, is based on application of MMP method. In this approach there is used a property of the following stochastic polynomials:

$$l_{sn}(\vec{x}/\vartheta)=nk_0(\vartheta)+\sum_{i=1}^{s}k_i(\vartheta)\sum_{v=1}^{n}x_v^i\ , \tag{6}$$

where $k_0(\vartheta)=\int\limits_a^{\vartheta}\sum\limits_{i=1}^{s}\left[h_i(\vartheta)\alpha_i(\vartheta)\right]d\vartheta\,,\quad k_i(\vartheta)=\int\limits_a^{\vartheta}h_i(\vartheta)d\vartheta\,,\ \ i=\overline{1,s}$ (7a,b)

The mathematical expectation $E\{l_{sn}(\vec{x}/\vartheta)\}$, treated as a function of $\vartheta$ assumes the maximum at the true value point of this parameter.

Values of the parameter $\vartheta$ belongs to some interval $(a,b)$. If the stochastic polynomial of the form (6) will be maximized with use a parameter $\vartheta$ which has a change-point (step change from value $\vartheta_0$ to value $\vartheta_1$ ), then we can build a polynomial form statistics:

$$P_r^{(s)}(\vartheta_0,\vartheta_1)=rk_0(\vartheta_0)+\sum_{i=1}^{s}k_i(\vartheta_0)\sum_{v=1}^{r}x_v^i+(n-r)k_0(\vartheta_1)+\sum_{i=1}^{s}k_i(\vartheta_1)\sum_{v=r+1}^{n}x_v^i\ , \tag{8}$$

which will have a maximum in a neighborhood of the true value $\tau$ of the change-point. Thus the general algorithm of applying MMP method for finding the estimation of the change-point $\tau$ can be formulated as follow

$$\hat{\tau}\ =\arg\max_{1\le r\le n-1} P_r^{(s)}(\vartheta_0,\vartheta_1). \tag{9}$$

### V. Polynomial estimation of the change-point of mean value

It is known from [10] that the estimate of the average $\theta$ obtained by MMPl method using a polynomial of order $s=1$ coincides with the form (1) of the linear estimate MM. Hence the synthesis of polynomial algorithms for estimating the change-point of this parameter is justified only for orders $s\ge 2$ . At a order $s=2$ polynomial estimate of the mean value can be found by solving the following quadratic equation:

$$\gamma_3\theta^2-\left[2\gamma_3\frac{1}{n}\sum_{v=1}^{n}x_v-\sigma(2+\gamma_4)\right]\theta-\sigma(2+\gamma_4)\frac{1}{n}\sum_{v=1}^{n}x_v+\gamma_3\left[\frac{1}{n}\sum_{v=1}^{n}(x_v)^2-\sigma^2\right]\Bigg|_{\theta=\hat{\theta}}=0\,. \tag{10}$$

The analysis of eq. (10) shows that the estimated value of $\hat{\theta}_{s=2}$ depends on coefficients of skewness $\gamma_3$ and kurtosis $\gamma_4$ . If the values of these parameters are equal to zero, then distribution is the normal (Gaussian) one. In this case the polynomial estimate (10) reduces to the classical estimate of the form (1). It is shown in [10] that the use of eq. (10) ensures higher accuracy (decrease the variance) of estimate compared with the estimate (1). The asymptotic value of this estimate (for $n\to\infty$ ) is given by the following formula:

$$g_2(\gamma_3,\gamma_4) = 1 - \frac{\gamma_3^2}{2+\gamma_4}. \qquad (10a)$$

Using the analytical expressions (5) and (7) one can easily find that, for order $s=2$, the coefficients maximizing the selected stochastic polynomial of the form (6) in a neighborhood of the true value of the parameter $\theta$ are the following:

$$k_0(\theta) = \frac{\sigma^3}{6\Delta_2}\left[2\gamma_3\theta^3 + 3(2+\gamma_4)\sigma\theta^2 - 6\gamma_3\sigma^2\theta\right], \quad k_1(\theta) = \frac{\sigma^3}{\Delta_2}\left[\gamma_3\theta^2 + (2+\gamma_4)\sigma\theta\right], \quad k_2(\theta) = -\frac{\sigma^3}{\Delta_2}\gamma_3\theta \qquad (11a\text{-}c)$$

where $\Delta_2 = \sigma^6\left(2+\gamma_4-\gamma_3^2\right)$.

In the presence of a priori information about the the mean values of $\theta_0$ before and $\theta_1$ after of change-point and under the condition $\theta_1 > \theta_0$, for the order of the polynomial $s=2$ the statistic (8) can be expressed as follows:

$$\begin{aligned} P_r^{(2)}(\theta_0,\theta_1) = (n-r)&\left[\frac{1}{3}\gamma_3\left(\theta_1^3-\theta_0^3\right) + \frac{1}{2}\sigma(2+\gamma_4)\left(\theta_1^2-\theta_0^2\right) - \sigma^2\gamma_3(\theta_1-\theta_0)\right] \\ &+ \left[\gamma_3\left(\theta_1^2-\theta_0^2\right) + \sigma(2+\gamma_4)(\theta_1-\theta_0)\right]\sum_{v=r+1}^{n} x_v - \gamma_3(\theta_1-\theta_0)\sum_{v=r+1}^{n} x_v^2. \end{aligned} \qquad (12)$$

In the case, when a priori information about the mean values $\theta_0$ and $\theta_1$ are unknown, polynomial evaluation of change-point moment (as in the classical case) can be found by replacing the unknown values of these parameters by their posterior estimates of the form (4). These estimates are formed for each potential change-point. Thus, for $s=2$ an adaptive algorithm for for estimating the time of change-point $\hat{\tau}$ based at MMPl method can be formulated as follows:

$$\begin{aligned} \hat{\tau} = \arg\max_{1\le r\le n-1} &\left\{ r\left[\frac{4}{3}\gamma_3\hat{\theta}_{0,r}^3 + \frac{3}{2}\sigma(2+\gamma_4)\hat{\theta}_{0,r}^2 - \left[\sigma^2 + \sum_{v=1}^{r}(x_v)^2\right]\gamma_3\hat{\theta}_{0,r}\right]\right\} + \\ &+ (n-r)\left[\frac{4}{3}\gamma_3\hat{\theta}_{1,r}^3 + \frac{3}{2}\sigma(2+\gamma_4)\hat{\theta}_{1,r}^2 - \left[\sigma^2 + \sum_{v=r+1}^{n}(x_v)^2\right]\gamma_3\hat{\theta}_{1,r}\right]\Bigg\}. \end{aligned} \qquad (13)$$

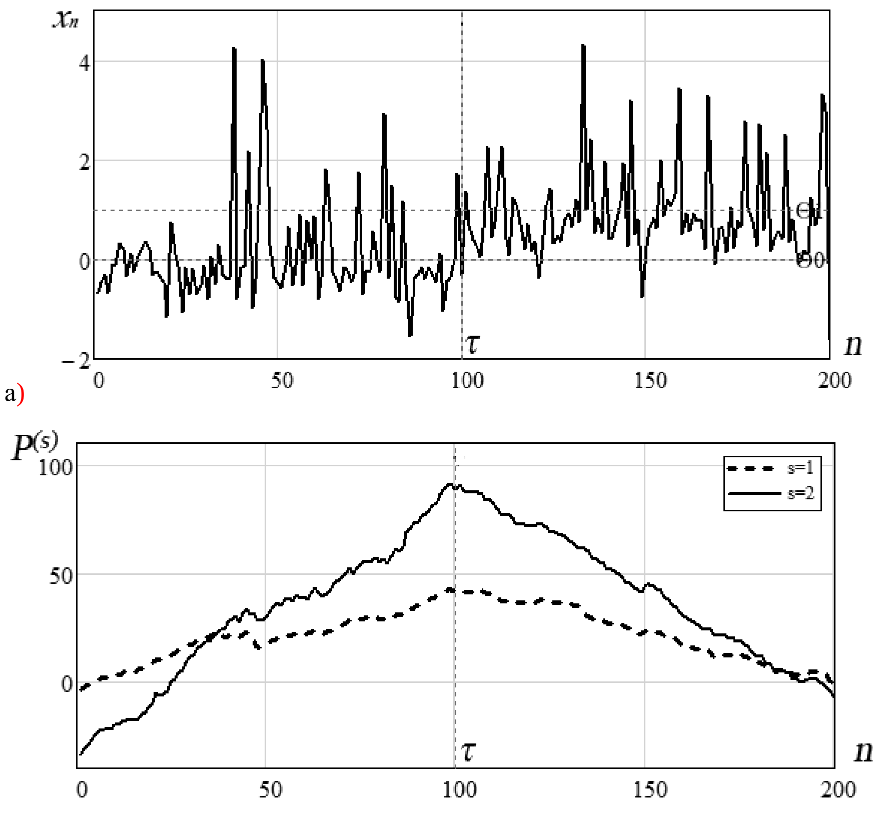
The analysis of the structure of polynomial statistics (12) and (13) confirms again the fact that, for $s=2$, the use MMPl is justified only in the case of an asymmetry ( $\gamma_3 \neq 0$ ) of the distribution of the statistical data.

## VI. Statistical modeling of a posteriori estimate of change-point

Based on results of above considerations, a software package in a software environment MATLAB, has been developed. It allows to perform the statistical modeling of the proposed semi-parametric estimation procedures, applied to the estimation of the mean value and variance of the change-points of non-Gaussian random sequences. Both, single and multi- experiments (in the sense of the Monte Carlo method) can be simulated. The accuracy obtained by classical and proposed polynomial algorithms for experimental data can be also compared.

In Fig. 1 there are presented the results for an numerical example obtained by estimation procedures for the mean value of the change-point of average values $\theta_0 = 0$ and $\theta_1 = 1$ of the non-Gaussian sequence (Fig. 1a), where $\sigma = 1$, $\gamma_3 = 2$ and $\gamma_4 = 5$. The calculations were performed using the classical version (2) of the algorithm of a posteriori estimation by MMP method (coinciding with MMPl if $s=1$) as well by polynomial algorithm (12) of MMP for $s=2$. The results presented in Fig. 1,b clearly confirm the potentially higher precision obtained by polynomial statistics for $s=2$, since the maximum of the corresponding function is strongly marked, as compared with the smoothed form of the statistic for $s=1$.

Results of the single experiment do not allow to compare adequately the accuracy of the statistical estimation algorithms. As a comparative criterion of efficiency, the ratio of variances of the estimates of the change-point is used. That can be obtained by a series of experiments with the same initial values of the model parameters. It should be noted that theoretically the results of statistical algorithms of a posteriori estimation of the change-point can depend on a various factors, including e.g.: the relative value of the mean jump at the change-point, the probabilistic nature (values of coefficients of higher order cumulants) of non-Gaussian random sequences, the presence of an a priori information about the values variable of parameters. Furthermore, the accuracy of estimations of the change-point depends on the chosen number *n* of the sample and on the accuracy of the variance estimates, i.e., on the number of experiments *m* performed under the same initial conditions.

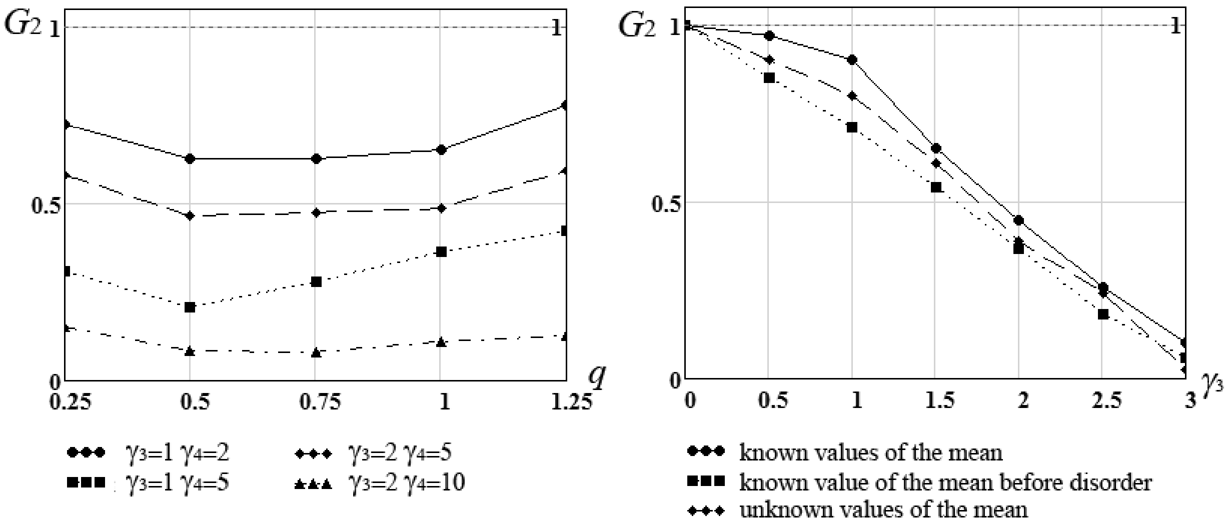


a)

b)

Figure 1. Example of a posteriori estimation of the change-point of the mean value.

The results of statistical modeling for $n = 200$ and $m = 2000$ are shown in Figure 2. $G_2$ is the ratio of variances of the change-point estimates obtained by MMPl method with the polynomial orders: $s = 2$ and $s = 1$ (Normal statistics) respectively. The value of $G_2$ characterizes the relative increase of accuracy. Figure 2a shows the dependence of $G_2$ on the relative values of the jump at the change-point $q = (\theta_1 - \theta_0)/\sigma$, obtained with different coefficients of skewness $\gamma_3$ and kurtosis $\gamma_4$. Figure 2b presents the dependence of $G_2$ on $\gamma_3$ (if $\gamma_4 = 10$ and $q = 0.5$), obtained under different a priori information about the mean values of random sequences before and after the change.



a)

b)

Figure 2 - The experimental values of $G_2$ coefficients, which show the reduction of variance of estimates of the change-point of mean value, obtained when MMPl method is used.

Analysis of these and many other experimental results confirm the theoretical results concerning the effectiveness of the polynomial method in the change-point estimation. It turns out that the relative growth of accuracy is roughly the same for different formulations of the problem, related to the presence or absence of an a priori information about values of the variable parameter. The improvement does not significantly depend on the relative magnitude of the jump at the change-point. It is determined primarily by the order of the "non-gaussian" of the process, which numerically is expressed as absolute values of the of higher-order cumulant coefficients.

## VII. Conclusions

Results of the research lead to the general conclusion about the potentially high efficiency of the implementation of the polynomial maximization method to the synthesis of simple adaptive algorithms for estimating of change-points of parameters of stochastic processes of non-Gaussian character of statistical data.

Obtained theoretical results allowed to develop a fundamentally new approach to the construction of semi-parametric algorithms for a posterior estimation of the change-point. This approach is based on the application of stochastic polynomials.

Among many possible directions of further research one should mention the following:

- increase of the degree of the stochastic polynomial, which is necessary to get more effective solutions, especially for non-Gaussian sequences with symmetrical distributions;
- analysis of the dependence of the accuracy of determining the parameters of non-Gaussian model (higher-order statistics) on the stability of polynomial algorithms for a posteriori estimation of the change-points;
- synthesis of polynomial algorithms for estimating the change-point with respect to other parameters (e.g., dispersion or correlation and regression coefficients), or in case where the values of several parameters are changed simultaneously (e.g., the mean value and the variance, etc.).

## References

[1] Chen J., Gupta A. K. (2012). *Parametric statistical change point analysis*. Birkhaeuser, p.273

[2] Reeves J., Chen J., Wang X. L., Lund R., and Lu Q. (2007). A review and comparison of change-point detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46 (6), 900-915

[3] Wang Y., Wu C., Ji Z., Wang B., and Liang Y. (2011). Non-parametric change-point method for differential gene expression detection. *PLoS ONE*, 6 (5), e.20060.

[4] Yamanishi K., Takeuchi J., Williams G., and Milne P. (2000). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.320-324,.

[5] Liu S., Yamada M., Collier N., & Sugiyama M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, vol.43, p.72- 83.

[6] Brodsky B. and Darkhovsky B. (1993) *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

[7] Lokajicek T, Klima K. A. (2000), First arrival identification system of acoustic emission (ae) signals by means of a higher-order statistics approach. *Measurement Science and Technology*. Vol. 17, 2461-2466.

[8] Yih-Ru Wang. (2008). The signal change-point detection using the high-order statistics of log-likelihood difference functions. *International Acoustics, Speech and Signal Processing,. ICASSP 2008. IEEE International Conference*, 4381-4384.

[9] Constantinos S. Hilas, Ioannis T. Rekanos, Paris Ast. Mastorocostas. (2013). Change point detection in time series using higher-order statistics: a heuristic approach, *Mathematical Problems in Engineering*, vol. 2013 , Article ID 317613, 10 pages.

[10] Kunchenko Y. (2002), *Polynomial Parameter Estimations of Close to Gaussian Random variables*. Shaker Verlag, Aachen Germany.

[11] Hinkley D. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*. vol. 57. No.1. p. 1- 17.