Nonparametric Change Point Detection in Time Series Using Dempster–Hill procedure

Dmitriy Klyushin [1] and Irina Martynenko [2]

[1] Taras Shevchenko National University of Kyiv, Ukraine, 03680, Kyiv, Akademika Glushkova Avenue, 4D
[2] National University of Life and Environmental Sciences of Ukraine, Ukraine, 03041, Kyiv, Henerala Rodimtseva, 19, build.1

Abstract

The paper describes a nonparametric test for recognizing the point of change in the time series, before and after which the values of the time series obey different distributions. The test is based on the Matveychuk–Petunin scheme, which is a generalization of the Bernoulli scheme using Dempster–Hill procedure. To recognize the point of change in the time series, a simplified Klyushin–Petunin homogeneity criterion is used, based on an exact confidence interval. The test works equally well with samples that do not have ties, as well as with samples having ties. It allows both online and offline implementations. The test compares segments of time series with high accuracy with a significance level of no more than 0.05. The sensitivity and stability of the proposed test is higher than that of its classical counterparts. The test provides high accuracy of recognition of two heterogeneous random samples for both the location shift hypothesis and the scale shift hypothesis. The proposed approach has wide practical applications in all areas where time series arise.

Keywords [1]

Time series, change point, Dempster–Hill procedure, Klyushin–Petunin test, Bernoulli scheme

1. Introduction

The problem of finding change points in time series has now become ubiquitous. It arises, for example, in medical applications in which it is necessary to continuously monitor the vital signs of patients. This task is typical for technological processes monitoring also. Early recognition of changes in the distribution of time series values makes it possible to identify and prevent unfavorable situations, including deterioration in the condition of patients, disruption of the flow of technological processes, industrial accidents, etc. Therefore, the development of accurate and stable algorithms for the appearance of change points in time series is an urgent task.

The problem of finding points of change in the time series is posed as follows: to find points before and after which the values of the time series obey different distributions. To do this, it is necessary to test the hypothesis that the distributions of random values of the time series in adjacent intervals are identical. If this hypothesis is rejected, the point separating these intervals is called the change point. The paper describes a new approach to finding change points in a time series based on an exact confidence interval.

Change point recognition methods are divided into online and offline methods. Online methods find change points by analyzing data streams in real time. Offline methods detect change points by analyzing the time series as a whole. For an overview of the corresponding algorithms, see [1]. In this paper, we consider one-dimensional random variables. Since multidimensional time series are widespread in various subject areas, there are many methods for finding transition points in

Information Technology and Implementation (IT&I-2022), November 30 - December 02, 2022, Kyiv, Ukraine
EMAIL: dmytroklyushin@gmail.com (D. Klyushin),dokmed5@gmail.com (I. Martynenko)
ORCID: 0000-0003-4554-1049 (D. Klyushin), 0000-0003-0249-4887 (I. Martynenko)

multivariate data streams. An overview of modern methods for finding transition points in multidimensional time series are published in [2, 3].

Our approach use the Dempster–Hill procedure (aka Hill Assumprion A(n) or Nonparametric Predictive Inference) [4], that is thoroughly investigated and applied for solving various problems in papers of F. Coolen (see, for example, [5–8]) and V. Vovk (see, for example, [9–11]).

Coolen, Coolen-Maturi, and Alqifari [5] presented nonparametric predictive inference for future order statistics and joint and conditional probabilities for events involving multiple future order statistics. The authors shown the use of predictive probabilities for order statistics in statistical inference. Bakera, Coolen-Maturi, and Coolen [6] introduced nonparametric predictive inference (NPI) for stock returns and presented the inference on future stock returns, illustrating the proposed NPI methods by historical stock market data. Yin, Coolen, and Coolen-Maturi [8] provided an exploration of the statistical methods based on imprecise probabilities for accelerated life testing, applying nonparametric predictive inference. Alqifari and Coolen [7] considered robustness of Nonparametric Predictive Inference (NPI), in particular inference involving future order statistics. The authors introduced new concepts for assessing the robustness of statistical procedures to the NPI and demostrated that most of their nonparametric inferences had good robustness to small changes in the data. Vovk et al [9] derived predictive distributions that are valid under a nonparametric assumption using applied conformal prediction. The authors introduced and explored predictive distribution functions that always guarantee coverage for i.i.d. observations. Their algorithm generalizes the classical Dempster-Hill predictive distributions. Vovk et al. [10] proposed schemes based on exchangeability martingales. Their method is general and may be applied to any prediction algorithm. Vovk [11] described a universal probability forecasting systems, i.e. a system that is consistent for any distribution, provided that the observations are i.i.d., and proved the existence of universal conformal predictive systems.

In opposite to papers mentioned above, we construct our approach on the Matveychuk–Petunin and Jonson–Kotz models [12–15] that are generalized Bernoulli schemes.

2. Homogeneity and change-point detection test

Consider two samples $x=(x_1,x_2,...,x_n)$ and $y=(y_1,y_2,...,y_n)$ drawn fro the distributions F_1 and F_2, respectively. The null hypothesis H_0 states that $F_1=F_2$. The alternative hypothesis is $F_1 \neq F_2$. The Matveychuk–Petunin and Jonson–Kotz models [12–15] allow construction of a two-sided confidence interval (p_1,p_2) with a given the significance level for both the true and false null hypothesis H_0.

Let $x_{(1)},x_{(2)},...,x_{(n)}$ be variance series constructed using the sample x . If H_0 is true and the sample x obeys an exchangeable continuous distribution, then the Hill's assumption [4] states that

$$P\left(x \in \left(x_{(i)},x_{(j)}\right)\right)=\frac{j-i}{n+1},\ j<i, \tag{1}$$

If the null hypothesis H_0 is false, then the probability of the random event $A_{ij}=\left\{x \in \left(x_{(i)},x_{(j)}\right)\right\}$ significantly deviates from (1). To estimate this deviation we construct $N=(n-1)n/2$ confidence limits $I_{ij}=\left(p_{ij}^{(1)},p_{ij}^{(2)}\right)$ for the binomial proportion p_{ij} corresponding to given significance level βusing various formulas [16]. Since these intervals have different coverage probability and lengths, the most natural choice is to use an exact confidence interval, like the Clopper–Pearson interval [17]. It allows avoiding problems connected with the varying coverage probability and selection of parameters. Let L be the number of intervals I_{ij} containing p_{ij} and $\rho(x,y)=L/N$ is the relative frequency of the random event $B=\left\{p_{ij} \in I_{ij}\right\}$ with the probability $1-\beta$. Using the arguments described above, we can construct the confidence interval I for the probability $p(B)$ with the significance level that is less than 0.05. The decision rule of the Klyushin–Petunin test is formulated in the following way: if the confidence interval I contains 0.95 then the hypothesis H_0 is accepted, otherwise it is rejected. The statistics $\rho(x,y)$ is a heterogeneous measure of the samples x and y.

As far as N can be quite large, the original version of the test may request quite long computations. Therefore, it is desirable to simplify the test. We propose do not use all intervals $\left(x_{(i)},x_{(j)}\right)$ but randomly choose the fixed number M of such intervals.

Consider the process of construction exact confidence interval for binomial proportion based on the 3σ-rule. Let x be a unimodal random value. Then, the 3σ-rule holds [18]

$$p\left(\left|x-m(x)\right|\geq 3\sigma(x)\right)\leq\frac{4}{81}$$

where $m(x)$ is the mean and $\sigma(x)$ is the standard deviation . Therefore, the coverage probability of the confidence interval $\left(m(x)-3\sigma(x),m(x)+3\sigma(x)\right)$ is greater than 0.95

In the classical Bernoulli model we have

$$a=m(x)-3\sigma(x)=np+\frac{1}{2}-3\sqrt{npq+\frac{1}{2}},\quad b=m(x)+3\sigma(x)=np+\frac{1}{2}+3\sqrt{npq+\frac{1}{2}},$$

Therefore, the coverage probability of the confidence interval (a,b) follows the binomial distribution, i.e. the significance level of the confidence interval

$$I=\left(np+\frac{1}{2}-3\sqrt{npq+\frac{1}{2}},np+\frac{1}{2}+3\sqrt{npq+\frac{1}{2}}\right)$$

does not exceed 0.05.

Let us re-state the random event $x\in I$ as follows:

$$\left|x-np\right|\leq\frac{1}{2}+3\sqrt{npq+\frac{1}{2}}.$$

Therefore, in the Bernoulli model we have

$$P\left(\left|h-p\right|\leq\frac{1}{2n}+\frac{3}{n}\sqrt{npq+\frac{1}{12}}\right).$$

To construct the exact confidence interval for the binomial proportion p on the relative frequency h in the Bernoulli model consisting of n independent trials introduce two functions depending of

$$p\in[0,1]:\varphi(p)=\left|h-p\right| \text{ and } \quad \tilde{\psi}(p)=\sqrt{np(1-p)+\frac{1}{12}}.$$

Denote

$$\tilde{\psi}(p)=\sqrt{np(1-p)+\frac{1}{12}}.$$

The graph $\tilde{\psi}(p)$ is the upper half of an ellipse E passing through the points

$$A=\left(\frac{1}{2n}\left(n+\sqrt{\frac{n}{3}+n^2}\right),0\right),\ B=\left(\frac{1}{2},\sqrt{\frac{1}{12n}+\frac{1}{4}}\right),\ C=\left(\frac{1}{2n}\left(n-\sqrt{\frac{n}{3}+n^2}\right),0\right),\ D=\left(\frac{1}{2},-\sqrt{\frac{1}{12n}+\frac{1}{4}}\right)$$

with the center $\left(\frac{1}{2},0\right)$. The graph of $\psi(p)$ is the restriction of the graph of $\tilde{\psi}(p)$ on the segment $[0,1]$ stretching or shrinking the graph by $\frac{3}{n}$ and shifting it by $\frac{1}{2n}$.

Therefore, the graph of the function $\psi(p)$ which does not depend on h is an arc of ellipse Γ passing through the points $(0,\psi(0))$, $\left(\frac{1}{2},\psi\left(\frac{1}{2}\right)\right)$, $(1,\psi(1))$, such that the function $\psi(p)$ reach the minimum at the point $p=\frac{1}{2}$ and it is symmetrical with respect to this point.

The lower confidence limit p_1 is a root of the quadratic equation

$$\left(1+\frac{9}{n}\right)p^2-\left(\frac{9}{n}-\frac{1}{n}+2h\right)p+h^2-\frac{h}{n}-\frac{1}{2n^2}=0. \tag{2}$$

If $h>\psi(0)=\frac{1}{2n}+\frac{3}{n\sqrt{12}}$, then the lower confidence limit p_1 is the least root of (2). If $h\le\psi(0)$, then $p_1=1$.

The upper confidence limit p_2 is a root of the quadratic equation

$$\left(1+\frac{3}{n}\right)p^2-\left(\frac{3}{n}+\frac{1}{n}+2h\right)p+h^2+\frac{h}{n}-\frac{1}{2n^2}=0.$$

If $1-h>\psi(1)$, then the upper confidence limit p_2 is the largest root of (3). If $1-h<\psi(1)$, then $p_2=1$.

For the generalized Bernoulli model similar reasoning gives the following quadratic equation for lower confidence limit:

$$\left(1+\frac{9(m+n+1)}{(n+2)m}\right)p^2+\left(\frac{1}{m}-\frac{9(m+n+1)}{(n+2)m}-2h\right)p+h^2-\frac{h}{m}-\frac{1}{2m^2}=0 \tag{3}$$

If $h>\frac{1}{2m}+\frac{3}{m\sqrt{12}}=\gamma$, then the lower confidence limit p_1 for the generalized Bernoulli model is the least root of (3). If $h\le\gamma$, then $p_1=0$.

Similar, the upper confidence limit p_2 for the generalized Bernoulli model is the root of the equation

$$\left(1+\frac{9(m+n+1)}{(n+2)m}\right)p^2-\left(\frac{1}{m}+\frac{9(m+n+1)}{(n+2)m}+2h\right)p+h^2+\frac{h}{m}-\frac{1}{2m^2}=0 \tag{4}$$

If $1-h>\gamma$, then the upper confidence limit p_2 is the largest root of (4). If $1-h\le\gamma$, then $p_2=1$. By virtue of the previous results the significance level of the confidence interval does not exceed 0.05.

3. Comparison of the sensitivity of versions of Klyushin–Petunin test

In [19] we compared the sensitivity and precision of the Klyushin–Petunin test based on the Wilson confidence interval with the sensitivity and precision of the Kolmogorov–Smirnov test and Wilcoxon test. Now, we compare the sensitivity of the Klyushin–Petunin test, Klyushin–Petunin exact confidence interval based on the 3σ-rule with complete selection of the intervals $\left(x_{(i)},x_{(j)}\right)$ and its simplified version when we use only given number of randomly selected intervals $\left(x_{(i)},x_{(j)}\right)$. In the simplified version we do not make exhaust selection of the intervals $\left(x_{(i)}, x_{(j)}\right)$ but just randomly select 100 intervals, using the practical observation that the relative frequency almost exactly approximates the probability after 100 trials [20]. We generated samples (n = 40) drawn from the distributions which have the same location and the different scale, the same scale and the different locations, and the different scales and locations. Hereinafter we use the following notation: N(μ, σ) is the normal distribution, where μ is the mean and σ is the standard deviation, U(a, b) is the uniform distribution on an interval (a,b), LN(μ, σ) is the lognormal distribution, E(λ) is the exponential distribution with the parameter λ, and G(k, Θ) is the gamma distribution with parameters k and Θ.

Consider the segment of the time series $(x_1, x_2, ...)$. The change point of this time series is the point x_m such that $(x_1, x_2, ..., x_m)$, $m < n$ has the distribution F_1 and $(x_{m+1}, x_{m+2}, ...)$ has the different distribution F_2. We propose to find a change point in the following way. Consider the sample $(x_1, x_2, ..., x_k)$ and a sliding segment $(x_i, x_{i+1}, ..., x_{i+k})$ where i = 1, …, n. As i increases the sliding window sample becomes "contaminated" by the elements of the second sample. Ideally, when we reach a change point, the homogeneity measure attains its minimum value, and when the sliding window moves across the change point the homogeneity measure increases. Therefore, the graph of the homogeneity measure shows a saw tooth pattern. The Klyushin–Petunin homogeneity measure is monotonically decreasing before the change points and monotonically increasing after the change point. In Table 1 we show the result of the comparison of the sensitivity of the various version of the Klyushin–Petunin test. If the test detect a change-point earlier than its counterparts, it is considered as more sensitive. 6In Table 1 the order numbers of the contaminants detected by the Klyushin–Petunin test when we consider all the intervals $\left(x_{(i)}, x_{(j)}\right)$ (original version), complete exact Klyushin–Petunin test and simplified exact version 100 with 5%- significance level for various distributions are represented. The change-point in Table 1 is such point x_k that the test accepts the hypothesis H_0 for samples $(x_1, x_2, ..., x_k) \in F_1$ and $(x_{k+1}, x_{k+2}, ..., x_n) \in F_2$, $k \le n$.

Table 1

Order numbers of change-points

Test	Tests		
	Original version	Complete exact version	Simplified exact version
N(0,1)–N(3,1)	15	17	16
N(0,1)–N(2,1)	16	19	18
N(0,1)–N(1,1)	22	30	21
N(0,1)–N(0,2)	23	24	33
N(0,1)–N(0,3)	18	17	21
N(0,1)–N(0,4)	22	17	20
LN(0,1)–LN(3,1)	12	14	18
LN(0,1)–LN(2,1)	15	24	19
LN(0,1)–LN(1,1)	25	26	20
LN(0,1)–LN(0,4)	24	28	22
LN(0,1)–LN(0,3)	19	18	29
LN(0,1)–LN(0,2)	29	33	31
U(0,1)–U(2,3)	13	12	14
U(0,1)–U(1,2)	12	14	12
U(0,1)–U(0.5,1.5)	22	27	17
E(1)–E(4)	21	17	24
E(1)–E(3)	22	26	25
E(1)–E(2)	27	29	24
G(1,2)–G(4,1)	15	14	17
G(1,2)–G(4,2)	17	15	15
G(1,2)–G(2,2)	20	17	24

Remember, that this fact does not effect on the precision of the change point detection because despite the detection of a contamination the homogeneity measure monotonically decreases until the left end of the sliding window attain the change point. After this the Klyushin–Petunin homogeneity measure becomes monotonically increasing.

For instance, when the first segment $\left(x_1, x_2, ..., x_{40}\right)$ has the distribution N(0,1) and the second segment $\left(x_{41}, x_{42}, ..., x_{80}\right)$ has the distribution N(3,1), the sample $\left(x_1, x_2, ..., x_{40}\right)$ is considered contaminated when $m > 15$ according to the complete Klyushin-Petunin original test (see Table 1). It is easy to see, that the Klyushin–Petunin test is more stable than its counterparts in all considered cases. If the entry of Table 1 is 40 then the corresponding test did nor reject the null hypothesis H_0.

The Table 1 shows that the Klyushin–Petunin test is more sensitive for shifted distributions with the different means and the same standard deviation (N(0,1) vs N(1,1), N(2,1), and N(3,1), LN(0,1) vs Lognogmal(1,1), LN(2,1), and LN(3,1), U(0,1) vs U(2,3), U(1,2), and U(0.5,1.5)) than its counterparts. For exponential and gamma distributions the exact Klyushin–Petunin test is in average more sensitive.

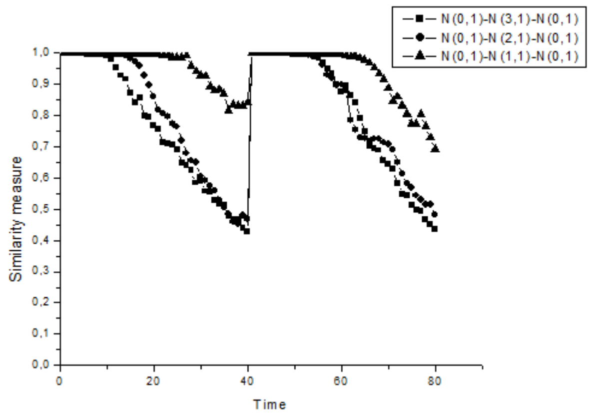

Figure 1: P-statistics between samples from different normal distributions, different locations

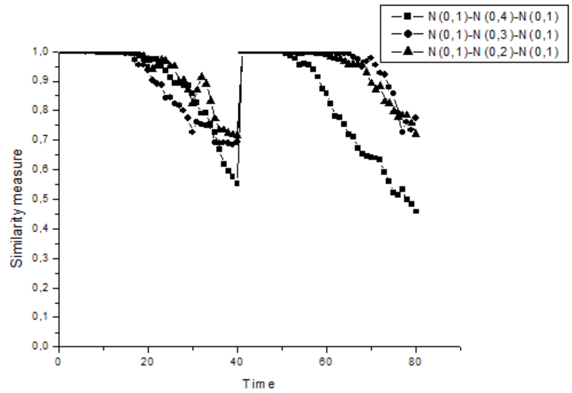

Figure 2: P-statistics between samples from different normal distributions, different scales

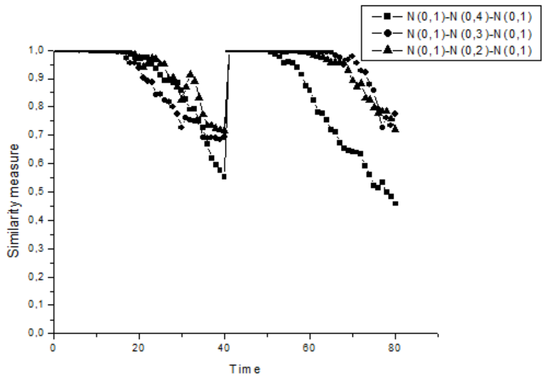

Figure 3: P-statistics between samples from different uniform distributions

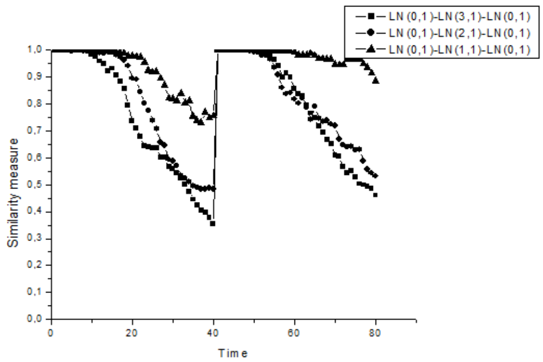

Figure 4: P-statistics between samples from different lognormal distributions, different locations

The Table 1 shows that the Klyushin–Petunin test is more sensitive for distributions with the the means and the different standard deviation (N(0,1) vs N(0,4), N(0,3), and N(0,2), LN(0,1) vs Lognogmal(0,4), LN(0,3), and LN(0,2) than its counterparts. The second important result shown in the Table 1 is the fact that the original Klyushin–Petunin test is more robust than its versions in all cases. The combination of the high sensitivity and robustness makes it the effective tool for test heterogeneity and change-point detection. In addition, Fig. 1–5 demonstrate monotonic property of the *p*-statistics and the fact that a sew-like form of its graphs where "a tooth of sew is a change point of a time series, allows exactly detect change points.

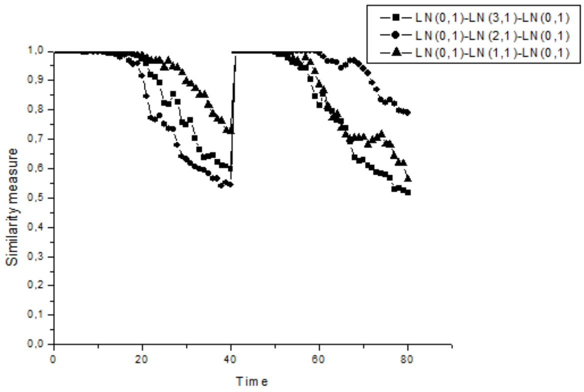

Figure 5: P-statistics between samples from different lognormal distributions, different scales

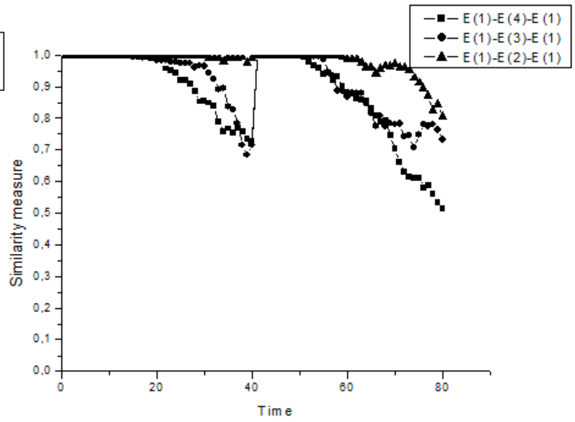

Figure 6: P-statistics between samples from different exponential distributions

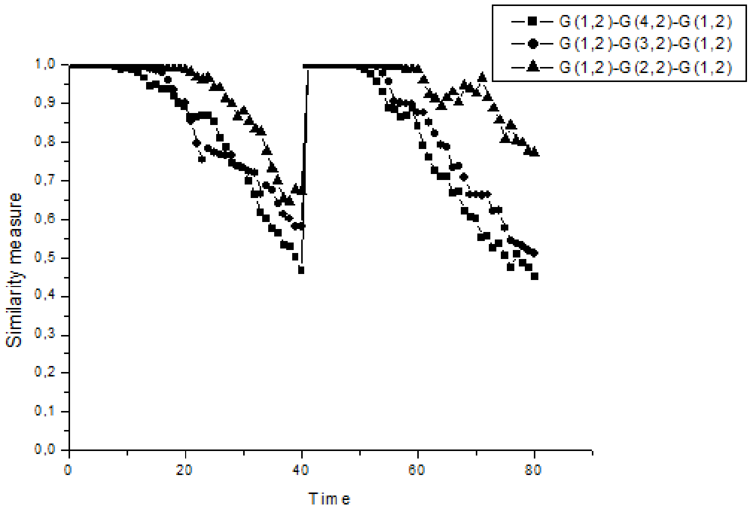

Figure 7: P-statistics between samples from different gamma distributions

4. Conclusion

We considered a nonparametric test for recognizing the point of change in the time series, before and after which the values of the time series obey different distributions. This test uses the Matveychuk–Petunin scheme, which is a generalization of the Bernoulli scheme using Dempster–Hill procedure. To recognize the point of change in the time series, we uses the original Klyushin–Petunin test, the exact Klyushin–Petunin test and simplified Klyushin–Petunin homogeneity test based on the proposed exact confidence interval. All the tests compared segments of time series with high accuracy with a significance level of no more than 0.05. The sensitivity and stability of the proposed tests is higher than that of its classical counterparts. The tests provide high accuracy of recognition of two heterogeneous segments for both the location shift hypothesis and the scale shift hypothesis. The proposed approach has wide practical applications in all areas where time series arise.

The original Klyushin–Petunin test based on the Wilson confidence interval is the most sensitive, robust and accurate for almost all considered distributions. The modifications of this test using the exact Klyushin–Petunin confidence interval has the same precision, require less computation, but are less robust. Therefore, they could be used as tools for detection change points in data streams in situations when the speed of computations is more important than the robustness. Nevertheless, future work will focus on improving the robustness of the proposed test and investigating the multivariate case.

5. References

[1] C. Truong, L. Oudre, N. Vayatis, A review of change point detection methods, CoRR abs/1801.00718 (2018), http://arxiv.org/abs/1801.00718.

[2] C. Truong, L. Oudre, N. Vayatis, Selective review of offline change-point detection methods, Signal Processing 167 (2020) 107299. doi:10.1016/j.sigpro.2019.107299.

[3] S. Aminikhanghahi, D. J. Cook, A Survey of Methods for Time Series Change Point Detection, Knowl. Inf. Syst. 51 (2017) 339–367. doi:10.1007/s10115-016-0987-z.

[4] H. N. Alqifari, F. P. A. Coolen, Robustness of Nonparametric Predictive Inference for Future Order Statistics, J Stat Theory Pract 12 (2019) 13. https://doi.org/10.1007/s42519-018-0011-x

[5] 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.

[6] R. M. Bakera, T. Coolen-Maturi, F. P. A. Coolen. Nonparametric Predictive Inference for Stock Returns, Journal of Applied Statistics 44 (2017) 1333–1349. doi:10.1080/02664763.2016.1204429.

[7] Y.-C. Yin, F. P. A. Coolen, T. Coolen-Maturi. An imprecise statistical method for accelerated life testing using the power-Weibull model, Reliability Engineering & System Safety 167 (2017) 158–167. doi:10.1016/j.ress.2017.05.045.

[8] F. P. A. Coolen, T. Coolen-Maturi, H. N. Alqifari, Nonparametric predictive inference for future order statistics, Communications in Statistics – Theory and Methods, 47 (2018) 2527–2548. doi:10.1080/03610926.2017.1342834.

[9] V. Vovk, J. Shen, V. Manokhin, Min-ge Xie, Nonparametric predictive distributions based on conformal prediction, Machine Learning 108 (2019) 445-474.

[10] V. Vovk, I. Petej, I. Nouretdinov, E. Ahlber, L. Carlsson,, A. Gammerman,. Retrain or not retrain: Conformal test martingales for change-point detection, in: L. Carlsson, Z. Luo, G. Cherubin, K. Nguyen (Eds), Proceedings of Machine Learning Research 152 (2021) 191–210. doi:10.48550/arXiv.2102.10439

[11] V. Vovk, Universal predictive systems. Pattern Recognition. 126 (2022) 108536. doi:10.1016/j.patcog.2022.108536.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] R. Andrushkiw, D. Klyushin, Yu. Petunin, and M. Savkina, The exact confidence limits for unknown probability Bernoulli models, Proceedings of the International Conference on Information Technology Interfaces, ITI, 2005, pp. 164–168. doi: 10.1109/ITI.2005.1491116.

[18] D. Vysochanskii, Yu. Petunin, Justification of the 3-sigma rule for unimodal distributions, Theory of Probability and Mathematical Statistics, 21 (1980) 25–37.

[19] 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

[20] H. Freudenthal, The empirical law of large numbers or the stability of frequencies. Educational Studies in Mathematics 4 (1972), 484–490.