Available online at www.sciencedirect.com

ScienceDirect

Mathematics and Computers in Simulation 102 (2014) 24–38

MATHEMATICS AND COMPUTERS IN SIMULATION

www.elsevier.com/locate/matcom

Original article

Approximating a class of goodness-of-fit test statistics

M.V. Alba Fernández [a,*], M.D. Jiménez Gamero [b], S. Castillo Gutiérrez [a]

[a] *Dpto. Estadística e Investigación Operativa, Universidad de Jaén, 23071 Jaén, Spain*
[b] *Dpto. Estadística e Investigación Operativa, Universidad de Sevilla, 41012 Sevilla, Spain*

Received 27 October 2011; received in revised form 17 August 2012; accepted 4 April 2013
Available online 23 August 2013

Abstract

A class of goodness-of-fit tests is considered. The test statistic of each test in this class is an L_2-norm of the difference between the empirical characteristic function associated with a random sample and an estimator of the characteristic function of the population in the null hypothesis. Because it is not always possible to give an easily computable analytic expression of the test statistic, a numerical integration formula is given to approximate it. The approximation is built by considering a piecewise quadratic Taylor expansion. The null distribution of the resultant test statistic is consistently estimated by means of a bootstrap estimator. A simulation study is carried out to illustrate the accuracy of the numerical approximation, the goodness of the bootstrap estimator of the null distribution and the power of the test. Applications to real data sets are also provided.

Keywords: Empirical characteristic function; Goodness-of-fit; Taylor approximation; Bootstrap distribution estimator; Consistency

1. Introduction

Most statistical methods assume certain distributional hypothesis on the random mechanism generating the data. In common cases, the conclusions of the analysis are rather sensitive to these distributional assumptions. So a crucial aspect of any data analysis is to check if the data support such distributional assumptions, that is to say, testing goodness-of-fit (gof). Since the characteristic function (cf) characterizes a probability law, there is an increasing number of gof tests whose test statistic measures deviations between the empirical characteristic function (ecf) and the cf in the null hypothesis. Examples are the tests in Epps and Pulley [7] for testing goodness-of-fit to the univariate normal distribution, in Baringhaus and Henze [5] for testing goodness-of-fit to the multivariate normal distribution, in Gürtler and Henze [8] and Matsui and Takemura [10] for the Cauchy distribution, in Meintanis [12] for the logistic distribution, in Matsui and Takemura [11] for the stable symmetric distributions, among many others. All the cited tests belong to a class of tests, that has been studied in Jiménez-Gamero et al. [9]. Next we describe such class.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random d-dimensional vectors with common distribution function (df) F, for some fixed integer $d \geq 1$. To test the composite null hypothesis,

$$H_0: \text{ the law of } X_1 \quad \in \mathcal{F},$$

* Corresponding author at: Dpto. de Estadística e Investigación Operativa, Universidad de Jaén, Paraje Las Lagunillas s.n., 23071 Jaén, Spain. Tel.: +34 953212142; fax: +34 953212034.

E-mail address: mvalba@ujaen.es (M.V. Alba Fernández).

0378-4754/$36.00
http://dx.doi.org/10.1016/j.matcom.2013.04.025

where $\mathcal{F}$ is a parametric family, $\mathcal{F} = \{F(x;\theta),\ x \in \mathbb{R}^d,\ \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$, we consider the following test function

$$\Psi = \begin{cases} 1 & \text{if} \quad D_{n,w}(\hat{\theta}) \geq d_{n,w,\alpha}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $d_{n,w,\alpha}$ is the $1-\alpha$ percentile of the null distribution of the test statistic $D_{n,w}(\hat{\theta})$,

$$D_{n,w}(\hat{\theta}) = \int |c_n(t) - c_0(t;\hat{\theta})|^2 w(t)\, dt, \tag{2}$$

$c_n(t)$ is the ecf,

$$c_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(\mathrm{i}t'X_j),$$

$c_0(t;\theta) = R_0(t,\theta) + \mathrm{i}I_0(t,\theta)$ is the cf of $F(x;\theta)$, $\hat{\theta}$ is a consistent estimator of θ, $w(t)$ is a density function on $\mathbb{R}^d$, which may depend on θ or not, for any complex number z, $z = a + \mathrm{i}b$ with $\mathrm{i} = \sqrt{-1}$ and $a, b \in \mathbb{R}$, $|z|^2 = a^2 + b^2$, and, from now on, an unspecified integral denotes integration over the whole space $\mathbb{R}^d$. The presence of $w(t)$ in the expression of $D_{n,w}(\hat{\theta})$ renders the integral in (2) finite and it also gives a readily computable closed form to $D_{n,w}(\hat{\theta})$ for suitable choices of w. Specifically, (see Lemma 1 in [9])

$$D_{n,w}(\hat{\theta}) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} h(X_j, X_k;\hat{\theta}),$$

where

$$\begin{aligned} h(x, y;\theta) &= u(x-y) - u_0(x;\theta) - u_0(y;\theta) + u_{00}(\theta), \\ u_0(x;\theta) &= \int u(x-y)\, dF(y;\theta), \quad u_{00}(\theta) = \int u(x-y)\, dF(x;\theta)\, dF(y;\theta), \end{aligned} \tag{3}$$

and $u(t)$ is the real part of the cf of a random vector with density function w, that is, $u(t) = \int cos(x't)w(x)\,dx$. Two problems arise with the test function Ψ defined in (1). The first one is the calculation of the critical point $d_{n,w,\alpha}$, because the exact null distribution of the test statistic $D_{n,w}(\hat{\theta})$ is unknown. A classical way to overcome this problem is by approximating the null distribution of the test statistic by means of its asymptotic null distribution. When H_0 is true, $D_{n,w}(\hat{\theta})$ converges in law to a linear combination of independent χ^2_1 variates, where the weights in this linear combination depend in a very complicated way on the unknown true value of the parameter θ, and thus they are unknown. Therefore, the asymptotic null distribution of $D_{n,w}(\hat{\theta})$ does not provide a useful approximation. Fortunately, the critical point $d_{n,w,\alpha}$ can be consistently approximated by a parametric bootstrap.

The second and perhaps the only serious problem with the test (1) is the calculation of the test statistic $D_{n,w}(\hat{\theta})$ itself, because for some parametric families the quantities u_0 and u_{00} in (3) may not have a closed or simple expression. The aim of this paper is to try to solve this drawback by using a numerical integration formula to approximate $D_{n,w}(\hat{\theta})$. To apply it we will assume that the integral in the expression of $D_{n,w}(\hat{\theta})$ is not over the whole space $\mathbb{R}^d$, but over a hypercube $I = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d]$, $a_i, b_i \in \mathbb{R}$, $a_i < b_i$, $1 \leq i \leq d$.

Note that this assumption is not a very severe restriction since we can always do a change of variable in such a way that the new variables range in a hypercube. Even if we do not do the change of variable, the assumption is still not very restrictive because as $|c_n(t) - c_0(t;\hat{\theta})|^2 \leq 4$ and $w(t)$ is a density function on $\mathbb{R}^d$, for every $\varepsilon > 0$ there exist a rectangle $I = I(\varepsilon)$ such that $\int_I w(t)\,dt \geq 1 - \varepsilon/4$ and thus $|\int |c_n(t) - c_0(t;\hat{\theta})|^2 w(t)\,dt - \int_I |c_n(t) - c_0(t;\hat{\theta})|^2 w(t)\,dt| \leq \varepsilon$.

The asymptotic null distribution of the resultant test statistic is again a linear combination of independent χ^2_1 variates, with the weights depending on the true value of θ, and hence the asymptotic null distribution does not yield a useful approximation to the null distribution of the test statistic. It will be shown that the bootstrap gives a consistent estimator of the null distribution.

The paper is organized as follows. In Section 2 a numerical integration formula to approximate $D_{n,w}(\hat{\theta})$ is given. The asymptotic null distribution of the resultant test statistic is derived in Section 3. A consistent bootstrap distribution estimator of the null distribution of the test statistic is studied in Section 4. The consistency against fixed alternatives of