Ensembles of detectors for online detection of transient changes

Artemov, Alexey, Burnaev, Evgeny

Alexey Artemov, Evgeny Burnaev, "Ensembles of detectors for online detection of transient changes," Proc. SPIE 9875, Eighth International Conference on Machine Vision (ICMV 2015), 98751Z (8 December 2015); doi: 10.1117/12.2228369

SPIE.

Event: Eighth International Conference on Machine Vision, 2015, Barcelona, Spain

Ensembles of Detectors for Online Detection of Transient Changes

Alexey Artemov[2], Evgeny Burnaev[1]
[1]Institute for Information Transmission Problems (Kharkevich Institute) RAS
[2]Yandex Data Factory

ABSTRACT

Classical change-point detection procedures assume a change-point model to be known and a change consisting in establishing a new observations regime, i.e. the change lasts infinitely long. These modeling assumptions contradicts applied problems statements. Therefore, even theoretically optimal statistics in practice very often fail when detecting transient changes online. In this work in order to overcome limitations of classical change-point detection procedures we consider approaches to constructing ensembles of change-point detectors, i.e. algorithms that use many detectors to reliably identify a change-point. We propose a learning paradigm and specific implementations of ensembles for change detection of short-term (transient) changes in observed time series. We demonstrate by means of numerical experiments that the performance of an ensemble is superior to that of the conventional change-point detection procedures.

Keywords: change-point detection, transient changes, ensemble, aggregation, blending, logistic regression

1. INTRODUCTION

Change-point detection is a study of methods for identification of changes in the probability distribution of an observed stochastic process. Problems where such changes occur span many applied areas and include automatic video surveillance based on motion features [1], intrusion detection in computer networks [2], anomaly detection in data transmission networks [3], fault detection in vehicle control systems [4], detection of onset of an epidemic [5], drinking water monitoring [6] and many others. In all of these applications, sensors monitoring the environment take observations that undergo a change in a distribution in response to changes in the environment. As long as the behavior of the observations is consistent with the normal state, one is content to let the process continue. If the state changes, then one is interested in detecting the change as soon as possible while minimizing false detections.

A common problem encountered during the design of a change-point detection procedure is the lack of information about the change-point model. In many practical situations, we're no longer within a classical change-point detection framework where the change-point model is assumed to be known. In addition, in many cases the change does not consist in establishing the new observations regime, i.e. the change does not last infinitely. Ignorance of these facts can drop the performance of a change-point detection procedure dramatically.

However, in practice many change-point detectors possessing different characteristics can be used simultaneously. The motivation for considering an ensemble of change-point detection procedures is based on the difference in statistical properties of its constituent detectors. When applied to data of various nature (such as data with heavy-tailed distributions, strong spatial or temporal correlations), these algorithms behave slightly differently, yet, every algorithm tries to extract as much information about the change-point as possible. By using sample paths of their statistics as features and learning an appropriate decision rule (an ensemble) we can obtain an increase in change-points detection efficiency and in overall robustness of a change-point detection procedure.

2. ENSEMBLES OF PROCEDURES FOR CHANGE-POINT DETECTION

2.1 Classical change-point detection procedures

Let the observed process $X = \left(X_t\right)_{t\in\mathbb{N}_0}$ have a probability density $p_\infty(\cdot)$ when X is *in-control* during the period $\mathscr{T}_\infty$ and an another probability density $p_0(\cdot)$ when X is *out-of-control* during $\mathscr{T}_0$. Conventionally is it assumed that $\mathscr{T}_\infty = [0,\theta)$ and $\mathscr{T}_0 = [\theta,\infty)$, where θ is the (unknown) change-point time. However, in many practical situations the change has inherently finite duration or must be detected within a prescribed detection window [6]. Therefore, we consider the situation of *a transient change* where $\mathscr{T}_\infty = [0,\theta)\cup[\theta+\Delta,\infty)$ and $\mathscr{T}_0 = [\theta,\theta+\Delta)$, where both change-point time θ and change duration Δ are unknown.

Eighth International Conference on Machine Vision (ICMV 2015), edited by Antanas Verikas, Petia Radeva, Dmitry Nikolaev, Proc. of SPIE Vol. 9875, 98751Z · © 2015 SPIE
CCC code: 0277-786X/15/$18 · doi: 10.1117/12.2228369

A change-point detection procedure is a stopping rule defined according to the sample path of a certain process $S = \left(S_t\right)_{t\in\mathbb{N}_0}, S_t = S(\mathbf{X}_t)$ where $\mathbf{X}_t$ denotes observations history up to the moment t: $\mathbf{X}_t = \{X_s, 0 \leqslant s \leqslant t\}$. Observations are stopped at time τ_S which is the first hitting time of S to some predefined level h: $\tau_S = \inf\{t \in \mathbb{N}_0 : S_t \geqslant h\}$.

Let $\mathbb{E}_\infty$ and $\mathbb{E}_0$ be the expectations with respect to $p_\infty(\cdot)$ and $p_0(\cdot)$ respectively. We further denote by $\mathbb{E}_\theta$ the expectation with respect to the density $p_\theta(\cdot)$, corresponding to the case where the change occurs at time θ. The key statistics in change-point detection problems are the likelihood ratio $L_t = p_0(X_1,\ldots,X_t)/p_\infty(X_1,\ldots,X_t)$ and its logarithm $Z_t = \log L_t$. It is often assumed that random variables $X_1,\ldots,X_t$ are independent and identically distributed in absense of a change-point; this implies that their joint density $p_\theta(X_1,\ldots,X_t)$ has the form $p_\theta(X_1,\ldots,X_t) = \prod_{k=1}^t f_\theta(X_k)$ where $\theta \in \{\infty, 0\}$. Therefore it is often convenient to consider per-sample likehood ratios $l_k = f_0(X_k)/f_\infty(X_k)$ and $\zeta_k = \log l_k$ s. t. $L_t = \prod_{k=1}^t l_k$ and $Z_t = \sum_{k=1}^t \zeta_k$. We briefly describe the structure of the change-point detection procedures and associated processes below.

The process $S = \left(S_t\right)_{t\in\mathbb{N}_0}$ for the **Shewhart control chart** [7] is defined by the relation

$$S_t = \sum_{k=t-K}^{t} \zeta_k, \qquad t \in \mathbb{N}_0, \tag{1}$$

where K is some predefined variable (the size of the sliding window). The **cumulative sum** (CUSUM) procedure [8] is based on the process $T = \left(T_t\right)_{t\in\mathbb{N}_0}$ defined according to the recursive relation

$$T_t = \max\left\{0, \max_{1\leqslant\theta\leqslant t} \sum_{k=\theta}^{t} \zeta_k\right\} = \max(0, T_{t-1} + \zeta_t), \qquad T_0 = 0, \qquad t \in \mathbb{N}_0, \tag{2}$$

The **Shiryaev-Roberts (SR) procedure** [9, 10] consists in computing the statistic $\psi = \left(\psi_t\right)_{t\in\mathbb{N}_0}$ defined by

$$\psi_t = \sum_{\theta=1}^{t} \prod_{k=\theta}^{t} l_k = (1+\psi_{t-1})l_k, \qquad \psi_0 = 0, \qquad t \in \mathbb{N}_0, \tag{3}$$

A slightly different procedure is the **changepoint** procedure originally proposed for detection of a change in a level of a gaussian sequence [11]. The process $S = \left(S_t\right)_{t\in\mathbb{N}_0}$ defined by this procedure is

$$S_t = \max_{t-K\leqslant\theta\leqslant t} \frac{\overline{X}^t_{\theta+1} - \overline{X}^\theta_{t-K}}{((t-\theta)^{-1} + (\theta - t + K)^{-1}))^{1/2} W^{1/2}}, \qquad t \in \mathbb{N}_0, \tag{4}$$

where $\overline{X}^j_i = \sum_{k=i}^j X_k/(j-i)$, $W = \left[\sum_{k=t-K}^{\theta}\left(X_k - \overline{X}^\theta_{t-K}\right)^2 + \sum_{k=\theta+1}^{t}\left(X_k - \overline{X}^t_{\theta+1}\right)^2\right] \Big/ (K-2)$, and K is the procedure parameter (the size of the sliding window). The change-point detection algorithms mentioned above make no specific assumptions regarding the change time θ. When the change time θ is a random variable assuming the values $0, 1, \ldots$ with probabilities

$$\mathrm{P}(\theta = 0) = \pi, \qquad \mathrm{P}(\theta = t \,\big|\, \theta > 0) = p(1-p)^{t-1}, \qquad t = 1, 2, \ldots,$$

where π and $p \in [0,1]$ are model parameters, a Bayesian change detection algorithm can be considered. This procedure is based on the **posterior probability** process $\pi = \left(\pi_t\right)_{t\in\mathbb{N}_0}$ defined by the recursive relations [9]

$$\pi_t = \varphi_t/(1+\varphi_t), \qquad \varphi_t = (p+\varphi_{t-1})l_t/(1-p), \qquad t \in \mathbb{N}_0. \tag{5}$$

2.2 Ensemble-based change-point detection

Let $\Pi_1,\ldots,\Pi_n$ denote n change-point detection procedures, where each procedure Π_k prescribes to stop observations at time τ_k which is the first hitting time of some process $S^k = \left(S^k_t\right)_{t\in\mathbb{N}_0}$ to a level $h_k > 0$: $\tau_k = \inf\{t \in \mathbb{N}_0 : S^k_t \geqslant h_k\}$. We further consider a set of *signals* $s^k = \left(s^k_t\right)_{t\in\mathbb{N}_0}, k = 1,\ldots,n$ defined by $s^k_t = S^k_t/h_k, t \in \mathbb{N}_0$. We call the procedure A an *ensemble* if its stopping time τ_A is defined as the first hitting time of some process $a = \left(a_t\right)_{t\in\mathbb{N}_0}$ to a specified point $h_\mathrm{A} > 0$: $\tau_\mathrm{A} = \inf\{t \in \mathbb{N}_0 : a_t \geqslant h_\mathrm{A}\}$, where

$$a_t = \psi(\boldsymbol{\lambda}; \mathbf{S}^1_t, \ldots, \mathbf{S}^n_t), \tag{6}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ ($d \geq n$) and $\mathbf{S}^k_t = \{s^k_s, 0 \leqslant s \leqslant t\}$ is the history of the signal $s^k = \left(s^k_t\right)_{t\in\mathbb{N}_0}$ up to the time t, $k = 1,\ldots,n$. Note that the specific ensemble is completely defined by the choice of the "aggregation function" $\psi(\cdot)$. Its parameters $\boldsymbol{\lambda} \in \mathbb{R}^d$

Dataset	Type of data	Parameters subject to change	Change time θ and change duration Δ	Change magnitude
WhiteNoise	Uncorrelated Gaussian Noise	mean	random, $\theta \sim U(200,800)$, $\Delta \sim U(5,100)$	random, $\mu \sim U(0.1,2.0)$
Fractal	Fractional Gaussian Noise			
Cauchy	Uncorrelated Cauchy Noise			
GARCH1	GARCH$(1,1)$ process	α_1, β_1		random, $\alpha_1 \sim U(.4,.8)$, $\beta_1 \sim U(.1,.2)$,
ARMA-AR	ARMA(10, 3) process	AR terms φ_i		random
ARMA-MA	ARMA(10, 3) process	MR terms θ_j		random
GARCH1 + ARMA	GARCH$(1,1)$ + ARMA(10, 3) process	$\alpha_1, \beta_1, \varphi_i, \theta_j$		random

Table 1: Summary of the considered artificial datasets

can be *learned* to optimize a certain performance measure (introduced below). Several examples of specific choices of the aggregation function are presented below.

The *majority voting-based ensemble* is defined by the following obvious choice of the aggregation function

$$\psi_{\mathrm{MAJ}}(\boldsymbol{\lambda}; \mathbf{S}_t^1, \ldots, \mathbf{S}_t^n) = \frac{2}{n} \sum_{k=1}^{n} \mathbb{1}_{\{s_t^k \geqslant 1\}}(t). \tag{7}$$

By setting $h_{\mathrm{A}} = 1$ we obtain a stopping rule prescribing to stop observations at the first time when the number $n^+ = \sum_{k=1}^{n} \mathbb{1}_{\{s_t^k \geqslant 1\}}(t)$ of "votes" submitted in favor of the change-point exceeds the number $n^- = n - n^+$ of "votes" against the change-point. When a change occurs, the information pre-filtered by the detectors is accumulated over time to form a signal used to raise an alarm. However, in practice the accumulation of this information might be slow due to model misspecification or subtle change-point "magnitude". Therefore it might be useful to employ the entire history $\mathbf{S}_t^k$ of a signal $S^k = \left(S_t^k\right)_{t \in \mathbb{N}_0}$ (as opposed to only the current value s_t^k) to detect the change. For this, we consider two ensembles employing the values of signals with lags up to p. The *weighted voting scheme with lag* is based on the function

$$\psi_{\mathrm{WEIGHT}-p}(\boldsymbol{\lambda}; \mathbf{S}_t^1, \ldots, \mathbf{S}_t^n) = \sum_{j=0}^{p} \sum_{k=1}^{n} \lambda_{kj} s_{t-j}^k. \tag{8}$$

It is natural to set $h_{\mathrm{A}} = 1$ for this procedure as such "normalization" could be achieved through the choice of $\lambda_{kj}, k = 1, \ldots, n, j = 0, \ldots, p$. The last ensemble we consider in this work is a *logistic regression-based* classifier for which the aggregation function could be written as

$$\psi_{\mathrm{LOG}-p}(\boldsymbol{\lambda}; \mathbf{S}_t^1, \ldots, \mathbf{S}_t^n) = \sigma\Big(\sum_{j=0}^{p} \sum_{k=1}^{n} \lambda_{kj} s_{t-j}^k - \lambda_0 \Big). \tag{9}$$

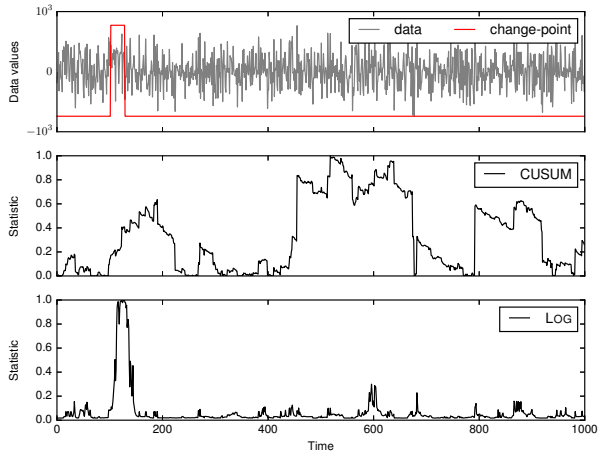
The value $a_t = \psi_{\mathrm{LOG}-p}(\boldsymbol{\lambda}; \mathbf{S}_t^1, \ldots, \mathbf{S}_t^n)$ of the ensemble statistic can be interpreted as a posterior probability of a change-point given the observations history $\mathbf{X}_t = \{X_s, 0 \leqslant s \leqslant t\}$ up to the moment t. Note that for this ensemble the threshold h_{A} must be chosen to belong to the interval $(0,1)$.

3. QUALITY MEASURES FOR ENSEMBLE LEARNING AND EVALUATION

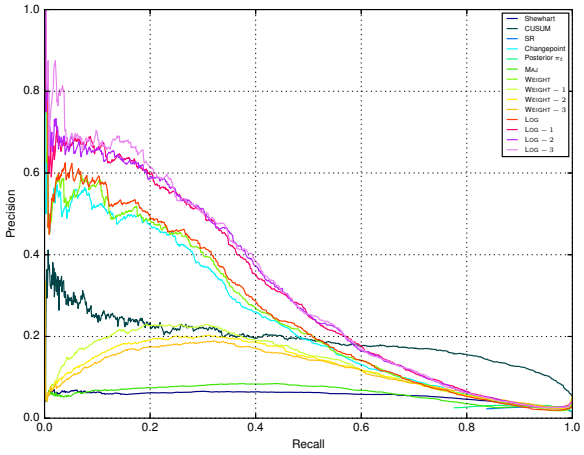
3.1 Performance measures for change-point detection procedures

One of the most widely used performance measures for characterization of efficiency of change-point detection procedures is the average detection delay $\mathrm{ADD}(\tau) = \mathbb{E}_\theta\big[\tau - \theta | \tau > \theta\big]$ at a given average run length to a false alarm $\mathrm{ARL}(\tau) = \mathbb{E}_\infty \tau$. Another approach is to estimate the average detection delay $\mathrm{ADD}(\tau)$ at a given level of the probability of a false alarm $\mathrm{PFA}(\tau) = \mathbb{P}_\infty(\tau < \theta)$. We note, however, that these measures are unable to characterize situations where missed detections can occur due to the finite nature of the change; to do so, we use two more performance measures described below. The first measure is defined by

$$\mathbf{F}(h, c_\infty, c_0) = c_\infty \mathbb{E}_\infty \left[\frac{\int \mathbb{1}_{\{a_t \geqslant h\}}(t) \mathbb{1}_{\mathscr{T}_\infty}(t) dt}{\int \mathbb{1}_{\mathscr{T}_\infty}(t) dt} \right] + c_0 \mathbb{E}_0 \left[\frac{\int \mathbb{1}_{\{a_t < h\}}(t) \mathbb{1}_{\mathscr{T}_0}(t) dt}{\int \mathbb{1}_{\mathscr{T}_0}(t) dt} \right], \tag{10}$$



(a) Sample trajectories for the CUSUM algorithm and for the LOG−0 ensemble against the ground truth



(b) Precision-recall curves for five detectors and considered ensembles

Figure 1: The performance of the ensemble algorithms for the Cauchy dataset

where the values of expectations $\mathbb{E}_\infty[\cdot]$ and $\mathbb{E}_0[\cdot]$ correspond to relative durations of a false alarm and a false silence signals, respectively, and c_∞ and c_0 have the meaning of losses associated to unit durations of false alarm and false silence, respectively. The second measure is the standard *precision-recall curve* used in the area of information retrieval.

3.2 Learning ensemble parameters

Let $\mathscr{X} = \{(X^i, Y^i), i = 1, \ldots, N\}$ be the labeled data where each point $(X^i, Y^i) \in \mathscr{X}$ is a pair, its first component $X^i = \left(X_t^i\right)_{t\in\mathbb{N}_0}$ being a sample path of the observations, and its label $Y^i = \left(Y_t^i\right)_{t\in\mathbb{N}_0}$ being an out-of-control state indicator: $Y_t^i = \mathbb{1}_{\mathscr{T}_0^i}(t)$. We formulate the problem of learning the parameters $\boldsymbol{\lambda} \in \mathbb{R}^n$ of an ensemble as an optimization problem $\mathbf{F}(h, c_\infty, c_0) \to \inf_{\boldsymbol{\lambda}\in\mathbb{R}^n}$. As the performance measure $\mathbf{F}(h, c_\infty, c_0)$ introduced in (10) utilizes a non-differentiable indicator function and therefore cannot be optimized using standard function minimization approaches, we introduce for it an empirical approximation $\widehat{\mathbf{F}}_{\mathrm{D}}(h, c_\infty, c_0)$ defined by

$$\widehat{\mathbf{F}}_{\mathrm{D}}(h, c_\infty, c_0) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{c_\infty}{T_\infty^i}\sum_{t\in\mathscr{T}_\infty^i}\sigma(a_t - h) + \frac{c_0}{T_0^i}\sum_{t\in\mathscr{T}_0^i}\sigma(h - a_t)\right\} \tag{11}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the logistic function. Note now that the function $\widehat{\mathbf{F}}_{\mathrm{D}}(h, c_\infty, c_0)$ is differentiable w.r.t. ensemble parameters $\boldsymbol{\lambda}$ and can therefore be optimized using gradient descent.

4. NUMERICAL EXPERIMENTS

We performed an extensive numerical evaluation of the developed approach using a number of simulated datasets differing by the distribution of the observations and the parameters subject to a change, as summarized in Table 1. We generated 1024 independent sample paths of each data type for training an ensemble and 1024 independent sample paths for validation. For each sample path, the total length of the time series is $T = 1000$.

We present the results of our evaluation in terms of the area under precision-recall curve (AUC) metric in Table 2. For almost every dataset, except for the GARCH + ARMA one, our approach presents an improvement over the classical change-point detection procedures. An example of sample paths for the CUSUM and LOG−0 ensemble and for the Cauchy dataset is presented on Figure 1, (a). It is easy to note that the CUSUM algorithm performs poorly for this kind of data while the ensemble algorithm is capable to use its underlying signals efficiently and correctly detect the change-point. On Figure 1, (b), we present precision-recall curves for all of the considered detectors and ensembles.

	WhiteNoise	Fractal	Cauchy	GARCH1	ARMA-AR	ARMA-MA	GARCH1 + ARMA
Shewhart	77.52	24.44	05.45	32.08	19.80	76.37	40.00
CUSUM	61.11	30.70	19.24	59.88	28.74	89.90	**75.44**
SR	22.22	6.11	.40	50.06	24.17	7.15	72.72
Changepoint	60.62	45.42	24.18	21.94	13.03	57.15	22.98
Posterior π_t	27.38	7.76	.66	53.60	29.00	35.69	74.58
MAJ	62.13	24.62	6.11	47.80	28.74	**92.71**	67.01
WEIGHT $-$ 0	71.73	38.94	25.08	55.62	24.94	79.48	67.23
WEIGHT $-$ 1	71.58	38.97	13.46	57.65	29.60	91.92	71.28
WEIGHT $-$ 2	73.89	39.63	12.29	56.57	**30.45**	91.13	69.28
WEIGHT $-$ 3	73.25	38.98	11.61	**57.83**	26.24	90.70	72.11
LOG $-$ 0	77.25	**48.64**	25.90	51.35	23.29	87.72	68.35
LOG $-$ 1	76.27	36.20	31.03	50.03	23.49	88.47	65.97
LOG $-$ 2	78.01	39.74	31.30	49.35	27.99	88.93	66.43
LOG $-$ 3	**78.85**	40.31	**32.24**	49.08	27.99	88.88	66.77

Table 2: Area under the precision-recall curve for the considered detectors and ensembles for the artificial datasets

5. CONCLUSION

In this work, we proposed the novel methodology for ensembling the change-point detection procedures. By introducing a new performance measure and a simple numerical algorithm for its optimization, we demonstrated how to learn the parameters of an ensemble for a specific type of data. We further presented the results of numerical evaluation of ensemble performance for simulated datasets of various nature. The efficiency of an ensemble remains comparable to that of the classical change detection algorithms when measured according to classical schemes (ADD at a given level of ARL and ADD at a given level of PFA). However, efficiency measures borrowed from the area of information retrieval (relative accuracy of state identification and precision-recall measures) indicate a significant increase in performance of the ensemble compared to the classical methods. Application of the proposed approach to real problems from [12] demonstrated significantly increased efficiency and robustness of the obtained change-point detectors compared to conventional ones.

Acknowledgement: The second author was supported by RFBR grants No. 13-01-00521 and No. 13-01-12447.

REFERENCES

[1] P. Duc-Son, S. Venkatesh, M. Lazarescu, S. Budhaditya. "Anomaly detection in large-scale data stream networks", Data Mining and Knowledge Discovery, 28(1), p. 145–189 (2014)
[2] A. Tartakovsky, B. Rozovskii, R. Blažek, H. Kim. "A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods", IEEE Transactions on Signal Processing, 54(9), p. 3372–3381 (2006)
[3] P. Casas, S. Vaton, L. Fillatre, I. Nikiforov. "Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements", Computer Networks, 54(11), p. 1750–1766 (2010)
[4] D. Malladi, J. Speyer. "A generalized Shiryayev sequential probability ratio test for change detection and isolation", IEEE Transactions on Automatic Control, 44(8), p. 1522–1534 (1999)
[5] I. MacNeill, Y. Mao. "Change-point analysis for mortality and morbidity rate", Applied Change Point Problems in Statistics, p. 37–55 (1995)
[6] B. Guépié, L. Fillatre, I. Nikiforov. "Sequential Detection of Transient Changes", Sequential Analysis, 31(4), p. 528–547 (2012)
[7] W. Shewhart, "Economic control of quality of manufactured product", Bell System Technical Journal, 9(2), p. 364–389 (1931)
[8] E. Page. "Continuous inspection schemes", Biometrika, 41(1), p. 100–115 (1954)
[9] A. Shiryaev, "On optimum methods in quickest detection problems", Theory of Probability & Its Applications, 8(1), p. 22–46 (1963)
[10] S. Roberts. "A comparison of some control chart procedures", Technometrics, 8(3), p. 411–430 (1966)
[11] A. Sen, M. Srivastava. "Some one-sided tests for change in level", Technometrics, 17(1), p. 61–64 (1975)
[12] A. Artemov, E. Burnaev. "Nonparametric Decomposition of Quasi-periodic Time Series for Change-point Detection", Proceedings of the ICMV-2015 conference (2015)