Independence Test for Continuous or Discrete Random Variables

Jesús E. García and Verónica A. González-López

June 27, 2018

The proposal

- Given a sample of (X, Y), the test is based on the **size of the longest increasing subsequence** of the permutation which maps the ranks of the X observations to the ranks of the Y observations.

The proposal

- Given a sample of (X, Y), the test is based on the **size of the longest increasing subsequence** of the permutation which maps the ranks of the X observations to the ranks of the Y observations.
- We identify the independence assumption between X and Y with the space of permutation equipped with the uniform distribution and we show the exact distribution of the size of the longest increasing subsequence.

Defining the statistic

Example

- sample $\{(x_i, y_i)\}_{i=1}^{n}$: :
  $\{(4.1, 3.2), (1.1, 3.5), (2.51, 4.17), (3.61, 3.18), (1.8, 2.86)\}$,
- sort in increasing order in relation to the sample $\{x_i\}_{i=1}^{n}$
  $\{(1.1, 3.5), (1.8, 2.86), (2.51, 4.17), (3.61, 3.18), (4.1, 3.2)\}$,
- replace the x_i value with its rank in the sequence:
  $\{(1, 3.5), (2, 2.86), (3, 4.17), (4, 3.18), (5, 3.2)\}$,
- replace each y_i with its rank in the $\{y_i\}_{i=1}^{n}$:
  $\{(1, 4), (2, 1), (3, 5), (4, 2), (5, 3)\}$.
- 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 statistic in a graphic

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

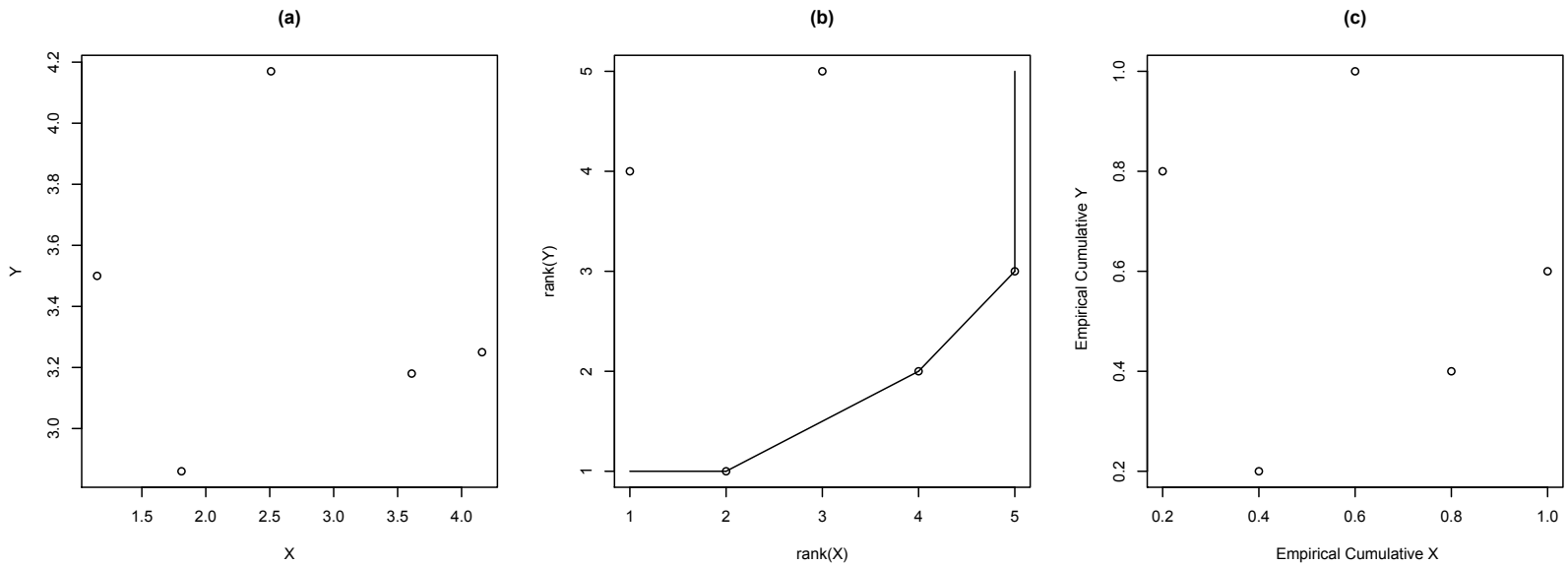


Figure: Dispersion's graphic and permutation. (a) is the dispersion plot for the sample, **(b) represents the permutation defined by the sample, the solid line shows the longest increasing subsequence**, (c) shows the empirical copula of the sample.

The size of the longest increasing subsequence is 3.

Assumption

Assumption 1:

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

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

- $(x_1, y_1), \cdots, (x_n, y_n)$ is a paired sample of size n of (X, Y).

Assumption

Assumption 1:

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

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

- $(x_1, y_1), \cdots, (x_n, y_n)$ is a paired sample of size n of (X, Y).
- $H_0 : X$ and Y are independent

Definition

Let $\mathcal{S}_n$ denote the group of permutations of $\{1, \cdots, n\}$. If $\pi \in \mathcal{S}_n$ we call $l_n(\pi)$ the length of the longest increasing subsequence of π.

Definition

Let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be replications of (X, Y) with continuous marginal distributions, we denote by L_n the random variable, $L_n = l_n(\pi_{\mathcal{D}})$ where $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ and $\pi_{\mathcal{D}}$ is the permutation which assigns $\pi(rank(X_i)) = rank(Y_i)$, $i = 1, \ldots, n$.

- In the example, the value of the statistic was $l_5(\pi_{\mathcal{D}}) = 3$.

Exact distribution

Theorem

Let (X, Y) be under Assumption 1. If $(S_n, \mathcal{U})$ is the space $\mathcal{S}_n$ with uniform distribution $\mathcal{U}$. Then, for $k = 1, 2 \ldots, n$, if $p_k^n = Prob(L_n = k)$,

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

where $V_n(k, m)$ is the set of shapes of standard Young tableaux of order n, having k columns and m rows and $N(W)$ is the number of standard Young tableaux with shape W, computed by Frame et al. (1954).

Asymptotic distribution

Under the same conditions of previous theorem, Baik et al. (1999) prove that if χ is a random variable with Tracy Widom distribution, then, when $n \to \infty$,

$$\chi_n = \frac{L_n - 2\sqrt{n}}{n^{1/6}} \to \chi \text{ in distribution.} \tag{2}$$

Defining JL_n statistic

Consider the set

$$\mathcal{D}^{diag} = \left\{ (U_i, V_i),\ i = 1, \ldots, n \ :\ |U_i - V_i| \leq cn^{5/6} \right\},$$

define $L_n^{diag} = l_{n^{diag}}(\pi_{\mathcal{D}^{diag}})$, with $n^{diag} = \#\left(\mathcal{D}^{diag}\right)$, $U_i = rank(X_i)$ and $V_i = rank(Y_i)$, for $i = 1, \ldots, n$.

- See Johansson (2000): typical deviations of a maximal path from the diagonal $U = V$ are of order $n^{5/6}$.

Definition

Let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be replications of (X, Y) with continuous marginal distributions,

$$JL_n := \frac{1}{n^{diag}} \sum_{(u,v) \in \mathcal{D}^{diag}} L_n^{diag}(u, v),$$

where $L_n^{diag}(u, v) = l_{n^{diag}-1}(\pi_{\mathcal{D}^{(u,v)}})$ with $\mathcal{D}^{(u,v)} = \mathcal{D}^{diag} \setminus \{(u, v)\}$, for each $(u, v) \in \mathcal{D}^{diag}$.

Defining JLM_n statistic

Definition

Let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be replications of (X, Y) with continuous marginal distributions,

$$JLM_n := \max\{JL_n, JL_n^-\},$$

with JL_n and JL_n^- given by the previous Definition applied over $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ and $(X_1, -Y_1), (X_2, -Y_2), ..., (X_n, -Y_n)$, respectively.

Simulation study

- For each joint distribution and sample sizes $20, 40, 60, 80, 100$ we simulated 5000 samples, and computed the p-values.

Simulation study

- For each joint distribution and sample sizes $20, 40, 60, 80, 100$ we simulated 5000 samples, and computed the p-values.
- Denote by $\left\{(X_i^j, Y_i^j)\right\}_{i=1}^{n}$ the j simulated sample of size $n = 20, 40, 60, 80, 100$. Where $j = 1, \cdots, 5000$.

Simulation study

- For each joint distribution and sample sizes $20, 40, 60, 80, 100$ we simulated 5000 samples, and computed the p-values.
- Denote by $\left\{(X_i^j, Y_i^j)\right\}_{i=1}^{n}$ the j simulated sample of size $n = 20, 40, 60, 80, 100$. Where $j = 1, \cdots, 5000$.
- Given a level α we calculate the empirical power as being, $\frac{\#\left\{j\text{:p-value}\left(\left\{(X_i^j, Y_i^j)\right\}_{i=1}^{n}\right) \leq \alpha\right\}}{5000}$ where p-value$\left(\left\{(X_i^j, Y_i^j)\right\}_{i=1}^{n}\right)$ denotes the p-value associated with the sample j, $\left\{(X_i^j, Y_i^j)\right\}_{i=1}^{n}$ to test the hypothesis.

We consider two main situations in which the samples show low correlation.

- (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 types of dependence with null expected correlation coefficients.

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 ρ and external radius of 1.

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

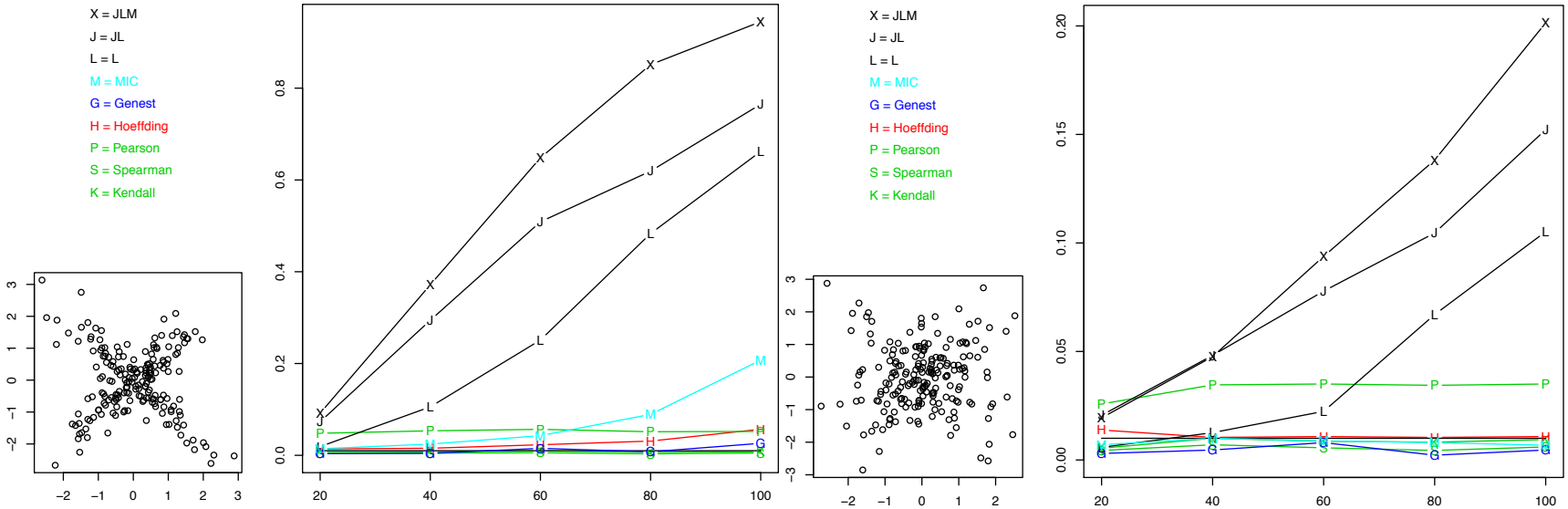


Figure: The left panel is the scatter plot of a sample size 200 of D1. The right panel is sample size vs. empirical significance level (α=0.01) for the same distribution. Left $\rho = 0.9$, Right $\rho = 0.7$.

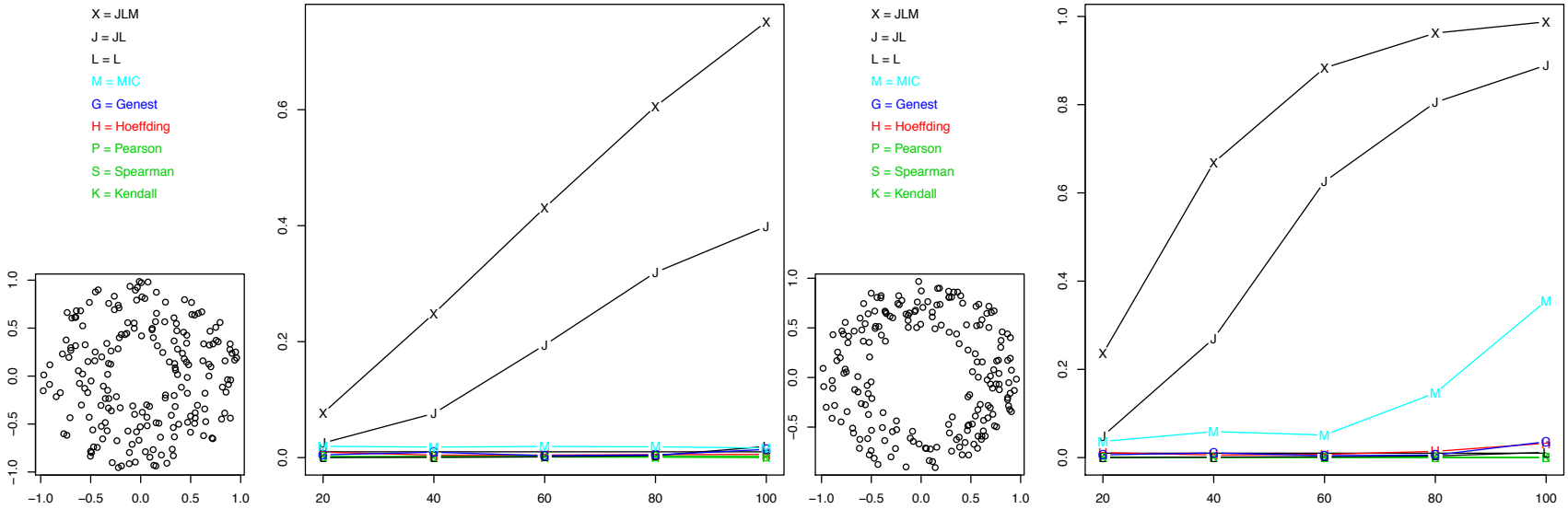


Figure: The left panel is the scatter plot of a sample size 200 of D2. The right panel is sample size vs. empirical significance level (α=0.01) for the same distribution. Left $\rho = 0.3$, Right $\rho = 0.5$.

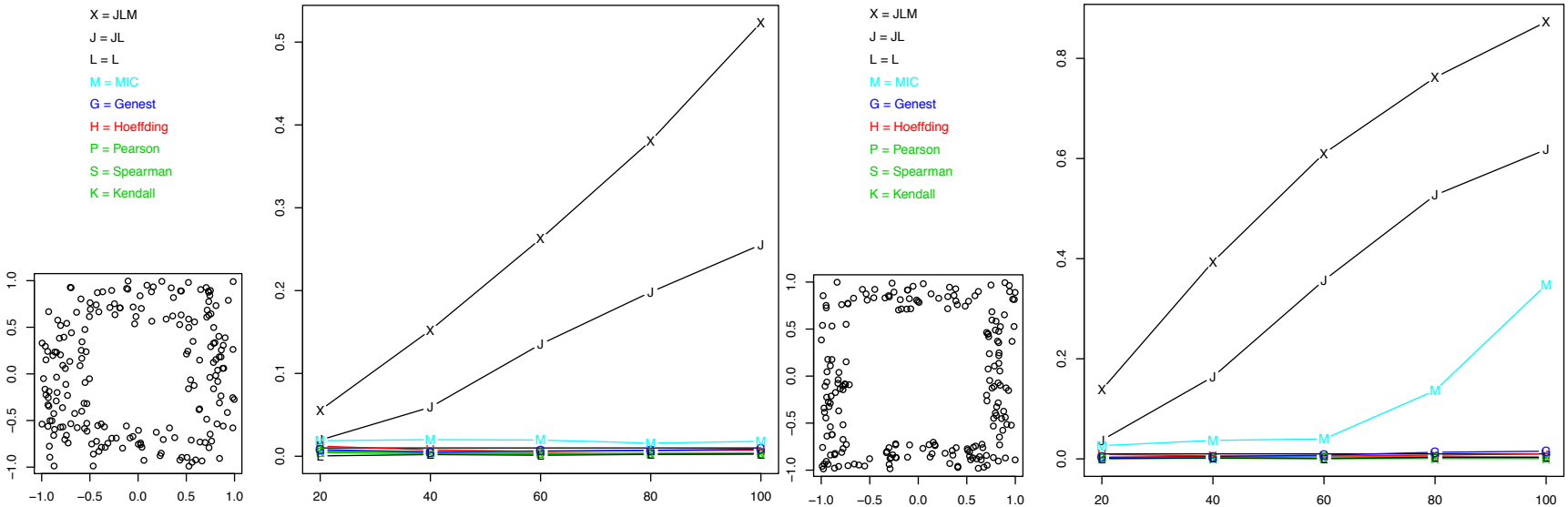


Figure: The left panel is the scatter plot of a sample size 200 of D3. The right panel is sample size vs. empirical significance level (α=0.01) for the same distribution. Left $\rho = 0.5$, Right $\rho = 0.7$.

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 $\underline{0} = (0, 0)$, $16I = \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

D5- Mixture of two bivariate Normal distributions, $\gamma\%$ independent with standard deviation 4 and $(1-\gamma)\%$ dependent with standard deviation 0.5 and correlation $\rho = 0.95$,
$(X, Y) \sim \gamma N_2(\underline{0}, 16I) + (1-\gamma) N_2(\underline{0}, \Sigma)$, 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;
$(X, Y) \sim 0.75 C_C(\cdot, \cdot | -0.1) + 0.25 C_C(\cdot, \cdot | 10)$.

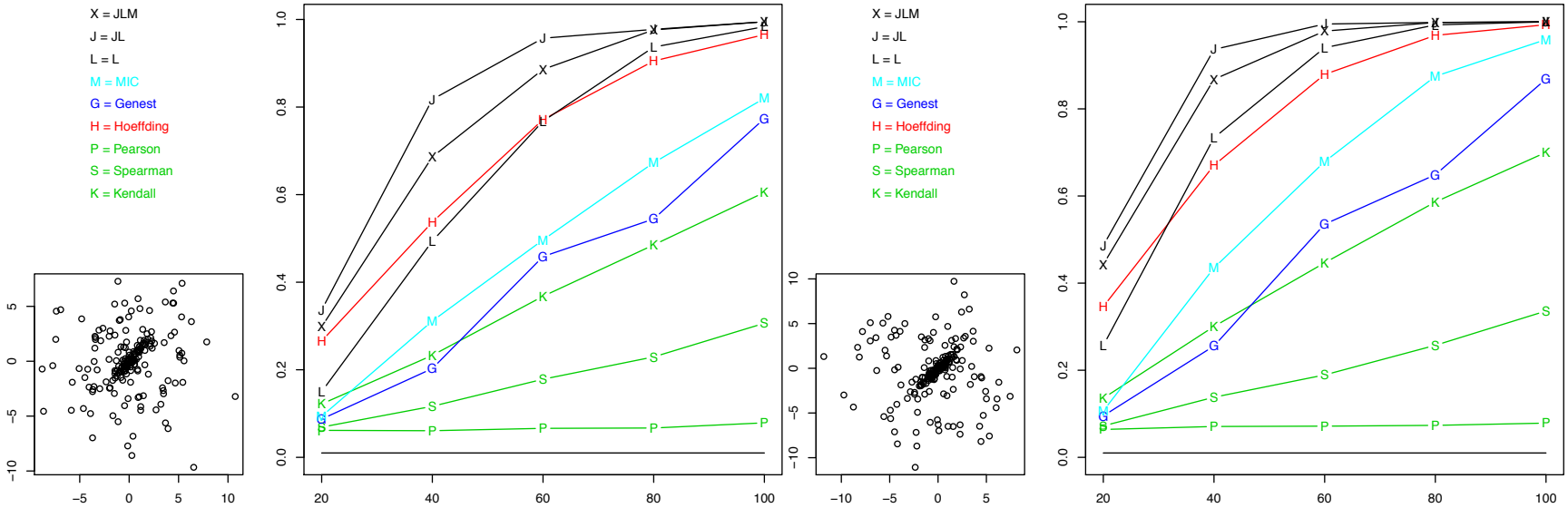


Figure: The left panel is the scatter plot of a sample size 200 of D4. The right panel is sample size vs. empirical significance level (α=0.01) for the same distribution. Left $\rho = 0.9$, Right $\rho = 0.95$.

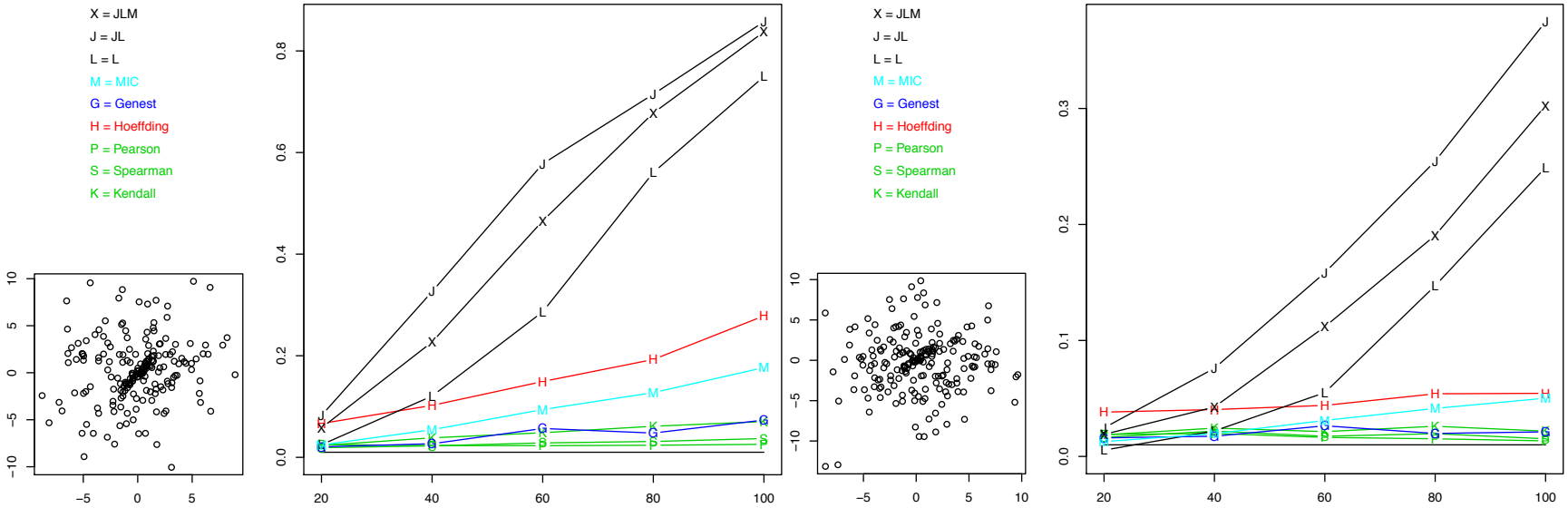


Figure: The left panel is the scatter plot of a sample size 200 of D5. The right panel is sample size vs. empirical significance level (α=0.01) for the same distribution. Left $\gamma = 0.75$, Right $\gamma = 0.85$.

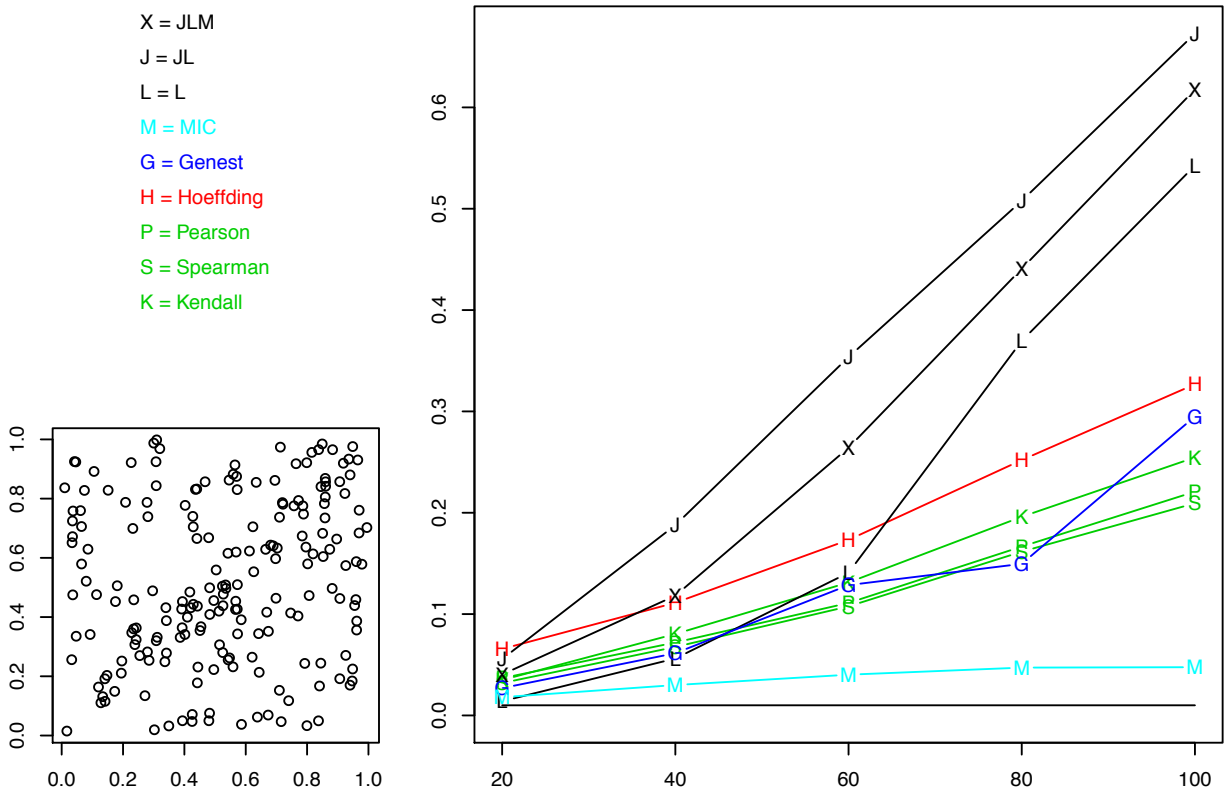


Figure: The left panel is the scatter plot of a sample size 200 of D6. The right panel is sample size vs. empirical significance level (α=0.01) for the same distribution.

Not continuous case

- The statistic is the size of the longest nondecreasing subsequence (LNDS) among the ranks of the observations.

Not continuous case

- The statistic is the size of the longest nondecreasing subsequence (LNDS) among the ranks of the observations.
- The distribution of the statistic depend of the distribution of the marginals.

Not continuous case

- The statistic is the size of the longest nondecreasing subsequence (LNDS) among the ranks of the observations.
- The distribution of the statistic depend of the distribution of the marginals.
- We will use bootstrap to simulate the distribution of statistic for the test.

Horseshoe Crabs from Agresti (2002)

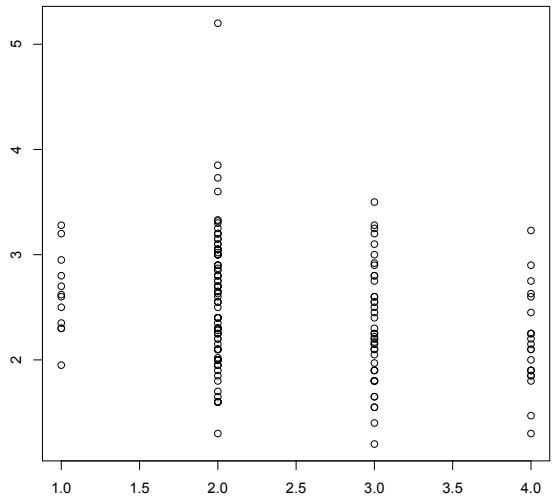


Figure: Color (C) versus Weight (Wt), explanatory variables in Horseshoe Crabs data set. Color: light medium (code=1), medium (code=2), dark medium (code=3), and dark (code=4), Weight: in Kg.

The data set is composed by several attributes of 173 female crabs.

Vineyard from Simonoff (1996)

Grape yields for each of 52 rows of a vineyard on an island in Lake Erie, for harvests in 1989, 1990 and 1991. The yield is measured as a number of lugs, where a lug is a basket used to carry the grapes.

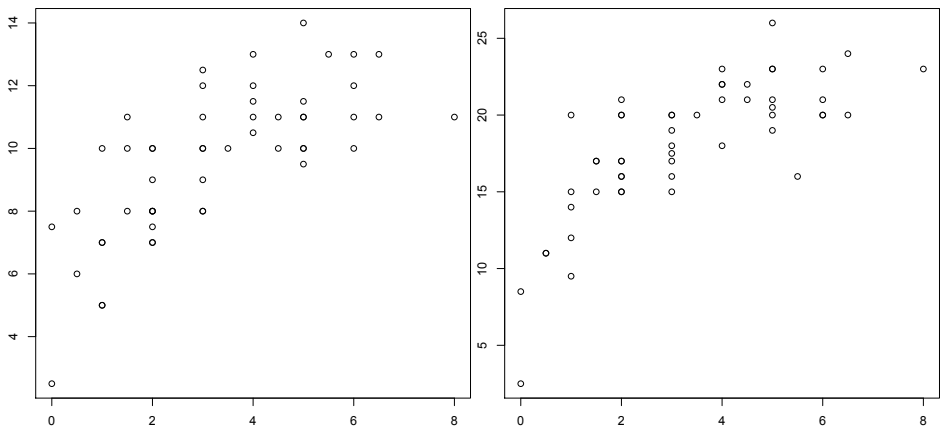


Figure: Left: Lugs in 1989 versus Lugs in 1990. Right: Lugs in 1989 versus Lugs in 1991.

The Statistical Procedure

Definition

i. Let $q_1, \ldots, q_n$ be a sequence in $\mathbb{R}$, $q_{i_1}, \ldots, q_{i_k}$ is a non-decreasing subsequence of $q_1, \ldots, q_n$ if $1 \leq i_1 < \ldots < i_k \leq n$ and $q_{i_1} \leq q_{i_2} \leq \ldots \leq q_{i_k}$.

ii. Given a size n sequence $Q = \{q_1, \ldots, q_n\}$, we call as $Ind_n(Q)$ the length of the longest non-decreasing subsequence of Q.

The Statistical Procedure

Definition

Consider (X, Y) a random vector with distribution function H. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be replications of (X, Y),

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

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

Example

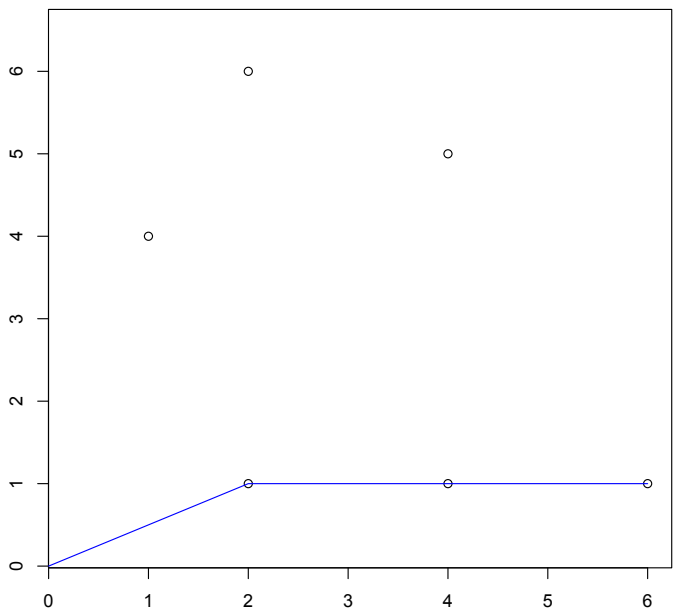
We show the graphical construction of LND_n.

x_i	y_i	rank(x_i)	rank(y_i)
1.2	2.0	1	4
1.4	1.5	2	1
1.4	2.7	2	6
1.5	2.3	4	5
1.5	1.5	4	1
2	1.5	6	1

This data defines a $Q_{\mathcal{D}} = \{4, 1, 6, 5, 1, 1\}$. The maximal subsequence non-decreasing is $\{1, 1, 1\}$

Example

given by the trajectory $(0,0) - (2,1) - (4,1) - (6,1)$ from the empirical copula of the data set, displayed in figure 28. This means that the variable LND_6 has value equal to 3 in this data set.



Jackknife version of the test

Definition

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be replications of (X, Y). We define

$$JLND_n = \frac{1}{n} \sum_{(u,v)\in\mathcal{D}} LND_n(u, v),$$

where $LND_n(u, v) = Ind_{n-1}(Q_{\mathcal{D}^{(u,v)}})$ with $\mathcal{D}^{(u,v)} = \mathcal{D} \setminus \{(u, v)\}$, for each $(u, v) \in \mathcal{D}$.

p-value

Definition

The estimated two-sided p-value for the statistical test with null hypothesis of independence against an alternative hypothesis of not independence between X and Y is defined by,

$$\min\left\{1, \max\left\{2\hat{F}_{JLND_n}(jlnd_0), 2(1-\hat{F}_{JLND_n}(jlnd_0))\right\}\right\},$$

where $jlnd_0$ is the observed value of $JLND_n$ in the sample and $\hat{F}_{JLND_n}$ is the bootstrap estimation of the cumulative distribution function of $JLND_n$, under the null hypothesis.

Bootstrap procedure to estimate $\hat{F}_{JLND_n}$

- B size n resamples of $X_1, X_2, \ldots, X_n$.

Bootstrap procedure to estimate $\hat{F}_{JLND_n}$

- B size n resamples of $X_1, X_2, \ldots, X_n$.
- Generating $X^b = (X_1^b, X_2^b, \ldots, X_n^b)$ for $b = 1, 2, \ldots, B$.

Bootstrap procedure to estimate $\hat{F}_{JLND_n}$

- B size n resamples of $X_1, X_2, \ldots, X_n$.
- Generating $X^b = (X_1^b, X_2^b, \ldots, X_n^b)$ for $b = 1, 2, \ldots, B$.
- The same procedure is performed with $Y_1, Y_2, \ldots, Y_n$, obtaining $Y^b = (Y_1^b, Y_2^b, \ldots, Y_n^b)$ for $b = 1, 2, \ldots, B$.

Bootstrap procedure to estimate $\hat{F}_{JLND_n}$

- B size n resamples of $X_1, X_2, \ldots, X_n$.
- Generating $X^b = (X_1^b, X_2^b, \ldots, X_n^b)$ for $b = 1, 2, \ldots, B$.
- The same procedure is performed with $Y_1, Y_2, \ldots, Y_n$, obtaining $Y^b = (Y_1^b, Y_2^b, \ldots, Y_n^b)$ for $b = 1, 2, \ldots, B$.
- Then, for each b define $\mathcal{D}^b = \left\{(X_i^b, Y_i^b)\right\}_{i=1}^n$ and from that sample compute $JLND_n^b$.

Bootstrap procedure to estimate $\hat{F}_{JLND_n}$

- B size n resamples of $X_1, X_2, \ldots, X_n$.
- Generating $X^b = (X_1^b, X_2^b, \ldots, X_n^b)$ for $b = 1, 2, \ldots, B$.
- The same procedure is performed with $Y_1, Y_2, \ldots, Y_n$, obtaining $Y^b = (Y_1^b, Y_2^b, \ldots, Y_n^b)$ for $b = 1, 2, \ldots, B$.
- Then, for each b define $\mathcal{D}^b = \left\{(X_i^b, Y_i^b)\right\}_{i=1}^n$ and from that sample compute $JLND_n^b$.
- $$\hat{F}_{JLND_n}(q) = \frac{\#\{b \,:\, JLND_n^b \leq q\}}{B}.$$

Bootstrap procedure to estimate $\hat{F}_{JLND_n}$

- B size n resamples of $X_1, X_2, \ldots, X_n$.
- Generating $X^b = (X_1^b, X_2^b, \ldots, X_n^b)$ for $b = 1, 2, \ldots, B$.
- The same procedure is performed with $Y_1, Y_2, \ldots, Y_n$, obtaining $Y^b = (Y_1^b, Y_2^b, \ldots, Y_n^b)$ for $b = 1, 2, \ldots, B$.
- Then, for each b define $\mathcal{D}^b = \left\{(X_i^b, Y_i^b)\right\}_{i=1}^n$ and from that sample compute $JLND_n^b$.
- $$\hat{F}_{JLND_n}(q) = \frac{\#\{b \,:\, JLND_n^b \leq q\}}{B}.$$
- We also estimate the expected value of $JLND_n$ under the hypothesis of independence,

$$\hat{E}(JLND_n) = \frac{\sum_{b=1}^{B} JLND_n^b}{B}.$$

Applications to the Vineyard example

We apply the new test to two pairs of *lugs*: years 1989-1990 and 1989-1991, also we show to compare, the results obtained from the Pearson's Chi-squared test.

Table: Pairs of *lugs* from vineyard dataset

Data	lugs 89 vs lugs 90	lugs 89 vs lugs 91
p-value (JLND)	**0.002**	**0.002**
Statistic (JLND)	29.54 (indep=21.735)	29.42(indep=20.682)
p-value (χ^2)	0.5798	0.03932
Statistic (χ^2)	190.3922 (df=195)	273.4034 (df = 234)

Applications to the Crabs example

Table: Color (C) versus weight (Wt).

p-value (JLND)	**0.014**
Statistic (JLND)	107.38 (indep=121.404)
p-value (χ^2)	0.1913
Statistic (χ^2)	186.9582 (df = 171)

THANK YOU

Bibliography


Agresti, A. (2002). Categorical data analysis. John Wiley & Sons.


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.


García, J. E., & González-López, V. A. (2014). Independence tests for continuous random variables based on the longest increasing subsequence. *Journal of Multivariate Analysis.* 127, 126-146.


García, J. E., & González-López, V. A., Nelsen, R. B. (2016) The Structure of the Class of Maximum Tsallis-Havrda-Chavt Entropy Copulas. *Entropy,* 18(7), 264.


García, J. E., & González-López, V. A. (2016) Independence Test for Sparse Data. *AIP Conference Proceedings.* 13th international conference of numerical analysis and applied mathematics, v. 1738, 140002


García, J. E., & González-López, V. A., Simis, M., Terranova, T., Battistella, L.R. (2016) A New Test in the Longest Increasing Subsequence Family of Independence Tests. *Stochastic Modelling and Applications,* 20(1), 35-53.


Genest, C. and Remillard, B. (2004). Tests of Independence or Randomness Based on the Empirical Copula Process. *Test* **13** 335-369.


Johansson, Kurt. (2000). Transversal fluctuations for increasing subsequences on the plane. *Probability theory and related fields* 116(4), 445-456.


Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M. and Sabeti, P. C. (2011). Detecting Novel Associations in Large Data Sets. *Science* 334, 1518-1524.


Schensted, C. (1961). Longest increasing and decreasing sub-sequeces. *Canad. J. Math.* **13** 179-191.