Computational Statistics and Data Analysis 92 (2015) 97–114

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A test for equality of two distributions via jackknife empirical likelihood and characteristic functions

Zhi Liu [a], Xiaochao Xia [b], Wang Zhou [c,*]

[a] Department of Mathematics, University of Macau, Macau

[b] College of Mathematics and Statistics, Chongqing University, Chongqing, China

[c] Department of Statistics and Applied Probability, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 7 December 2014
Received in revised form 18 June 2015
Accepted 23 June 2015
Available online 30 June 2015

Keywords:
Jackknife empirical likelihood
Two-sample test
Equality of distributions
Characteristic function
Normal limit

ABSTRACT

The two-sample problem: testing the equality of two distributions is investigated. A jackknife empirical likelihood (JEL) test is proposed through incorporating characteristic functions, which reduces to a two-sample U-statistic. When the dimension of data is fixed, the nonparametric Wilks's theorem for the proposed JEL ratio statistics is established. When the dimension diverges with the sample size at a moderate rate, $p = o(n^{1/3})$, it is proved that under some mild conditions the normalized JEL ratio statistic has a standard normal limit. Moreover, when the dimension exceeds the sample size, $p > n$, an alternative version of JEL test is proposed. It is verified that under the null hypothesis this alternative version of JEL test has an asymptotical chi-squared distribution with two degrees of freedom. Some numerical results via simulation study and an analysis of a microarray dataset are presented and both confirm theoretical results empirically.

1. Introduction

In statistics, there is an enduring interest in knowing if two multivariate populations share the same distributions or certain distributional features, such as mean (vector) and co-variance (matrix). The questions of this kind have also been pursued and studied in many other fields. It frequently arises in the case-control study in biomedicine. For example, suppose that there are two groups of patients infected with a particular disease and the patients in one group are given to a new medicine and another is kept original treatment intact. Researchers are interested in exploring whether the new medicine has any better effect or not. The data for this treatment can be collected by trials, which usually measure the expression levels of genes of patients. So it enables researchers to utilize some appropriate statistical approaches to identifying the latent causes.

In practice, it is not trivial to handle this two-sample problem. The main concern is two-fold. On one hand, it is difficult to directly generalize the classical tests such as the nonparametric Kolmogorov–Smirnov test, the Wald–Wolfowitz runs test and Wilcoxon rank test, which are commonly used in the univariate population, to deal with multivariate problems. Even if we apply these univariate tests to marginal models, the global significance based on all individual tests will seriously cause the false discovery rates (Benjamini and Hochberg, 1995) especially in the case of large number of observed variables. On the other hand, it is known that the statistics like Hotelling's T^2 statistics are commonly used to make inference on the multivariate means of populations. However, when the dimension is high, its performance such as empirical power behaves

* Corresponding author.
E-mail addresses: liuzhi@umac.mo (Z. Liu), xia_xiao_chao@126.com (X. Xia), stazw@nus.edu.sg (W. Zhou).

http://dx.doi.org/10.1016/j.csda.2015.06.004
0167-9473/

poorly due to the possible inconsistency of sample covariance matrix (Bai and Saranadasa, 1996). This stimulates statisticians to work out new procedures for testing certain distributional characteristics. Typically for testing the high dimensional mean vector and covariance matrix, there have been advances, see Bai and Saranadasa (1996), Chen and Qin (2010), Li and Chen (2012), Wang et al. (2013) and references therein.

However, the foregoing mentioned literatures more or less rely on the parametric forms of distribution functions. Recently, there is some progress on testing the equality of two distributions. Biswas and Ghosh (2014) proposed a nonparametric two-sample test based on the inter-point distances. They showed that the test is applicable to high dimensional datasets. Also, focusing on the inter-point distance, Liu and Modarres (2011) constructed a triangle test for testing the equality of two continuous distributions.

In this paper, instead of considering the distributions directly, we study the characteristic functions (cf.s) for such a testing problem. We apply the empirical likelihood method to test whether two cf.s are equal or not. Over the past two decades, the empirical likelihood method since originally proposed by Owen (1988) has been demonstrated to be a very powerful nonparametric way in the inference on parameters of interest. It has the merits such as imposing few assumptions on the error distribution in the regression model, producing narrower confidence intervals for unknown parameters and avoiding the variance estimation via self-normalizing. For more references one can refer to Owen (2001). Our primary purpose is to incorporate the jackknife empirical likelihood technique (Jing et al., 2009) to study the problem of equality of distributions of two (univariate, multivariate and high dimensional) populations. Differing from the aforementioned existing methods, the approach proposed in this paper is focusing on the usage of the cf.s, which is a vital portray for random variables since the distribution function of a random variable can be uniquely determined by its cf. and vice versa. To the best of our knowledge, this idea has not been well studied in the literature since the work of Chan et al. (2009) who investigated the empirical likelihood method based on cf. to estimate the parameters of Lévy distribution in financial modeling. The testing method proposed in Section 2 can be applied to deal with the moderate large dimensional problem in the sense that p^3/n goes to zero as the dimension p and the sample size n tend to infinity, which is empirically validated by simulation studies. Moreover, we also consider the case where the dimension is larger than sample size, and we propose an alternative simple testing procedure, of which the JEL statistic has a chi-squared limiting distribution with two degrees of freedom, which is free of dimension p as long as n goes to infinity.

The rest of paper is organized as follows. In Section 2, our proposed methodology including three aspects: univariate, multivariate and high dimensional cases is stated and main theoretical results including nonparametric Wilks's theorems are established. Simulation studies and an analysis for a microarray data are presented in Section 3. A concluding remark is given in Section 4. All proofs are relegated to Appendix.

2. Methodology

2.1. Univariate case

Suppose that X and Y are two populations having distributions F and G, respectively. An important issue in hypothesis testing is to consider the following problem:

$$\mathrm{H}_0 : F = G \quad \text{versus} \quad \mathrm{H}_1 : F \neq G. \tag{1}$$

In practice, we can only collect two groups of sample observations $\{X_i, 1 \leq i \leq n_1\}$ and $\{Y_j, 1 \leq j \leq n_2\}$ from X and Y, respectively, where n_1 and n_2 are the corresponding sample sizes.

As stated before, the distribution of X is mutually determined by its cf. So testing (1) is equivalent to testing whether $\phi(t) = \psi(t)$ for all $t \neq 0$, where $\phi(t)$ and $\psi(t)$ are the cf.s of X and Y, respectively. Hence, the problem (1) is reformulated by testing the hypothesis

$$\mathrm{H}_0 : \phi(t) = \psi(t) \quad \text{for all } t \neq 0 \quad \text{versus} \quad \mathrm{H}_1 : \phi(t) \neq \psi(t) \quad \text{for some } t \neq 0. \tag{2}$$

By the definition of cf., we have $E(e^{itX} - e^{itY}) = 0$ for all $t \neq 0$ under the null hypothesis, which is equivalent to testing whether $E(\sin(tX) - \sin(tY)) = 0$ and $E(\cos(tX) - \cos(tY)) = 0$ hold simultaneously for all $t \neq 0$. If we define

$$U = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i, Y_j), \tag{3}$$

where $h(X, Y) = (h^{(1)}(X, Y), h^{(2)}(X, Y))^T$ with $h^{(1)}(X, Y) = \sin(tX) - \sin(tY)$ and $h^{(2)}(X, Y) = \cos(tX) - \cos(tY)$, then we have

$$\theta := EU = 0.$$

Clearly U is a Wilcoxon–Mann–Whitney statistic which is a two-sample U-statistic (Lee, 1990). Here it is difficult to implement Owen's empirical likelihood directly due to constrained nonlinear optimization, since the maximization problem usually reduces to solving a number (dependent on n_1 and n_2) of simultaneous equations.

Now, we readily formulate a test by applying the jackknife empirical likelihood approach which is proposed in Jing et al. (2009) to handle U-statistics. This method reduces the nonlinear optimization to a linear problem. To be more specific, we