Change Point Detection with Copula Entropy based Two-Sample Test

Jian MA*

Hitachi China Research Laboratory

Abstract

Change point detection is a typical task that aim to find changes in time series and can be tackled with two-sample test. Copula Entropy is a mathematical concept for measuring statistical independence and a two-sample test based on it was introduced recently. In this paper we propose a nonparametric multivariate method for multiple change point detection with the copula entropy-based two-sample test. The single change point detection is first proposed as a group of two-sample tests on every points of time series data and the change point is considered as with the maximum of the test statistics. The multiple change point detection is then proposed by combining the single change point detection method with binary segmentation strategy. We verified the effectiveness of our method and compared it with the other similar methods on the simulated univariate and multivariate data and the Nile data.

Keywords: Change Point Detection; Copula Entropy; Two-Sample Test; Non-Parametric Method

1 Introduction

Change point detection is a typical task that aim to find single or multple changes in time series. The detection can be offline or online and the time series can be univariate or multivariate. In this paper, we focus on offline multivariate multiple change point detection. Many algorithms have been proposed for the task, see [1, 2, 3, 4] for the reviews on this topic. Change point detection can be widely applied to natural, social, or industrial systems where abrupt changes happen.

Two-sample test is a common problem of hypothesis testing in statistics. It is to test the hypothesis whether two samples are from a same distribution. There are many two-sample test based on different mathematical concepts. A typical way of defining test statistic for testing is based on the measures of statistical independence in two samples, such as kernels base measure [5], mutual information [6].

Copula Entropy (CE) is a recently defined mathematical concept for measuring statistical independence [7]. It is proved to be equivalent to mutual

*Email: majian@hitachi.cn

information in information theory. A nonparametric method for estimating it was also proposed [7]. Recently, CE has been applied to two-sample test [8], in which the test statistic is defined as the difference between CEs of two hypotheses.

There are several work on change point detection with copulas. Xiong and Cribben [9] proposed a method for estimating change points with Vine copula and applied it to fMRI data. Bücher et al [10] proposed a change point detection method based on empirical copula process. Stark and Otto [11] proposed to test structural changes in multivariate time series in copula-base dependence measures, such as Spearman's ρ and quantile dependencies.

In this paper, we propose to use CE-based two-sample test for multiple change point detection. The idea is simple: first transforming the change point detection problem into a group of CE-based two-sample tests on every points of time series and then find the change point as that with the maximum of the test statistics. A multiple change point detection problem can be solved by combining the single change point detection method with binary segmentation strategy. Since the CE-based two-sample test is nonparametric and multivariate, the proposed change point detection method is also nonparametric and multivariate. We verified the effectiveness of the proposed method and compared it with the other similar methods on both simulated and real data in this paper.

This paper is organized as follows: Section 2 introduces copula entropy and the two-sample test based on it, Section 3 presents the proposed methods on single and multiple change point detection, experiments with simulated and real data will be presented in Section 4 and Section 5 respectively, followed by some discussion in Section 6, and finally we conclude the paper in Section 7.

2 Methodology

2.1 Copula Entropy

Copula theory is a probabilistic theory on representation of multivariate dependence [12, 13]. According to Sklar's theorem [14], any multivariate density function can be represented as a product of its marginals and copula density function (cdf) which represents dependence structure among random variables.

With copula theory, Ma and Sun [7] defined a new mathematical concept, named Copula Entropy, as follows:

Definition 1 (Copula Entropy)**.** *Let* $\mathbf{X}$ *be random variables with marginals* $\mathbf{u}$ *and copula density function* c*. The CE of* $\mathbf{X}$ *is defined as*

$$H_c(\mathbf{x}) = -\int_{\mathbf{u}} c(\mathbf{u}) \log c(\mathbf{u}) d\mathbf{u}. \tag{1}$$

A non-parametric estimator of CE was also proposed in [7], which composed of two simple steps:

1. estimating empirical copula density function;

2. estimating the entropy of the estimated empirical copula density.

The empirical copula density in the first step can be easily derived with rank statistic. With the estimated empirical copula density, the second step is essentially a problem of entropy estimation, which can be tackled with the KSG estimation method [15]. In this way, a non-parametric method for estimating CE was proposed in [7].

2.2 Two-sample test with CE

CE has been applied to solve the two-sample test problem [8]. Given two samples $\mathbf{X}_1 = \{X_{11}, \cdots, X_{1m}\} \sim P_1, \mathbf{X}_2 = \{X_{21}, \cdots, X_{2n}\} \sim P_2$, the null hypothesis for two sample test is

$$H_0 : P_1 = P_2, \tag{2}$$

and the alternative is

$$H_1 : P_1 \neq P_2. \tag{3}$$

where $\mathbf{X}_1, \mathbf{X}_2 \in R^d$ and P_1, P_2 are the corresponding probability distribution functions.

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and Y_0, Y_1 be two labeling variables for the two hypotheses respectively that $Y_1 = (0_1, \cdots, 0_m, 1_1, \cdots, 1_n)$ and $Y_0 = (1_1, \cdots, 1_{m+n})$. So the CE between $\mathbf{X}$ and Y_i can be calculated as

$$H_c(\mathbf{X}; Y_i) = H_c(\mathbf{X}, Y_i) - H_c(\mathbf{X}). \tag{4}$$

Then the test statistic for H_0 is defined as the difference between the CEs of the two hypotheses, as follows:

$$T_{ce}(X_1, X_2) = H_c(\mathbf{X}, Y_0) - H_c(\mathbf{X}, Y_1). \tag{5}$$

It is easy to know T_{ce} will be a small value if H_0 is true and a large value if H_1 is true.

The test statistic in (5) can be easily estimated from data by estimating the two terms in it with the non-parametric estimator of CE. Since the CE estimator is non-parametric, the estimator of the test statistic can be applied to any cases without assumptions. Another merit of such estimator of the test statistic is hyperparameter-free.

3 Proposed Method

3.1 Single change point detection

In this section, we first propose a method for single change point detection based on the above two-sample test. The idea is simple: for a time series, the CE-based two-sample test is conducted on two sub-series devided by each point of time series and the point associated with the maximal test statistic of these tests is the change point.

Given a time series $\mathbf{X} = \{x_1, \ldots, x_n\}, x_i \in R^d$, the single change point detection problem can be formulated as follows:

$$i = \arg\max_{i \in [1, n-1]} T_{ce}(X_1, X_2), \tag{6}$$

where $T_{ce}(X_1, X_2)$ is the statistic of the CE-based two-sample test on the samples $X_1 = \{x_1, \ldots, x_i\}$ and $X_2 = \{x_{i+1}, \ldots, x_n\}$.

Table 1: Parameters (mean μ and variance δ) of the normal distributions in univariate time series simulations.

	μ_1	δ_1	μ_2	δ_2	μ_3	δ_3	μ_4	δ_4
mean	0	1	5	1	10	1	3	1
mean-var	0	1	5	3	10	1	3	10
var	0	1	0	10	0	5	0	1

3.2 Multiple change point detection

The multiple change point detection can be transformed as a group of the above single change point detection problems with binary segmentation strategy. For a time series data, we first detect a change point with the above single change point detection method, and if detected, then the whole time series is separated into two segments before and after the detected change point. Such detection is continued on the two segments such derived till no change point can be detected on all the following such derived segments.

In the proposed method, a threshold on the test statistic is set for judging whether there is a change point in each segment. A change point is detected if its associated maximal test statistic is larger than the threshold. By means of a threshold, our method can estimate the number of multiple change points automatically.

Our method is based on the two-sample test in Section 2.2. Since the CE-based test is nonparametric and multivariate, the proposed method is also nonparametric and multivariate, and can be applied to any cases without assumptions.

4 Simulations

4.1 Experiments

We conducted simulation experiments to test the proposed method. In each simulation, a univariate or multivariate time series data with several change points was first simulated and our method was then applied to the simulated data to detect these change points. Each time series was composed of four sub-series generated from four different distributions with length as 50 points, which means there are three change points at $[51, 101, 151]$.

For univariate time series, all the sub-series were generated with normal distribution with different mean and variance. We simulated three typical cases of change points: different means, different means and variances, and different variances. The parameters of the normal distributions in the three cases are listed in Table 1.

For multivariate time series, the sub-series were generated with bi-variate normal distributions with different mean and covariances first. We simulated three typical cases as well: different means, different means and covariances, and different covariances. We also simulated a group of sub-series with bi-variate normal distributions and bi-variate copula functions. The copula functions here are frank copula ($\theta = 0.9$) and normal copula ($\rho = 0.3$) both with normal distribution ($\mu = 0$ and $\delta = 2$) and exponential distribution (rate=0.5) as

Table 2: Parameters of the normal distributions or copula functions of the four sub-series in multivariate time series simulations.

	1	2	3	4
mean	$\mu = (0, 0)$, $cov = 0.2$	$\mu = (10, 10)$, $cov = 0.2$	$\mu = (5, 5)$, $cov = 0.2$	$\mu = (1, 0)$, $cov = 0.2$
mean-var	$\mu = (0, 0)$, $cov = 0.2$	$\mu = (10, 10)$, $cov = 0.8$	$\mu = (5, 5)$, $cov = 0.1$	$\mu = (1, 0)$, $cov = 0.9$
mean-var	$\mu = (0, 0)$, $cov = 0.2$	$\mu = (0, 0)$, $cov = 0.8$	$\mu = (0, 0)$, $cov = 0.1$	$\mu = (0, 0)$, $cov = 0.9$
copula	$\mu = (0, 0)$, $cov = 0.2$	frank copula	$\mu = (5, 5)$, $cov = 0.1$	normal copula

marginals. The parameters of the simulations in the three cases are list in Table 2.

We compared our method with traditional change point detection methods. In univariate three cases, our method and the three methods for detecting changes in mean, in mean and variance, and in variance were compared respectively. The binary segmentation strategy [16] were adopted in these three compared methods. In multivariate cases, our method was compared with the kernel change point detection method [17]. The penalty parameter of the kernel method was tuned to obtain the best possible results.

In the experiments, the implementation of the CE-based two-sample test in the **R** package `copent`[18] was used. The thresholds for the test statistics is 0.13 in all the experiments, except for the multivariate case with different covariance case the threshold is 0.05. The compared method in univariate cases was those implemented in the **R** package `changepoint`[19]. The kernel method implemented in the **R** package `ecp`[20] was used. The codes of the experiments are available at `https://github.com/majianthu/cpd`.

4.2 Results

The simulation results on univariate and multivariate time series data are presented in Table 3 and 4 respectively. For the univariate data, our method detected all the change points in different means, different means and variances, and different variances cases, as the compared method did. For the multivariate data, both our method and the kernel method work well in different means, different means and variances cases. In different variances case, our method detected two right change points (48,102) with additional false positives while the kernel method detected only false positives after hyperparameter tuning. In the copula function case, our method detected one change point (155) while the kernel method cannot detect any change point. There are two false positives in the different variances case of univariate data (9) and in the different means case of multivariate case (18) but both test statistics of these false positives are smaller than those of the right change points.

Table 3: Detected change points in univeriate time series simulations.

	Our method	Compared method
mean	52,101,151	50,100,150
mean-var	52,101,151	50,100,150
var	9,50,100,151	50,99,150

Table 4: Detected change points in multivariate time series simulations.

	Our method	Kernel method
mean	51,101,151,18	1,51,101,151,201
mean-var	51,101,151	1,51,101,151,201
var	14,48,102,162,169	1,46,59,80,157,159,201
copula	155	1,201

5 Real data

We verified the effectiveness of the proposed method on the Nile data, a well-known benchmark for change point detection [21], which contains the time series measurement of the annual flow of the river Nile at Aswan from 1871 to 1970 with an apparent decreasing change happened at 1898.

We applied the single change point detection method on the Nile data. The results is shown in Figure 1, from which it can be learned that our method successfully detect the right point where the change of river flow happened and the test statistic reach it maximum as well.

6 Discussion

We proposed a method for multiple change point detection with CE-based two-sample test. The power of the proposed method was tested on the simulated and real data. In the simulated experiments, we compared the proposed method with different methods on univariate and multivariate data. Since the proposed method is nonparametric and multivariate, it can be applied directly to all the cases. As contrast, different compared methods should be used for each case of the univariate data.

Our method has one hyperparameter for the threshold on the test statistic. However, in the experiments, only one value (0.13) was used for all the cases, except for the different variances of the multivariate data. It works so well that we did not have to tune it too much. As contrast, the kernel method should tune its penalty parameter frequently for each case to detect the right change points out. This advantage of our method is because that CE is rigorously defined and model-free, and hence the test statistic of the two-sample test based on it is comparable for all the cases.

There are several false positives in the simulation results of our method. However, they can be avoided easily by means of setting larger threshold of the test statistic.

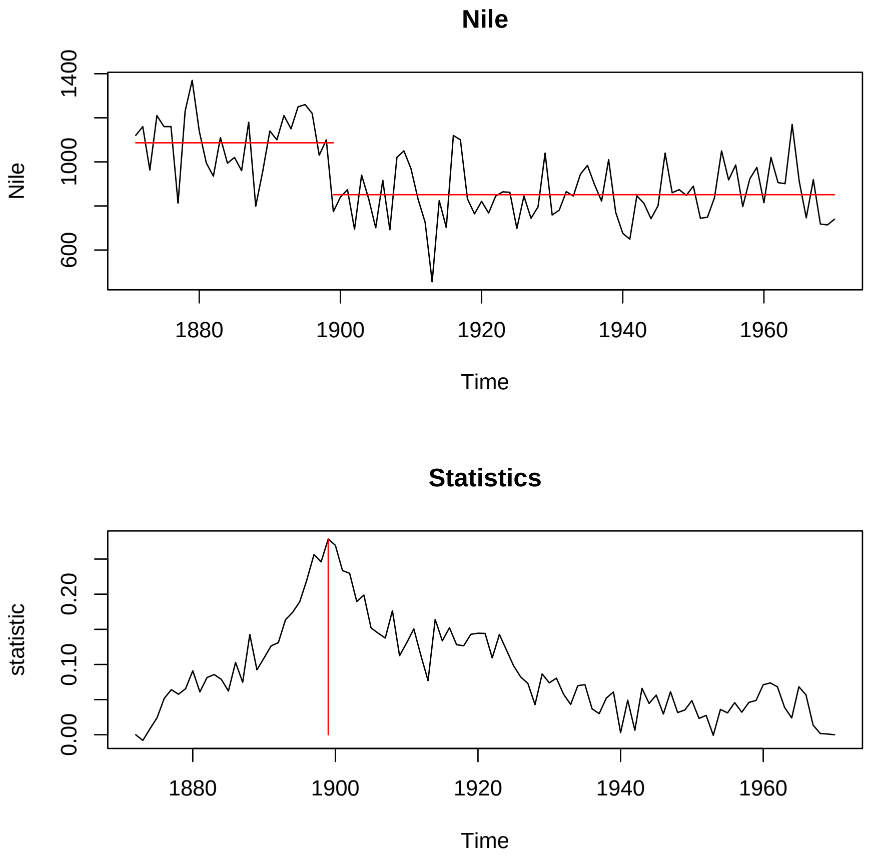

Figure 1: Experimental results on the Nile data.

7 Conclusions

We propose a nonparametric multivariate method for multiple change point detection with the CE-based two-sample test. The single change point detection is first proposed as a group of two-sample tests on every points of time series data and the change point is considered as with the maximum of the test statistics. The multiple change point detection is then proposed by combining the single change point detection method with binary segmentation strategy. We verified the effectiveness of our method and compared it with the other similar methods on the simulated univariate and multivariate data and the Nile data.

References

[1] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[2] Samaneh Aminikhanghahi and Diane J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, May 2017.

[3] Jaxk Reeves, Jien Chen, Xiaolan L. Wang, Robert Lund, and Qi Qi Lu. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900 – 915, 2007.

[4] Yue S. Niu, Ning Hao, and Heping Zhang. Multiple Change-Point Detection: A Selective Overview. *Statistical Science*, 31(4):611 – 623, 2016.

[5] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[6] Apratim Guha and Tom Chothia. A two sample test based on mutual information. *Calcutta Statistical Association Bulletin*, 66(1-2):39–54, 2014.

[7] Jian Ma and Zengqi Sun. Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1):51–54, 2011.

[8] Jian Ma. Two-sample test with copula entropy. *arXiv preprint arXiv:2307.07247*, 2023.

[9] Xin Xiong and Ivor Cribben. Beyond linear dynamic functional connectivity: A Vine copula change point model. *Journal of Computational and Graphical Statistics*, 32(3):853–872, 2023.

[10] Axel Bücher, Ivan Kojadinovic, Tom Rohmer, and Johan Segers. Detecting changes in cross-sectional dependence in multivariate time series. *Journal of Multivariate Analysis*, 132:111–128, 2014.

[11] Florian Stark and Sven Otto. Testing and dating structural changes in copula-based dependence measures. *Journal of Applied Statistics*, page 1–19, Nov 2020.

[12] Roger B Nelsen. *An introduction to copulas.* Springer Science & Business Media, 2007.

[13] Harry Joe. *Dependence modeling with copulas.* CRC press, 2014.

[14] Abe Sklar. Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231, 1959.

[15] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[16] A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974.

[17] Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20(162):1–56, 2019.

[18] Jian Ma. copent: Estimating copula entropy and transfer entropy in R. *arXiv preprint arXiv:2005.14025*, 2021.

[19] Rebecca Killick and Idris A. Eckley. changepoint: An R package for change-point analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.

[20] Nicholas A. James and David S. Matteson. ecp: An R package for non-parametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62(7):1–25, 2015.

[21] George W. Cobb. The problem of the Nile: Conditional solution to a changepoint problem. *Biometrika*, 65(2):243–251, 08 1978.