COMSTA: 6068 Model 3Gsc pp. 1–20 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Fast goodness-of-fit tests based on the characteristic function

Q1 M. Dolores Jiménez-Gamero [a], Hyoung-Moon Kim [b,*]

[a] Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Sevilla, Spain

[b] Department of Applied Statistics, Konkuk University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 17 July 2014
Received in revised form 18 December 2014
Accepted 25 March 2015
Available online xxxx

Keywords:
Characteristic function
Goodness-of-fit
Weighted bootstrap
Consistency
Skew-*t* distribution

ABSTRACT

A class of goodness-of-fit tests whose test statistic is an L_2 norm of the difference of the empirical characteristic function of the sample and a parametric estimate of the characteristic function in the null hypothesis, is considered. The null distribution is usually estimated through a parametric bootstrap. Although very easy to implement, the parametric bootstrap can become very computationally expensive as the sample size, the number of parameters or the dimension of the data increase. It is proposed 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. In the numerical examples carried out, the estimated type I errors are close to the nominal values. The asymptotic properties are similar to those of the parametric bootstrap but, from a computational point of view, it is more efficient.

1. Introduction

Since the characteristic function (cf) characterizes the distribution of a random variable and the empirical characteristic function (ecf) converges to the population cf, many goodness-of-fit (gof) tests are based on measuring discrepancies between the ecf and an estimator of the cf of the population in the null hypothesis. Specifically, let $X_1, X_2, \dots, X_n$ be independent and identically distributed (iid) random d-dimensional vectors, for some fixed integer $d \geq 1$. For testing the composite null hypothesis

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

where $\mathcal{F}$ is a parametric family,

$$\mathcal{F} = \{F(x;\theta),\ x \in \mathbb{R}^d,\ \theta \in \Theta\}, \quad \Theta \subseteq \mathbb{R}^p,$$

$F(x;\theta)$ stands for the cumulative distribution function (cdf) and θ is assumed to be unknown, we consider the test $\Psi = \Psi(X_1, \dots, X_n)$,

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

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

$$D_{n,w}(\hat{\theta}_n) = n \int |c_n(t) - c(t;\hat{\theta}_n)|^2 w(t) dt. \qquad (2)$$

* Corresponding author.
E-mail addresses: dolores@us.es (M.D. Jiménez-Gamero), hmkim@konkuk.ac.kr (H.-M. Kim).

http://dx.doi.org/10.1016/j.csda.2015.03.015
0167-9473/

Please cite this article in press as: Jiménez-Gamero, M.D., Kim, H.-M., Fast goodness-of-fit tests based on the characteristic function. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.03.015

In (2), $c_n(t) = n^{-1}\sum_{j=1}^{n}\exp(it'X_j)$ is the ecf of the sample, $c(t;\theta)$ denotes the characteristic function of $F(x;\theta)$, $\hat{\theta}_n$ is a consistent estimator of θ and $w(t)$ is a finite measure (a density function often) on $\mathbb{R}^d$, which may depend on θ. Also, in what follows, an unspecified integral denotes integration over the whole space $\mathbb{R}^d$. Some general properties of the test Ψ have been studied in Jiménez-Gamero et al. (2009). Some special cases of (1) are the tests in Epps and Pulley (1983), Baringhaus and Henze (1988), Gürtler and Henze (2000), Meintanis (2004), Klar and Meintanis (2005), Epps (2005), Matsui and Takemura (2005), Matsui and Takemura (2008), Fragiadakis and Meintanis (2011), among others.

In spite of the good statistical properties of the test Ψ (it is consistent against fixed alternatives and able to detect local alternatives converging to the null at the rate $n^{-1/2}$), from a practical point of view, it possesses certain computational difficulties. A main problem is the calculation of the critical point $d_{n,w,\alpha}$, because the exact null distribution of the tests statistic $D_{n,w}(\hat{\theta}_n)$ is unknown. Since in most cases the asymptotic null distribution does not provide a useful approximation, it is usually consistently approximated by a parametric bootstrap (PB).

Although very easy to implement, the PB can become very computationally expensive as the sample size, the number of parameters or the dimension of the data increase. This problem is not specific to the test Ψ, the same problem arises when one instead consider a test based on comparing the empirical cdf and a parametric estimator of the cdf under the null hypothesis. To overcome this difficulty for gof tests based on the empirical cdf, Kojadinovic and Yan (2012b) have proposed to approximate the null distribution of the test statistics by a computationally more efficient estimator, obtained by using a weighted bootstrap (WB), in the sense of Burke (2000). In view of the good properties of the WB in Kojadinovic and Yan (2012b), it is also expected to work well for approximating the null distribution of the test statistics considered in this paper. Because of this reason, the purpose of this paper is to investigate, both theoretically and empirically, the use of a WB for approximating the null distribution of the test statistic (2).

Since

$$D_{n,w}(\hat{\theta}_n) = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n} h(X_j, X_k; \hat{\theta}_n), \tag{3}$$

where

$$\begin{aligned} &h(x,y;\theta) = u(x-y) - u_0(x;\theta) - u_0(y;\theta) + u_{00}(\theta),\\ &u_0(x;\theta) = \int u(x-y)dF(y;\theta), \qquad u_{00}(\theta) = \int u(x-y)dF(x;\theta)dF(y;\theta), \end{aligned} \tag{4}$$

and $u(t)$ is the real part of the cf of a random vector with density function w, that is, $u(t) = \int\cos(x't)w(x)d(x)$, the test statistic $\frac{1}{n}D_{n,w}(\hat{\theta}_n)$ resembles a degree-2 V-statistic. In the statistical literature there are several papers dealing with the consistency of the WB distribution estimator of U-statistics and V-statistics. Let $X_1,\ldots,X_n$ be iid and let

$$V_n(h) = \frac{1}{n^2}\sum_{1\le j,k\le n} h(X_j, X_k)$$

be a degree-2 V-statistic. Assume that it is degenerate, that is, $E\{h(X_1,x)\} - E\{h(X_1,X_2)\} = 0$. Delhing and Mikosch (1994) (see also Hušková and Janssen, 1993) showed that if $\xi_1,\ldots,\xi_n$ are iid with $E(\xi_1) = 0$ and $var(\xi_1) = 1$, independent of $X_1,\ldots,X_n$, then the conditional distribution, given $X_1,\ldots,X_n$, of

$$\frac{1}{n}\sum_{1\le j,k\le n} h(X_j, X_k)\xi_j\xi_k$$

consistently estimates that of $nV_n(h)$. In the light of this result, one may be tempted to estimate the null distribution of $D_{n,w}(\hat{\theta}_n)$ by means of the conditional distribution, given $X_1,\ldots,X_n$, of

$$W = \frac{1}{n}\sum_{1\le j,k\le n} h(X_j, X_k; \hat{\theta}_n)\xi_j\xi_k. \tag{5}$$

It will be shown that this is not the case, since the presence of $\hat{\theta}_n$ has an effect on the asymptotic null distribution of $D_{n,w}(\hat{\theta}_n)$ that it is not captured by the conditional distribution of W. We instead proceed as follows. In a first step we note that, when H_0 is true,

$$D_{n,w}(\hat{\theta}_n) = D_{1,n,w}(\theta) + o_P(1),$$

where $\frac{1}{n}D_{1,n,w}(\theta) = \frac{1}{n^2}\sum_{1\le j,k\le n} h^c(X_j, X_k;\theta)$ is a degenerate degree-2 V-statistic and θ is the limit of $\hat{\theta}_n$. Now we could imitate the procedure in Delhing and Mikosch (1994) to get a consistent distribution estimator of $D_{n,w}(\hat{\theta}_n)$ by considering the conditional distribution, given $X_1,\ldots,X_n$, of

$$\frac{1}{n}\sum_{1\le j,k\le n} h^c(X_j, X_k;\theta)\xi_j\xi_k.$$

A problem with this approach is that the kernel h^c depends on the unknown true value of θ. Moreover, it will be seen that even the expression of h^c can be hardly computed. As it will be seen later, h^c can be often approximated by an easily

Please cite this article in press as: Jiménez-Gamero, M.D., Kim, H.-M., Fast goodness-of-fit tests based on the characteristic function. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.03.015