symmetry

MDPI

Article

Random Permutations, Non-Decreasing Subsequences and Statistical Independence

Jesús E. García and Verónica A. González-López *

Department of Statistics, University of Campinas, Sérgio Buarque de Holanda, 651, Campinas 13083-859, São Paulo, Brazil; jg@ime.unicamp.br

* Correspondence: veronica@ime.unicamp.br

Received: 1 August 2020; Accepted: 21 August 2020; Published: 26 August 2020

Abstract: In this paper, we show how the longest non-decreasing subsequence, identified in the graph of the paired marginal ranks of the observations, allows the construction of a statistic for the development of an independence test in bivariate vectors. The test works in the case of discrete and continuous data. Since the present procedure does not require the continuity of the variables, it expands the proposal introduced in *Independence tests for continuous random variables based on the longest increasing subsequence (2014)*. We show the efficiency of the procedure in detecting dependence in real cases and through simulations.

Keywords: symmetric group; permutations; hypothesis tests

1. Introduction

In this article, we use an expanded structure of the symmetric group $\mathcal{S}_n$, over the set of permutations from $\{1,\ldots,n\}$ to $\{1,\ldots,n\}$, to develop a dependence detection procedure in bivariate random vectors. The procedure is based on identifying the longest non-decreasing subsequence (LNDSS) detected in the graph of the paired marginal ranks of the observations. It records the size of the subsequence and verifies the chances that it has to occur in the expanded space of $\mathcal{S}_n$, under the assumption of independence between the variables. The procedure does not require assumptions about the type of the two random variables being tested, such as being both discrete, both continuous or a mixed structures (discrete-continuous).

When we face the challenge of deciding whether the independence between random variables can be discarded, it is necessary to establish the nature of the variables, whether they are continuous or discrete. For continuous random variables, we have several procedures, for example, Hoeffding's test and those based on dependence's coefficients (Spearman's coefficient, Pearson's coefficient, Kendall's coefficient, etc.). Instead, for the discrete case, the options are few, the most popular is Pearson's Chi-squared test. Also, the tests based on Kendall and Spearman coefficients going through corrections that consider ties can be used to test for independence between two discrete and ordinal variables (see [1,2]). In general, recommended for small sample sizes. Moreover, some derivations of the Chi-squared statistic have been projected to test independence between two nominal variables, as is the case of the Cramér's V statistic, see [3].

The goal of this article is to show an independence test, developed from the notion of the LNDSS among the ranks of the observations, see [4]. The main notion was introduced previously in [5] with a different implementation from the one proposed in this paper. The alterations proposed in this paper aim to improve the procedure's performance. This methodology works without limitation on the type of the two random variables being tested, which can be continuous/discrete.

The existence of ties in a dataset cast doubts about the use of matured tests for continuous variables, see, for instance, [6] for a discussion on this issue. The use of procedures preconized for continuous random variables, in cases with repetitions in the observations due to the precision used to record the data, may have unforeseen consequences on the performance of the procedure. If the ties are eliminated, the use of asymptotic distributions can be compromised, if the ties are considered (by means of some correction), the control of type 1 and 2 errors can be put at risk (increasing the false positives/negatives of the procedure). Another frequent situation is when one of the variables is continuous, and the other is discrete. For some test of independence, problems may arise from this situation forcing the practitioner to apply some arbitrary data categorization. Under this picture, one of the most popular procedures is the Pearson's Chi-squared statistic. The traditional tests are based on some of the following statistics Pearson's Chi-squared, likelihood ratio [1], and Zelterman's [7] for the case in which the number of categories is too large for the available sample size. Moreover, Zelterman's [7] do not work well when one of the variables is continuous. [1] shows several examples of independent data, where Pearson's Chi-squared, likelihood ratio, and Zelterman statistics fail. In [1] is shown that to be reliable, those tests require that each cell in the frequency table should have a minimal (non zero) frequency, which can depend on the total size of the data set. It is shown in [7], that in some situations, with a large number of factors, Pearson's Chi-squared statistic will behave as a normal random variable with summaries as variance and mean that are unassociated to the Chi-squared distribution, even with large sample size. Those situations are similar to the case of continuous random variables registered with limited precision, which in fact, is similar to a discrete random variable with a large number of categories producing sparseness (or sparse tables).

This article is organized as follows. Section 2 introduces the formulation of the test showing the new strategy, in comparison with the implemented in [5]. Section 3 simulates different situations showing the performance of the procedure. The purpose is to show situations in which the statistic proposed in this paper is efficient in detecting dependence. We consider in the simulations settings concentrating points in the diagonals, the variables being continuous or discrete. We also consider perturbations of such situations, which will show the maintenance or loss of power of the test developed here. Section 4 applies the new procedure to real data, and Section 5 presents the final considerations.

2. The Procedure

We start this section with the construction of the test's statistics. For that, we introduce the LNDSS notion.

Definition 1. *Given the set* $Q = \{q_1, \ldots, q_n\}$ *of cardinality* n *such that* $q_i \in \mathbb{R}, \forall i \in \{1, \ldots, n\}$,

i. *the subsequence* $\{q_{i_1}, \ldots, q_{i_k}\}$ *of* Q *is a non-decreasing subsequence of* Q *if* $1 \leq i_1 < \cdots < i_k \leq n$ *and* $q_{i_1} \leq q_{i_2} \leq \cdots \leq q_{i_k}$;
ii. *the length of a subsequence verifying i. is k*;
iii. $lnd_n(Q) = \max_k\{1 \leq k \leq n : \{q_{i_1}, \ldots, q_{i_k}\} \in S_n\}$, *where* S_n *is the set of subsequences of* Q *verifying i.*

$lnd_n(Q)$ (item iii., Definition 1) is the length of the LNDSS of Q. Here, consider two illustrations of Definition 1. Suppose that $Q = \{1.3, 0.2, 2, 2.1, 1.2\}$. Then the LNDSS are $\{1.3, 2, 2.1\}$ and $\{0.2, 2, 2.1\}$, then $lnd_5(Q) = 3$. Consider now a collection Q with replications $Q = \{1.5, 2.4, 1.1, 2.4, 3, 3.1\}$, so the LNDSS is $\{1.5, 2.4, 2.4, 3, 3.1\}$ and $lnd_6(Q) = 5$.

Using the next Definition we adapt this notion to the context of random samples.

Definition 2. *Consider* (X, Y) *a random vector with joint cumulative distribution function* H, *let* $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ *be independent realizations of* (X, Y), *we denote by* LND_n *the random variable built from iii. of Definition 1 as*

$$LND_n = lnd_n\left(Q_{\mathcal{D}}\right),$$

where $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}$ *and* $Q_{\mathcal{D}} = \{q_{rank(X_i)} = rank(Y_i),\ i = 1, \ldots, n\}$.

Remark 1.

i. Note that without the presence of ties, the set $Q_{\mathcal{D}}$ *is a particular case of all the permutations of the values in the set* $\{1, \ldots, n\}$.

ii. With ties, there is more than one way of defining ranks. We apply the minimum rank notion. For example, the sample $6.1, 2.1, 5.3, 4.7, 5.5, 6.2, 5.3, 4.7$ *has ranks* $7, 1, 4, 2, 6, 8, 4, 2$.

If we consider $\mathcal{S}_n = \{\pi$ permutations such that $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}\}$, the subset $Q_{\mathcal{D}}$ given by Definition 2 and without ties is a specific case of the finite set $\mathcal{S}_n$. Also, $\mathcal{S}_n$ is an algebraic group if it is considered operating with the law of composition among the possible permutations. Given two permutations π_1, π_2 the composition between them results when applying $\pi_2 * \pi_1$ from right to left, it means first applying π_1 and to its result applying π_2, that composition also is a permutation. The law of composition is associative, with an identity element and with the existence of an inverse element for each member of $\mathcal{S}_n$. By Definition a symmetric group defined over any set is the group whose elements are all the bijections from the set to itself, then $\mathcal{S}_n$ is the symmetric group of the set $\{1, \ldots, n\}$ since, it is composed by all the bijections from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$. Since $\{1, \ldots, n\}$ is finite, the bijections are permutations.

Through the next example, we show the construction of the LNDSS in a set $Q_{\mathcal{D}}$ related to fictional observations.

Example 1. *Table 1 shows an artificial data with* $n = 6$ *and already ordered in terms of the magnitude of* x_i *values. We show the graphical construction of* LND_n,

Table 1. Artificial data set, and its marginal ranks.

x_i	y_i	Rank (x_i)	Rank (y_i)
5.3	10.2	1	5
5.3	9.3	1	1
6.1	9.3	3	1
6.1	10.1	3	3
7.1	10.1	5	3
7.3	11.0	6	6

This data defines a $Q_{\mathcal{D}} = \{5, 1, 1, 3, 3, 6\}$. *The maximal non-decreasing subsequence is* $\{1, 1, 3, 3, 6\}$ *given by the trajectory* $(0, 0) - (1, 1) - (3, 1) - (3, 3) - (5, 3) - (6, 6)$ *from the plot between the ranks of the observations, shown in Figure 1. The value of* LND_6 *for this example is 5. We note that the indicated trajectory refers to the correspondence of* $1 \to 1, 3 \to 1, 3 \to 3, 5 \to 3, 6 \to 6$, *which is no longer a permutation in the traditional sense since, it allows repetition both in the domain and in the image.*

Remark 2. *Note that the construction of the statistic* LND_n *is symmetric in the sense that if we exchange the roles of* X *and* Y, *we obtain the same result. Formally, this characteristic is a consequence of the following property. Consider a sample* $\{(X_i, Y_i)\}_{i=1}^{n}$ *and the increasing set of indexes* $\{I_1, \ldots, I_k\} \subseteq \{1, \ldots, n\}$ *such that the trajectory* $(X_{I_1}, Y_{I_1}) - (X_{I_2}, Y_{I_2}) - \cdots - (X_{I_k}, Y_{I_k})$ *constitutes a non-decreasing subsequence (as illustrated*

by Example 1), this occurs if and only if $X_{I_i} \leq X_{I_{i+1}}$ *and* $Y_{I_i} \leq Y_{I_{i+1}}$, $1 \leq i \leq k-1$, *then the trajectory* $(Y_{I_1}, X_{I_1}) - (Y_{I_2}, X_{I_2}) - \cdots - (Y_{I_k}, X_{I_k})$ *constitutes a non-decreasing subsequence also.*

The example shows that the procedure operates in an extended space of the symmetric group $\mathcal{S}_n$. Below we show a motivation to identify the dependence by trajectories such as those used by Definition 2 and exemplified in Figure 1. The dependence on a bivariate vector can be represented by the ranks of the observations; let's see a simple motivation.

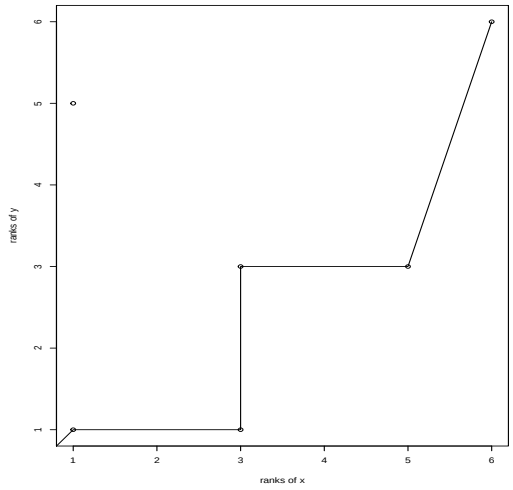

Figure 1. The LNDSS of $Q_{\mathcal{D}}$, from Table 1.

We see on the left of Figure 2 an apparent relationship between the random variables, this illusion of relationship disappears in the graph on the right, since when computing the ranks of the observations, the marginal stochastic structure is neutralized, showing the dependence between X and Y. And, in this case, X and Y are independent, since they have been generated in this way. On the other hand, if the variables X and Y were dependent, Figure 2 on the right should expose a pattern, and traces of it would be captured by the LND_n notion.

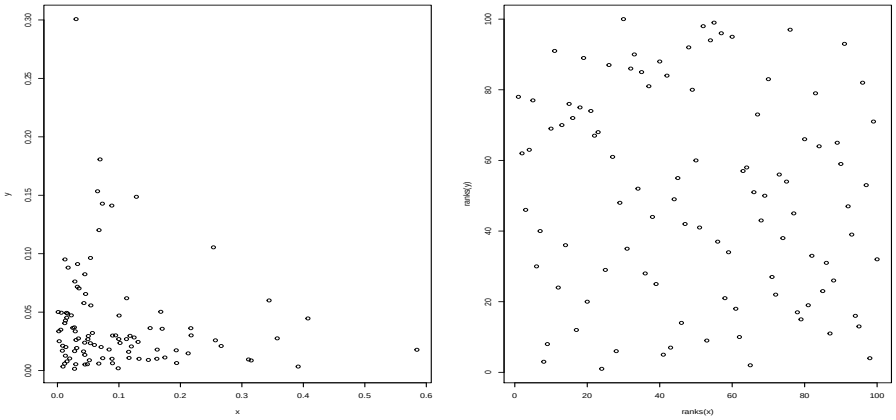

Figure 2. (**left**) X vs. Y. (**right**) $ranks(X)$ vs. $ranks(Y)$. The values of X and Y are simulated from two independent exponential distributions, $\lambda = 10$ for X and $\lambda = 20$ for Y, $n = 100$.

The formulation of the conjectures of independence between the random variables is then given by

$$\begin{aligned} H_0 &: X \text{ and } Y \text{ are independent} \\ H_1 &: X \text{ and } Y \text{ are dependent.} \end{aligned} \tag{1}$$

Here follows the test's statistic build from Definition 2.

Definition 3. *Let* $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ *be replications of* (X, Y). *Define* $JLND_n = \frac{1}{n}\sum_{(u,v)\in\mathcal{D}} LND_n(u,v)$, *where* $LND_n(u,v) = lnd_{n-1}(Q_{\mathcal{D}^{(u,v)}})$ *as given by Definition 2, and* $\mathcal{D}^{(u,v)} = \mathcal{D} \setminus \{(u,v)\}$, *with* $(u,v) \in \mathcal{D}$.

That is, we consider the notion given by the Definition 2 for each set $\mathcal{D}^{(u,v)}$, which include the entire sample except one, allowing to build $Q_{\mathcal{D}^{(u,v)}}$. Then, we define $LND_n(u,v)$ and, the test statistic is the average between all the cases $LND_n(u,v)$. Next we introduce the most frequent formulation of estimation of the two-sided p-value in a context such as that given by the $JLND_n$ statistic.

Definition 4. *The estimator of the two sided p-value for the statistical test of independence between X and Y (see (1)) is defined by,*

$$\min\left\{2\hat{F}_{JLND_n}(jlnd_0)I_{\left\{\hat{F}_{JLND_n}(jlnd_0)\leq\frac{1}{2}\right\}} + 2(1-\hat{F}_{JLND_n}(jlnd_0))I_{\left\{\hat{F}_{JLND_n}(jlnd_0)>\frac{1}{2}\right\}}, 1\right\},$$

where $jlnd_0$ *is the value of* $JLND_n$ *calculated in the sample, see Definition 3.* $\hat{F}_{JLND_n}$ *is the empirical cumulative distribution function of* $JLND_n$, *under independence, and* I_A *is the indicator function of the set* A.

In the following subsection, we analyze the performance of two proposals to estimate F_{JLND_n}, one introduced in [5] and the other proposed by this paper.

2.1. F_{JLND_n} Estimates

$\hat{F}_{JLND_n}$ can be estimated by using bootstrap, for instance see [5]. Denote this kind of estimation as $\hat{F}^B_{JLND_n}$. The procedure to buid $\hat{F}^B_{JLND_n}$ under H_0 hypothesis is replicated here. Let be B a positive and integer value, we compute B size n resamples with replacement of $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ separately, since we assume that H_0 is true. That is, we generate $X_1^b, X_2^b, \ldots, X_n^b$ for $b = 1, 2, \ldots, B$, resampling from $X_1, X_2, \ldots, X_n$, and, we generate $Y_1^b, Y_2^b, \ldots, Y_n^b$ for $b = 1, 2, \ldots, B$, resampling from $Y_1, Y_2, \ldots, Y_n$. Then, for each b define $\mathcal{D}^b = \left\{(X_i^b, Y_i^b)\right\}_{i=1}^n$ and from that sample compute the notion $JLND_n$, from Definition 3, say $JLND_n^b$. Then, if $|A|$ denotes the cardinal of A, set

$$\hat{F}^B_{JLND_n}(q) = \frac{|\{b \,:\, JLND_n^b \leq q\}|}{B}. \tag{2}$$

In Table 2, we show the performance of the $JLND_n$' s test based on the computation of the p-value (Definition 4) according to the Bootstrap technique, given by Equation (2). We generated n independent pairs of discrete Uniform distributions from 1 to m, and we computed in 1000 simulations, the proportion of them showing a p-value (Definition 4) $\leq \alpha$, indicating the rejection of H_0. Such a proportion is expected to be close to α, in order to control type 1 error. As we can see, when increasing the number of categories m, the α level is no longer respected, since the registered proportion always exceeds α. In order to improve the control of type 1 error, in this paper is proposed an alternative way to estimate F_{JLND_n}. The Bootstrap method described above and used in [5] can be modified in order to avoid the removal of any of the observations, following the

strategy of swapping them. We consider $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ separately, given $B \in \mathbb{Z}$, for each $b \in 1, \ldots, B$ consider a permutation $\pi^b : \{1, \ldots, n\} \to \{1, \ldots, n\}$ and define $X_{\pi^b(1)}, \ldots, X_{\pi^b(n)}$. Similarly, consider a permutation $\sigma^b : \{1, \ldots, n\} \to \{1, \ldots, n\}$ and define $Y_{\sigma^b(1)}, \ldots, Y_{\sigma^b(n)}$. Then, for each b define $\mathcal{D}^{\pi^b,\sigma^b} = \left\{\left(X_{\pi^b(i)}, Y_{\sigma^b(i)}\right)\right\}_{i=1}^{n}$ and from that sample compute the notion $JLND_n$, from Definition 3, say $JLND_n^{\pi^b\sigma^b}$. Then, set

$$\hat{F}^{B,\pi,\sigma}_{JLND_n}(q) = \frac{|\{b \ : \ JLND_n^{\pi^b\sigma^b} \leq q\}|}{B}. \tag{3}$$

Bootstrap generates the estimate by Equation (2), it considers samples with replacement, which tends to increase the number of ties. For example, if the original sample has no ties, the Bootstrap procedure tends to create ties, leading to longer non-decreasing subsequences. The permutation-based procedure that allows the formulation of Equation (3) lacks such a tendency, and this principle seems to be a more suitable strategy.

Table 2. The proportion of p-value $\leq \alpha$ computed from Definition 4 and Equation (2), in 1000 simulations of size n, of two independent and discrete Uniform distributions in $\{1, \ldots, m\}$. On top, results for $\alpha = 0.01$, on bottom results for $\alpha = 0.05$.

	n	$m = 10$	$m = 20$	$m = 50$	$m = 100$
$\alpha = 0.01$	20	0.013	0.021	0.022	0.032
	40	0.021	0.038	0.037	0.041
	60	0.025	0.033	0.043	0.050
	80	0.019	0.040	0.053	0.050
	100	0.028	0.034	0.044	0.059
	n	$m = 10$	$m = 20$	$m = 50$	$m = 100$
$\alpha = 0.05$	20	0.084	0.089	0.112	0.100
	40	0.091	0.105	0.134	0.143
	60	0.104	0.114	0.148	0.149
	80	0.095	0.124	0.139	0.159
	100	0.113	0.111	0.125	0.143

In Table 3, we show the performance of the $JLND_n$' s test based on the computation of the p-value (Definition 4) according to Equation (3). We implement the same settings used in Table 2, also we include simulations for $m = 2, 3, 4, 5$. The impact of Equation (3) allows better control of the type 1 error, we see that in most cases the proportion does not exceed α and when it does it remains close to α.

Returning to the construction of the hypothesis test (Equation (1)), we note that the hypothesis H_0 is used in the construction of both types of estimates of the cumulative distribution, Equations (2) and (3). For both cases, the observed values $\{(X_i, X_i)\}_{i=1}^{n}$ are treated separately, as being independent $\{X_i\}_{i=1}^{n}$ by one side and $\{Y_i\}_{i=1}^{n}$ for other side. Then, the distribution of the length of the LNDSS, under H_0, is estimated by both procedures, which allows computing the evidence against H_0 given by the observed value in the originally paired $\{(X_i, Y_i)\}_{i=1}^{n}$ sample and applying Definition 4. Moreover, the type 1 error control refers to the ability of a procedure to reject H_0 under its validity. In other words, it represents an unwanted situation, which we must control. In the study presented by Tables 2 and 3, based on the two ways of estimating the cumulative distribution of $JLND_n$ (under H_0) by Equations (2) and (3) respectively, we see that fixed a level α, Equation (3) offers better performance than Equation (2) since it maintains type 1 error at pre-established levels. For this reason, the test based on the statistic $JLND_n$ with the implementation given by Equation (3) is more advisable in practice.

The following section describes the behavior of the test in different simulated situations, in order to identify its strengths and weaknesses.

Table 3. The proportion of p-value $\leq \alpha$ computed from Definition 4 and Equation (3), in 1000 simulations of size n, of two independent and discrete Uniform distributions in $\{1, \ldots, m\}$. On top, results for $\alpha = 0.01$, on bottom results for $\alpha = 0.05$.

	n	$m=2$	$m=3$	$m=4$	$m=5$	$m=10$	$m=20$	$m=50$	$m=100$
	20	0.004	0.006	0.008	0.014	0.011	0.007	0.007	0.004
	40	0.004	0.009	0.005	0.008	0.007	0.008	0.010	0.011
$\alpha = 0.01$	60	0.005	0.004	0.009	0.007	0.006	0.004	0.010	0.014
	80	0.006	0.005	0.010	0.011	0.012	0.009	0.006	0.008
	100	0.005	0.011	0.011	0.009	0.007	0.009	0.012	0.008
	n	$m=2$	$m=3$	$m=4$	$m=5$	$m=10$	$m=20$	$m=50$	$m=100$
	20	0.021	0.032	0.041	0.047	0.038	0.048	0.042	0.043
	40	0.019	0.038	0.030	0.043	0.030	0.046	0.045	0.037
$\alpha = 0.05$	60	0.029	0.041	0.042	0.044	0.040	0.032	0.056	0.044
	80	0.039	0.031	0.045	0.046	0.051	0.046	0.046	0.052
	100	0.031	0.041	0.048	0.049	0.052	0.054	0.053	0.053

3. Simulations

To investigate the performance of the $JLND_n$-based procedure, we will aim to determine the rejection ability of the procedure in scenarios with dependence. Our research focuses on the procedure that uses the Equation (3) to compute the p-value, given the justification of Section 2.1. We begin our study considering discrete distributions that we describe below and some mixtures or disturbances of them.

We take discrete uniform distributions on different regions, consider m, b and a fixed values such that $m, b, a \in \mathbb{Z}_{>0}$, and set

i. $D1(m,a)$: Uniform on $A = \{(x,y) \in \{1,\ldots,m\}^2 : |x-y| \leq a\}$;
ii. $D2(m,a)$: Uniform on $A = \{(x,y) \in \{1,\ldots,m\}^2 : |x-y| \leq a$ or $|x+y-m-1| \leq a\}$;
iii. $D3(m,a,b)$: Uniform on $A = \{(x,y) \in \{1,\ldots,m\}^2 : |x-y| \leq a$ or $|x-y+b| \leq a$ or $|x-y-b| \leq a$ or $|x-y-2b| \leq a$ or $|x-y+2b| \leq a\}$.

The performance of the distributions i.- iii. is illustrated in Figure 3, using $m = 20, a = 1$ and $b = 6$. Denote by $U(m)$ the Uniform distribution on $A = \{(x,y) \in \{1,\ldots,m\}^2\}$, given $p \in [0,1]$ consider now the next three mixture of distributions

iv. $M1(m,a) : pD1(m,a) + (1-p)U(m)$;
v. $M2(m,a) : pD2(m,a) + (1-p)U(m)$;
vi. $M3(m,a,b) : pD3(m,a,b) + (1-p)U(m)$.

where the notation $pD1(m,a) + (1-p)U(m)$ represents that the bivariate vector is given in proportion p by the distribution $D1(m,a)$ and in proportion $(1-p)$ by the distribution $U(m)$. Note that if $p = 1$ we recover the distributions $Di, i = 1,2,3$. From Tables 4–6, settings have projected that increase the number of categories of the discrete Uniform distribution $U(m)$, and also increase the parameters a and b.

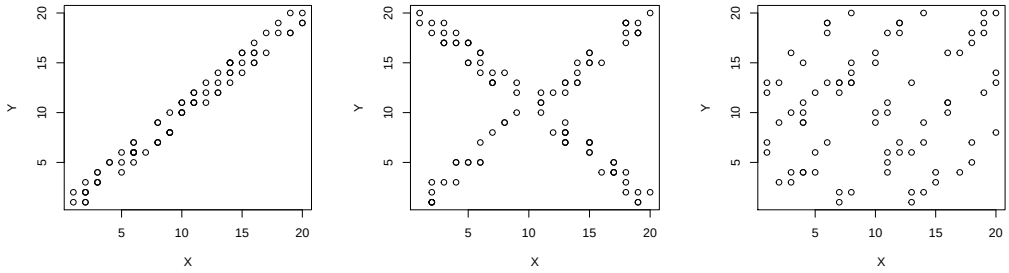

Figure 3. (**left**) $D1(20,1), n = 80$, (**middle**) $D2(20,1), n = 80$, (**right**) $D3(20,1,6), n = 80$.

Tables 4–6 show the rejection rates, obtained through 1000 simulations of samples of size n. We inspect the distributions $Di, i = 1,2,3$ and the mixtures $Mi, i = 1,2,3$ with $p = 0.8$. What is sought is to obtain high proportions evidencing the control of type 2 error.

Table 4. The proportion of p-value $\leq \alpha$ computed from Definition 4 and Equation (3), in 1000 simulations of size n. On top, results for $\alpha = 0.01$, on bottom results for $\alpha = 0.05$. $m = 20$, $a = 1$, $b = 6$.

	$p = 1.0$			$p = 0.8$		
n	$D1$	$D2$	$D3$	$M1$	$M2$	$M3$
20	1.000	0.349	0.028	0.994	0.179	0.018
40	1.000	0.798	0.050	1.000	0.568	0.034
$\alpha = 0.01$ 60	1.000	0.983	0.136	1.000	0.858	0.078
80	1.000	0.999	0.252	1.000	0.963	0.109
100	1.000	1.000	0.352	1.000	0.990	0.181
	$p = 1.0$			$p = 0.8$		
n	$D1$	$D2$	$D3$	$M1$	$M2$	$M3$
20	1.000	0.537	0.101	1.000	0.366	0.064
40	1.000	0.906	0.177	1.000	0.757	0.125
$\alpha = 0.05$ 60	1.000	0.993	0.306	1.000	0.934	0.192
80	1.000	1.000	0.468	1.000	0.985	0.250
100	1.000	1.000	0.601	1.000	0.997	0.368

Table 5. The proportion of p-value $\leq \alpha$ computed from Definition 4 and Equation (3), in 1000 simulations of size n. On top, results for $\alpha = 0.01$, on bottom results for $\alpha = 0.05$. $m = 50$, $a = 2$, $b = 12$.

	$p = 1.0$			$p = 0.8$		
n	$D1$	$D2$	$D3$	$M1$	$M2$	$M3$
20	1.000	0.431	0.044	0.993	0.229	0.024
40	1.000	0.902	0.172	1.000	0.719	0.079
$\alpha = 0.01$ 60	1.000	0.989	0.374	1.000	0.924	0.180
80	1.000	1.000	0.615	1.000	0.986	0.342
100	1.000	0.999	0.762	1.000	0.998	0.515
	$p = 1.0$			$p = 0.8$		
n	$D1$	$D2$	$D3$	$M1$	$M2$	$M3$
20	1.000	0.610	0.126	0.998	0.404	0.097
40	1.000	0.949	0.368	1.000	0.837	0.196
$\alpha = 0.05$ 60	1.000	0.997	0.608	1.000	0.963	0.370
80	1.000	1.000	0.823	1.000	0.993	0.571
100	1.000	1.000	0.918	1.000	0.999	0.711

Table 6. The proportion of p-value $\leq \alpha$ computed from Definition 4 and Equation (3), in 1000 simulations of size n. On top, results for $\alpha = 0.01$, on bottom results for $\alpha = 0.05$. $m = 100$, $a = 5$, $b = 30$.

	$p = 1.0$			$p = 0.8$		
	n	$D1$	$D2$	$D3$	$M1$	$M2$	$M3$
$\alpha = 0.01$	20	1.000	0.409	0.038	0.997	0.179	0.024
	40	1.000	0.865	0.137	1.000	0.623	0.063
	60	1.000	0.984	0.292	1.000	0.884	0.141
	80	1.000	0.999	0.473	1.000	0.969	0.247
	100	1.000	1.000	0.655	1.000	0.991	0.394
	$p = 1.0$			$p = 0.8$		
	n	$D1$	$D2$	$D3$	$M1$	$M2$	$M3$
$\alpha = 0.05$	20	1.000	0.597	0.110	1.000	0.345	0.090
	40	1.000	0.933	0.300	1.000	0.771	0.194
	60	1.000	0.996	0.520	1.000	0.949	0.326
	80	1.000	1.000	0.715	1.000	0.990	0.468
	100	1.000	1.000	0.848	1.000	0.998	0.620

As expected, for distribution $D1$ the procedure $JLND_n$ shows maximum performance, for all sample sizes and variants of m and a. For distribution $D2$, the performance of the procedure $JLND_n$ improves and reaches maximum performance as the sample size increases, for all variants of m and a. For distribution $D3$, we noticed a deterioration in the performance of the test when compared to the other two cases $D1$ and $D2$, despite this, the procedure responds adequately to the sample size, increasing its ability to detect dependence with increasing sample size.

Mi is a distribution that results from disturbing Di, so it makes sense to compare the effect of the disturbance, which in the illustrated cases is 20% from $U(m)$. For the distribution $M1$ the $JLND_n$-based procedure shows optimal performance, as occurs in the case $D1$. In cases $M2$ and $M3$, there is a deterioration in the performance of the procedure $JLND_n$ when compared to $D2$ and $D3$, respectively. Despite this, within the framework given by $M2$, we see that the good properties of the procedure are preserved when the sample size is increased.

In the following simulations, we investigate the dependence between discrete and continuous variables. The types explored are denoted by $D4$ and $D5$, Figure 4 illustrates the cases. Consider $m \in \mathbb{Z}_{>0}$,$a \in \mathbb{R}_{>0}$ and set

vii. $D4(m,a)$: Uniform distribution on $A = \{(x,y) \in \{1,\ldots,m\} \times [0,m+1] : |x-y| \leq a\}$;
viii. $D5(m,a)$: Uniform on $A = \{(x,y) \in \{1,\ldots,m\} \times [0,m+1] : |x-y| \leq a \text{ or } |x+y-m-1| \leq a\}$.

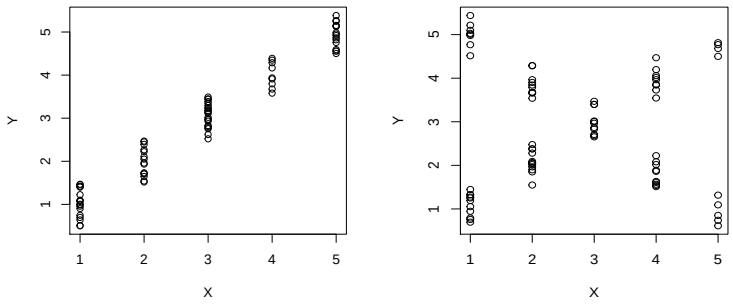

Figure 4. (**left**) $D4(5,0.5), n = 80$, (**right**) $D5(5,0.5), n = 80$.

Denote by $W(m)$ the Uniform distribution on $A = \{(x,y) \in \{1,\ldots,m\} \times [0,m+1]\}$ given $p \in [0,1]$ consider now the next two mixture of distributions

ix. $M4(m,a) : pD4(m,a) + (1-p)W(m)$;

x. $M5(m,a) : pD5(m,a) + (1-p)W(m)$.

Note that when using $p = 1$ in ix (or x) we recorver $D4$ (or $D5$).

Tables 7 and 8 show the performance of 1000 simulations of size n, from $M4(m,0.5)$ in Table 7, $M5(m,0.5)$ in Table 8. To the left of each Table (with $p = 1$), are simulated cases similar to the illustrated in Figure 4, $D4$ and $D5$. Table 7 shows that in the case of distribution $D4$, the procedure is very efficient and, we see that when the distribution is disturbed (by including 20% from W, to the right of Table 7) the procedure maintains its efficiency in detecting dependence. In relation to the distribution $D5$, we see from Table 8 that two effects occur, the one produced by the sample size n and the one produced by the value of m. By increasing n and m the procedure gains power quickly. The same effect is observed in the $M5$ distribution ($D5$ disturbance), with a certain deterioration in the power of the test.

Table 7. The proportion of p-value $\leq \alpha$ computed from Definition 4 and Equation (3), in 1000 simulations of size n. On top, results for $\alpha = 0.01$, on bottom results for $\alpha = 0.05$. Distribution $M4$ with $a = 0.5$.

		$p = 1.0$				$p = 0.8$			
	n	$m = 2$	$m = 3$	$m = 5$	$m = 10$	$m = 2$	$m = 3$	$m = 5$	$m = 10$
$\alpha = 0.01$	20	1.000	1.000	1.000	1.000	0.713	0.936	0.985	0.996
	40	1.000	1.000	1.000	1.000	0.993	1.000	1.000	1.000
	60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$p = 1.0$				$p = 0.8$			
	n	$m = 2$	$m = 3$	$m = 5$	$m = 10$	$m = 2$	$m = 3$	$m = 5$	$m = 10$
$\alpha = 0.05$	20	1.000	1.000	1.000	1.000	0.882	0.977	0.999	0.999
	40	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000
	60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	80	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 8. The proportion of p-value $\leq \alpha$ computed from Definition 4 and Equation (3), in 1000 simulations of size n. On top, results for $\alpha = 0.01$, on bottom results for $\alpha = 0.05$. Distribution $M5$ with $a = 0.5$.

		$p = 1.0$			$p = 0.8$		
	n	$m = 3$	$m = 5$	$m = 10$	$m = 3$	$m = 5$	$m = 10$
$\alpha = 0.01$	20	0.070	0.209	0.446	0.031	0.105	0.209
	40	0.211	0.634	0.926	0.118	0.374	0.708
	60	0.446	0.910	0.996	0.224	0.655	0.945
	80	0.611	0.977	1.000	0.364	0.852	0.994
	100	0.728	0.999	1.000	0.528	0.949	0.999
		$p = 1.0$			$p = 0.8$		
	n	$m = 3$	$m = 5$	$m = 10$	$m = 3$	$m = 5$	$m = 10$
$\alpha = 0.05$	20	0.161	0.385	0.638	0.112	0.239	0.388
	40	0.358	0.791	0.971	0.231	0.578	0.828
	60	0.612	0.970	0.999	0.384	0.810	0.973
	80	0.736	0.995	1.000	0.529	0.951	0.998
	100	0.850	0.999	1.000	0.671	0.981	1.000

The $JLND_n$ statistic is built in the graph of the paired ranks of the observations, and it is given by the size of the LNDSS found in this graph (see Figure 1). The proposal induces a region where this statistic can found evidence of dependence, in the diagonal of the graph. The simulation study points that the detection power of the procedure occurs in situations with an increasing pattern in the direction in which the $JLND_n$ statistic is built. Even more, the concomitant presence of increasing patterns and decreasing patterns does not necessarily nullify the detection capacity of the procedure, since the statistic $JLND_n$ is formulated considering the expanded $\mathcal{S}_n$ space provided with the uniform distribution. See Tables 4–8 in which we observe that by increasing the sample size, the detection capacity of $JLND_n$ is preserved. Also, looking at the right side of the tables already cited, we verify the robustness of the procedure, when inspecting cases with a concentration of points in the diagonals and suffering contamination, if the sample size grows.

In the next section, we apply the test to real data and compare our results with other procedures.

4. Applying the Test in Real Data

As it has already been commented, in some data sets, we have ties, produced by the precision used in data collection. This is the case of the *wine* data set (from the *glus* R-package), composed of 178 observations. For example, consider the cases (i) *Alcohol* vs. *Flavonoids* (see Figure 5, left) and (ii) *Flavanoids* vs. *Intensity* (see Figure 5, right). For each case (i) and (ii) both variables are continuous but recorded with a precision of two decimal places. We use known procedures in the area of continuous variables. For all the computations is used the R-project software environment. The "hoeffd" function in the "Hmisc" package is used to compute the p-value in the case of Hoeffding's test. The "cor.test" function in the "stat" package is used to compute the p-value for Pearson, Spearman and Kendall tests, see also [8]. Finally, we use the "indepTest" function, from the "copula" package to compute the "Copula" test.

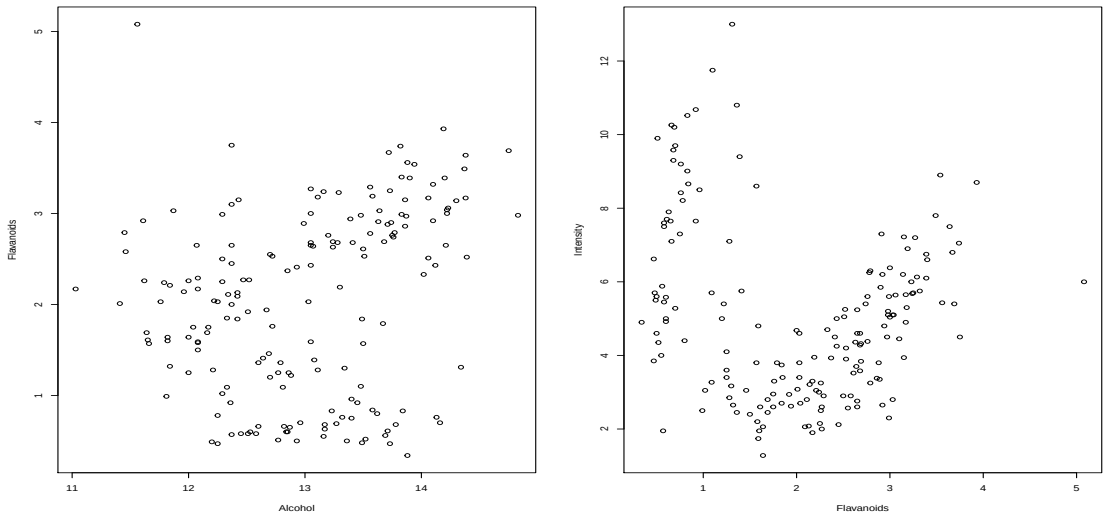

Figure 5. (**left**): *Alcohol* vs. *Flavanoids*. (**right**): *Flavanoids* vs. *Intensity*. Variables coming from *wine* data set from *gclus* R-package.

In case (i) of Figure 5 (left) all the procedures report p-value less than 0.02. Using $JLND_n$ ($jlnd_0 = 31.843$) we obtain p-value = 0.0160 and p-value = 0.0004, applying Equations (2) and (3), respectively. That is, $JLND_n$-based procedures detect dependence without the possible contraindications that the other procedures have, since we see ties in the dataset.

From the appearance of the scatter plot (Figure 5, right), it is understandable that the tests based on the Spearman and Kendall coefficients show difficulties in recording dependence, see Table 9. We also see that the other procedures capture the signs of dependence as well as the one proposed in this paper ($jlnd_0 = 29.904$). In both situations (cases (i) and (ii)) the only procedure, without contraindication, with significant p-value to reject H_0 is $JLND_n$.

Table 9. p-value of Copula's test, Hoeffding's test, $JLND_n$'s test ($B = 5000$); p-value and coefficient of Kendall's test, Pearson's test and Spearman's test. Case (ii) *Flavanoids* vs. *Intensity*.

	Copula	**Hoeffding**	**Spearman**	**Pearson**	**Kendall**	$JLND_n$ **Equation (2)**	$JLND_n$ **Equation (3)**
p-value	0.0005	0.0000	0.5695	0.0214	0.5713	0.0380	0.0044
coefficient			−0.0429	−0.1724	0.0287		

We inspect also the dependence between the variables *Duration*: duration of the eruption and *Interval*: time until following eruption, both measures in minutes, corresponding to 222 eruptions of the Old Faithful Geyser during August 1978 and August 1979. The data is coming from [9] and it is a traditional data set used in regression analysis with the aim of predicting the time of the next eruption using the duration of the most recent eruption (see [10]).

Figure 6 clearly shows the high number of ties, which compromises procedures designed for continuous variables. We have run the $JLND_n$ test ($jlnd_0 = 63.797$), using various values of B, B = 1000, 2000, 5000, 10,000. In all cases the p-value is less than 0.00001 and using both versions to estimate the cumulative distribution, Equations (2) and (3). Then the hypothesis of independence between *Duration* and *Interval* is rejected.

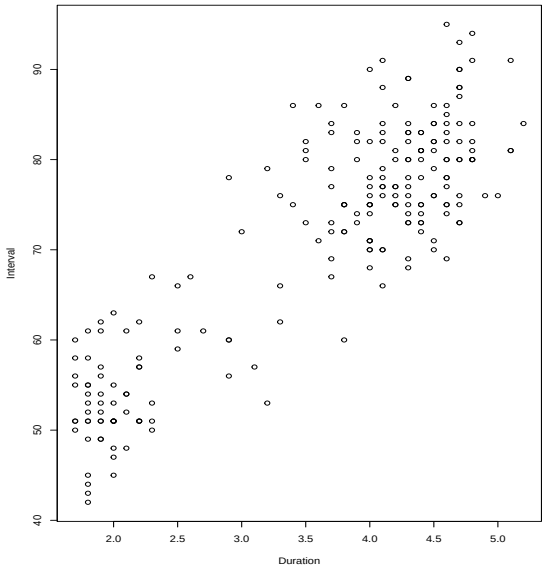

Figure 6. *Duration* vs. *Interval* of *geyser* data set [9].

The data set, *cdrate* is composed by 69 observations given in the 23 August 1989, issue of *Newsday*, it consists of the three-month certificate of deposit rates *Return on CD* for 69 Long Island banks and thrifts. The variables are *Return on CD* and *Type* = 0 (bank), 1 (thrift), source: [9]. Table 10 shows the data arranged

based on the values of the attribute *Return on CD* and divided into the two cases of the variable *Type*. That table shows sparseness, an issue reported in the literature, that compromises the performance of tests Pearson's Chi-squared based (Table 11), see [1].

Table 10. Data set *cdrate* from [9] organized by attributes *Return on CD* and *Type* = 0 (bank), 1 (thrift).

Return on CD	Type = 0	Type = 1	Return on CD	Type = 0	Type = 1	Return on CD	Type = 0	Type = 1
7.51	0	1	8.15	0	1	8.49	0	3
7.56	1	0	8.17	1	0	8.50	1	9
7.57	1	0	8.20	0	1	8.51	1	0
7.71	1	0	8.25	0	2	8.52	0	1
7.75	0	1	8.30	1	2	8.55	1	0
7.82	2	0	8.33	2	1	8.57	1	0
7.90	1	1	8.34	0	1	8.65	2	0
8.00	7	3	8.35	0	2	8.70	0	1
8.05	2	0	8.36	0	1	8.71	1	0
8.06	1	0	8.40	1	6	8.75	0	1
8.11	1	0	8.45	0	1	8.78	0	1

Table 11. Independence tests between *Return on CD* and *Type* of *cdrate* data set from [9]. In Equations (2) and (3), $B = 5000$.

Test	χ^2	$JLND_n$ (Equation (2))	$JLND_n$ (Equation (3))
p-value	0.0558	0.0320	0.0068
Statistic's value	45.6450 (df = 32)	50.275	50.275

In Table 11 we see the results for testing H_0. We see that according to $JLND_n$'s test we must reject H_0, which seems to be confirmed by Figure 7. Figure 7 comparatively shows the performance of variable *Return on CD*, for the two values of variable *Type*.

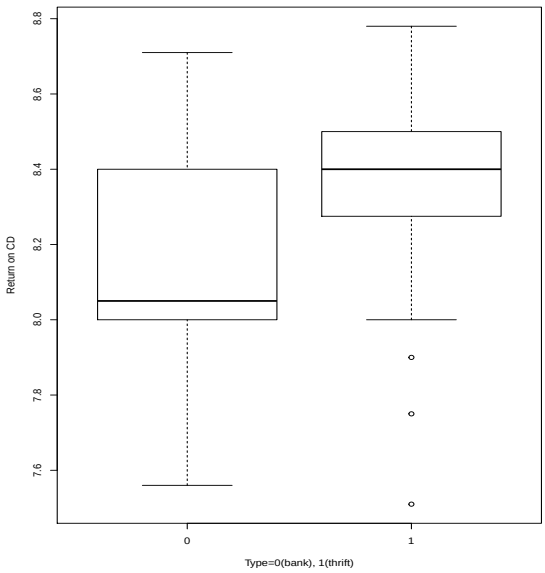

Figure 7. Boxplots of *Return on CD* by variable *Type*. See *cdrate* data set [9].

We conclude this section with a case of the *wine* data set, *Class* vs. *Alcohol*. Figure 8 shows the relationship in which we wish to verify whether independence can be rejected. *Class* registers 3 possible

values and *Alcohol* has been registered with low precision, which leads to observing ties. The observed value of $JLND_n$ is $jlnd_0 = 99.438$. The p-value given by the Equations (2) and (3) indicate the rejection of H_0. By the Equation (2) we obtain a p-value = 0.0004 and by the Equation (3) we obtain a p-value < 0.00001.

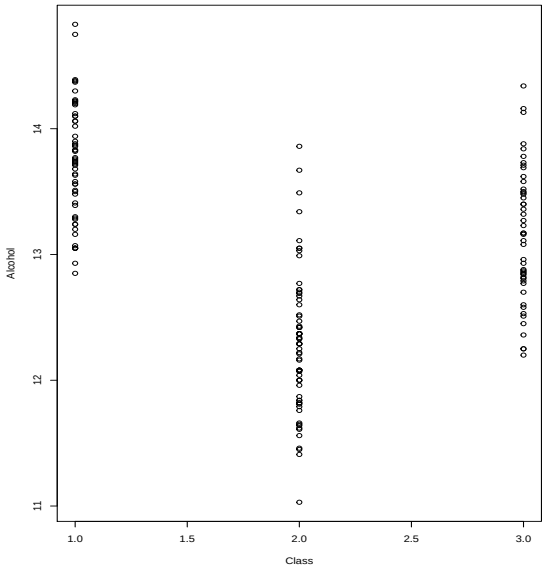

Figure 8. *Class* vs. *Alcohol*, see *wine* data set from *gclus* R-package.

We note that in the cases of Figure 5 we have verified that the test $JLND_n$ through Equation (3) offers lower p-value than the version given by Equation (2). In the cases of Figures 6 and 8, it may simply be an effect of computational precision. For the other cases, it is necessary to take into account that the Bootstrap version, by tending to create more ties, shows a tendency to underestimate the cumulative distribution, in other words, $\hat{F}^B_{JLND_n}(q) \leq F_{JLND_n}(q)$ where $F_{JLND_n}(\cdot)$ is the true cumulative distribution. Due to the increasing tendency shown by the cases addressed (see Figure 5), it is expected that the observed value of the statistic $JLND_n$, $jlnd_0$, in each case is positioned in the upper tail of the distribution, which leads to the p-value be given by $2(1 - \hat{F}^B_{JLND_n}(jlnd_0))$, see Definition 4. As a consequence $2(1 - \hat{F}^B_{JLND_n}(jlnd_0)) > 2(1 - F_{JLND_n}(jlnd_0))$. With the proposal made through Equation (3), we seek to correct the underestimation, since it does not favor the proliferation of ties. Which would explain the relationship between the p-value.

5. Concluding Remarks

In this article, we investigate the performance of the $JLND_n$ statistic to identify dependence on bivariate random vectors from a paired sample of size n. The procedure requires identifying the LNDSS that can be found on the graph between the marginal ranks of the paired observations, see Definitions 1 and 2. The goal is to compare the length of such subsequence (Definition 3) with the length of all possible subsequences, under the assumption of independence. This means, imposing an uniform distribution on the expanded $\mathcal{S}_n$ space. For the formulation of the procedure, it is required to estimate the distribution of the statistic $JLND_n$, under the assumption of independence and, in this paper it is given by Equation (3) (see also Definition 4). The estimation proposed in this paper shows an improved performance compared with the one given in [5], see Section 2.1. The concept, *longest non-decreasing subsequence*, allows us to build a tool without restrictions over the type of variable, continuous or discrete in which it can be applied.

From the simulation study we confirm that the detection power of the procedure occurs in situations with an increasing pattern from left to right and from bottom to top, which is the direction in which the $JLND_n$ statistic is sought (see Figure 1). The observations can be associated with continuous or discrete variables, not affecting the power of the test. The concomitant presence of increasing patterns and decreasing patterns does not necessarily nullify the detection capacity of the procedure if the size of the samples is big enough. We also verify the robustness of the procedure when inspecting cases that suffer contamination that could conceal the dependence. See Tables 4–8, we use different real data sets that expose the versatility of the procedure to reject independence in situations such as (a) in the presence of ties, (b) in the presence of sparseness, (c) in mixed situations.

Author Contributions: Conceptualization, J.E.G. and V.A.G.-L.; methodology, J.E.G. and V.A.G.-L.; software, J.E.G. and V.A.G.-L.; validation, J.E.G. and V.A.G.-L.; formal analysis, J.E.G. and V.A.G.-L.; investigation, J.E.G. and V.A.G.-L.; data curation, J.E.G. and V.A.G.-L.; writing–review and editing, J.E.G. and V.A.G.-L. Both authors have read and agreed to the published version of the manuscript.

Funding: No funds were received for the execution of this work.

Acknowledgments: The authors wish to thank the two referees for their many helpful comments and suggestions on an earlier draft of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Agresti, A. *Categorical Data Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2002.
2. Kendall, M.G. The treatment of ties in ranking problems. *Biometrika* **1945**, *33*, 239–251. [CrossRef] [PubMed]
3. Cramir, H. *Mathematical Methods of Statistics*; Princeton U. Press: Princeton, NJ, USA, 1946; Volume 500.
4. Romik, D. *The Surprising Mathematics of Longest Increasing Subsequences*; Cambridge University Press: New York, NY, USA, 2015; Volume 4.
5. García, J.E.; González-López, V.A. Independence test for sparse data. *AIP Conf. Proc.* **2016**, *1738*, 140002.
6. García, J.E.; González-López, V.A. Independence tests for continuous random variables based on the longest increasing subsequence. *J. Multivar. Anal.* **2014**, *127*, 126–146. [CrossRef]
7. Zelterman, D. Goodness-of-fit tests for large sparse multinomial distributions. *J. Am. Stat. Assoc.* **1987**, *82*, 624–629. [CrossRef]
8. Hollander, M.; Wolfe, D. *Nonparametric Statistical Methods*; John Wiley & Sons: New York, NY, USA, 1973; pp. 185–194.
9. Simonoff, J.S. *Smoothing Methods in Statistics*; Springer: New York, NY, USA, 1996.
10. Weisberg, S. *Applied Linear Regression*, 4th ed.; John Wiley & Sons: Minneapolis, MN, USA, 2005; Volume 528.