Testing Symmetry with Copula Entropy based Two-Sample Test

Authors: Jian Ma, Jian Ma

Date: 2025-05-18T06:10:31+00:00

Abstract

We propose a distribution-free method for testing symmetry of distribution with copula entropy based two-sample test. The test statistic is defined as the statistic of two-sample test on symmetric transformation of distribution and its nonparametric estimation method is proposed. Three simulation experiments with beta distribution, asymmetric Laplace distribution, and bimodal normal distribution are conducted to verify the effectiveness of the proposed method and to compare it with the existing methods in the field. Experimental results show that the proposed method can measure the symmetry of the simulated distributions effectively and performs best among others.

Full Text

Preamble

Testing Symmetry with Copula Entropy based Two-Sample Test
Jian Ma∗
Hitachi China Research Laboratory

Abstract

We propose a distribution-free method for testing symmetry of distribution with a copula entropy based two-sample test. The test statistic is defined as the statistic of a two-sample test on a symmetric transformation of the distribution, and its nonparametric estimation method is proposed. Three simulation experiments with beta distribution, asymmetric Laplace distribution, and bimodal normal distribution are conducted to verify the effectiveness of the proposed method and to compare it with existing methods in the field. Experimental results show that the proposed method can measure the symmetry of the simulated distributions effectively and performs best among others.

Keywords: Symmetry Test; Copula Entropy; Two-Sample Test; Distribution Free; Non-Parametric Method

1 Introduction

Symmetry is an important property of probability distributions, and symmetry of distribution is an assumption of many statistical models and methods. Tests for symmetry are therefore of great importance in statistics and its various applications. The first such test dates back to the 18th century by Arbuthnot [1]. There are many existing methods for testing symmetry [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14].

Two-sample test is a common problem of hypothesis testing in statistics. It is used to test 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 test statistics for testing is based on measures of statistical independence in two samples, such as kernel-based measures [15] and mutual information [16].

Copula Entropy (CE) is a recently defined mathematical concept for measuring statistical independence [17]. It is proved to be equivalent to mutual information in information theory. A nonparametric method for estimating CE was also proposed [17]. Recently, CE has been applied to two-sample test [18], in which the test statistic is defined as the difference between CEs of two hypotheses. A method for change point detection based on this two-sample test was also proposed recently [19]. CE was also proposed to test multivariate normality in [20].

In this paper, we propose to test symmetry of distribution with a CE-based two-sample test. The idea is to first derive a distribution from the target distribution via symmetry transformation, and then test the equality of these two distributions with a CE-based two-sample test. Since the CE-based two-sample test is distribution-free and universally applicable, so is the symmetry test based on it. Another merit of the proposed method is that the test statistic can be estimated nonparametrically.

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 testing symmetry, simulation experiments with three typical asymmetrical distribution families are performed in Section 4, followed by discussion in Section 5, and finally we conclude the paper in Section 6.

2.1 Copula Entropy

Copula theory is a probabilistic theory on representation of multivariate dependence [21, 22]. According to Sklar's theorem [23], any multivariate density function can be represented as a product of its marginals and copula density

function (cdf), which represents the dependence structure among random variables.

With copula theory, Ma and Sun [17] 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 $H_c(x) = -\int c(u) \log c(u) du$.

A non-parametric estimator of CE was also proposed in [17], 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 with rank statistics. 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 [24]. In this way, a non-parametric method for estimating CE was proposed in [17].

2.2 Two-sample test with CE

CE has been applied to solve the two-sample test problem [18]. 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 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 Y_0, Y_1 be two labeling variables for the two hypotheses respectively, such 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 $H_c(\mathbf{X}; Y_i) = H_c(\mathbf{X}, Y_i) - H_c(\mathbf{X})$.

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

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

It is easy to see that T_{tst} 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 an estimator of the test statistic is that it is hyperparameter-free.

3 Method for testing symmetry

Given a sample $\mathbf{X} = \{X_i, i = 1, \cdots, N\}$ generated from probability density function $p(x) \in \mathbb{R}$. The problem is to test if p is symmetric based on X_i. The null hypothesis is $H_0 : p$ is symmetric, and the alternative is $H_1 : p$ is asymmetric.

We propose to test symmetry of distribution function p with a CE-based two-sample test. Let $\hat{\mathbf{X}}$ be the sample derived by subtracting the mean of $\mathbf{X}$ from $\mathbf{X}$. Then the problem is transformed into testing whether $\hat{\mathbf{X}}$ and $-\hat{\mathbf{X}}$ are from the same distribution. So the statistic for testing symmetry of p is defined as:

$$T_{sym}(\mathbf{X}) = T_{tst}(\hat{\mathbf{X}}, -\hat{\mathbf{X}}).$$

$T_{sym} = 0$ if p is symmetric and $T_{sym} > 0$ if p is asymmetric. The more asymmetric, the larger T_{sym}.

The proposed method consists of two simple steps: 1) deriving $\hat{\mathbf{X}}$ from $\mathbf{X}$; 2) calculating T_{sym} by performing a two-sample test on $\hat{\mathbf{X}}$ according to (8). Since T_{tst} can be estimated non-parametrically, so can T_{sym}.

4.1 Experiments

In this section we conduct simulation experiments to test the effectiveness of the proposed method and compare it with existing ones.

Three simulation experiments on different asymmetric distribution families are designed. The asymmetric distribution families used in the experiments are:

Beta distribution: Beta distribution is a family of continuous probability functions defined on the interval $[0, 1]$ or $(0, 1)$ with two parameters $a, b > 0$ which control the shape of functions, as follows: $f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1 - x)^{b-1}$.

Asymmetric Laplace distribution: The asymmetric Laplace distribution is a continuous probability function that generalises the Laplace distribution. It is defined as $\delta(k + k^{-1})$ where $s = \text{sgn}(x - \mu)$ and μ, δ are location and scale parameters and k is an asymmetry parameter. $e^{-(x-\mu)ks/\delta}, f(x; \mu, \delta, k) =$

Bimodal normal distribution: Bimodal normal distribution is a continuous probability distribution derived by mixing two normal distributions $X_1 \sim N(\mu_1, \delta_1)$ and $X_2 \sim N(\mu_2, \delta_2)$ with a proportion parameter p.

In the simulation with Beta distribution, 9 samples are simulated by letting $a = 0.1, \cdots, 0.9$ and $b = 0.9, \cdots, 0.1$. In the simulation with asymmetric Laplace distribution, 9 samples are simulated with $\mu = 0$, $\delta = 1$ and $k = 0.1, \cdots, 0.9$. In the simulation with bimodal normal distribution, 9 samples are simulated with $\mu_1 = 0$, $\mu_2 = 5$, $\delta_1 = \delta_2 = 1$ and $p = 0.1, \cdots, 0.9$. The size of all samples is

300. All three experiments simulate distributions changing from asymmetry to symmetry then to asymmetry again.

In the simulations, the R packages `stats`, `ald`, and `FamilyRank` are used for simulating Beta distribution, asymmetric Laplace distribution, and bimodal normal distribution respectively.

In the experiments, we compare the proposed method with existing ones in the field. The R package `copent` [25] is used as the implementation of the CE-based two-sample test. The methods implemented in the R package `symmetry` are used for comparison, including: - **MI**: The Mira test statistic [9] - **CM**: The Cabilio–Masaro test statistic [10] - **MGG**: The Miao, Gel and Gastwirth test statistic [6] - **B1**: The b1 test statistic [8] - **KS**: The Kolmogorov–Smirnov test statistic [8] - **SGN**: The Sign test statistic [8] - **WCX**: The Wilcoxon test statistic [8] - **FM**: The characterization-based test statistic [11] - **RW**: The Rothman-Woodroofe test statistic [12] - **BHI**: The Litvinova test statistic [13] - **BHK**: The Baringhaus and Henze supremum-type test statistic [4] - **BH2**: The Baringhaus-Henze test statistic [4] - **MOI and MOK**: The Milošević and Obradović test statistics [7] - **NAI and NAK**: The Nikitin and Ahsanullah test statistics [14] - **K2 and K2U**: The Božin, Milošević, Nikitin and Obradović Kolmogorov type statistics based on V- and U-statistics respectively [5] - **NAC1, NAC2, BHC1 and BHC2**: The Allison and Pretorius test statistics

The code for the simulation experiments is available at https://github.com/majianthu/symmetry.

4.2 Results

Experimental results are shown in Figure 1 [Figure 1: see original paper], Figure 2 [Figure 2: see original paper], and Figure 3 [Figure 3: see original paper] respectively.

It can be learned from the results of the experiments with Beta distributions and asymmetric Laplace distributions that the statistics of the proposed method can reflect the change of distribution symmetry, while only 4 of the compared methods succeed in the first simulation and 13 of the compared methods succeed in the second simulation. For the simulation with bimodal normal distribution, the statistics of the proposed method present results with two peaks, which means the distributions are symmetric at first, middle, and last. The 'BHC1' and 'BHC2' methods present results with one peak which is close to the situation of the simulation.

5 Discussion

From the simulation results, one can learn that the proposed method presents more reasonable results than the compared methods. Especially in the simulation with bimodal normal distributions where the distribution is symmetric in the middle and close to symmetry at first and last as the proportion of two normal distributions changes by the parameter p, only the statistics of our method

measure the symmetry of the simulated distributions correctly.

In this paper, our method only deals with univariate distributions. However, it can be easily generalized to multivariate distributions by testing symmetry in high-dimensional spaces. It can also be generalized to test more complicated symmetry since the symmetry considered in this paper is the most basic one. Exchangeability, a special case of distribution symmetry, can also be tested with a CE-based two-sample test in a similar way.

6 Conclusions

We propose a method for testing symmetry of distribution with a CE-based two-sample test. The test statistic is defined as the statistic of a two-sample test on symmetric transformation of distributions, and its nonparametric estimation method is proposed. Three simulation experiments with beta distribution, asymmetric Laplace distribution, and bimodal normal distribution are conducted to verify the effectiveness of the proposed method and to compare it with existing methods in the field. Experimental results show that the proposed method can measure the symmetry of the simulated distributions effectively and performs best among others.

References

[1] John Arbuthnot. II. an argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. by dr. john arbuthnot, physitian in ordinary to her majesty, and fellow of the college of physitians and the royal society. *Philosophical Transactions of the Royal Society of London*, 27(328):186–190, 1710.

[2] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[3] James S Allison and Charl Pretorius. A monte carlo evaluation of the performance of two new tests for symmetry. *Computational Statistics*, 32(4):1323–1338, 2017.

[4] L Baringhaus and N Henze. A characterization of and new consistent tests for symmetry. *Communications in statistics-theory and methods*, 21(6):1555–1566, 1992.

[5] Vladimir Božin, Bojana Milošević, Ya Yu Nikitin, and Marko Obradović. New characterization-based symmetry tests. *Bulletin of the Malaysian Mathematical Sciences Society*, 43:297–320, 2020.

[6] Weiwen Miao, Yulia R Gel, and Joseph L Gastwirth. A new test of symmetry about an unknown median. In *Random walk, sequential analysis and related topics: A festschrift in honor of Yuan-Shih Chow*, pages 199–214. World Scientific, 2006.

[7] Bojana Milošević and Marko Obradović. Characterization based symmetry tests and their asymptotic efficiencies. *Statistics & Probability Letters*, 119:155–162, 2016.

[8] Bojana Milošević and Marko Obradović. Comparison of efficiencies of some symmetry tests around an unknown centre. *Statistics*, 53(1):43–57, 2019.

[9] Antonietta Mira. Distribution-free test for symmetry based on bonferroni's measure. *Journal of Applied Statistics*, 26(8):959–972, 1999.

[10] Paul Cabilio and Joe Masaro. A simple test of symmetry about an unknown median. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 349–361, 1996.

[11] Andrey Feuerverger and Roman A Mureika. The empirical characteristic function and its applications. *The annals of Statistics*, pages 88–97, 1977.

[12] Daniel Gaigall. Rothman–Woodroofe symmetry test statistic revisited. *Computational Statistics & Data Analysis*, 142:106837, 2020.

[13] VV Litvinova. New nonparametric test for symmetry and its asymptotic efficiency. *Vestnik St. Petersburg University Mathematics*, 34(4):12–14, 2001.

[14] Ya Yu Nikitin and Mohammad Ahsanullah. New U-empirical tests of symmetry based on extremal order statistics, and their efficiencies. In *Mathematical Statistics and Limit Theorems: Festschrift in Honour of Paul Deheuvels*, pages 231–248. Springer, 2015.

[15] 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.

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

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

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

[19] Jian Ma. Change point detection with copula entropy based two-sample test. arXiv preprint arXiv:2403.07892, 2024.

[20] Jian Ma. Multivariate normality test with copula entropy. arXiv preprint arXiv:2206.05956, 2022.

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

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

[23] 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.

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

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

Figure 1

Simulation results with Beta distributions.

Figure 2

Simulation results with asymmetric Laplace distributions.

Figure 3

Simulation results with bi-modal normal distributions.

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