Journal of Multivariate Analysis 127 (2014) 126–146

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Independence tests for continuous random variables based on the longest increasing subsequence

Jesús E. García, V. A. González-López [*]

Department of Statistics, University of Campinas, Rua Sérgio Buarque de Holanda, 651, Campinas, São Paulo. CEP 13083-859, Brazil

ARTICLE INFO

Article history:
Received 13 March 2013
Available online 3 March 2014

AMS 2000 subject classifications:
62G10
62G30

Keywords:
Longest increasing subsequence
Test for independence
Copula

ABSTRACT

We propose a new class of nonparametric tests for the supposition of independence between two continuous random variables X and Y. Given a size n sample, let π be the permutation which maps the ranks of the X observations on the ranks of the Y observations. We identify the independence assumption of the null hypothesis with the uniform distribution on the permutation space. A test based on *the size of the longest increasing subsequence of* π (L_n) is defined. The exact distribution of L_n is computed from Schensted's theorem (Schensted, 1961). The asymptotic distribution of L_n was obtained by Baik et al. (1999). As the statistic L_n is discrete, there is a small set of possible significance levels. To solve this problem we define the JL_n statistic which is a jackknife version of L_n, as well as the corresponding hypothesis test. A third test is defined based on the JLM_n statistic which is a jackknife version of the longest monotonic subsequence of π. On a simulation study we apply our tests to diverse dependence situations with null or very small correlations where the independence hypothesis is difficult to reject. We show that L_n, JL_n and JLM_n tests have very good performance on that kind of situations. We illustrate the use of those tests on two real data examples with small sample size.

1. Introduction

Call Ω the space of the univariate, continuous cumulative distributions. Let (X, Y) be a random vector with unknown joint cumulative distribution H and univariate marginal distributions F and G respectively, $F \in \Omega$, $G \in \Omega$. Suppose that $(x_1, y_1), \dots, (x_n, y_n)$ is a paired sample of size n of (X, Y). Set

$$H_0 : X \text{ and } Y \text{ are independent.} \tag{1}$$

A test is constructed with no extra assumption (other than continuity) about the form of the marginal distributions (marginal free test). The procedure is based on the size of the longest increasing subsequence of the random permutation defined by the paired sample and denoted by L_n. Theorem 3.1 shows how to compute the exact distribution of L_n and it is a straightforward application of Schensted's theorem and Frame et al.'s theorem, see Schensted [12] and Frame et al. [6]. In addition, we proposed two test statistics denoted briefly by JL_n and JLM_n, respectively. JL_n is a Jackknife version of L_n while JLM_n is based on the size of the longest monotonic subsequence.

The power of these tests is compared with those of various existing tests by simulation. This new class of tests is rank-based, therefore, it will be compared with other rank-based procedures for testing independence as the nonparametric tests Kendall, Spearman and Hoeffding and the independence test from Genest et al. [7], denoted here by Genest's test.

[*] Corresponding author.
E-mail addresses: jg@ime.unicamp.br (J.E. García), veronica@ime.unicamp.br (V. A. González-López).

http://dx.doi.org/10.1016/j.jmva.2014.02.010

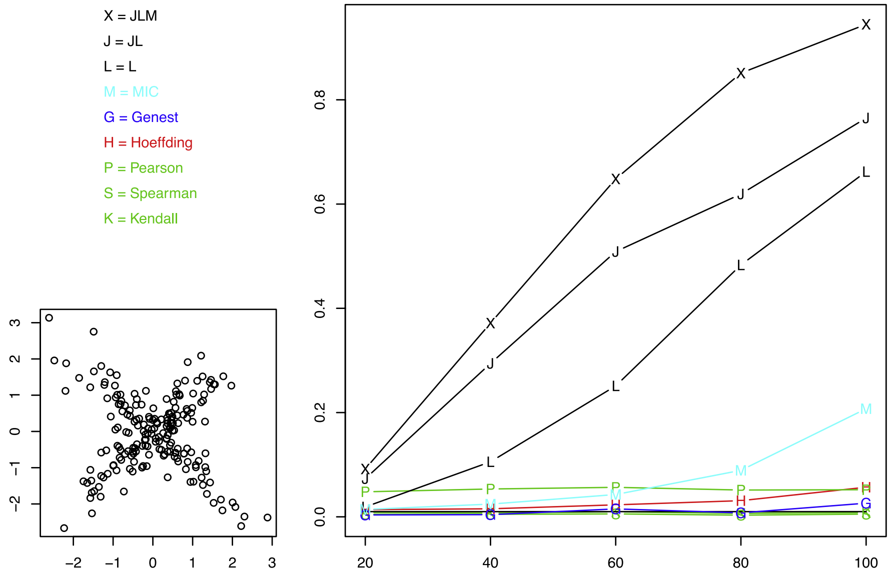

Fig. 1. The left figure is the scatter plot of a sample (size $= 200$) from a mixture 50–50 of two bivariate Normal distributions, with correlation 0.9 and -0.9 respectively (distribution D1 from Section 4). The right figure shows the plot of the sample size vs. the empirical power (level 0.01) for the same distribution.

We include also the MIC test, based on the maximal information coefficient, from Reshef et al. [11]. In addition we include Pearson's test for its well known performance in the normal case. In the case of Kendall's test, Spearman's test, Hoeffding's test and Pearson's test, each methodology estimates the association between X and Y and computes a test of the association being zero. They use different measures of association, all of them in the interval $[-1, 1]$ with 0 indicating no association/correlation. The asymptotic Genest's test consist on computing the approximate p-values of the test statistic with respect to the empirical distribution obtained by simulation. For the MIC test, the p-value of a given MIC score is computed by selecting a probability δ of false rejection, creating a set of $\frac{1}{\delta} - 1$ surrogate datasets, and comparing the MIC of the real data with the MIC scores of the surrogate datasets. To compute the p-values for Kendall, Spearman and Pearson methods, we use the "cor.test" function, available in the "stat" package from R-project. Details about each test may be found in Hollander et al. [9]. In the case of Hoeffding's test, to compute the p-values, we use the "hoeffd" function, available in the "Hmisc" package from R-project. For Genest's test we use the "indepTest" function, available in the "copula" package from R-project. For the MIC test was used the support program given in http://www.exploredata.net/.

We performed a simulation study with different conditions. For example, we use a mixture 50–50 of two bivariate Normal distributions, with correlation ρ and $-\rho$ respectively (zero expected correlation). In this case L_n, JL_n and JLM_n were competitive and markedly more powerful than the other six tests considered. Fig. 1, on the left, shows a scatter plot for a sample (size $= 200$) of this mixture when $\rho = 0.9$ and Fig. 1, on the right, shows the sample size versus the empirical power (level 0.01). The other tests do not detect the dependence for any sample size.

This situation illustrates the usefulness of our proposal, we will explore more situations like that, in Section 4.2.

We applied the tests based on the longest increasing subsequence to two real datasets, both with small sample sizes considering that for bigger sample sizes there exists very efficient procedures designed for asymptotic situations. The first dataset was provided by Professor Dalia Chakrabarty, researcher in the School of Physics and Astronomy, University of Nottingham. It consist on two measures, the projected radius and the radial velocity for 30 Globular Clusters around the galaxy NGC 3379 (see Chakrabarty [5]). The second dataset appears on "VGAM" (package from R-project), named "coalminers". The data is about coal-miners who are smokers without radiological pneumoconiosis, classified by age, breathlessness and wheeze.

We adapted and implemented (in C language) the algorithm provided by Zoghbi et al. [13]. We use that algorithm to compute the exact probability of L_n, in the case of $n \leq 100$. For $n > 100$ the asymptotic distribution of L_n, obtained by Baik et al. [3] can be used and we show how to use it in our test, in Section 3. Nevertheless, the exact probability could be calculated for $n > 100$ also. The probabilities for JL_n and JLM_n were estimated by simulation. The tests and simulations were implemented in the R-project environment (LIStest package).

Section 2 provides the main concepts and the definition of the test statistic. In Section 3 we calculate the distribution of the test statistic, proposed here. In Theorem 3.1 is shown the exact distribution of the test statistic under the independence assumption, by a direct application of results from Schensted [12] and Frame et al. [6]. Section 4 is devoted to show the capacity to detect dependence of each test statistic introduced here. Through simulations, we discuss each one of the test statistics, face to face with several dependence situations. We apply the test, to real datasets in Section 4.3. In the Appendix A we include the proof of Theorem 3.1.