

A non-parametric test of independence *

Jesús E. García

Verónica A. González-López †

September 30, 2011

Abstract

We propose a new class of nonparametric tests for the supposition of independence between two continuous random variables X and Y . Given a sample of (X, Y) , the tests are based on the size of the longest increasing subsequence (from now on L.I.S.) of the permutation which maps the ranks of the X observations to the ranks of the Y observations. We identify the independence assumption between the two continuous variables with the space of permutation equipped with the uniform distribution and we show the exact distribution of the L.I.S. We calculate the distribution for several sample sizes. Through a simulation study we estimate the power of our tests for diverse alternative hypothesis under the null hypothesis of independence. We show that our tests have a remarkable behavior when the sample has very low correlation and visible (or hidden) dependence. We illustrate the use of our test on two real data examples with small sample sizes.

Keywords: Test for independence; copula; natural sorting over permutation spaces; the longest increasing subsequence.

1 Introduction

Call Ω the space of the univariate, cumulative and continuous distributions.

Assumption 1 *Let (X, Y) be a random vector with unknown joint cumulative distribution H and univariate marginal distributions F and G , $F \in \Omega$, $G \in \Omega$.*

Suppose that $(x_1, y_1), \dots, (x_n, y_n)$ is a paired sample of size n of (X, Y) . We want to test the hypothesis

$$H_0 : X \text{ and } Y \text{ are independent.} \quad (1.1)$$

A test is constructed with no extra assumption (other than continuity) about the form of the marginal distributions. Let $rank(x_i)$ ($rank(y_i)$) be the position occupied by x_i (y_i) in the sample $\{x_j\}_{j=1}^n$ ($\{y_j\}_{j=1}^n$), the test statistic depends on the rank order of the observations. The procedure is based on the size of the longest increasing subsequence (L.I.S.) of the random permutation defined by the paired samples and we proposed two statistics, denoted briefly by L_n and JL_n . The second one is constructed by doing Jackknife in L_n . The power of our tests are compared with those of various existing tests by simulation. Five independence tests are selected for this comparative process, namely, Pearson, Kendall,

*This work is partially supported by CNPq Edital Universal (485999/2007-2), project: "Padrões rítmicos, domínios prosódicos e modelagem probabilística em corpora do português" and CNPq Edital Universal (476501/2009-1), project: "Sistemas estocásticos com interação de alcance variável".

†Departamento de Estatística - Instituto de Matemática Estatística e Computação Científica - Universidade Estadual de Campinas. Rua Sergio Buarque de Holanda, 651. Cidade Universitária - Barão Geraldo. CEP 13083-859, Campinas, SP, Brazil. emails: jg@ime.unicamp.br and veronica@ime.unicamp.br

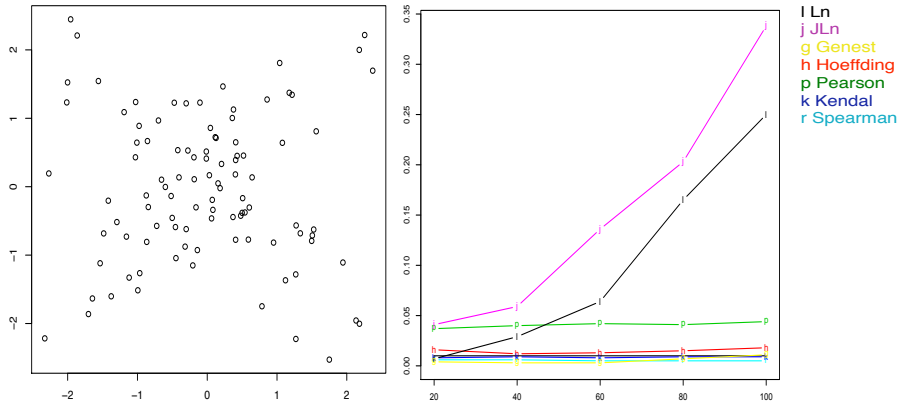


Figure 1: The picture on the left is the scatter plot of a sample size 100 of a mixture (50-50) of two bivariate Normal distributions, one with correlation 0.8 and the other -0.8 (distribution D1 in Section 3). The picture on the right is sample size vs. empirical power (level 0.01) for the same distribution.

Spearman, Hoeffding and the independence test from Genest et al. (2004), denoted here by Genest test. The Pearson test was selected by its well known performance in the normal case, the next other three are nonparametric. In the case of the first four, each methodology estimates the association between paired samples and computes a test of the value 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 test consist on computing the approximate p-values of the test statistic with respect to the empirical distribution obtained by simulation. To compute the p-values for each one of the four first 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. (1973). In the case of Hoeffding test, to compute the p-values, we use the “hoeffd” function, available in the “Hmisc” package from R-project. And for the Genest test we use the “indepTest” function, available in the “copula” package from R-project.

In our simulations, the Hoeffding test has a good power but at the expense of not controlling the significance level. Our L_n test underestimates the nominal level while our JL_n test control it by construction. Assuming Normal joint distribution and linear dependence between the Normal random variables, our tests have lower powers compared to the other tests which are designed for that case. For the case in which the joint distribution is non Normal, we performed a simulation study with different conditions. For example, we use a mixture (50-50) of two bivariate Normal distributions, one with correlation ρ and the other $-\rho$. In this case our procedures were competitive and more powerful than the other five tests considered. The left picture on figure 1 shows a scatter plot for a sample size 100 of this mixture when $\rho = 0.8$, the picture on the right shows the sample size vs the empirical power (level 0.01). We can see that for sample sizes going from 20 to 100 our tests are the only ones to show any increase in the empirical power, the Hoeffding test only detect the dependence for larger sample sizes. According to our simulation study our tests have a remarkable behavior in mixture cases, in which the samples are composed by two subsamples, one subsample is coming from a strongly correlated distribution and the other subsample is coming from a weakly correlated distribution. The mixture have a high proportion of the second subsample, resulting in a weakly correlated sample.

We applied our test to two real data sets. The first one appearing 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. Some of the quantities that we test are showed in figure 2(left). The picture (on the left) shows the plot between counts with breathlessness and wheeze (BW) and counts with no breathlessness but wheeze (nBW). Each point was took according to 9 age groups. The Spearman test gives a p-value equal to 0.2125, ignoring the dependence between the variables. Our L_n test gives a p-value equal to 0.099 and our JL_n test gives a p-value of 0.096. This result suggests

evidence against the hypothesis 1.1. The second data set was provided by Professor Dalia Chakrabarty researcher in

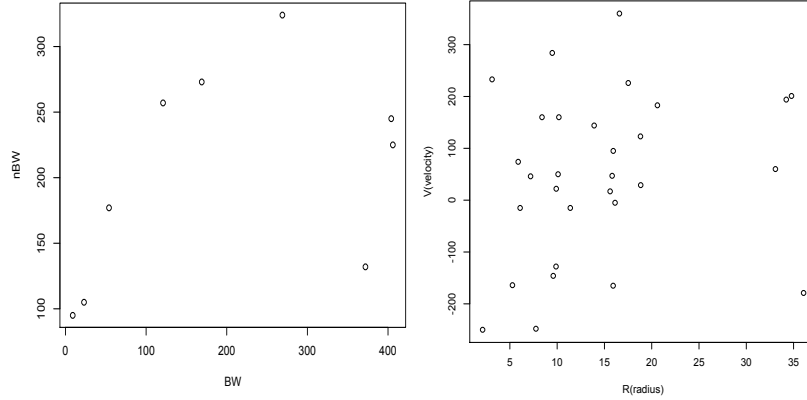


Figure 2: The left plot shows breathlessness and wheeze (BW) vs no breathlessness but wheeze (nBW). The right plot shows the radius of a galactic particle in the 2-D image (R) vs the 1-D velocity (V) of a galactic particle that sits at radius R .

the School of Physics and Astronomy, University of Nottingham. It consist on two measures (R and V) for 30 Globular Clusters around the galaxy NGC 3379, showed in figure 2 (on the righth). R is the radius of a galactic particle and V is the velocity of a galactic particle that sits at radius R . The Spearman correlation between R and V is around 0.27 and the p-value of Spearman test is equal to 0.1390. Hence the dependence between R and V is not detected by the Spearman test and detected by our L_n test which rejects the hypothesis 1.1 at level 0.05, with a p-value equal to 0.0203 while our JL_n test rejects it with a p-value equal to 0.018.

In order to explore the behavior of the L_n statistic, we compute in R-project the exact probability values of L_n for sample size $n < 100$. For $n \geq 100$ we use the asymptotic distribution of L_n , from a result by Baik et al. (1999), and we show how to use it in our test in Section 2. In Section 2 we calculate the distribution of our test statistic. In theorem 2.1 we show the exact distribution of the test statistic under the independence assumption. In Subsection 2.5 we apply our test and in Section 3 we show the effectivity of our proposal, using simulations and we discuss the results.

2 The longest increasing subsequence (L.I.S.)

We identify a sample of size n with a specific permutation of n points π , as shown on the following example.

Example 2.1 Consider the size 4 sample

$$\{(4.1, 3.2), (1.1, 3.5), (2.51, 4.17), (3.61, 3.18), (1.8, 2.86)\}.$$

First, sort the sample $\{(x_i, y_i)\}_{i=1}^n$ in increasing order in relation to the sample $\{x_i\}_{i=1}^n$ and replace the x_i value with its rank in the sequence, this produces $\{(1, 3.5), (2, 2.86), (3, 4.17), (4, 3.18), (5, 3.2)\}$. Next, replace each y_i with its rank in the $\{y_i\}_{i=1}^n$ sequence, this produces $\{(1, 4), (2, 1), (3, 5), (4, 2), (5, 3)\}$. The permutation π related to this sample is defined by $\pi(1) = 4, \pi(2) = 1, \pi(3) = 5, \pi(4) = 2, \pi(5) = 3$. The longest increasing subsequence is $\{1, 2, 3\}$ see figure 3 (b).

Our test is based on the distribution of the size of the longest increasing subsequence (from now on L.I.S.) of a random permutation of n points, assuming uniform distribution on the random permutation space.

Definition 2.1 Let \mathcal{S}_n denote the group of permutations of $\{1, \dots, n\}$. If $\pi \in \mathcal{S}_n$, we say that $\pi(i_1), \dots, \pi(i_k)$ is an increasing subsequence in π if $1 \leq i_1 < \dots < i_k \leq n$ and $1 \leq \pi(i_1) < \pi(i_2) < \dots < \pi(i_k) \leq n$.

Definition 2.2 Given a random permutation $\pi \in \mathcal{S}_n$, we call $l_n(\pi)$ (or $ld_n(\pi)$) the length of the longest increasing (or decreasing) subsequence of π .

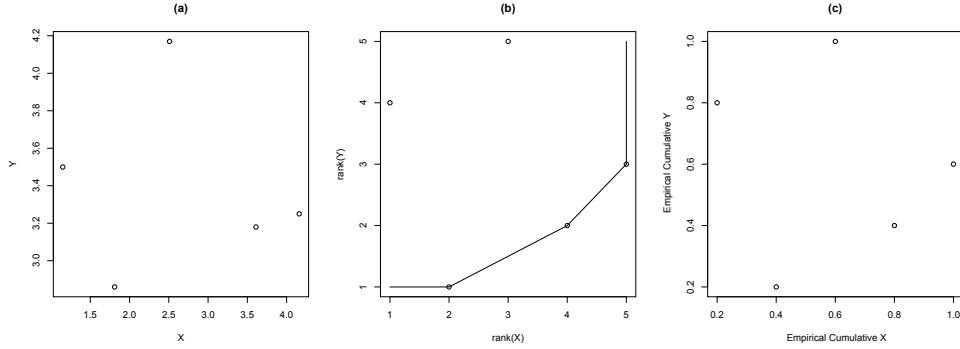


Figure 3: Dispersion's graphic and permutation (Example 2.1). (a) is the dispersion plot for the sample. (b) represents the permutation defined by the sample, the black line shows the L.I.S. (c) shows the empirical copula of the sample.

Example 2.2 (continued). On the Example 2.1 $l_5(\pi) = 3$, see figure 3.

On the next section we study the distribution of the length of the L.I.S., under the assumption of independence between X and Y .

Under the hypotheses (1.1) every possible permutation defined by a random sample of size n , $\{(x_i, y_i)\}_{i=1}^n$, has the same probability $1/n!$. With some abuse of terminology, let us call $\pi(\{(x_i, y_i)\}_{i=1}^n)$ the permutation defined by the sample as being $\pi(\text{rank}(x_i)) = \text{rank}(y_i)$, $i = 1, \dots, n$.

Definition 2.3 Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be replications of (X, Y) under the assumption 1, we denote by L_n the random variable,

$$L_n = l_n(\pi(\{(X_i, Y_i)\}_{i=1}^n)).$$

Let \mathcal{S}_n denote the group of permutations of $\{1, \dots, n\}$ and let $(\mathcal{S}_n, \mathcal{U})$ be \mathcal{S}_n with the uniform distribution \mathcal{U} , then for $k = 1, 2, \dots, n$, if $p_k^n = P(L_n = k)$,

$$p_k^n = \frac{\#\{\pi \in \mathcal{S}_n : l_n(\pi) = k\}}{n!} \quad (2.1)$$

Remark 1 $(\mathcal{S}_n, \mathcal{U})$ is the space of permutations $\pi(\{(X_i, Y_i)\}_{i=1}^n)$ where (X_i, Y_i) are i.i.d. with the same law of (X, Y) under hypotheses (1.1) and assumption 1.

2.1 Exact distribution of L_n in the case of independence

In order to compute p_k^n we introduce the Young Tableau.

Definition 2.4 A standard Young Tableau of order n is an arrangement of n distinct natural numbers in rows and columns so that the numbers in each row and in each column form increasing sequences, and so that there is an element of each row in the first column and an element of each column in the first row, and there are no gaps between numbers.

Schensted (1961) shows that the number of sequences of size n with a L.I.S. of size k and longest decreasing subsequence of length m , is the sum of the squares of the number of standard Young tableaux of identical shape, having k columns and m rows.

Let $V_n(k, m)$ be the set of shapes of Young tableaux of size n having k columns and m rows.

Theorem 2.1 Under the assumption 1, if X, Y are independent then,

$$p_k^n = \frac{1}{n!} \left[\sum_{m=1}^n \sum_{W \in V_n(k,m)} N(W)^2 \right] \quad (2.2)$$

where $V_n(k, m)$ is the set of shapes of standard Young tableaux of size n having k columns and m rows, $N(W)$ is the number of standard Young tableaux with shape W and p_k^n is given by equation (2.1).

Using theorem 2.1 we compute p_k^n for $1 \leq k \leq n$, $n < 100$. The table can be accessed from our LIStest package, implemented in R project.

2.2 Asymptotic distribution of L_n . Case of independence

The asymptotic distribution for random permutations, after appropriate centering and scaling, was first obtained by Baik et al. (1999). The Tracy-Widom distribution have cumulative distribution,

$$F_{TW}(t) = \exp \left(- \int_t^\infty (z-t)q^2(z)dz \right), \quad t \in \mathbb{R}. \quad (2.3)$$

Under the same conditions of theorem 2.1, Baik et al. (1999) prove that if χ is a random variable whose distribution function is F_{TW} . Then,

$$\chi_n = \frac{L_n - 2\sqrt{n}}{n^{1/6}} \rightarrow \chi \text{ in distribution, as } n \rightarrow \infty. \quad (2.4)$$

We calculate the asymptotic p-values using the R-package ‘‘RMTstat’’, specifying the parameter $\beta = 2$ in the cumulative function ‘‘ptw’’.

2.3 The L_n independence test

Let $(x_1, y_1), \dots, (x_n, y_n)$ be a paired sample of size n of (X, Y) , where the assumption 1 is true. The p-value for a statistical test with null hypothesis of independence against an alternative hypothesis of not independence between X and Y is defined in the following way.

Definition 2.5 We define the two-sided p-value by,

$$\min \left\{ 2F_{L_n}(l_0)I_{\{F_{L_n}(l_0) \leq \frac{1}{2}\}} + 2(1 - F_{L_n}(l_0))I_{\{F_{L_n}(l_0) > \frac{1}{2}\}}, 1 \right\}$$

where l_0 is the observed value of L_n in the sample, F_{L_n} is the cumulative distribution function, $F_{L_n}(l_0) = \sum_{k=1}^{l_0} p_k^n$ (see equation (2.2)) and I_E denotes the indicator function of set E .

2.4 A variation of the L_n independence test

Significance test that are based on discrete statistics are conservative. That is the case with the L_n independence test. Figure 4 (a) shows $F_{L_{80}}$ which is the L_n cumulative distribution function for $n = 80$. For example, $F_{L_{80}}(x)$ jumps from 0.0078 for $x = 11$ to 0.057 for $x = 12$, because of this, if we want to test at level 0.05 for unilateral alternative, we will reject the independence when the unilateral p-value is ≤ 0.0078 and the exact level 0.05 can not be achieved.

To mitigate this problem we define the following variation of the L_n statistic which is inspired on the jackknife procedure.

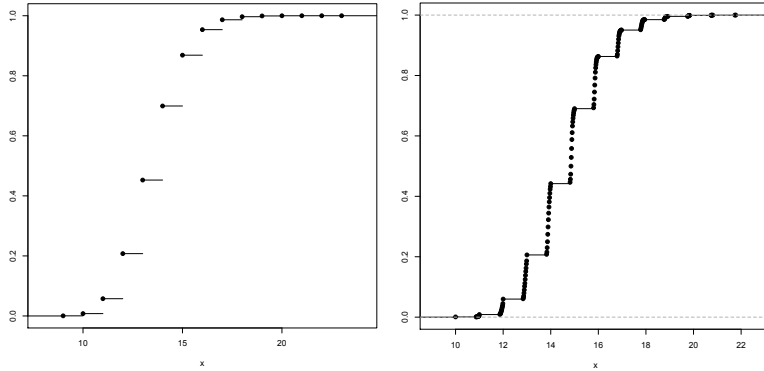


Figure 4: (a) L_n cumulative distribution function for $n = 80$ (left). (b) JL_n cumulative distribution function for $n = 80$ (right).

Definition 2.6 Let $(x_1, y_1), \dots, (x_n, y_n)$ be a paired sample of size n of (X, Y) , under the assumption 1. We define the JL_n statistic as

$$JL_n = \frac{1}{n} \sum_{i=1}^n L_n \left((x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_n, y_n) \right)$$

Figure 4 (b) shows $F_{JL_{80}}$ which is JL_n cumulative distribution function for $n = 80$. We can see that even if this statistic is still discrete the number of steps in the function has grown. In this case, $F_{JL_{80}}(x)$ jumps from 0.0449 for $x = 11.99$ to 0.0598 for $x = 12$, if we want to test at level 0.05 with p-values ≤ 0.0449 we reject the hypothesis 1.1.

Remark 2 Another interesting characteristic of the L_n (and JL_n) statistic is that, as seen in Aldous et al. (1995), under the hypothesis 1.1, $\lim_{n \rightarrow \infty} \frac{E(L_n)}{\sqrt{n}} = 2$, in other words, under the assumption of independence, when n grows L_n grows like $2\sqrt{n}$.

In section 3 we will use this two characteristics to explain the behavior of our test on simulated data.

2.5 Applications

2.5.1 Application 1, “coalminers” data

We can see in figure 2 (left) the plot between the variable BW (counts with breathlessness and wheeze) and nBW (counts with no breathlessness but wheeze) the Spearman correlation was 0.467 and because of that the Spearman correlation test cannot reject the independence, with a p-value equal to 0.2125. We show in Table 1 the p-values for all the test considered on this paper. We observe that our test L_n (and JL_n) was overcome by the Hoeffding test (with a lesser p-value equal to 0.0426), nevertheless we note that the Hoeffding test for small and moderate sample size tends to reject the hypothesis 1.1 (see the results from simulations in table 2).

Test	Spe	Ken	Hoe	Pea	Gen	L_n	JL_n
p-value	0.2125	0.1802	0.0426	0.2567	0.1314	0.0994	0.0962

Table 1: Testing hypothesis 1.1 (BW vs nBW)

Since the figure 2 (left) exposes a increasing tendency, we can test some unilateral hypotheses, when it make sense. For the tests based on specific measure, we test “measure > 0 ”; for the L_n test we test $L_n > M_0$, where M_0 is the mode of the distribution under the independence assumption. We show in table 3 the results in which case our tests L_n and

Sample size	10	20	30	40	50	60	70	80	90	100
Power	0.11	0.07	0.069	0.072	0.062	0.06	0.063	0.057	0.055	0.054

Table 2: Empirical power of Hoeffding test at nominal level 0.05. Independent $N(0,1)$.

Table 3: Unilateral hypotheses

Test	Spe	Ken	Pea	L_n	JL_n
p-value	0.1063	0.0901	0.1283	0.0497	0.0480

JL_n show better performance, rejecting the independence in favor of the increasing tendency. In the same data set we can track other combination of variables which performance is similar to the described (BW vs nBW), they are (nBW vs BnW), were BnW denotes the counts with breathlessness but no wheeze.

2.5.2 Application 2, “Globular Clusters” data

The dataset is composed by a sample of globular clusters (GC) around the galaxy NGC 3379. The NGC 3379 is the brightest elliptical galaxy in the constellation Leo and it is know to have a supermassive black hole. The measures are R vs V to the galaxy for 30 GCs, where in x axis, R represents the radius of a galactic particle in the 2-D image and in y axis, V represents the 1-D velocity of a galactic particle that sits at radius R in the system. While conceptually the dependence between R and V is obvious, it is not detected by Pearson, Kendall, Spearman, Hoeffding and Genest tests, only our tests L_n (with a p-value around 0.02) and JL_n (with a p-value around 0.018) are capable to show that fact. The Spearman correlation between R and V is equal to 0.2766.

Test	Spe	Ken	Hoe	Pea	Gen	L_n	JL_n
p-value	0.1390	0.0900	0.1331	0.2760	0.1953	0.0203	0.0180

In other situations our test can be useful in order to reject the independence. For example, it is the case with the data appearing on [Simonoff \(1999\)](http://pages.stern.nyu.edu/~jsimonof/SmoothMeth/Data/Tab/ethanol.tab). It can be obtained at <http://pages.stern.nyu.edu/~jsimonof/SmoothMeth/Data/Tab/ethanol.tab>. The data relate two quantities (88 paired observations), the concentration of nitric oxides (Nitric oxide) in engine exhaust to the equivalence ratio (Equivalence ratio), a measure of richness of the air ethanol mix, for burning ethanol in a single-cylinder automobile test engine. The Spearman test gives a p-value equal to 0.1896, ignoring the dependence between the variables. Our L_n test gives a p-value equal to 0.02 and our JL_n test gives a p-value of 0.0078. This result suggests evidence against the hypothesis 1.1. We note that in that case the Hoeffding test is the champion with a p-value of 1e-08 and as second best option the Genest’ test with a p-value equal to 0.0005.

3 Simulation

To compare the power of our tests against the Pearson, Kendall, Spearman, Hoeffding and Genest tests, we carried out a simulation study in which for each test we estimate the power function for different sample sizes and diverse joint distributions. For each joint distribution and sample sizes 40, 60, 80, 100 we simulated 5000 samples, and computed the p-values. Our simulation was implemented in the R-project. Denote by $\left\{ (X_i^j, Y_i^j) \right\}_{i=1}^n$ the j simulated sample of size $n = 40, 60, 80, 100$. Where $j = 1, \dots, 5000$. Given a level α we calculate the empirical power as being,

$$\frac{\# \left\{ j : \text{p-value} \left(\left\{ (X_i^j, Y_i^j) \right\}_{i=1}^n \right) \leq \alpha \right\}}{5000} \quad (3.1)$$

where $\text{p-value}(\{(X_i^j, Y_i^j)\}_{i=1}^n)$ denotes the p-value associated with the sample j , $\{(X_i^j, Y_i^j)\}_{i=1}^n$. For the L_n p-values we use the theorem 2.1 and definition 2.5; for the JL_n p-values we applied the definition 2.6 and definition 2.5, for sample sizes $n < 100$. For $n = 100$, each p-value is given by the definition 2.5 and by the asymptotic distribution (equation (2.4)). In this way the standard deviation of the estimator is upper bounded by 0.007.

3.0.3 Dependence with zero or very small correlation/association coefficients.

We consider two main situations in which the samples show low correlation/association. (a) Visible dependence; (b) hidden dependence. In the first group we explore distributions with the following x-y plot shapes (i) a cross, (ii) a ring, (iii) a square. All of them are indisputable types of dependence with null correlation/association coefficients. We note that the Pearson, Kendall and Spearman tests are not consistent for the hypothesis 1.1 explaining its poor performance in almost all the cases exposed in this section.

For case (a) we implemented the following joint distributions,

D1- Mixture of two bivariate Normal distributions with variances 1 and correlations ρ and $-\rho$; $(X, Y) \sim \frac{1}{2}N_2(\underline{0}, \Sigma_1) + \frac{1}{2}N_2(\underline{0}, \Sigma_2)$, where $\underline{0} = (0, 0)$, $\Sigma_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$.

D2- Uniform ring centered in 0 with internal radius of 0.95 and external radius of 1.

D3- Uniform distribution on $\{[-1, 1] \times [-1, 1]\} \setminus \{[-0.95, 0.95] \times [-0.95, 0.95]\}$ (Border of a square).

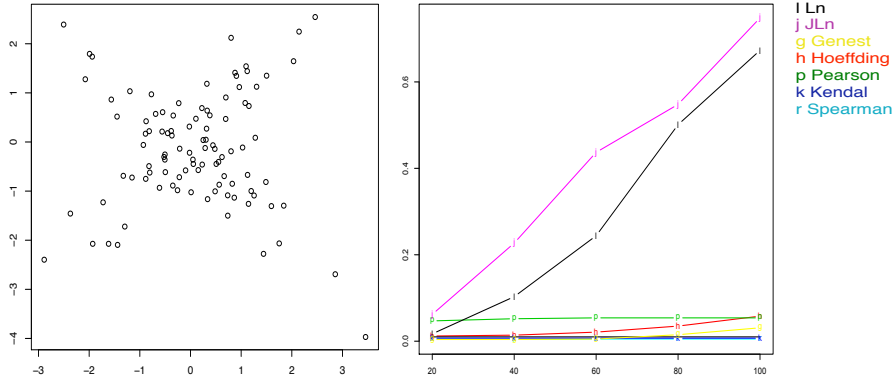


Figure 5: The picture on the left is the scatter plot of a sample size 100 of D1 when $\rho = 0.9$. The picture on the right is sample size vs. empirical power (level 0.01) for the same distribution.

As we remarked in section 2, $E(L_n)/\sqrt{n} \rightarrow 2$ when $n \rightarrow \infty$, in other words we expect L_n to be close to $2\sqrt{n}$. Figure 5 shows a sample of D1 when $\rho = 0.9$ and the empirical powers (level 0.01) for that distribution. In this case, the L_n and JL_n tests have the highest empirical power. Looking at the X shape of the sample graphic, it is clear that the L.I.S. for that distribution will be realized mostly with points on the positive correlated part of the mixture, because of this for ρ close to 1, the size of the L.I.S. is close to $n/2$, while in the independent case, the L.I.S. size is approximately $2\sqrt{n}$. As a consequence the L_n (and JL_n) tests detect the dependence. Table 4 shows the tendencies of the empirical power under distribution D1 with correlation coefficients taking values 0.7, 0.8 and 0.9. The JL_n (and L_n) test achieves the higher values in the empirical power, when the sample size grows. On distribution D2 (see figure 6) Hoeffding, L_n and JL_n give the best results. The Genest test shows a increasing power for sample sizes bigger than 60, but in our study it maintain the power under the L_n , JL_n and Hoeffding power tests. For distribution D3 all the tests fail. In D3, the estimated size of the L.I.S. is close to $2\sqrt{n}$ (see the copula in the figure 7), while in the independent case it is close to $2\sqrt{n}$. The

ρ	n	Spe	Ken	Hoe	Pea	Gen	L_n	JL_n
0.7	40	0.006	0.009	0.014	0.030	0.003	0.013	0.032
	60	0.006	0.009	0.012	0.033	0.004	0.024	0.075
	80	0.006	0.009	0.012	0.033	0.008	0.061	0.079
	100	0.007	0.010	0.013	0.033	0.007	0.104	0.144
0.8	40	0.006	0.009	0.012	0.040	0.003	0.029	0.059
	60	0.005	0.008	0.013	0.042	0.003	0.064	0.136
	80	0.005	0.009	0.015	0.041	0.007	0.165	0.203
	100	0.005	0.009	0.018	0.044	0.011	0.25	0.338
0.9	40	0.004	0.007	0.014	0.052	0.004	0.103	0.227
	60	0.005	0.007	0.021	0.054	0.005	0.244	0.437
	80	0.005	0.007	0.035	0.054	0.015	0.501	0.548
	100	0.005	0.007	0.058	0.054	0.031	0.673	0.748

Table 4: Empirical power (level 0.01) for distribution D1

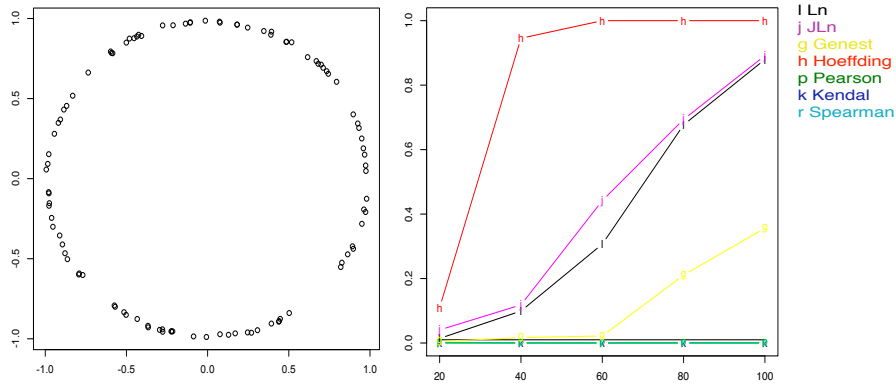


Figure 6: The picture on the left is the scatter plot of a sample size 100 of D2. The picture on the right is sample size vs. empirical power (level 0.01) for the same distribution.

n	Spe	Ken	Hoe	Pea	L_n	Gen	JL_n
40	0.000	0.000	0.945	0.001	0.099	0.017	0.118
60	0.000	0.000	1.000	0.000	0.307	0.022	0.442
80	0.000	0.000	1.000	0.001	0.677	0.211	0.694
100	0.000	0.000	1.000	0.001	0.880	0.356	0.891

Table 5: Empirical power function for an uniform ring centered in 0 with internal radius of 0.95 and external radius of 1 (D2) at level 0.01

behavior of the L.I.S. is clear looking to the copula in the figure 7. We have 4 rectangular blocks. Inside each block we find around $n/4$ points independently distributed in the block. Consequently, the L.I.S. of each block will have size close to $2\sqrt{n/4} = \sqrt{n}$. From the copula plot, it is evident that the L.I.S. of the sample will be realized mostly in one of two forms, it can be the union of the L.I.S. on the left block plus the L.I.S. on the top block or it can be the union of the L.I.S. on the bottom block plus the L.I.S. on the right block. Any way, we have a L_n taking values around $2\sqrt{\frac{n}{4}} + 2\sqrt{\frac{n}{4}} = 2\sqrt{n}$. We obtain then, that the JL_n and L_n tests achieve low empirical power. The Hoeffding test achieves the bigger power but it is very small (around 0.055 for $n = 100$).

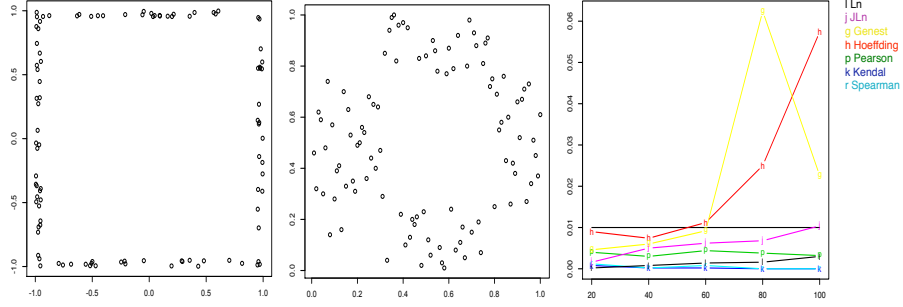


Figure 7: The picture on the left is the scatter plot of a sample size 100 of D3. The picture on the middle is the empirical copula. The picture on the right is sample size vs. empirical power (level 0.01) for the same distribution.

n	Spe	Ken	Hoe	Pea	L_n	Gen	JL_n
40	0.000	0.000	0.007	0.003	0.001	0.006	0.005
60	0.001	0.000	0.011	0.004	0.001	0.009	0.006
80	0.000	0.000	0.025	0.004	0.002	0.062	0.007
100	0.000	0.000	0.057	0.003	0.003	0.023	0.010

Table 6: Empirical power function for D3, at level 0.01

For case (b) we implemented the following joint distributions,

D4- Mixture of two bivariate Normal distributions one independent with standard deviation 4 and the other dependent with standard deviation 1 and correlation ρ , $(X, Y) \sim \frac{1}{2}N_2(\underline{0}, 16I) + \frac{1}{2}N_2(\underline{0}, \Sigma)$, where $\underline{0} = (0, 0)$, $16I = \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix}$

$$\text{and } \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

D5- Mixture of two bivariate Normal distributions, $\alpha\%$ independent with standard deviation 4 and $(1 - \alpha)\%$ dependent with standard deviation 0.5 and correlation $\rho = 0.95$, $(X, Y) \sim \alpha N_2(\underline{0}, 16I) + (1 - \alpha)N_2(\underline{0}, \Sigma)$, where $\underline{0} = (0, 0)$, $16I = \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 0.25 & \rho \\ \rho & 0.25 \end{bmatrix}$.

D6- Mixture of two bivariate Clayton's copulas one with parameter -0.1 and the other with parameter equal to 10; $(X, Y) \sim 0.75C_C(\cdot, \cdot | -0.1) + 0.25C_C(\cdot, \cdot | 10)$.

Table 7 shows the tendencies of the empirical powers under distribution D4 with correlation coefficients taking values 0.9, 0.95, 0.97 and 0.99. The JL_n (and L_n) test achieves the higher values in the empirical power. Table 8 shows the empirical power for distribution D5 with $\alpha = 0.8$ and 0.85. The L_n and JL_n tests have better empirical powers. Figure 8 and figure 9 show the empirical powers for D4 with $\rho = 0.95$ and D5 for $\alpha = 0.8$, respectively. As an explanation for our tests good powers, we recall our first remark in section 2, L_n and JL_n statistic go to regions of very low probability when the dependence between the random variables increase the L.I.S. size even in a small quantity with respect to the values of L.I.S for the independence case. For D4 and D5, the mixture with the high correlated data (even if it have low variance) is detected because of the significative increment on the L.I.S. size. We observe, through figure 10 and table 9 that for D6 distribution in which case the Spearman correlation was about 0.15, our L_n (and the JL_n version) achieves the best power, the Spearman, Kendall and Pearson test can not detect the dependence because the sample proportion (25%) which has strong correlation (around 0.94) is smaller that the sample proportion (75%) having negative and small

ρ	n	Spe	Ken	Hoe	Pea	Gen	L_n	JL_n
0.9	40	0.119	0.241	0.538	0.065	0.225	0.225	0.686
	60	0.169	0.365	0.761	0.070	0.297	0.297	0.911
	80	0.233	0.481	0.900	0.072	0.536	0.536	0.975
	100	0.288	0.591	0.963	0.076	0.784	0.784	0.995
0.95	40	0.136	0.299	0.677	0.067	0.261	0.737	0.895
	60	0.192	0.443	0.882	0.068	0.360	0.941	0.980
	80	0.258	0.577	0.969	0.073	0.629	0.994	0.997
	100	0.336	0.687	0.994	0.077	0.876	1.000	1.000
0.97	40	0.142	0.324	0.739	0.069	0.278	0.861	0.960
	60	0.195	0.483	0.938	0.067	0.389	0.985	0.994
	80	0.272	0.635	0.989	0.074	0.684	1.000	1.000
	100	0.346	0.752	0.999	0.077	0.915	1.000	1.000
0.99	40	0.145	0.370	0.841	0.063	0.308	0.979	0.998
	60	0.213	0.553	0.973	0.073	0.435	1.000	1.000
	80	0.288	0.710	0.998	0.072	0.734	1.000	1.000
	100	0.367	0.813	1.000	0.077	0.937	1.000	1.000

Table 7: Empirical power (level 0.01) for distribution D4

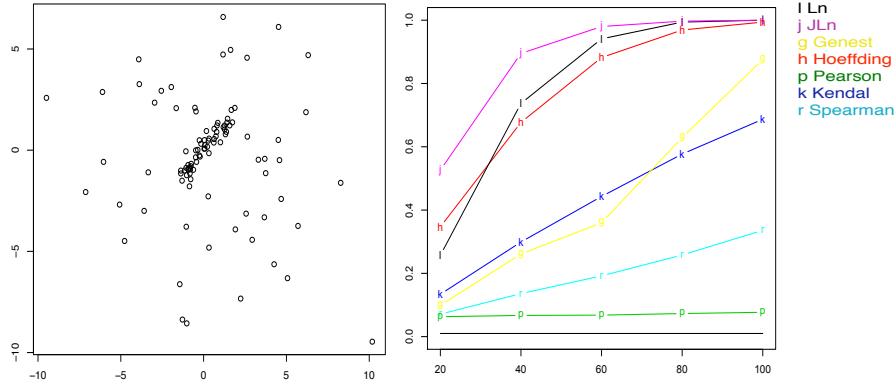


Figure 8: The picture on the left is the scatterplot of a sample size 100 of D4 when $\rho = 0.95$. The picture on the right is sample size vs. empirical power (level 0.01) for the same distribution.

α	n	Spe	Ken	Hoe	Pea	L_n	Gen	JL_n
0.85	40	0.017	0.021	0.042	0.017	0.044	0.015	0.092
	60	0.014	0.019	0.047	0.016	0.123	0.023	0.252
	80	0.017	0.022	0.055	0.019	0.301	0.016	0.330
	100	0.015	0.021	0.066	0.019	0.488	0.020	0.563
0.80	40	0.021	0.029	0.071	0.022	0.109	0.020	0.210
	60	0.022	0.034	0.085	0.021	0.296	0.037	0.494
	80	0.022	0.036	0.125	0.021	0.582	0.025	0.636
	100	0.019	0.034	0.147	0.018	0.766	0.034	0.838

Table 8: Empirical power function for D5, $\alpha = 0.85$ (upper table) and 0.8 (lower table), at level 0.01

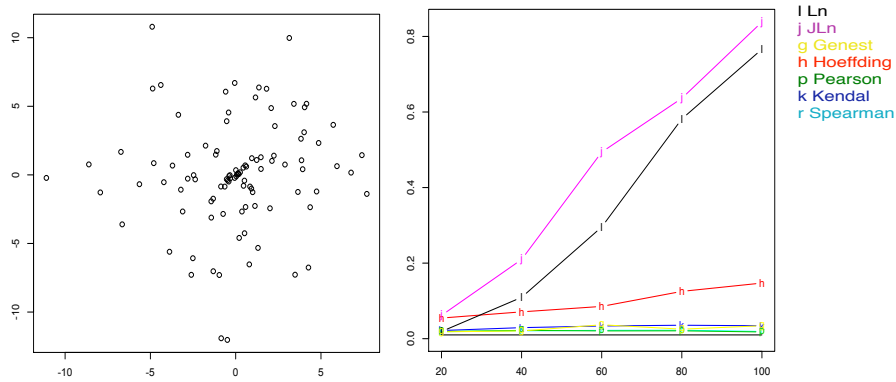


Figure 9: The picture on the left is the scatter plot of a sample size 100 of D5 with $\alpha = 0.8$. The picture on the right is sample size vs. empirical power (level 0.01) for the same distribution.

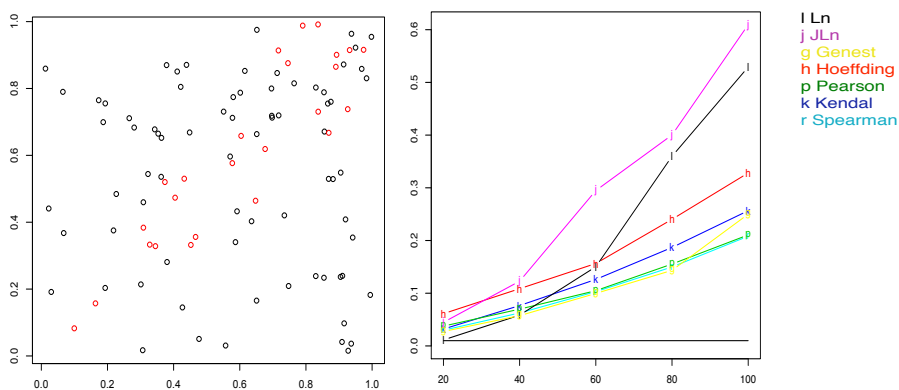


Figure 10: The picture on the left is the scatter plot of a sample size 100 of D6, in red we show the points coming from Clayton copula with parameter equal to 10. The picture on the right is sample size vs. empirical power (level 0.01) for the same distribution.

	Spe	Ken	Hoe	Pea	L_n	Gen	JL_n
40	0.063	0.077	0.109	0.070	0.058	0.057	0.123
60	0.103	0.126	0.156	0.105	0.150	0.100	0.295
80	0.150	0.187	0.240	0.155	0.360	0.143	0.400
100	0.208	0.257	0.328	0.211	0.529	0.250	0.609

Table 9: Empirical power function for a mixture 25%-75% of Clayton copulas with parameters 10 and -0.1 respectively (D6) at level 0.01

correlation (about -0.17). This means that 25% it is not sufficient to detect dependence using a test based on that specific coefficients.

4 Conclusions

In this work we develop a new nonparametric independence test for two continuous random variables. We find the exact distribution for the test statistics, we compare our test with Pearson, Kendall, Spearman Genest and Hoeffding tests using simulations and we apply our test in two real data examples, where our test shows better performance than the others.

For the sample sizes considered on our study, our procedures, L_n and JL_n , where the only ones to detect dependence for distributions D1 and D5 with large power. For distributions D4 and D6 our procedures had the larger power for detecting the dependence, for moderate and bigger sample sizes the Hoeffding test power follows ours. For D2 the best test was Hoeffding followed by our procedures and finally for distribution D3 all the test fails for the sample sizes considered, the Hoeffding test seems to be starting to detect some dependence but need larger sample sizes.

Our tests detect dependence through the existence of runs in the rank graphic. This means that if the data appear to be independent for the most part of the graphic, except for the presence of some sub-region of the graphic in which some significant run could be detected then our tests show evidence against the independence. Another property of our procedure that is necessary to remark is the ability to deal with moderate (even small) sample sizes. In all the cases in which our test works well, our tests have the highest power for sample sizes bigger than 40, this property added to the capacity of to control the significance level (through the JL_n implementation) put our procedure in an advantaged position in relation to Hoeffding test, as is strongly exposed in the simulation study and in the applications.

References

- ALDOUS, D. and DIACONIS, P. (1995). Hammersley's interacting particle process and longest increasing subsequence. *Probab. Theory Related Fields* **103** 199-213. 1355056
- BAER, R. M. and BROCK, P. (1968). Natural Shorting Over Permutation Spaces. *Math. Comp.* **22** 385-410.
- BAIK, J., DEIFT, P. and JOHANSSON, K. (1999). On The Distribution of the Length of the Longest Increasing Subsequence of Random Permutations. *J. Amer. Math. Soc.* **12** 1119-1178.
- FRAME, J. S., ROBINSON, B. and THRALL, R. M. (1954). The hook graphs of the symmetric group. *Canad. J. Math.* **6** 316-324.
- GENEST, C. and REMILLARD, B. (2004). Tests of Independence or Randomness Based on the Empirical Copula Process. *Test* **13** 335-369.
- HASTINGS, S. P., McLEOD, J. B. (1980). A boundary value problem associated with the second Painlevé transcendent and the Korteweg de Vries equations. *Arch. Ration. Mech. Anal.* **73** 31-51.
- HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19** 546-57.
- HOLLANDER, M., WOLFE, D. (1973). *Nonparametric Statistical Methods*. New York: John Wiley & Sons. 185-194.
- LOGAN, B. F. and SHEPP, L. A. (1977). A variational problem for random Young tableaux. *Adv. Math.* **26** 206-222.
- MIKUSINSKI, P., SHERWOOD, H. and TAYLOR, M. D. (1991-92). The Fréchet Bounds Revisited. *Real Anal. Exchange.* **17** 759-764.
- NELSEN, R. B. (1999). *An Introduction to Copulas*, Springer Verlag, New York.
- ODLYZKO, A. M. and RAINS, E. M. (2000). On longest increasing subsequence in random permutations. Analysis, Geometry, Number Theory: The Mathematics of Leon Ehrenpreis, E. L. Grinberg, S. Berhanu, M. Knopp, G. Mendoza, and E. T. Quinto, eds., Amer. Math. Soc., Contemporary Math. Number 251.
- SCHENSTED, C. (1961). Longest increasing and decreasing sub-sequences. *Canad. J. Math.* **13** 179-191.
- SIMONOFF, J., S. (1996). Smoothing Methods in Statistics. *Springer Series in Statistics*.