Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 137 (2017) 390–410
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original articles

Fast tests for the two-sample problem based on the empirical characteristic function

M.D. Jiménez-Gamero[a], M.V. Alba-Fernández[b,*], P. Jodrá[c], I. Barranco-Chamorro[a]

[a] *Dpto. Estadística e Investigación Operativa, Universidad de Sevilla, 41012 Sevilla, Spain*
[b] *Dpto. Estadística e Investigación Operativa, Universidad de Jaén, 23071 Jaén, Spain*
[c] *Dpto. Métodos Estadísticos, Universidad de Zaragoza, 50018 Zaragoza, Spain*

Received 29 October 2015; received in revised form 29 April 2016; accepted 16 September 2016
Available online 29 September 2016

Abstract

A class of tests for the two-sample problem whose test statistic is an L_2 norm of the difference of the empirical characteristic functions of the samples is considered. The null distribution can be estimated by means of bootstrap or permutation procedures. Although very easy to implement, such procedures can become computationally expensive as the sample size or the dimension of the data increase. This paper proposes to approximate the null distribution through a weighted bootstrap. The method is studied both theoretically and numerically. It provides a consistent estimator of the null distribution. The asymptotic properties are similar to those of the bootstrap and permutation estimators but, from a computational point of view, the weighted bootstrap estimator is more efficient. The proposed approach is also applied to the two-sample location problem and to the k-sample problem.

Keywords: Characteristic function; Two-sample problem; Weighted bootstrap; Consistency

1. Introduction

Let X and Y be two random vectors taking values in $\mathbb{R}^d$, for some fixed $d \in \mathbb{N}$, with cumulative distribution functions (CDF) F_X and F_Y, respectively. This paper considers the problem of testing for the equality of both distributions. The null hypothesis is stated as

$$H_0 : F_X(x) = F_Y(x), \quad \forall x \in \mathbb{R}^d \quad \Longleftrightarrow \quad C_X(t) = C_Y(t), \quad \forall t \in \mathbb{R}^d,$$

where C_X and C_Y are the characteristic functions (CF) of X and Y, respectively.

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples from X and Y, with sizes n and m, respectively. The problem of testing whether two samples come from the same population has generated a considerable amount of research in Statistics. Many different approaches have been proposed to deal with this problem (see, for example, the references in Meintanis [16] and Alba-Fernández et al. [2]). To the best of our knowledge, one of the

* Correspondence to: Dpto. de Estadística e I.O., Universidad de Jaén, Paraje Las Lagunillas s.n., 23071 Jaén, Spain.
E-mail address: mvalba@ujaen.es (M.V. Alba-Fernández).

http://dx.doi.org/10.1016/j.matcom.2016.09.007

more general approaches was studied in Meintanis [16] (see also Anderson et al. [3], Henze et al. [9], Alba-Fernández et al. [1,2]), based on the use of the following test function for testing H_0,

$$\Phi_{n,m} = \begin{cases} 1, & \text{if } D_{n,m} \geq d_{n,m,\alpha}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where

$$D_{n,m} = \int \left| C_{X,n}(t) - C_{Y,m}(t) \right|^2 dG(t),$$

G is a CDF defined on $\mathbb{R}^d$, $C_{X,n}(t)$ and $C_{Y,m}(t)$ denote the empirical characteristic functions (ECF) associated with the samples,

$$C_{X,n}(t) = \frac{1}{n} \sum_{j=1}^{n} \exp(\mathrm{i}t'X_j), \qquad C_{Y,m}(t) = \frac{1}{m} \sum_{l=1}^{m} \exp(\mathrm{i}t'Y_l),$$

the prime denotes transpose and $d_{n,m,\alpha}$ is the $1-\alpha$ percentile of the null distribution of $D_{n,m}$. Here by "general approach" we mean that it can be applied to random vectors of arbitrary nature (continuous, discrete or mixed) and that the validity of the properties of $\Phi_{n,m}$ has been stated under the assumption

$$\frac{n}{n+m} \longrightarrow \tau \in (0,1), \tag{2}$$

as $n, m \to \infty$, which clearly is minimal.

To decide when to reject the null hypothesis H_0, the critical point $d_{n,m,\alpha}$ must be calculated or, equivalently, the p-value corresponding to the observed value of the test statistic. To this end, we need to know the null distribution of $D_{n,m}$. In general, this task is quite difficult, so, in most cases, one has to approximate it. A usual way to do this is by considering its limiting null distribution. Since the test statistic $D_{n,m}$ is not asymptotically distribution free, Meintanis [16] proposed to approximate its null distribution by means of permutation and bootstrap procedures.

Although very easy to implement, such procedures can become very computationally expensive as the sample size or the dimension of the data increase. Because of this reason, the aim of this paper is to investigate other ways of approximating the null distribution of $D_{n,m}$ that maintain the good properties of the above cited procedures (easy to implement, yield consistent null distribution estimators) and are more efficient from a computational point of view. Specifically, this paper studies the weighted bootstrap (WB) in the sense of Burke [4]. This method has been previously suggested in Kojadinovic and Yan [14] and Ghoudi and Rémillard [8], to approximate the null distribution of goodness-of-fit tests based on the empirical CDF, in Jiménez-Gamero and Kim [12], to approximate the null distribution of goodness-of-fit tests based on the ECF, and in Quessy and Éthier [17], for the two-sample problem for dependent data, among others. In view of the good properties of the WB in these and other papers, it is also expected to work well for approximating the null distribution of the test statistics considered in this paper. The purpose of the current study is to investigate, both theoretically and empirically, the use of the WB for approximating the null distribution of the test statistic $D_{n,m}$.

The paper is organized as follows. Section 2 shows that the WB provides a consistent approximation to the null distribution of the test statistic $D_{n,m}$. Section 3 studies the WB approximation for the null distribution of statistics designed for testing hypotheses that are related to H_0 and that are based on comparing ECFs. More precisely, the two-sample location problem and the k-sample problem are considered. Some practical matters related to the calculation of the bootstrap, permutation and WB approximations are dealt with in Section 4. The finite sample performance of the WB approximation, including comparison to the bootstrap and permutation approximations, is numerically studied by means of some simulation experiments, which are presented in Section 5. The main findings of this paper are briefly summarized in Section 6. For the sake of clarity, all proofs are deferred to the last section.

Before ending this section, we introduce some notation: all vectors are column vectors; for any vector v, v_j denotes its jth coordinate; $1_n \in \mathbb{R}^n$ has all its components equal to 1; for any complex number $x = a + ib$, $|x| = \sqrt{a^2+b^2}$; for any complex function $f(t)$, $\mathrm{Re}f(t)$ and $\mathrm{Im}f(t)$ denote the real and the imaginary parts of f, respectively, that is to say, $f(t) = \mathrm{Re}f(t) + \mathrm{iIm}f(t)$; P_0, E_0 and Cov_0 denote probability, expectation and covariance, respectively, by assuming that the null hypothesis is true; P_*, E_* and Cov_* denote the conditional probability law, expectation