Nonparametric Test for Change Point Detection in Time Series

Dmitriy Klyushin [1], Irina Martynenko [2]

[1] *Taras Shevchenko National University of Kyiv, prospect Glushkova 4D, Kyiv, 03680, Ukraine*
[2] *Academy of Labour Social Relations and Tourism, 3-A, Kiltseva doroha, Kyiv, 03187, Ukraine*

Abstract

We describe new nonparametric tests for detection change points of the time series, which are the points dividing the time series into segments of random values obeying different distributions. The proposed tests are based on the Dempster–Hill theory and generalizations of the Bernoulli scheme. To recognize the change points, simplified versions of the Klyushin–Petunin homogeneity test are proposed. The significance level of these tests does not exceed 0.05, and the accuracy is comparable to the original version. The tests have high sensitivity and specificity of recognizing the heterogeneity of two random samples with different means and the same variances or equal means but different variances. The described tests can be useful in a wide variety of areas from healthcare (for example, when analyzing time series generated by pulse oximeters) to IoT devices in industry and in everyday life.

Keywords [1]

Time series, regression analysis, modeling and prediction, nonparametric statistics

1. Introduction

The modern epoch is characterized by the intensive development of information networks and IoT technologies. There are more and more information sources that generate large amounts of sequential data that form time series which need to be processed in a short time. The problem has been exacerbated by the COVID-19 outbreak, which has raised questions related to the processing of signals from medical devices, tracking tools, etc. For example, a change in the distribution of random variables representing the level of oxygen in the blood of a patient with COVID-19 may require an immediate response from clinicians, which means that the change point in the time series generated by the pulse oximeter must be found in real time.

A time series is considered homogeneous if all its segments consist of random values that obey the same distribution. A change point is a point before and after which the values of the time series obey different distributions. Consequently, the problem of detecting the change point can be reduced to testing of samples homogeneity in adjacent segments of time series.

Change point detection methods are divided into online and offline. Online methods analyze fragments of a time series in real time. Offline methods analyze a complete time series from the first to the last point. A fairly detailed review of methods for finding change points in time series is given in [1]. Despite the fact that we restrict ourselves to one-dimensional time series, it may be easily generalized on multivariate case (see a survey in [2]).

Since the problem of finding a change point can be reduced to the problem of samples homogeneity, it should be mentioned that such tests are divided into parametric and nonparametric tests. The former tests are based on the assumption that the time series data obey a specific (e.g., Gaussian distribution) or parameterized distribution [3–8]. The last tests do not depend on assumptions about the shape of the distribution, using only the most general properties, for example,

MoMLeT+DS 2021: 3rd International Workshop on Modern Machine Learning Technologies and Data Science, June 5, 2021, Lviv-Shatsk, Ukraine
EMAIL: dokmed5@gmail.com (D. Klyushin); martia@rambler.ru (I. Martynenko)
ORCID: 0000-0003-4554-1049 (D. Klyushin); 0000-0003-0249-4887 (I. Martynenko)

continuity [9–19]. Online methods for finding change points in time series are considered in [20–24]. These methods analyze individual fragments of a time series, and not the entire series from beginning to end in opposite to offline methods. This allows us not to accumulate data in large storages. The purpose of the article is to outline new nonparametric online methods for recognizing change points in a time series. The originality of the proposed methods lies in the simplification of the Klyushin–Petunin criterion [25] without loss of accuracy.

2. Simplified versions of Klyushin–Petunin test for homogeneity

Consider the problem of testing homogeneity of two samples which is equivalent to the problem of change point detection. Usually, to solve this problem the Kolmogorov–Smirnov test and the Mann–Whitney–Wilcoxon test are used. Recently, we proposed to use the Klyushin–Petunin test [25] and demonstrated its high performance [26].

Suppose that samples $x=(x_1,x_2,...,x_n)$ and $y=(y_1,y_2,...,y_m)$ are drawn from the distributions F_1 and F_2, respectively. The null hypothesis states that x and y obey the same distribution, i.e. $F_1=F_2$. The alternative hypothesis states that $F_1\neq F_2$.

Standard non-parametric two-sample homogeneity tests (Kolmogorov–Smirnov, Mann–Whitney–Wilcoxon, etc.) produce only one-sided confidence limits with the given significance level. This means that if the test statistics exceeds a given threshold the samples are considered as homogeneous, otherwise the answer is uncertain. Therefore, it is desirable to construct tests with two-sided confidence limits for a test statistics. Such tests is the Klyushin–Petunin homogeneity test [25] based on the Matveychuk–Petunin model [27–29]. It allow constructing a two-sided confidence interval with a given the significance level for both the true and false null hypothesis and estimating the probability of types I and II errors.

2.1. The Dempster–Hill theory and *p*-statistics

Let $x=(x_1,x_2,...,x_n)\in F_1$, $y=(y_1,y_2,...,y_n)\in F_2$, and $x_{(0)}=-\infty, x_{(1)},x_{(2)},...,x_{(n)},x_{(n+1)}=+\infty$ be corresponding variance series. The null hypothesis H_0 states that $F_1=F_2$. If the null hypothesis is true, then due to the Dempster–Hill assumption [30] we have

$$p_{ij}^{(k,n)}=P\left(y_k\in\left(x_{(i)},x_{(j)}\right)\right)=\frac{j-i}{n+1}, i<j. \tag{1}$$

Corollary 1. Selecting random natural numbers i and j such that $i<j$ N times, for a sufficiently large N we have

$$p_{ij}^{(k,n)}=P\left(y_k\in\left(x_{(i)},x_{(j)}\right)\right)\approx\frac{j-i}{n+1}, i<j.$$

Corollary 2. Selecting random natural numbers N times $(N<n)$ from (1) we have

$$p_{ij}^{(k,n)}=P\left(y_k\in\left(x_{(1)},x_{(j)}\right)\right)=\frac{j-1}{n+1}.$$

If $F_1\neq F_2$, then $p_{ij}^{(k,n)}$ significantly deviates from $\frac{j-i}{n+1}$. Therefore, we must estimate a difference between observed relative frequency $h_{ij}^{(k,n)}$ of the event $y_k\in\left(x_{(i)},x_{(j)}\right)$. To do this, we use a confidence limits $\left(p_{ij}^{(1)},p_{ij}^{(2)}\right)$ for the binomial proportion $p_{ij}^{(k,n)}$ with given significance level β. Let L_{xy} be the number of the intervals $\left(p_{ij}^{(1)},p_{ij}^{(2)}\right)$ containing p_{ij}, $h_{xy}=\frac{2L_{xy}}{n(n-1)}$ be the proportion of the intervals $\left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ containing $p_{ij}^{(k,n)}$, $h_{yx} = \frac{2L_{yx}}{m(m-1)}$ is the proportion of the intervals $\left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ containing $p_{ij}^{(k,m)}$ in the scheme where the samples x and y are switched. Then, *p*-statistics [25] is

$$\rho(x.y) = \frac{1}{2}\left(h_{xy} + h_{yx}\right). \tag{2}$$

It is the probability that the samples x and y are homogeneous, i.e. they are drawn from the same distribution.

2.2. The Klyushin–Petunin test and its simplifications

The original version of the Klyushin–Petunin test use $N = \frac{(n-1)n}{2}$ confidence intervals $\left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$. Let L be the number of intervals $\left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ containing $p_{ij}^{(k,n)}$ and (2) is the relative frequency of the random event $\left\{p_{ij} \in \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)\right\}$ with the probability $1-\beta$. Construct the confidence interval I_n for the probability of the event $\left\{p_{ij} \in \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)\right\}$ with the given significance level (for definiteness, in this paper we use the Wilson confidence interval [31]). The decision rules may be formulated in the following way:

1) Original version. If the confidence interval I_n covers 0.95 then the null hypothesis is accepted, otherwise it is rejected.
2) First simplified version. Generate N random natural numbers i and j such that $j > i$, where $N \geq 30$. Compute $\rho(x, y)$ using only intervals $\left(x_{(i)}, x_{(j)}\right)$ with above mentioned i and j, and the confidence interval I_n. If I_n covers 0.95 then the null hypothesis is accepted, otherwise it is rejected.
3) Second simplified version. Compute $\rho(x, y)$ using only intervals $\left(x_{(1)}, x_{(j)}\right)$ and the confidence interval I_n. If I_n covers 0.95 then the null hypothesis is accepted, otherwise it is rejected.

3. Numerical experiments

As far as, there are many formulas for confidence intervals for binomial proportions [31], we must justify the choice. To compare *p*-statistics based on different confidence intervals we used samples of 100 and 300 numbers and parameterized distributions. The step for parameters was set to 0.1. To avoid bias of pseudorandom number generators, we have conducted 10 experiments, and results were averaged. We have considered eight methods:

I. Clopper–Pearson interval.
II. Bayes's interval.
III. Wilson interval with corrections for continuity.
IV. Wilson interval without corrections for continuity.
V. Normal approximation with corrections for continuity.
VI. Normal approximation without corrections for continuity.
VII. Arcsine transformations with corrections.
VIII. Arcsine transformations without corrections.

In order to compare the statistics, samples from the following general categories were considered: 1) same variance and different means; 2) same mean and different variances; 3) both different means and variances. Samples from a parametric family of hypothetical distribution with a sample from the general population with a reference distribution were compared (Fig. 1–6).

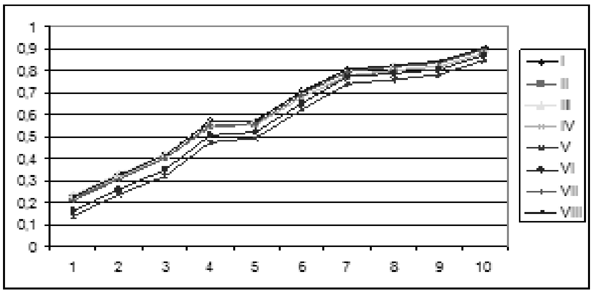


Figure 1: N(0, α), size = 100

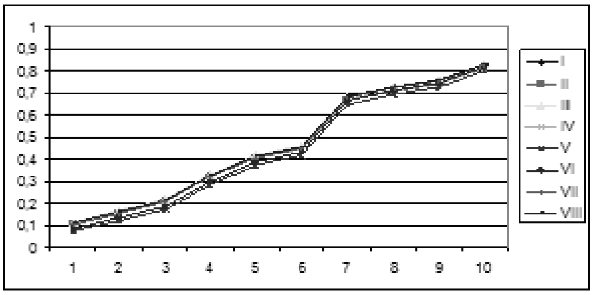


Figure 2: N(0, α), size = 300

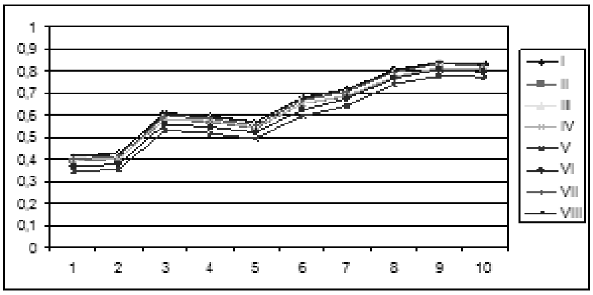


Figure 3: N(α, 1), size = 100

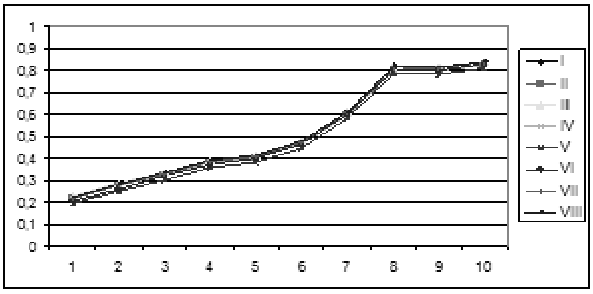


Figure 4: N(α, 1), size = 300

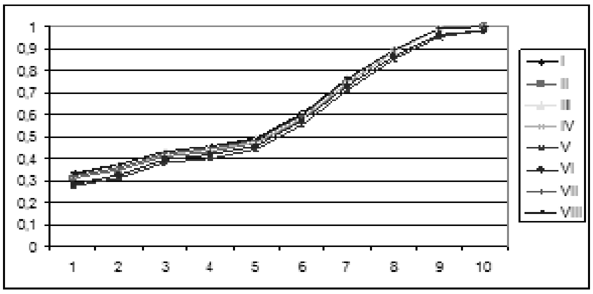


Figure 5: αU(0, 1)+ (1–α)U(1/2, 3/2), size = 100

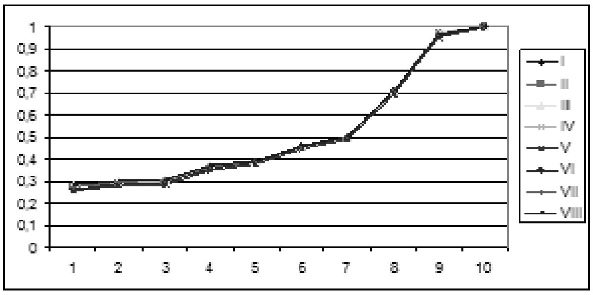
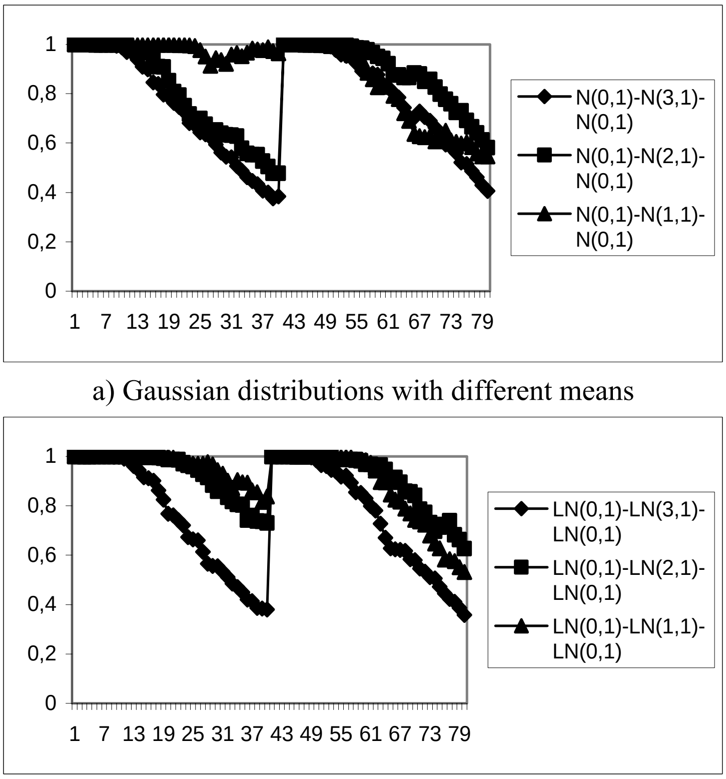


Figure 6: αU(0, 1)+ (1–α)U(1/2, 3/2), size = 300

Analyzing the Figures 1-6, we see that the choice of the confidence intervals I_{ij} does not effect to the final result: homogeneity measures are equivalent. Therefore, the family of homogeneity measures based on Dempster–Hills theory has high performance both for small and large samples, and the *p*-statistics may be considered as independent of the selection of the type of the confidence interval.

In [26] we have shown that in the context of change point detection the sensitivity and specificity of the Klyushin–Petunin test are higher than the sensitivity and specificity of the classical tests (Kolmogorov–Smirnov and Mann–Whitney–Wilcoxon). Now, we consider the simplified versions of the Klyushin–Petunin test and restricted ourselves with randomly selected 100 intervals $\left(x_{(i)}, x_{(j)}\right)$ following to the recommendations given in [32]. We generated samples containing 40 random numbers obeying distributions which have the same mean and the different variances, the different means and the same variance, and different means and different variances, and averaged results on 10 runs.

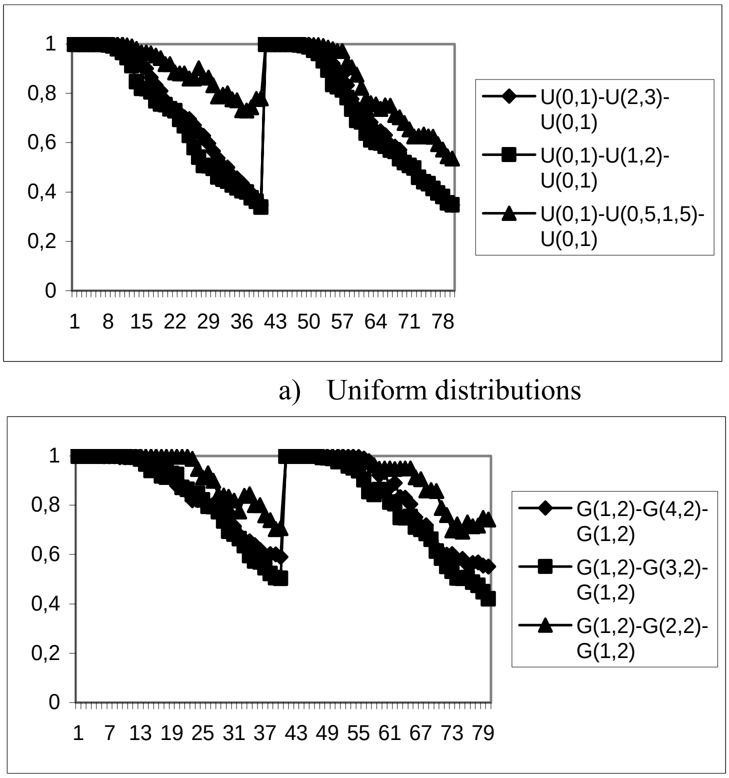
Here $N(\mu, \sigma)$ is the Gaussian distribution with the mean μ and the standard deviation σ, $U(a, b)$ is the uniform distribution on an interval (a, b), $LN(\mu, \sigma)$ is the lognormal distribution with the mean μ and the standard deviation σ, Exp(λ) is the exponential distribution with the parameter λ, Γ(α, β) is the gamma distribution with parameters α and β (Fig. 7–9). Consider a time series $x_1, x_2, ..., x_n, ...$. A change-point in the time series is a point x_m such that a sample $\left(x_1, x_2, ..., x_m\right)$ is drawn from a distribution F_1 and a sample $\left(x_{m+1}, x_{m+2}, ..., x_n\right)$ is drawn from a distribution $F_2 \neq F_1$. Let the sample $\left(x_1, x_2, ..., x_n\right)$ be fixing. Consider the sliding window $\left(x_i, x_{i+1}, ..., x_{i+n}\right)$, where $i = 1, ..., n.$ As i increases the sliding window “contaminated by the elements of the sample $\left(x_{n+1}, x_{n+2}, ..., x_{2n}\right)$.



a) Gaussian distributions with different means

b) Lognormal distributions with different means

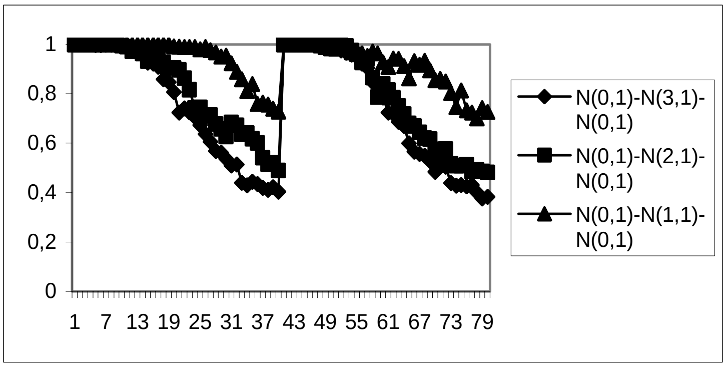
Figure 10: P-statistics for a) Gaussian and b) lognormal distributions



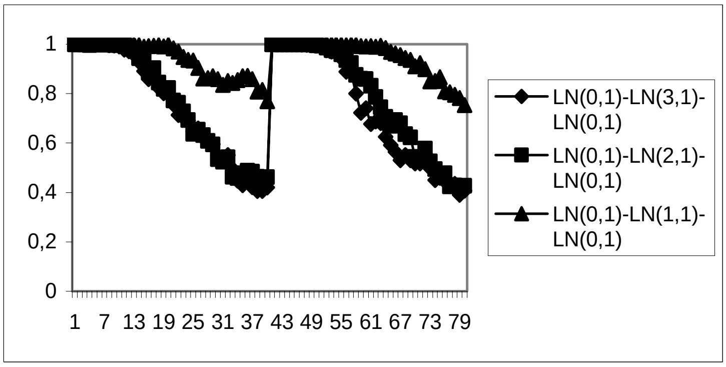
a) Uniform distributions

b) Gamma distributions

Figure 11: P-statistics for a) uniform and b) gamma distributions

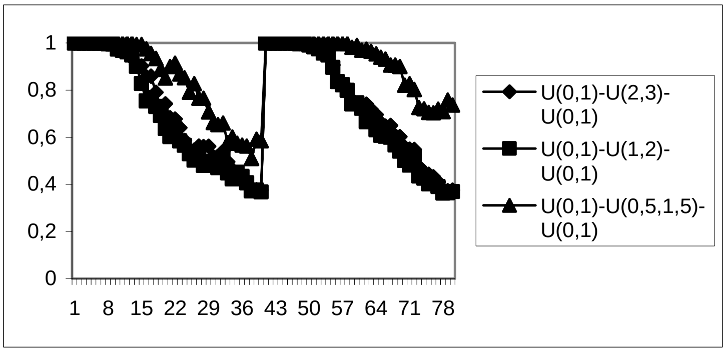


a) Gaussian distributions

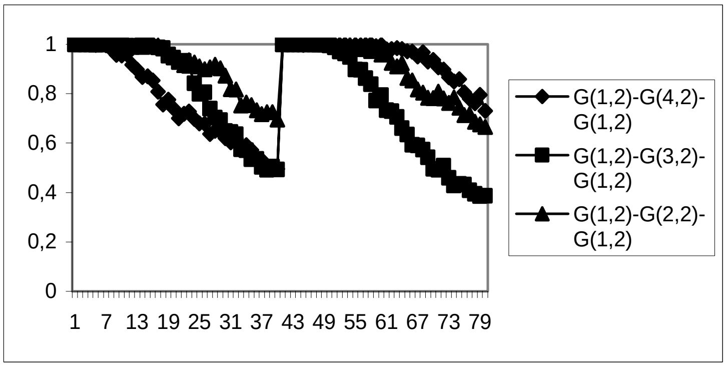


b) Lognormal distributions

Figure 12: Simplified P-statistics (first version) for a) Gaussian and b) lognormal distributions

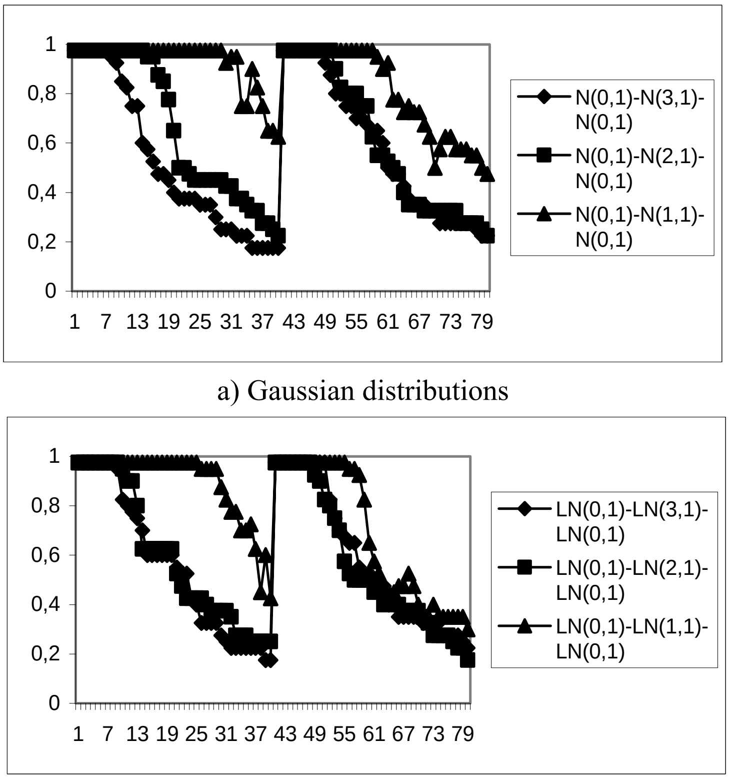


a) Uniform distributions



b) Gamma distributions

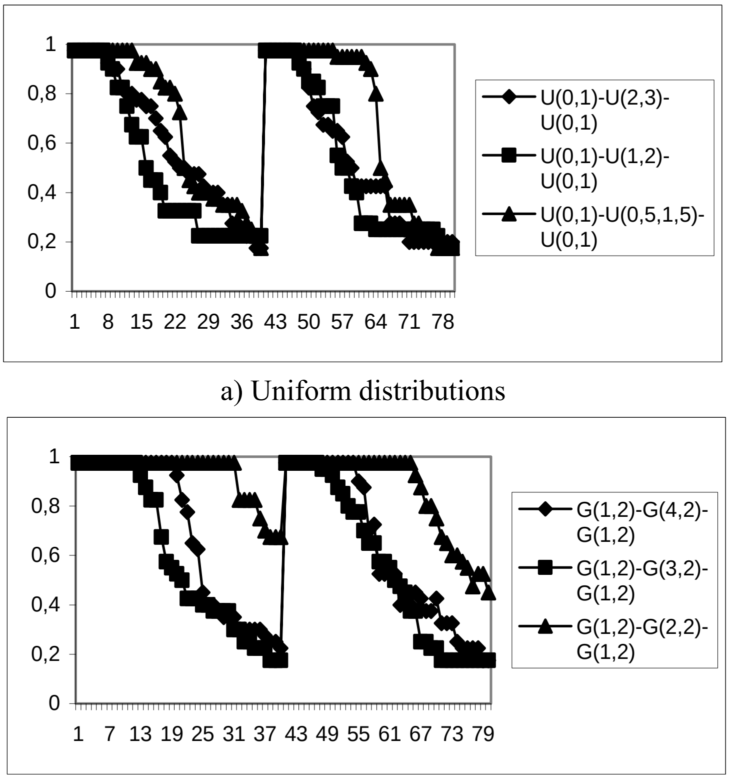
Figure 13: Simplified P-statistics (first version) for a) uniform and b) gamma distributions



a) Gaussian distributions

b) Lognormal distributions

Figure 14: Simplified P-statistics (second version) for a) Gaussian and b) lognormal distributions



a) Uniform distributions

b) Gamma distributions

Figure 15: Simplified P-statistics (second version) for a) uniform and b) gamma distribution functions

The *p*-statistics attains the minimum at a change point and increases when the sliding window moves further (see Figures 10–15). The sensitivity of versions of the Klyushin–Petunin test is shown in Table 1. The earlier a test detects the "contamination, the more sensitive it is.

Table 1

Sensitivity of versions of the Klyushin–Petunin test (order number of the detected change point)

Distribution	Original version	First version (N=100)	Second version (N=40)
N(0,1)–N(3,1)	12	14	8
N(0,1)–N(2,1)	14	17	15
N(0,1)–N(1,1)	22	29	28
N(0,1)–N(0,4)	15	12	14
N(0,1)–N(0,3)	26	17	14
N(0,1)–N(0,2)	27	32	18
LN(0,1)–LN(3,1)	12	12	9
LN(0,1)–LN(2,1)	23	12	10
LN(0,1)–LN(1,1)	28	21	28
LN(0,1)–LN(0,4)	14	13	9
LN(0,1)–LN(0,3)	18	17	9
LN(0,1)–LN(0,2)	25	21	28
U(0,1)–U(2,3)	12	11	6
U(0,1)–U(1,2)	12	11	6
U(0,1)–U(0.5,1.5)	16	16	12
Exp(1)–Exp(4)	17	14	14
Exp(1)–Exp(3)	18	18	29
Exp(1)–Exp(2)	22	37	32
Γ(1,2)–Γ (4,1)	17	10	18
Γ(1,2)– Γ (4,2)	15	18	10
Γ(1,2)– Γ (2,2)	22	19	30

If one test detects a change-point earlier that other do, it is considered as more sensitive. This fact does not affect the accuracy of the change point detection because despite of the detection of the contamination the p-statistics monotonically decreases to a change point at the left end of the sliding window. After this point, the *p*-statistics becomes monotonically increasing.

All the results are consistent. For example, when the first segment $(x_1, x_2, ..., x_{40})$ has the distribution N (0,1) and the second segment $(x_1, x_2, ..., x_{80})$ has the distribution N(3,1), the first sample is considered contaminated when $m > 16$ according to the original Klyushin–Petunin original test (see Table 1). When the change point is equal to 40 then the corresponding test is considered as failed. As expected, that the Klyushin–Petunin test in general is more robust and sensitive than its simplified versions.

The Table 1 shows that the Klyushin–Petunin test is sensitive both for distributions with different means and the same standard deviation (for example, N(0,1) vs N(1,4), and similar variants) and for distributions with the same means and the different standard deviations (for example, N(0,1) vs N(0,4), and similar variants). It also demonstrates the high performance for distributions with different means and standard deviations (exponential and gamma distributions). The less distributions differ from each other, the earlier change points are detected.

Table 1 demonstrates that the original Klyushin–Petunin test is most robust. The first and second simplified versions of the p-statistics are not monotonic. Due to the high sensitivity and robustness the original Klyushin–Petunin test may be considered as an effective method for testing samples heterogeneity and change-point detecting. The fact that first and second simplified versions are unstable is quite understandable, since they are based on the incomplete information (former) or random choice of intervals (latter). Therefore, despite on the complicated computations, the original version of the Klyushin–Petunin test is a preferred choice.

4. Conclusion

The accuracy, sensitivity and specificity of the simplified versions of the Klyushin–Petunin test are comparable with the original version of this test. However, the original version, despite of its computational complication, is more robust. All the versions of the Klyushin–Petunin test do not depend on assumptions about parameters of distributions and equally sensitive to difference between means and standard deviations of distributions. Their significance levels are less that 0.05. They do not require large storage for saving data. All the versions of the Klyushin–Petunin test are more effective than the Kolmogorov–Smirnov and Mann–Whitney–Wilcoxon tests for small samples (size less than 40).

REFERENCES

[1] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline changepoint detection methods, Signal Processing, Elsevier, 167 (2020) 107299. doi:10.1016/j.sigpro.2019.107299.hal-02442692

[2] S. Aminikhanghahi, D. J. Cook, "A survey of methods for time series change point detection, Knowledge and Information Systems, 51 (2017) 339–367. doi:10.1007/s10115-016-0987-z.

[3] E. Gombay, L. Horvath, An application of the maximum likelihood test to the change-point problem, Stochastic Processes and Their Applications, 50 (1994) 161–171. doi:10.1016/0304-4149(94)90154-6.

[4] E. Gombay, L. Horvath, "On the rate of approximations for maximum likelihood tests in change-point models, Journal of Multivariate Analysis, 56 (1996) 120–152. doi: 10.1006/jmva.1996.0007.

[5] G. Gurevich, "Retrospective parametric tests for homogeneity of data, Communications in Statistics. Theory and Methods, 36 (2007) 2841–2862. doi:10.1080/03610920701386968.

[6] B. James, K. James, and D. Siegmund, Tests for a change-point, Biometrika, 74 (1987) 71–83. doi: 10.1093/biomet/74.1.71.

[7] A. Vexler, G. Gurevich, "Average most powerful tests for a segmented regression, Communications in Statistics. Theory and Methods, 38 (2009) 2214–2231. doi: 10.1080/03610920802521208.

[8] G. Gurevich, A.Vexler, "Retrospective change point detection: from parametric to distribution free policies, Communications in Statistics. Simulation and Computation, 39 (2010) 1–22. doi:10.1080/03610911003663881.

[9] B. Brodsky, B. Darkhovsky, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series, Kluwer Academin Press, Dordrecht, Boston, 1993. doi:10.1007/978-94-015-8163-9.

[10] B. Brodsky, B. Darkhovsky, Non-Parametric Statistical Diagnosis: Problems and Methods, Springer, Netherlands, 2010. doi:10.1007/978-94-015-9530-8.

[11] B. Brodsky, Change-point analysis in nonstationary stochastic models, CRC Press, Boca Raton, 2017. doi:10.1201/9781315367989.

[12] J. Chen, A. Gupta, Parametric Statistical Change Point Analysis With Applications to Genetics, Medicine, and Finance, Birkhauser, 2012. doi:10.1007/978-0-8176-4801-5.

[13] D. Ferger, "On the power of nonparametric changepoint-tests, Metrika, 41 (1994) 277–292. doi: 10.1007/BF01895324.

[14] A. Pettitt, "A non-parametric approach to the change-point problem, Applied Statistics, 1979 28, 126–135. doi:10.2307/2346729.

[15] D. Wolfe, E. Schechtman, "Nonparametric statistical procedures for the changepoint problem, Journal of Statistical Planning and Inference, 9 1984 3896–3396. doi:10.1016/0378-3758(84)90013-2.

[16] E. Gombay, "U-statistics for sequential change detection, Metrika, 52 (2000) 113–145. doi:10.1007/PL00003980.

[17] E. Gombay, "U-statistics for change under alternatives, Journal of Multivariate Analysis, 78 2001 139–158. doi:10.1006/jmva.2000.1945.

[18] C. Zou, Y. Liu, P. Qin, and Z. Wang, “Empirical likelihood ratio test for the change-point problem, Statistics & Probability Letters, 77 (2007) 374–382. doi:10.1016/j.spl.2006.08.003.

[19] M. Holmes, I. Kojadinovic, and J. Quessy, “Nonparametric tests for change-point detection a la Gomabay and Hovath, Journal of Multivariate Analysis, 115 (2013) 16–32. doi:10.1016/j.jmva.2012.10.004.

[20] P. Fearnhead, Z. Liu, “On line inference for multiple change point problems, Journal of Royal Statistical Society, Series B, 69 (2007) 203–213, doi:10.1111/j.1467-9868.2007.00601.x.

[21] Y. Mei, “Sequential change-point detection when unknown parameters are present in the pre-change distribution, The Annals of Statistics, 34 (2006) 92–122. doi:10.1214/009053605000000859.

[22] D. Siegmund, Sequential analysis, Springer Series in Statistics. Springer-Verlag, New York, 1985. doi:10.1007/978-1-4757-1862-1.

[23] H. Poor, O. Hadjiliadis, Quickest Detection, Cambridge University Press, Cambridge, 2009. doi:10.1017/CBO9780511754678.

[24] A. Tartakovsky, B. Rozovskii, R. Blazek, and H. Kim, A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods, IEEE Trans. Signal Process, 54 (2006) 3372–3382. doi: 10.1109/TSP.2006.879308.

[25] D. Klyushin, Yu. Petunin, “A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples, Ukrainian Mathematical Journal, 55 2003 181–198. doi:10.1023/A:1025495727612.

[26] D. Klyushin, I. Martynenko, “Nonparametric Test for Change-Point Detection in Data Stream, 2020 IEEE Third International Conference on Data Stream Mining and Processing (DSMP), 2020, pp. 281–286. doi:10.1109/DSMP47368.2020.9204193.

[27] S. Matveichuk, Yu. Petunin, “A generalization of the Bernoulli model occurring in order statistics. I., Ukrainian Mathematical Journal, 42 1990 459–466. doi:10.1007/BF01071335.

[28] S. Matveichuk, Yu. Petunin, “A generalization of the Bernoulli model occurring in order statistics. II., Ukrainian Mathematical Journal, 43 (1991) 728–734. doi:10.1007/BF01058940.

[29] N. Johnson, S. Kotz, “Some generalizations of Bernoulli and Polya-Eggenberger contagion models, Statistical Papers, 32 (1991) 1–17. doi:10.1007/BF02925473.

[30] B. Hill, “Posterior distribution of percentiles: Bayes' theorem for sampling from a population, Journal of the American Statistician Association, 63 1968 677–691. doi:10.1080/01621459.1968.11009286.

[31] A. Pires, C. Amado, Interval Estimators for a Binomial Proportion: Comparison of Twenty Methods, REVSTAT-Statistical Journal, 6 (2008) 165–197. doi:10.1080/01621459.1968.11009286.

[32] H. Freudenthal, “The ‘empirical law of large numbers’ or ‘The stability of frequencies’, Educ Stud Math, 4 (1972) 484–490. doi:10.1007/BF00567002.