Change Point Detection with Copula Entropy based Two-Sample Test

Authors: Jian Ma, Jian Ma

Date: 2024-03-01T00:00:00+00:00

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.

Full Text

Preamble

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 aims to find changes in time series and can be tackled using a two-sample test. Copula Entropy is a mathematical concept for measuring statistical independence, and a two-sample test based on it has been introduced recently. In this paper, we propose a nonparametric multivariate method for multiple change point detection using the copula entropy-based two-sample test. Single change point detection is first proposed as a series of two-sample tests at every point of the time series data, with the

change point identified as the point with the maximum test statistic. Multiple change point detection is then proposed by combining the single change point detection method with a binary segmentation strategy. We verify the effectiveness of our method and compare it with other similar methods on simulated univariate and multivariate data, as well as the Nile dataset.

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

Introduction

Change point detection is a typical task that aims to find single or multiple 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 this task; see [1, 2, 3, 4] for reviews on this topic. Change point detection can be widely applied to natural, social, or industrial systems where abrupt changes occur.

The two-sample test is a common problem of hypothesis testing in statistics. It tests the hypothesis of whether two samples are from the same distribution. There are many two-sample tests based on different mathematical concepts. A typical way of defining a test statistic is based on measures of statistical independence between two samples, such as kernel-based measures [5] and 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 information in information theory. A nonparametric method for estimating it was also proposed in [7]. Recently, CE has been applied to the two-sample test [8], in which the test statistic is defined as the difference between CEs of two hypotheses. There are several works 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 testing structural changes in multivariate time series using copula-based dependence measures, such as Spearman' s and quantile dependencies.

In this paper, we propose using CE-based two-sample test for multiple change point detection. The idea is simple: first transform the change point detection problem into a series of CE-based two-sample tests at every point of the time series, then find the change point as that with the maximum test statistic. A multiple change point detection problem can be solved by combining the single change point detection method with a 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 verify the effectiveness of the proposed method and compare it with other similar methods on both simulated and real data.

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 for single and multiple change point detection, experiments with simulated and real data are presented in Sections 4 and 5 respectively, followed by discussion in Section 6, and finally we conclude the paper in Section 7.

2.1 Copula Entropy

Copula theory is a probabilistic theory for representing multivariate dependence [12, 13]. According to Sklar' s theorem [14], any multivariate density function can be represented as a product of its marginals and a copula density function (cdf) that represents the 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

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

A nonparametric estimator of CE was also proposed in [7], which consists of two simple steps: (1) estimating the 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 using rank statistics. With the estimated empirical copula density, the second step is essentially an entropy estimation problem, which can be tackled with the KSG estimation method [15]. In this way, a nonparametric 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 $\mathbf{X}_1 = \{X_{11}, \cdots, X_{1m}\} \sim P_1$ and $\mathbf{X}_2 = \{X_{21}, \cdots, X_{2n}\} \sim P_2$, the null hypothesis for the two-sample test is $H_0 : P_1 = P_2$, and the alternative is $H_1 : P_1 \neq P_2$, where $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^d$ and P_1, P_2 are the corresponding probability distribution functions.

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

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

$$T_{ce}(\mathbf{X}_1, \mathbf{X}_2) = H_c(\mathbf{X}, Y_0) - H_c(\mathbf{X}, Y_1).$$

It is easy to see that T_{ce} will be small if H_0 is true and large if H_1 is true. The test statistic in (5) can be easily estimated from the data by estimating the

two terms using the nonparametric estimator of CE. Since the CE estimator is nonparametric, the estimator of the test statistic can be applied to any case without assumptions. Another merit of this test statistic estimator is that it is hyperparameter-free.

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 divided by each point in the time series, and the point associated with the maximal test statistic among these tests is the change point.

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

$$\hat{i} = \arg \max_{i \in [1, n-1]} T_{ce}(\mathbf{X}_1, \mathbf{X}_2),$$

where $T_{ce}(\mathbf{X}_1, \mathbf{X}_2)$ is the statistic of the CE-based two-sample test on the samples $\mathbf{X}_1 = \{x_1, \dots, x_i\}$ and $\mathbf{X}_2 = \{x_{i+1}, \dots, x_n\}$.

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

3.2 Multiple change point detection

Multiple change point detection can be transformed into a series of the above single change point detection problems using a binary segmentation strategy. For a time series, we first detect a change point using the above single change point detection method, and if one is detected, the whole time series is separated into two segments before and after the detected change point. This detection process continues recursively on the resulting segments until no change points can be detected in any of the derived segments.

In the proposed method, a threshold on the test statistic is set to judge 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 using a threshold, our method can automatically estimate the number of change points. 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 case without assumptions.

4.1 Experiments

We conducted simulation experiments to test the proposed method. In each simulation, a univariate or multivariate time series with several change points was first generated, 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, each of length 50 points, which means there are three change points at positions 51, 101, and 151.

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

For multivariate time series, the sub-series were first generated from bivariate normal distributions with different means and covariances. 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 bivariate normal distributions and bivariate copula functions. The copula functions used were Frank copula ($\theta = 0.9$) and normal copula ($\rho = 0.3$), both with normal ($\mu = 0$ and $\delta = 2$) and exponential (rate $= 0.5$) marginals. The parameters of the simulations in these cases are listed in Table 2 .

We compared our method with traditional change point detection methods. In the three univariate cases, our method was compared with three methods for detecting changes in mean, changes in mean and variance, and changes in variance, respectively. The binary segmentation strategy [16] was adopted in these three comparison 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, we used the implementation of the CE-based two-sample test from the R package copent[18]. The threshold for the test statistics was 0.13 in all experiments, except for the multivariate case with different covariances, where the threshold was 0.05. The comparison methods in univariate cases were those implemented in the R package changepoint[19]. The kernel method implemented in the R package ecp[20] was used for comparison. The experimental code is available at https://github.com/majianthu/cpd.

4.2 Results

The simulation results on univariate and multivariate time series data are presented in Tables 3 and 4 , respectively. For the univariate data, our method detected all change points in the different means, different means and variances, and different variances cases, just as the comparison methods did.

For the multivariate data, both our method and the kernel method worked well in the different means and different means and variances cases. In the different variances case, our method detected two correct change points (48, 102) with additional false positives, while the kernel method detected only false positives even after hyperparameter tuning. In the copula function case, our method detected one change point (155), while the kernel method could not detect any

change point. There were two false positives: one in the different variances case of univariate data (9) and another in the different means case of multivariate data (18). However, the test statistics for these false positives were smaller than those for the true change points.

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

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

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

Case	Our method	Kernel method
mean-var	51, 101, 151, 18	51, 101, 151
copula	14, 48, 102, 162, 169	1, 51, 101, 151, 201
mean-var	1, 51, 101, 151, 201	1, 46, 59, 80, 157, 159, 201
copula	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 time series measurements of the annual flow of the Nile River at Aswan from 1871 to 1970, with an apparent decreasing change occurring around 1898. We applied the single change point detection method to the Nile data. The results are shown in Figure 1 [Figure 1: see original paper], from which we can see that our method successfully detected the correct point where the change in river flow occurred, and the test statistic reached its maximum there as well.

6 Discussion

We have proposed a method for multiple change point detection using a CE-based two-sample test. The effectiveness of the proposed method was tested on both 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 cases. In contrast, different comparison methods must be used for each case of the univariate data.

Our method has one hyperparameter: the threshold on the test statistic. However, in the experiments, only one value (0.13) was used for all cases, except for the different variances case in multivariate data. It worked so well that we

did not need to tune it extensively. In contrast, the kernel method required frequent tuning of its penalty parameter for each case to detect the correct change points. This advantage of our method arises because CE is rigorously defined and model-free, and hence the test statistic of the two-sample test based on it is comparable across all cases.

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

Figure 1 [Figure 1: see original paper]: Experimental results on the Nile data. (Top) Annual flow of the Nile River over time. (Bottom) Test statistic values over time.

7 Conclusions

We have proposed a nonparametric multivariate method for multiple change point detection using the CE-based two-sample test. Single change point detection is first proposed as a series of two-sample tests at every point of the time series data, with the change point identified as the point with the maximum test statistic. Multiple change point detection is then proposed by combining the single change point detection method with a binary segmentation strategy. We verified the effectiveness of our method and compared it with other similar methods on simulated univariate and multivariate data, as well as the Nile dataset.

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 répartition à n 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 nonparametric 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.

*Email: majian@hitachi.cn

Note: Figure translations are in progress. See original paper for figures.