Journal of Multivariate Analysis 140 (2015) 193–208

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On an independence test approach to the goodness-of-fit problem

Ludwig Baringhaus *, Daniel Gaigall

Institut für Mathematische Stochastik, Leibniz Universität Hannover, Postfach 60 09, D-30060 Hannover, Germany

A R T I C L E I N F O

Article history:
Received 10 October 2014
Available online 27 May 2015

AMS 2000 subject classifications:
62G10
62G09

Keywords:
Goodness-of-fit test
Independence test
Parametric bootstrap
Vapnik–Červonenkis class
U-process
Gamma distribution
Inverse Gaussian distribution

A B S T R A C T

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with distribution $\mathbb{F}$. Assuming that there are measurable functions $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ characterizing a family $\mathcal{F}$ of distributions on the Borel sets of $\mathbb{R}$ in the way that the random variables $f(X_1, X_2), g(X_1, X_2)$ are independent, if and only if $\mathbb{F} \in \mathcal{F}$, we propose to treat the testing problem $H : \mathbb{F} \in \mathcal{F},\ K : \mathbb{F} \notin \mathcal{F}$ by applying a consistent nonparametric independence test to the bivariate sample variables $(f(X_i, X_j), g(X_i, X_j))$, $1 \leqslant i, j \leqslant n, i \neq j$. A parametric bootstrap procedure needed to get critical values is shown to work. The consistency of the test is discussed. The power performance of the procedure is compared with that of the classical tests of Kolmogorov–Smirnov and Cramér–von Mises in the special cases where $\mathcal{F}$ is the family of gamma distributions or the family of inverse Gaussian distributions.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Let $\mathcal{G}$ be some general nonparametric class of non-degenerate distributions on the Borel sets of $\mathbb{R}$, and let $\emptyset \neq \mathcal{F} = \{\mathbb{F}(\cdot; \vartheta); \vartheta \in \Theta\} \subsetneqq \mathcal{G}$ be a parametric subfamily indexed by some parameter $\vartheta \in \Theta$, where $\Theta \neq \emptyset$ is a subset of $\mathbb{R}^d$, say. Let $X_1, \ldots, X_n, \ldots$ be real valued, independent and identically distributed random variables with unknown distribution $\mathbb{F} \in \mathcal{G}$. Let us express the fact that X_1 has distribution (function) $\mathbb{F}$ by $X_1 \sim \mathbb{F}$. On the basis of $X_1, \ldots, X_n$ we consider testing the hypothesis

$$H : \mathbb{F} \in \mathcal{F} \tag{1.1}$$

against the general alternative $K : \mathbb{F} \in \mathcal{G} \backslash \mathcal{F}$. Let us assume that there are measurable functions $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ characterizing the family $\mathcal{F}$ in the way that the random variables $f(X_1, X_2), g(X_1, X_2)$ are independent, if and only if $\mathbb{F} \in \mathcal{F}$.

Example 1.1. (1) The independence of $X_1 - X_2$ and $X_1 + X_2$ characterizes the family of normal distributions [5].

(2) Let the X_i be non-negative. Then $\frac{X_1}{X_1+X_2}$ and $X_1 + X_2$ are independent, if and only if $\mathcal{F}$ is the family of Gamma distributions $\mathrm{G}(\alpha, \lambda)$ with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ [22].

(3) Let the X_i be positive with finite moments $E(X_1^2)$ and $E(X_1^{-1})$. Then $\overline{X} = (X_1 + X_2)/2$ and $V = \frac{1}{2}\left(\frac{1}{X_1} + \frac{1}{X_2}\right) - \frac{1}{\overline{X}}$ are independent if and only if $\mathcal{F}$ is the family of inverse Gaussian distributions $\mathrm{IG}(\mu, \lambda)$ with densities given by $\sqrt{\frac{\lambda}{2\pi}} x^{-3/2} \exp\left(-\frac{\lambda(x-\mu)}{2\mu^2 x}\right)$, $x > 0$, for parameter values $\mu > 0$ and $\lambda > 0$ [19].

* Corresponding author.
E-mail addresses: lbaring@stochastik.uni-hannover.de (L. Baringhaus), gaigall@stochastik.uni-hannover.de (D. Gaigall).

http://dx.doi.org/10.1016/j.jmva.2015.05.013
0047-259X/© 2015 Elsevier Inc. All rights reserved.

(4) Let $0 < X_1 < 1$. Then the independence of $\frac{1-X_1}{1-X_1X_2}$ and $1 - X_1X_2$ characterizes the family of beta distributions [30].
(5) Let $X_1 > 0$ and let the distribution function of X_1 be strictly increasing. Then $\min(X_1, X_2)$ and $|X_1 - X_2|$ are independent if and only if X_1 has an exponential distribution [12, Theorem 3.3.1].

Given independent and identically distributed bivariate random vectors $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ with absolutely continuous distribution, the hypothesis of independence of Y_j and Z_j can simply and consistently be tested by using the Hoeffding–Blum–Kiefer–Rosenblatt independence criterion, that is by rejecting the hypothesis of independence for large values of

$$T_n = n \int (H_n(y, z) - F_n(y)G_n(z))^2 \, dH_n(y, z),$$

where

$$H_n(y, z) = \frac{1}{n} \sum_{j=1}^{n} I(Y_j \leqslant y, Z_j \leqslant z), \quad (y, z) \in \overline{\mathbb{R}}^2,$$

$$F_n(y) = H_n(y, \infty) = \frac{1}{n} \sum_{j=1}^{n} I(Y_j \leqslant y), \quad y \in \overline{\mathbb{R}},$$

$$G_n(z) = H_n(\infty, z) = \frac{1}{n} \sum_{j=1}^{n} I(Z_j \leqslant z), \quad z \in \overline{\mathbb{R}},$$

are the empirical distribution functions of the joint and the marginal sample variables. Defining for each $1 \leqslant j \leqslant n$

$$N_1(j) = \sum_{\nu=1}^{n} I(Y_\nu \leqslant Y_j, Z_\nu \leqslant Z_j), \qquad N_2(j) = \sum_{\nu=1}^{n} I(Y_\nu \leqslant Y_j, Z_\nu > Z_j),$$

$$N_3(j) = \sum_{\nu=1}^{n} I(Y_\nu > Y_j, Z_\nu \leqslant Z_j), \qquad N_4(j) = \sum_{\nu=1}^{n} I(Y_\nu > Y_j, Z_\nu > Z_j),$$

it turns out that

$$T_n = \frac{1}{n^4} \sum_{j=1}^{n} (N_1(j)N_4(j) - N_2(j)N_3(j))^2.$$

In what follows, let $n \geqslant 2$. We introduce the bivariate random vectors

$$(Y_{ij}, Z_{ij}) = \left(f(X_i, X_j), g(X_i, X_j)\right), \quad 1 \leqslant i, j \leqslant n, i \neq j, \tag{1.2}$$

and adopt the above approach replacing Y_j and Z_j by Y_{ij} and Z_{ij}, respectively. The resulting test statistic is

$$\mathcal{HBKR}_n = n \int \left(H_n(y, z) - F_n(y)G_n(z)\right)^2 dH_n(y, z),$$

where now

$$H_n(y, z) = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} I(Y_{ij} \leqslant y, Z_{ij} \leqslant z), \quad (y, z) \in \overline{\mathbb{R}}^2,$$

and

$$F_n(y) = H_n(y, \infty), \qquad G_n(z) = H_n(\infty, z), \quad y, z \in \overline{\mathbb{R}},$$

are empirical distribution functions of U-statistics structure. There is an alternative expression as before,

$$\mathcal{HBKR}_n = \frac{n}{(n(n-1))^5} \sum_{\substack{\mu,\nu=1 \\ \mu \neq \nu}}^{n} \left(N_1(\mu, \nu)N_4(\mu, \nu) - N_2(\mu, \nu)N_3(\mu, \nu)\right)^2$$

with obvious meaning of $N_i(\mu, \nu)$, $i = 1, 2, 3, 4$, i.e.,

$$N_1(\mu, \nu) = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} I(Y_{ij} \leqslant Y_{\mu\nu}, Z_{ij} \leqslant Z_{\mu\nu}),$$

$$N_2(\mu, \nu) = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} I(Y_{ij} \leqslant Y_{\mu\nu}, Z_{ij} > Z_{\mu\nu}),$$

Download English Version:

https://daneshyari.com/en/article/1145425

Download Persian Version:

https://daneshyari.com/article/1145425

Daneshyari.com