An independence test based on recurrence rates. An empirical study and applications to real data

Juan Kalemkerian\textsuperscript{a} and Diego Fernández\textsuperscript{b}

\textsuperscript{a}Centro de Matemática, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay; \textsuperscript{b}Facultad de Ciencias Económicas y Administración, Universidad de la República, Montevideo, Uruguay

\section*{ABSTRACT}
In this article we propose several variants to perform the independence test between two random elements based on recurrence rates. We will show how to calculate the test statistic in each one of these cases. From simulations we obtain that in high dimension, our test clearly outperforms, in almost all cases, the other widely used competitors. The test was performed on two data sets including small and large sample sizes and we show that in both cases the application of the test allows us to obtain interesting conclusions.

\section*{1. Introduction}
Let \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\) be an i.i.d. sample of \((X, Y)\), \(X \in S_X\) and \(Y \in S_Y\), where \(S_X\) and \(S_Y\) are metric spaces. Consider the hypothesis \(H_0\) which asserts that \(X\) and \(Y\) are independent random elements: this is the so called independence test. In Kalemkerian and Fernández (2020) a quick historical enumeration (although not complete) is made on the independence tests developed from (Galton 1888) to the present day. Not many independence test have been developed for random elements where \(X\) and \(Y\) lie in an arbitrary metric space. Recently, in Kalemkerian and Fernández (2020), an independence test based on recurrence rates was proposed. This test is based on the following simple idea: if \(X\) and \(Y\) are independent, then \(d_X(X_1, X_2)\) and \(d_Y(Y_1, Y_2)\) are independent for all i.i.d. \((X_1, Y_1), (X_2, Y_2)\) with the same distribution as \((X, Y)\), where \(d_X\) and \(d_Y\) are the distance functions between elements of \(S_X\) and \(S_Y\), respectively. The authors proposed working with an \(L^2\)-Cramér–von Mises functional. On the one hand, from the theoretical point of view, this test is interesting because it does not need any assumptions on the topological structure of the metric spaces (e.g. assuming that they are Banach spaces), and in Kalemkerian and Fernández (2020) can be found the first study where the mathematical properties of the recurrence rates were established. Marwan (Marwan 2008) gives a historical review of recurrence plots techniques, together with everything developed from them. However, the potential of these techniques has not yet been studied in depth from the point of view of mathematical statistics. On the other hand, from the practical point of view, this test is interesting because it can be used for \(X\) and \(Y\) lying in spaces of any dimension, and in Kalemkerian and Fernández (2020) a power comparison shows the very good performance of this test compared to others widely used for random variables and random vectors.
In the present article we will show, using a power comparison, that the independence test of
recurrence rates in high dimension outperforms the other competing tests in almost all cases, and
we will study the incidence of the distance functions considered \((d_X \text{ and } d_Y)\) in the performance
of the test. As expected, we will show that the test statistic in high dimension has some sensitivity
to the choice of the distance function, \(d_X\) or \(d_Y\). Also, we will propose and compare other func-
tionals to be taken into account, such as an \(L^1\)-Cramér–von Mises functional and a
Kolmogorov–Smirnov functional, and we will show how to compute the statistic in each case.
Lastly, we will present applications of the test to two real data sets. In one of them (subsection
6.2) we will use the test to detect regions where the temperature are dependent or independent in
a certain area of South America. The same for wind velocity.

The rest of this article is organized as follows. In Sec. 2 we define the procedure to test \(H_0\) vs
\(H_1\) proposed in Kalemkerian and Fernández (2020) and we present the definition of the test
statistics based on a functional of the \(L^1\)-Cramér–von Mises type, and the statistics based on a
functional of the Kolmogorov–Smirnov type, and three different distances to use in \(S_X\) and \(S_Y\). In Sec.
3, we show how to compute the different statistics presented in Sec. 2. In Sec. 4 we present a
simulation study that compares, under several alternatives, the powers of the different tests of
recurrence rates when varying the test statistics and the distance functions. In Sec. 5, we compare
the performance of the recurrence test of independence with others in high dimension and we
show that our test clearly outperforms the rest in almost all cases. In Sec. 6, we present two appli-
cations of the recurrence test of independence to meteorological data, one of them with small
sample size and the second on involving a huge data set, and we show the ability of the recur-
cence rate test to obtain interesting conclusions. Some concluding remarks are given in Sec. 7
and the proof of the validity of the formulas established in Sec. 3 can be found in Sec. 8. Finally,
in Sec. 9 we give the R-code used in real data set and simulations.

2. Test approach and different statistics to consider

Let \((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\) be i.i.d. samples of \((X, Y)\) where \(X \in S_X, Y \in S_Y, S_X\) and \(S_Y\) are
metric spaces. To simplify the notation and without risk of confusion, we will use the same letter
\(d\) for the distance function on both metric spaces, \(S_X\) and \(S_Y\).

Given the recurrence thresholds \(r, s > 0\), we define the recurrence rate for the sample of \(X\)
and \(Y\) as

\[
RR^X_n(r) := \frac{1}{n^2 - n} \sum_{i \neq j} 1\{d(X_i, X_j) < r\}, \quad RR^Y_n(s) := \frac{1}{n^2 - n} \sum_{i \neq j} 1\{d(Y_i, Y_j) < s\},
\]

respectively, and the joint recurrence rate for \((X, Y)\) as

\[
RR^X_{n,Y}(r, s) := \frac{1}{n^2 - n} \sum_{i \neq j} 1\{d(X_i, X_j) < r, \ d(Y_i, Y_j) < s\}.
\]

Recall that \(1\{d(X_i, X_j) < r\}\) means 1 when \(d(X_i, X_j) < r\) and 0 when \(d(X_i, X_j) \geq r\) (indicator
function), and analogously in the other cases. If we define \(p_X(r) := P(d(X_1, X_2) < r)\) the probability
that the distance between any two elements of the sample \(X\) is less than \(r\) and \(p_{X,Y}(r, s) :=
P(d(X_1, X_2) < r, \ d(Y_1, Y_2) < s)\) the joint probability that the distance between any two elements
of the sample \(X\) is less than \(r\) and any two elements of the sample \(Y\) is less than \(s\), the strong law
of large numbers for \(U\)-statistics (Hoeffding 1961) allows us to affirm that for any \(r, s > 0,\)
\begin{equation}
RR^X_{n,Y}(r, s) \overset{a.s.}{\rightarrow} p_{X,Y}(r, s) \quad \text{and} \quad RR^X_n(r) \overset{a.s.}{\rightarrow} p_X(r) \quad \text{and} \quad RR^Y_n(s) \overset{a.s.}{\rightarrow} p_Y(s).
\end{equation}

We want to test \(H_0 : X \text{ and } Y \text{ are independent},\) against \(H_1 : H_0 \text{ does not hold. If } H_0 \text{ is true, then}\)
\(p_{X,Y}(r, s) = p_X(r)p_Y(s)\) for all \(r, s > 0,\) and we expect that if \(n\) is large, \(RR^X_{n,Y}(r, s) \approx RR^X_n(r)RR^Y_n(s)\)
for any $r, s > 0$. In Kalemkerian and Fernández (2020) it is proposed to reject $H_0$ when $T_n > c$, where

$$T_n := n \int_0^{+\infty} \int_0^{+\infty} \left( RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s) \right)^2 dG(r,s),$$

(2)

where $c$ is a constant and $G$ is a properly chosen distribution function. It is enough that $G$ be a finite measure to ensure the existence of the integral that defines $T_n$. With this condition, all the theoretical results obtained in Kalemkerian and Fernández (2020) remains valid. Then, although we use the condition that $dG(r, s)$ is a density, actually $dG(r, s)$ is a weight function.

The calculation of $T_n$ will be seen in Sec. 3, and in Sec. 4 will be seen how to calculate the $p$-value of the test (which is enough to perform the test) and the value of $c$. Observe that $T_n$ is a functional of the $L^2$ Cramér–von Mises type applied to the process $\{E_n(r,s)\}_{r,s>0}$ where

$$E_n(r,s) := \sqrt{n} \left( RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s) \right).$$

(3)

The theoretical results established in Kalemkerian and Fernández (2020) about the process defined in (3), are valid for any distance functions $d_X$ and $d_Y$, and remains valid if we consider other continuous functionals such as an $L^1$-Cramér–von Mises type or one of Kolmogorov–Smirnov type.

If $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are i.i.d. in $S_X \times S_Y$, we will compare the $T_n$ statistic proposed in Kalemkerian and Fernández (2020), which we will call $T_n^{(2)}$, defined by

$$T_n^{(2)} := n \int_0^{+\infty} \int_0^{+\infty} \left( RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s) \right)^2 dG(r,s)$$

with the statistics defined as

$$T_n^{(1)} := \sqrt{n} \int_0^{+\infty} \int_0^{+\infty} \left| RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s) \right| dG(r,s)$$

and

$$T_n^{(\infty)} := \sqrt{n} \sup_{r,s>0} \left| RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s) \right|.$$

Observe that in the general case in which $X$ and $Y$ lie in metric spaces $(S_X, d_X)$ and $(S_Y, d_Y)$, the statistics $T_n^{(1)}, T_n^{(2)}$ and $T_n^{(\infty)}$ depend on the distance functions $d_X$ and $d_Y$. In Sec. 4 we will compare the power under several alternative tests based on $T_n^{(1)}, T_n^{(2)}$ and $T_n^{(\infty)}$ for different distance functions $d_X$ and $d_Y$.

In the case in which $X$ and $Y$ are discrete time series, we will use the classical $l^1$, $l^2$ and $L^\infty$ distances, that is, $d_X(x,x') = \sum_{n \geq 1} |x_n - x'_n|$, $d_X(x,x') = \sqrt{\sum_{n \geq 1} (x_n - x'_n)^2}$ and $d_X(x,x') = \sup_{n \geq 1} |x_n - x'_n|$ and analogously for $d_Y$. Analogously, when $X$ and $Y$ are continuous time series, we will use the classical $L^1, L^2, L^\infty$ distances, that is, $d_X(x,x') = \int_{-\infty}^{+\infty} |x(t) - x'(t)|dt$, $d_X(x,x') = \sqrt{\int_{-\infty}^{+\infty} (x(t) - x'(t))^2 dt}$ and $d_X(x,x') = \sup_{t \in \mathbb{R}} |x(t) - x'(t)|$. We will use the notation $T_n^{(i,j)}$ where $i,j = 1, 2, \infty$ for the statistic $T_n^{(i)}$ where the distance functions used are the $l^i$ (or $L^i$) distance. For example, $T_n^{(1,2)}$ means that we use the statistic $T_n^{(1)}$ and the distance considered for $S_X$ and $S_Y$ is $d_X(x,x') = \sqrt{\sum_{n \geq 1} (x_n - x'_n)^2}$ for the discrete case or $d_X(x,x') = \sqrt{\int_{-\infty}^{+\infty} (x(t) - x'(t))^2 dt}$ for the continuous case.
In all cases, as proposed in Kalemkerian and Fernández (2020), we will use a weight function \( G \) such that \( dG(r, s) = g_1(r)g_2(s)drds \) where \( g_1 \) and \( g_2 \) are \( g_1(z) = \varphi \left( \frac{z - \mu_X}{\sigma_X} \right) \) with \( \varphi \) being the density function of an \( N(0, 1) \) random variable and \( \mu_X = \mathbb{E}(d(X_1, X_2)), \sigma_X^2 = \mathbb{V}(d(X_1, X_2)) \) being \( X_1, X_2 \) independent random variables with the same distribution as \( X \). Analogously, \( g_2(t) = \varphi \left( \frac{t - \mu_Y}{\sigma_Y} \right) \).

In practice \( \mu_X \) and \( \sigma_X \) are unknown, but they can be estimated naturally by \( \hat{\mu}_X = \frac{1}{N} \sum_{i \neq j} d(X_i, X_j) \) and \( \hat{\sigma}_X^2 = \frac{1}{N} \sum_{i \neq j} (d(X_i, X_j) - \hat{\mu}_X)^2 \) where \( N = n(n - 1) \), and analogously with \( \hat{\mu}_Y \) and \( \hat{\sigma}_Y^2 \). Taking a function \( G \) such that \( dG(r, s) = g_1(r)g_2(s)drds \) simplifies the calculation of the statistic \( T_n^{(i)} \) for \( i = 1 \) and \( i = 2 \) as will be seen in the next section.

**3. Computing the statistics**

Given \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\) i.i.d. in \( S_X \times S_Y \), and chosen the weight functions \( g_1, g_2 \) to be used, the statistic \( T_n^{(1)} \) and \( T_n^{(\infty)} \) can be computed in the steps that indicated the following two propositions. The steps to calculate \( T_n^{(2)} \) can be found in Kalemkerian and Fernández (2020).

**Proposition 3.1. Calculation of \( T_n^{(1)} \).**

Step 1. Compute \( d(X_i, X_j) \) and \( d(Y_i, Y_j) \) for all \( i, j \in \{1, 2, 3, \ldots, n\} \) where \( i \neq j \) and put \( N = n(n - 1) \).

Step 2. Re-order \( \{d(X_i, X_j)\}_{i \neq j} \) as \( Z_1, Z_2, \ldots, Z_N \) such that \( Z_1 < Z_2 < \ldots < Z_N \) and \( \{d(Y_i, Y_j)\}_{i \neq j} \) as \( W_1, W_2, \ldots, W_N \) maintaining the same indexing as \( Z \)'s (that is, if \( d(X_i, X_j) = Z_h \) then \( d(Y_i, Y_j) = W_h \)).

Step 3. Compute the order statistics for \( W \)'s, that is, \( W_1 < W_2 < \ldots < W_N \).

Step 4. For each \( h, j \in \{1, 2, 3, \ldots, N - 1\} \) compute \( c(h, j) = \sum_{i=1}^{h-j} 1\{W_i < W_{j+1}\} \), that is, the number of elements of the vector \( (W_1, W_2, \ldots, W_h) \) that are less than \( W_{j+1} \) for \( h, j = 1, 2, 3, \ldots, N - 1 \).

Step 5. Compute

\[
T_n^{(1)} = \frac{\sqrt{n}}{N} \sum_{h=1}^{n-N} \left( G_1(Z_{h+1}) - G_1(Z_h) \right) \left( G_2(W_{j+1}^*) - G_2(W_j^*) \right) \left| c(h, j) - \frac{jh}{N} \right|.
\]

**Proposition 3.2. Calculation of \( T_n^{(\infty)} \).**

Step 1. Compute \( d(X_i, X_j) \) and \( d(Y_i, Y_j) \) for all \( i, j \in \{1, 2, 3, \ldots, n\} \) where \( i \neq j \) and put \( N = n(n - 1) \).

Step 2. Re-order \( \{d(X_i, X_j)\}_{i \neq j} \) as \( Z_1, Z_2, \ldots, Z_N \) such that \( Z_1 < Z_2 < \ldots < Z_N \) and \( \{d(Y_i, Y_j)\}_{i \neq j} \) as \( W_1, W_2, \ldots, W_N \) maintaining the same indexing as \( Z \)'s (that is, if \( d(X_i, X_j) = Z_h \) then \( d(Y_i, Y_j) = W_h \)).

Step 3. Compute the order statistics for \( W \)'s, that is, \( W_1 < W_2 < \ldots < W_N \).

Step 4. Compute the \( (N - 1) \times (N - 1) \) matrix \( C \) such that

\[
C_{ij} = \sum_{k=1}^{N} 1\{Z_k < Z_i, W_k < W_j^*\} - \frac{ij}{N}.
\]

Step 5. Compute

\[
T_n^{(\infty)} = \frac{\sqrt{n}}{N} \max_{i,j} C_{ij}.
\]
4. Simulation study

Recall that we have \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\) be an i.i.d. sample of \((X, Y)\) being \(n\) the sample size. When \(X\) and \(Y\) lie in high dimensional spaces (or a function spaces), it is interesting to analyze the performance of the test statistics \(T_n^{(1)}, T_n^{(2)}\) and \(T_n^{(\infty)}\) for different distance functions \(d_X\) and \(d_Y\). In this section we will compare the power of the nine test statistics \(T_n^{(i,j)}\) for \(i, j = 1, 2, \infty\) in the cases in which \(X\) and \(Y\) are discrete and continuous time series under several alternatives. In all cases we will use the same distance function for \(X\) and \(Y\), that is, if \(X\) and \(Y\) are discrete time series, then we will use \(l\) for both \(X\) and \(Y\) for \(j = 1, 2, \infty\), and analogously in the case in which \(X\) and \(Y\) are continuous time series. In all cases, every \(X\) and \(Y\) are the observation of a time series at 100 points, and the power (due to the computational cost) was calculated at the 5% level from 500 replications. That is, for every alternative we have simulated \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\) and perform the test 500 times, and we have calculate the proportion of times the test rejects the null hypothesis.

Every \(p\)-value was calculated by a permutation method, as suggested in Kalemkerian and Fernández (2020), for 100 replications. Roughly speaking, if \(H_0\) is true and we have \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\), then the distribution of \(T_n\) is equal to the distribution of \(T_n^\sigma\) calculated from \((X_{\sigma(1)}, Y_{\sigma(1)}), (X_{\sigma(2)}, Y_{\sigma(2)}), \ldots, (X_{\sigma(n)}, Y_{\sigma(n)})\) for any permutation \(\sigma\). Therefore, we can obtain a sample of the random variables \(T_n\) by changing a permutation \(\sigma\) and then, we can calculate approximately the \(p\)-value.

4.1. The discrete case

We have analyzed six dependency scenarios between \(X\) and \(Y\) and compared the power using the nine test considered: \(T_n^{(i,j)}\) where \(i, j = 1, 2, \infty\). On the one hand, we have considered \(X\) is AR(1) where \(\phi = 0.1\), which we call simply AR(0.1). On the other hand, \(X\) is ARMA(2, 1) with parameters \(\phi = (0.2, 0.5)\) and \(\theta = 0.2\). In both cases the length of \(X\) is 100 and we consider three possible \(Y\): \(X_1 = X^2 + 3\varepsilon\), \(Y_2 = \sqrt{|X|} + \sigma \varepsilon\) where \(\sigma^2\) means the variance of \(\sqrt{|X|}\) and \(Y_3 = \varepsilon X\). In all cases, \(\varepsilon\) is a standard Gaussian white noise \((N(0,1))\) independent of \(X\). Tables 1–4 do not show important differences between using \(T_n^{(2)}, T_n^{(1)}\) or \(T_n^{(\infty)}\). In Figure 1 we show the power as a function of sample size, where the statistic considered is \(T_n^{(2)}\), that is \(T_n^{(2,1)}, T_n^{(2,2)}\) and \(T_n^{(2,\infty)}\). The behavior of \(T_n^{(1)}\) and \(T_n^{(2)}\) is similar. Figure 1 suggest that the power increases as the distance function considered goes from \(d_{\infty}\) (\(L^\infty\) distance) to \(d_1\) (\(L^1\) distance). Also for the alternative \(Y = Y_2\), the statistic based on the \(L^\infty\) distance has difficulties in detecting the dependence between \(X\) and \(Y\) (which grows very slowly as \(n\) increases), while for \(n = 60\) the power of the test based on \(L^1\) or \(L^2\) distances is near unity.

4.2. The continuous case

In this subsection, we will take \(X\) to be a fractional Brownian motion with \(\sigma = 1\) observed in \([0, 1]\) (at times \(0, 1/100, 2/100, \ldots, 99/100\)) for \(H = 0.5\) (standard Brownian motion) and \(H = 0.7\). We consider seven dependence cases between \(X\) and \(Y\). We will consider the case in which \(Y\) is a fractional Ornstein–Uhlenbeck process driven by a Brownian motion \((X)\) for \(H = 0.5\) (\(Bm\)) and fractional Brownian motion for \(H = 0.7\) (\(Bm\)), which we call the OU and FOU processes, respectively. A particular linear combination of FOU, which we call FOU(2), and whose definition, theoretical development and simulations are found in Kalemkerian and León (2019) and Kalemkerian (2020), is a particular case of the models proposed in Arratia, Cabáñez, and Cabáñez (2016). More explicitly, the FOU process is defined by \(Y_t = \sigma \int_{-\infty}^{t} e^{-\lambda(t-s)} dX_s\) (where \(X = \{X_t\}\) is an fBm), and the FOU(2)
Table 1. Comparison of powers, at the 5% level, for the different tests, where $X$ is $AR(1)$ where $\phi = 0.1$ and $Y_1 = X^2 + 3\varepsilon$, $Y_j = \sqrt{|X|} + c\varepsilon$ where $c^2$ means the variance of $\sqrt{|X|}$, $Y_3 = \varepsilon X$ for sample size of $n=30$. In all cases, $\varepsilon$ is a standard Gaussian white noise $N(0,1)$ independent of $X$.

| $n = 30$ | $T_n^{(1,1)}$ | $T_n^{(1,2)}$ | $T_n^{(2,1)}$ | $T_n^{(2,2)}$ | $T_n^{(2,\infty)}$ | $T_n^{(\infty,1)}$ | $T_n^{(\infty,2)}$ | $T_n^{(\infty,\infty)}$ |
|----------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $Y = Y_1$ | 0.39           | 0.40           | 0.39           | 0.40           | 0.24           | 0.32           | 0.28           | 0.16           |
| $Y = Y_2$ | 0.45           | 0.22           | 0.10           | 0.71           | 0.52           | 0.69           | 0.19           | 0.05           |
| $Y = Y_3$ | 0.91           | 0.79           | 0.28           | 0.87           | 0.77           | 0.92           | 0.77           | 0.27           |

In bold we show for each alternative the maximum power.

Table 2. Comparison of powers, at the 5% level, for the different tests, where $X$ is $AR(1)$ where $\phi = 0.1$ and $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + c\varepsilon$ where $c^2$ means the variance of $\sqrt{|X|}$, $Y_3 = \varepsilon X$ for sample size of $n=50$. In all cases, $\varepsilon$ is a standard Gaussian white noise $N(0,1)$ independent of $X$.

| $n = 50$ | $T_n^{(1,1)}$ | $T_n^{(1,2)}$ | $T_n^{(2,1)}$ | $T_n^{(2,2)}$ | $T_n^{(2,\infty)}$ | $T_n^{(\infty,1)}$ | $T_n^{(\infty,2)}$ | $T_n^{(\infty,\infty)}$ |
|----------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $Y = Y_1$ | 0.90           | 0.92           | 0.74           | 0.54           | 0.59           | 0.47           | 0.49           | 0.57           |
| $Y = Y_2$ | 0.34           | 0.50           | 0.38           | 0.97           | 0.81           | 0.16           | 0.92           | 0.89           |
| $Y = Y_3$ | 1.00           | 0.94           | 0.64           | 1.00           | 0.94           | 0.61           | 0.99           | 0.92           |

In bold we show for each alternative the maximum power.

Table 3. Comparison of powers, at the 5% level, for the different tests, where $X$ is $ARMA(2, 1)$ where $\phi = (0.2, 0.5)$, $\theta = 0.2$ and $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + c\varepsilon$ where $c^2$ means the variance of $\sqrt{|X|}$, $Y_3 = \varepsilon X$ for sample size of $n=30$. In all cases, $\varepsilon$ is a standard Gaussian white noise $N(0,1)$ independent of $X$.

| $n = 30$ | $T_n^{(1,1)}$ | $T_n^{(1,2)}$ | $T_n^{(2,1)}$ | $T_n^{(2,2)}$ | $T_n^{(2,\infty)}$ | $T_n^{(\infty,1)}$ | $T_n^{(\infty,2)}$ | $T_n^{(\infty,\infty)}$ |
|----------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $Y = Y_1$ | 0.85           | 0.82           | 0.53           | 0.83           | 0.78           | 0.57           | 0.77           | 0.85           |
| $Y = Y_2$ | 0.36           | 0.27           | 0.08           | 0.84           | 0.78           | 0.34           | 0.49           | 0.34           |
| $Y = Y_3$ | 1.00           | 1.00           | 0.93           | 0.99           | 0.93           | 0.50           | 0.96           | 0.88           |

In bold we show for each alternative the maximum power.

Table 4. Comparison of powers, at the 5% level, for the different tests, where $X$ is $ARMA(2, 1)$ where $\phi = (0.2, 0.5)$, $\theta = 0.2$ and $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + c\varepsilon$ where $c^2$ means the variance of $\sqrt{|X|}$, $Y_3 = \varepsilon X$ for sample size of $n=50$. In all cases, $\varepsilon$ is a standard Gaussian white noise $N(0,1)$ independent of $X$.

| $n = 50$ | $T_n^{(1,1)}$ | $T_n^{(1,2)}$ | $T_n^{(2,1)}$ | $T_n^{(2,2)}$ | $T_n^{(2,\infty)}$ | $T_n^{(\infty,1)}$ | $T_n^{(\infty,2)}$ | $T_n^{(\infty,\infty)}$ |
|----------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $Y = Y_1$ | 0.98           | 0.98           | 0.82           | 0.96           | 0.97           | 0.82           | 0.96           | 0.96           |
| $Y = Y_2$ | 0.76           | 0.48           | 0.04           | 0.98           | 0.43           | 0.65           | 0.43           | 0.03           |
| $Y = Y_3$ | 1.00           | 1.00           | 0.74           | 1.00           | 1.00           | 0.77           | 1.00           | 1.00           |

In bold we show for each alternative the maximum power.

The process is defined by $Y_t = \frac{\lambda_1}{\lambda_2 - \lambda_1} \sigma \int_{-\infty}^{t} e^{-\lambda_1(t-s)} dX_s + \frac{\lambda_2}{\lambda_2 - \lambda_1} \sigma \int_{-\infty}^{t} e^{-\lambda_2(t-s)} dX_s$ (where $X = \{X_t\}$ is an fBm). When $H = 0.5$, we will call them simply the OU and OU(2) processes. Tables 5 and 6 gives us the power for $n=30$ and $n=50$, respectively, for the 7 alternatives. As expected, for values of $\sigma$ larger than 1, the dependence between $X$ and $Y$ is more difficult to detect, and we need to increase the sample size. The same occurs if we take $\lambda_1$ near to $\lambda_2$ in OU(2) and FOU(2). Tables 5 and 6 show, as in the discrete case, no substantial differences between the performance of the three statistics ($T_n^{(1,1)}$, $T_n^{(1,2)}$ or $T_n^{(\infty)}$). With respect to which distance between elements of $X$ and $Y$ is more appropriate, Tables 5 and 6 show that the $L^\infty$ distance performs poorly under the alternatives $Y_1$, $Y_2$ and $Y_3$, but better under alternatives $Y_4$, $Y_5$, $Y_6$ and $Y_7$. The performance of the $L^1$ and $L^2$ distances is similar throughout the 7 alternatives. Figure 2 expands the information given in Table 5 and Table 6 for the cases $Y_1$, $Y_2$ and $Y_3$ because it shows us the power for the statistics $T_n^{(i,j)}$ for $i,j = 1, 2, \infty$ for sample sizes of $n = 10$ to $n = 50$. Figure 2 show clearly that the $L^\infty$ distance has a poorer performance than the $L^1$ and $L^2$ distances. Figure 3 shows the power as a function of sample size for the statistic $T_n^{(2)}$ in the cases $Y_4$, $Y_5$, $Y_6$ and $Y_7$. The behavior of the statistics $T_n^{(1)}$ and $T_n^{(2)}$ is similar. Contrary to what happened in the cases $Y_1$, $Y_2$ and $Y_3$, the $L^\infty$ distance performs clearly better than the $L^1$ and $L^2$ distances. Also, Figure 3 show that the performance of the $T_n^{(2)}$ statistic increases as we
move from the use of the \(L^1\) distance to the use of the \(L^\infty\) distance. On the other hand, Figure 3 shows that the power in the case of the OU alternative is higher than for the OU(2) alternative (and the same for FOU versus FOU(2)), which is reasonable, because the dependence between \(X\) and \(Y\) is simpler in the OU (FOU) case than in the OU(2) (FOU(2)) case. Also, the power in the OU (OU(2)) case is higher than in the FOU (FOU(2)) case, which is to be expected because when \(H = 0.7\), the fractional Brownian motion has a long range dependence, therefore it is reasonable that the dependence between \(X\) and \(Y\) is more difficult to detect.

To conclude this section, observe that the test of independence based on recurrence rates has a power that grows as \(n\) grows for the 9 statistics considered, \(T_n^{(i,j)}\) for \(i,j = 1,2,\infty\) (as expected according to the theory developed in Kalemkerian and Fernández (2020)) in all the alternatives considered for both the discrete and the continuous cases. In most of the cases, the test has a power near to unity for moderately small sample sizes. Taking into account what was observed in this section, it can be said that there is no preference to use the test based on \(T_n^{(1)}, T_n^{(2)}\) or \(T_n^{(\infty)}\), but in the three cases, in general the performance is better as the function distance goes from the
In Table 5, we compare powers at the 5% level of the different tests, where $X \sim Bm$ in alternatives $Y_1, Y_2, Y_3, Y_4, Y_6$ and $X \sim fBm$ with $H = 0.7$ in alternatives $Y_5, Y_7$, where $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X| + \varepsilon}$, $Y_3 = iX + 3\varepsilon'$, $Y_4 = OU$, $Y_5 = FOU$, $Y_6 = OU(2)$, and $Y_7 = FOU(2)$ for sample size of $n = 30$. In all cases, $\varepsilon$ and $\varepsilon'$ are Gaussian white noises with $\sigma = 1$ such that $X, \varepsilon$ and $\varepsilon'$ are independent.

| $n = 30$ | $T_n^{(1)}$ | $T_n^{(1,2)}$ | $T_n^{(1,\infty)}$ | $T_n^{(2,1)}$ | $T_n^{(2,2)}$ | $T_n^{(2,\infty)}$ | $T_n^{(\infty,1)}$ | $T_n^{(\infty,2)}$ | $T_n^{(\infty,\infty)}$ |
|---|---|---|---|---|---|---|---|---|---|
| $Y = Y_1$ | 0.70 | 0.58 | 0.38 | 0.79 | 0.76 | 0.57 | 0.66 | 0.83 | 0.47 |
| $Y = Y_2$ | 0.44 | 0.43 | 0.27 | 0.51 | 0.52 | 0.22 | 0.51 | 0.54 | 0.22 |
| $Y = Y_3$ | 0.33 | 0.42 | 0.29 | 0.42 | 0.41 | 0.19 | 0.39 | 0.37 | 0.20 |
| $Y = Y_4$ | 0.69 | 0.79 | 0.92 | 0.58 | 0.67 | 0.96 | 0.43 | 0.56 | 0.54 |
| $Y = Y_5$ | 0.30 | 0.37 | 0.53 | 0.30 | 0.46 | 0.81 | 0.54 | 0.56 | 0.25 |
| $Y = Y_6$ | 0.16 | 0.21 | 0.15 | 0.21 | 0.31 | 0.74 | 0.17 | 0.16 | 0.18 |
| $Y = Y_7$ | 0.07 | 0.15 | 0.95 | 0.18 | 0.21 | 0.43 | 0.07 | 0.11 | 0.02 |

In bold we show for each alternative the maximum power.

In Table 6, we compare powers at the 5% level of the different tests, where $X \sim Bm$ in alternatives $Y_1, Y_2, Y_3, Y_4, Y_6$ and $X \sim fBm$ with $H = 0.7$ in alternatives $Y_5, Y_7$, where $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X| + \varepsilon}$, $Y_3 = iX + 3\varepsilon'$, $Y_4 = OU$, $Y_5 = FOU$, $Y_6 = OU(2)$, and $Y_7 = FOU(2)$ for sample size of $n = 50$. In all cases, $\varepsilon$ and $\varepsilon'$ are Gaussian white noises with $\sigma = 1$ such that $X, \varepsilon$ and $\varepsilon'$ are independent.

| $n = 50$ | $T_n^{(1)}$ | $T_n^{(1,2)}$ | $T_n^{(1,\infty)}$ | $T_n^{(2,1)}$ | $T_n^{(2,2)}$ | $T_n^{(2,\infty)}$ | $T_n^{(\infty,1)}$ | $T_n^{(\infty,2)}$ | $T_n^{(\infty,\infty)}$ |
|---|---|---|---|---|---|---|---|---|---|
| $Y = Y_1$ | 0.90 | 0.91 | 0.90 | 0.93 | 0.95 | 0.84 | 0.89 | 0.92 | 0.79 |
| $Y = Y_2$ | 0.70 | 0.78 | 0.46 | 0.82 | 0.80 | 0.44 | 0.63 | 0.85 | 0.33 |
| $Y = Y_3$ | 0.68 | 0.67 | 0.41 | 0.51 | 0.62 | 0.44 | 0.61 | 0.67 | 0.38 |
| $Y = Y_4$ | 1.00 | 0.86 | 1.00 | 0.75 | 0.92 | 0.99 | 0.74 | 0.70 | 0.31 |
| $Y = Y_5$ | 0.33 | 0.46 | 0.97 | 0.36 | 0.58 | 0.98 | 0.70 | 0.64 | 0.29 |
| $Y = Y_6$ | 0.40 | 0.48 | 0.46 | 0.39 | 0.58 | 0.95 | 0.46 | 0.55 | 0.78 |
| $Y = Y_7$ | 0.06 | 0.30 | 1.00 | 0.22 | 0.33 | 0.72 | 0.17 | 0.21 | 0.32 |

In bold we show for each alternative the maximum power.

$L^1$ ($l^1$) distance to the $L^\infty$ ($l^\infty$) distance in some cases, and in the opposite direction for other cases. Therefore, it can be suggested that one use the test statistic using the $L^1$ ($l^1$) or $L^2$ ($l^2$) distance and the $L^\infty$ ($l^\infty$) distance to cover both possibilities.

**5. Comparison with other tests in high dimension**

In Kalemkerian and Fernández (2020) the very good performance of the recurrence rates test for random variables and random vectors was shown. In this section, we will compare our test when $X$ and $Y$ lie in high dimensional spaces. According to what was shown in the previous section, we have considered the test using the $T_n^{(2,2)}$ and $T_n^{(2,\infty)}$ statistics. We will consider three competitors: the well known distance covariance test proposed in Székely, Rizzo, and Bakirov (2007) and adapted to perform better in high dimensions in Székely and Rizzo (2013), the Hilbert–Schmidt Information Criterion proposed in Gretton et al. (2007), and that proposed more recently in Fraiman, Moreno, and Vallejo (2017) based on random projections. Basically, this test is based on the idea of choosing $K$ pairs of random directions, and observing that if $X$ and $Y$ are independent, then the projections of $X$ and $Y$ in each one of $K$ pairs of directions are independent. This test is universally consistent. To perform this test, it is necessary to previously choose the number of pairs of projections ($K$), and then $K$ independence hypothesis tests are performed. If at least one of these tests rejects the hypotheses of independence, then $H_0$ is rejected. To work at the 5% level, in Fraiman, Moreno, and Vallejo (2017) it is proposed to use a Bonferroni correction, that is, to compute the proportion of $p$-values smaller than 0.05/$K$ to perform each one of the $K$ uni-dimensional tests. We will call the RPK test. In Table 7 we report a power comparison at the 5% level, when $X$ is a realization of a discrete time series of length 100 in three possible scenarios, where there are three alternatives for $Y$ in each scenario. The performance of RPK is very bad in these cases, and the power using the Bonferroni correction is 0. For this reason we present in Table 7 the power of the RPK test using 0.05/4 instead of 0.05/$K$ for $K = 100$ random
projections. Table 7 shows that our test based on $T_{n}^{(2,1)}$ outperforms the other tests in the 9 cases considered. Table 8 shows a comparison at the 5% level, of the powers in 12 scenarios in which $X$ and $Y$ are realizations of a continuous time series viewed at 100 equispaced points in $[0,1]$. In this table, we have considered the RPK test for $K = 5$ random projections and we have used the Bonferroni correction. We chose $K = 5$ projections because this is the value of $K$ for which the power of the RPK test reaches its maximum. Table 8 shows that our test based on $T_{n}^{(2,2)}$ or $T_{n}^{(2,\infty)}$ outperforms the other competitors in 6 scenarios, and the HSIC, DCOV and RPK tests have the best performance in two cases each.

6. Applications to real data

In this section we will see a couple of applications to meteorological data. Two applications to economic data can be found in Kalemkerian and Fernández (2019).
6.1. Temperature, humidity, wind, and evaporation

In this subsection we consider the meteorological data given in table 7.2 of Rencher (1995). The data are 46 observations grouped into 11 variables defined as follows:

- $Y_1$ = “maximum daily air temperature,”
- $Y_2$ = “minimum daily air temperature,”
- $Y_3$ = “integrated area under daily air temperature curve,”
- $Y_4$ = “maximum daily soil temperature,”
- $Y_5$ = “minimum daily soil temperature,”
- $Y_6$ = “integrated area under daily soil temperature curve,”
- $Y_7$ = “maximum daily relative humidity,”
- $Y_8$ = “minimum daily relative humidity,”
- $Y_9$ = “integrated area under daily humidity curve,”
- $Y_{10}$ = “total wind (in miles per day),” and
- $Y_{11}$ = “evaporation.”

We consider the vectors $Z_1 = (Y_1, Y_2, Y_3)$, $Z_2 = (Y_4, Y_5, Y_6)$, $Z_3 = (Y_7, Y_8, Y_9)$ and the variables $Z_4 = Y_{10}$ and $Z_5 = Y_{11}$.

Taking into account what was seen in the previous section, that there are no important differences between the use of $T_n^{(1)}$, $T_n^{(2)}$, or $T_n^{(\infty)}$ as a test statistic, we apply our independence test between couples of Z’s using $T_n^{(2,1)}$ as the test statistic. In Table 9 we show the $p$-values of our test in each case. In Figure 4 we show the dependogram of order 2 of the mutual independence test of the $Z_i$’s, that is, the critical values at 5% and 10% and the value of our statistic. The dependogram is a graphical tool to observe for several hypothesis test showing the obtained value of the statistic and the critical value of the rejection region, it was introduced by Genest and

Figure 3. Power at the 5\% level, under several alternatives for the statistic $T_n^{(2)}$ using the Manhattan distance ($T_n^{(2,1)}$ in black), Euclidean distance ($T_n^{(2,2)}$ in blue) and maximum ($T_n^{(2,\infty)}$ in red). OU alternative ($Y_4$) means that $X \sim \text{Bm}$, $Y \sim \text{OU}$ where $\lambda = 0.3$, FOU alternative ($Y_5$) means that $X \sim \text{fBm}$ (where $H=0.7$) and $Y \sim \text{FOU}$ where $\lambda = 0.3$, OU(2) alternative ($Y_6$) means that $X \sim \text{fBm}$ (where $H=0.7$), $Y \sim \text{FOU}(2)$ where $\lambda_1 = 0.3, \lambda_2 = 0.8$, OU(2) alternative ($Y_7$) means that $X \sim \text{fBm}$ (where $H=0.7$), $Y \sim \text{FOU}(2)$ where $\lambda_1 = 0.3, \lambda_2 = 0.8$.

6.1. Temperature, humidity, wind, and evaporation

In this subsection we consider the meteorological data given in table 7.2 of Rencher (1995). The data are 46 observations grouped into 11 variables defined as follows: $Y_1$ = “maximum daily air temperature,” $Y_2$ = “minimum daily air temperature,” $Y_3$ = “integrated area under daily air temperature curve,” $Y_4$ = “maximum daily soil temperature,” $Y_5$ = “minimum daily soil temperature,” $Y_6$ = “integrated area under daily soil temperature curve,” $Y_7$ = “maximum daily relative humidity,” $Y_8$ = “minimum daily relative humidity,” $Y_9$ = “integrated area under daily humidity curve,” $Y_{10}$ = “total wind (in miles per day),” and $Y_{11}$ = “evaporation.” We consider the vectors $Z_1 = (Y_1, Y_2, Y_3)$, $Z_2 = (Y_4, Y_5, Y_6)$, $Z_3 = (Y_7, Y_8, Y_9)$ and the variables $Z_4 = Y_{10}$ and $Z_5 = Y_{11}$.

Taking into account what was seen in the previous section, that there are no important differences between the use of $T_n^{(1)}$, $T_n^{(2)}$, or $T_n^{(\infty)}$ as a test statistic, we apply our independence test between couples of Z’s using $T_n^{(2,1)}$ as the test statistic. In Table 9 we show the $p$-values of our test in each case. In Figure 4 we show the dependogram of order 2 of the mutual independence test of the $Z_i$’s, that is, the critical values at 5% and 10% and the value of our statistic. The dependogram is a graphical tool to observe for several hypothesis test showing the obtained value of the statistic and the critical value of the rejection region, it was introduced by Genest and
Rémillard (2004). The approximate p-values and critical values were calculated under \( m = 1000 \) replications by a permutation method as has been suggested in Kalemkerian and Fernández (2020). The test concludes that \( Z_1, Z_2, Z_3 \) and \( Z_5 \) are pairwise dependent, but the wind (\( Z_4 \)) does not exhibit any dependence with any of the other variables. On the one hand, these conclusions are equivalent to those obtained in Bilodeau and Getsop (2017). On the other hand, we observe that in the cases in which the test reject \( H_0 \), the difference between the observed and the critical values is very large. The largest observed value is 0.2887, reached in the case \( Z_1, Z_2 \) (see Figure 4, in all cases in which our test reject \( H_0 \), the observed value of the statistic values is outside the figure). To end this subsection we can conclude that we have worked with a real data set that has a small sample size (\( n = 46 \)) and low dimension spaces, and we have seen that our test works well (in the sense of the reasonableness of the results) and it is interesting to see that when our test rejects the null hypothesis, it does so conclusively, since the observed value of the statistic greatly exceeds the critical value.

### 6.2. Temperature, westbound wind, eastbound wind

The purpose of this example is to use the independence test to detect geographic regions where there is independence in the study variable. If the study variable in two geographic regions (\( X \)
Table 8. Comparison at the 5% level of the powers of different independence tests considered in the case of continuous time series and different sample sizes. Bm and fbm denote a Brownian motion and fractional Brownian motion with $H = 0.7$. $\varepsilon$ and $\varepsilon'$ are independent white noises with $\sigma = 1$ (and independent of $X$). OU alternative means that $X \sim \text{Bm}$, $Y \sim \text{OU}$ where $\lambda = 0.3$, FOU alternative means that $X \sim \text{Bm}$ (where $H = 0.7$) and $Y \sim \text{FOU}$ where $\lambda = 0.3$, OU(2) alternative means that $X \sim \text{Bm}$, $Y \sim \text{OU(2)}$ where $\lambda_1 = 0.3, \lambda_2 = 0.8$, OU(2) alternative means that $X \sim \text{Bm}$ (where $H = 0.7$), $Y \sim \text{FOU(2)}$ where $\lambda_1 = 0.3, \lambda_2 = 0.8$. In OU, FOU, OU(2) and FOU(2) we have taken $\sigma = 1$.

| $X \sim \text{Bm}$ | $Y = X^2 + 3\varepsilon$ | $Y = \sqrt{|X|} + \varepsilon$ | $Y = \varepsilon X + 3\varepsilon'$ | $Y \sim \text{OU}$ | $Y \sim \text{OU(2)}$ | $X \sim \text{fbm}$ | $Y = X^2 + 3\varepsilon$ | $Y = \sqrt{|X|} + \varepsilon$ | $Y = \varepsilon X + 3\varepsilon'$ | $Y \sim \text{FOU}$ | $Y \sim \text{FOU(2)}$ | $X \sim \text{OU}(\lambda_1)$ | $Y \sim \text{OU}(\lambda_2)$ | $X \sim \text{FOU}(\lambda_1)$ | $Y \sim \text{FOU}(\lambda_2)$ |
|-------------------|------------------|------------------|------------------|-----------------|-----------------|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| $n$               | RPK              | HSIC             | DCOV             | $T_{1, 2, 2}^{(2, \infty)}$ | $T_{1, 2, 2}^{(2, \infty)}$ | $n$               | RPK              | HSIC             | DCOV             | $T_{1, 2, 2}^{(2, \infty)}$ | $T_{1, 2, 2}^{(2, \infty)}$ | $n$               | RPK              | HSIC             | DCOV             | $T_{1, 2, 2}^{(2, \infty)}$ | $T_{1, 2, 2}^{(2, \infty)}$ |
| 30                | 0.767            | 0.815            | 0.596            | 0.757            | 0.570            | 0.427            | 0.498            | 0.567            | 0.613            | 0.663            | 0.717            | 0.764            | 0.814            | 0.872            | 0.932            | 0.992            |
| 50                | 0.951            | 0.977            | 0.829            | 0.947            | 0.836            | 0.846            | 0.906            | 0.945            | 0.982            | 0.982            | 0.961            | 0.927            | 0.977            | 0.990            | 0.999            | 0.996            |
| 80                | 0.999            | 1.000            | 0.977            | 0.994            | 0.968            | 0.958            | 1.000            | 0.992            | 0.982            | 0.982            | 0.934            | 0.992            | 1.000            | 1.000            | 0.990            | 0.996            |

In bold the maximum power.

Table 9. $p$-values for the test between couples of the $Z$'s where $Z_1 = (Y_1, Y_2, Y_3)$, $Z_2 = (Y_4, Y_5, Y_6)$, $Z_3 = (Y_7, Y_8, Y_9)$, $Z_4 = Y_{10}$ and $Z_5 = Y_{11}$, being $Y_1$ = “maximum daily air temperature,” $Y_2$ = “minimum daily air temperature,” $Y_3$ = “integrated area under daily air temperature curve,” $Y_4$ = “maximum daily soil temperature,” $Y_5$ = “minimum daily soil temperature,” $Y_6$ = “integrated area under daily soil temperature curve,” $Y_7$ = “maximum daily relative humidity,” $Y_8$ = “minimum daily relative humidity”, $Y_9$ = “integrated area under daily humidity curve,” $Y_{10}$ = “total wind (in miles per day),” and $Y_{11}$ = “evaporation.”

| $Z_2$ | $Z_3$ | $Z_4$ | $Z_5$ |
|-------|-------|-------|-------|
| 0.000 | 0.000 | 0.109 | 0.000 |
| 0.000 | 0.000 | 0.394 | 0.000 |
| 0.373 | 0.000 | 0.000 | 0.000 |
| 0.403 | 0.000 | 0.000 | 0.000 |

In bold the maximum power.
and $Y$) are independent, it may be convenient to apply different models in each region to explain and predict the variable in the best possible way. In this subsection we will consider the data set formed by the forecast temperature ($T$), westbound wind ($U$) and eastbound wind ($V$) at 850 hPa (around 1200 m above sea level) from each day from January 2012 to December 2012. There are a total of 341 forecasts in each geographical point, due to the fact that 25 data points are missing. The numerical domain is shown in Figure 5 and consists of a total of $117 \times 75 = 8775$ geographical points where the daily forecasts are made. The red line separates the southern half from the northern half.

Figure 4. Critical values at 5\% (blue), 10\% (red) and observed values (black) for the pairwise independence test between the $Z$'s variables. In tests in which the null hypothesis is rejected, the observed value of the statistics is very large in relation to the critical value. The largest observed value is 0.2887, reached in the case $Z_1, Z_2$.

Figure 5. $117 \times 75 = 8775$ geographical points where the daily forecasts are made. The red line separates the southern half from the northern half.
geographical points. Latitud varies from 11°29'31" to 42°37'9" (south) and longitude from 40°58'32" to 79°1'28" (west). The time horizon of the forecasts is 24 h, and they are for 0:00 GMT hour of each day. The numerical simulations were obtained using the Weather Research and Forecasting (WRF) regional model, and the initial and lateral boundary conditions were obtained from the NCEP Global Forecast System, as in Cazes Boezio and Ortelli (2018). If we consider \( (U_i, V_i, T_i) \), \( (U_j, V_j, T_j) \), \( \ldots \), \( (U_{341}, V_{341}, T_{341}) \) where \( U_i, V_i, T_i \in \mathbb{R}^{8775} \) for all \( i = 1, 2, 3, \ldots, 3841 \), the \( p \)-values for the independence test between \( U \) and \( V \) is equal to zero, and so on for the test between \( U \) and \( T \), and \( V \) and \( T \). This is expected because for each geographical point \( i \), the variables \( U_i, V_i \) and \( T_i \) are pairwise dependent. Now we consider (for each day) every vector \( U \in \mathbb{R}^{8775} \) to be decomposed as \( U = (U_1, U_2, \ldots, U_{75}) \) where \( U_i \in \mathbb{R}^{117} \). In this form, each \( U_i \) represents the forecast of the 117 geographical points at latitude \( i \) and can be seen as a discretization of a curve at latitude \( i \), \( (U(i)) \). Here, \( i = 1 \) indicates the southeastern most latitude given in Figure 5 and \( i = 75 \) the northeastern most latitude. We consider the first 30 forecasts, corresponding to January 2012. In this way, we get a sample of 30 curves for each latitude \( i \), and we will test the mutual independence between \( U_i \) and \( U_j \) for \( i = 1, 2, 3, \ldots, 38 \) and \( j = 76 - i \). Observe that varying \( i \) from 1 to 38, we have the southern half of the curves while with \( j = 76 - i \) we have the...
northern half. We decompose $T$ and $V$ analogously. It is to be expected that, at least for small values of $i$, the variables $U_i$ and $U_j$ would be independent, due to the geographical distance, and the same for the variables $V$ and $T$. In Figures 6–9 we show the dependograms for the independence test, using $T_n^{(2,1)}$ and $T_n^{(2,\infty)}$ statistics, between $U_i$ and $U_{76-i}$ for each $i = 1, 2, ..., 38$ and the same for the variables $V$, $T$ and the other combinations between $U$, $V$ and $T$. In Figure 6, similar results between $T_n^{(2,1)}$ and $T_n^{(2,\infty)}$ are shown. However, in the case of $U_i$ and $U_j$, $T_n^{(2,\infty)}$ detects the dependence in more cases than $T_n^{(2,1)}$. Both tests show that when $i$ and $j = 76 - i$ are close, then the variables $U_i$ and $U_j$ are dependent. The same occurs with $V_i$, $V_j$ and $T_i$, $T_j$. Also the geographical region in which the vectors are dependent is longer for $T$ than for $U$ and $V$. Figure 7 shows that $T_n^{(2,1)}$ performs better than $T_n^{(2,\infty)}$ because for $i \geq 32$ the test based on $T_n^{(2,1)}$ detects a dependence for both cases: $U_i$, $V_j$ and $V_i$, $U_j$. Figures 8 and 9 show that the tests based on $T_n^{(2,1)}$ and $T_n^{(2,\infty)}$ perform similarly. Also, still being geographically close, the vectors $U_i$ and $T_j$ are independent. However both tests detect a dependence between $T_i$ and $U_j$ for $i = 21$ to $i = 27$ (Figure 8). Figure 9 shows that in most cases, $T_i$ and $V_j$ are dependent, while for $V_i$ and $T_j$ the test does not detect a dependence except for the cases in which $i$ and $j$ are close.

From Figures 6–9 together, we can conclude that for $i \geq 33$ the each one of the variables considered in this study are dependent, that is $U_i$ and $U_j$, $U_i$ and $V_j$, and analogous the other cases. In Figure 10 we show the geographical region in which there is mutual independence (or dependence) between the variables considered in Figures 6–9 (in blue the “dependent” region). To end

Figure 7. Comparison between dependograms for $T_n^{(2,1)}$ and $T_n^{(2,\infty)}$ between $U$ and $V$. Top left, the dependogram for each $i = 1, 2, ..., 38$ for the test $H_0: U_i$ and $V_j$ are independent ($j = 76 - i$). In blue the critical value at 5% level, in black the observed value of the statistic. Analogously the other dependograms included in the figure.
this subsection we can conclude that from the considered data set, our independence test based on recurrence rates, help us to obtain a region in which all the variables in study are mutually dependent.

7. Final comments

In this section, we want to comment on the relationship between the hypothesis test proposed in our work and other studies related to interrelationship, coupling and interdependence. In our hypothesis test we start from $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ iid (that is independent and identically distributed), therefore $X_1, X_2, ..., X_n$ are independent and distributed as certain $X$, and $Y_1, Y_2, ..., Y_n$ are independent and distributed as certain $Y$, and we want to test $H_0 : X$ and $Y$ are independent, vs $H_0$ does not hold. The works concerning, coupling, interdependence and inter-relationship see for instance (Romano, Thiel, and Kurths 2004, Romano et al. 2009, 2007; Hirata and Aihara 2010; Zou et al. 2011; Feldhoff et al. 2012; Goswami et al. 2012; Ramos et al. 2017), starts from two time series $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$ and anlize the possible dependence between them. That is, the dependence or relation between $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_m)$. Being $X$ and $Y$ time series, in general $X_1, X_2, ..., X_n$ are dependent variables and the same for $Y_1, Y_2, ..., Y_m$.

Figure 8. Comparison between dependograms for $T^{(2,1)}_n$ and $T^{(2, \infty)}_n$ between $U$ and $T$. Top left, the dependogram for each $i = 1, 2, ..., 38$ for the test $H_0 : U_i$ and $T_j$ are independent ($j = 76 - i$). In blue the critical value at 5% level, in black the observed value of the statistic. Analogously the other dependograms included in the figure.
8. Conclusions

In this article we have proposed several variants to perform the independence test based on recurrence rates. We have shown how to calculate the test statistic in each one of these cases.

Figure 9. Comparison between dependograms for $T_{n}^{(2,1)}$ and $T_{n}^{(2,\infty)}$ between $V$ and $T$. Top left, the dependogram for each $i = 1, 2, ..., 38$ for the test $H_{0}: V_{i}$ and $T_{j}$ are independent ($j = 76 - i$). In blue the critical value at 5% level, in black the observed value of the statistic. Analogously the other dependograms included in the figure.

Figure 10. In blue the geographical region in which there is mutual dependence between the variables considered in Figures 6–9. The red line separates the southern half from the northern half.

8. Conclusions

In this article we have proposed several variants to perform the independence test based on recurrence rates. We have shown how to calculate the test statistic in each one of these cases.
When $X$ and $Y$ lie in high dimensional spaces, we have shown that the test performs better as the distance function considered goes from the $L^1(I^1)$ distance to the $L^\infty(I^\infty)$ distance in some cases and in the opposite direction in other cases. Therefore, the test statistic using the $L^1(I^1)$ distance and the $L^\infty(I^\infty)$ distances to cover both possibilities can be suggested. From simulations we obtain that in high dimension, our test clearly outperforms the competitors and widely used tests in almost all the alternatives considered. The test was performed on two data sets including small and large sample sizes and we have shown that in both cases the application of the test allows us to obtain interesting conclusions. Taking this together with the simulations presented in Kalemkerian and Fernández (2020), we can conclude that the independence test based on recurrence rates has very good performance for random variables, random vectors, and also for random elements lying in high dimensional spaces.

9. Proofs

**Proof of Proposition 3.1.** We re-order $d(X_i, X_j)$ with $(i, j) \in I^n$ in the form $Z_1, Z_2, \ldots, Z_n$. Assume that $Z_1 < Z_2 < \ldots < Z_n$, and we will use $W_1, W_2, \ldots, W_N$ to denote the values of $d(Y_i, Y_j)$ using the same indexing. We also write $W'_1, W'_2, \ldots, W'_N$ for the order statistics of $W's$.

\[
\int_0^{+\infty} \int_0^{+\infty} \left( R_{n,Y}^Y(r,s) - R_{n}^Y(r)R_{n}^Y(s) \right) g_1(r)g_2(s) \, dr \, ds = \frac{1}{N} \int_0^{+\infty} g_2(s)ds \times \int_0^{+\infty} \sum_{i \neq j} 1 \{ d(X_i, X_j) < r, d(Y_i, Y_j) < s \} - \frac{1}{N} \sum_{i \neq j} 1 \{ d(X_i, X_j) < r \} \sum_{h \neq k} 1 \{ d(Y_h, Y_k) < s \} \left| g_1(r) \right| \, dr 
\]

Observe that

\[
= \sum_{h=1}^{N-1} \left[ \int_0^{+\infty} \sum_{i=1}^{N} 1 \{ Z_i < r, W_i < s \} \right] - \frac{1}{N} \sum_{i=1}^{N} 1 \{ Z_i < r \} \sum_{j=1}^{N} 1 \{ W_j < s \} \left| g_1(r) \right| \, dr 
\]

Then, (4) is equal to

\[
= \frac{1}{N} \sum_{h=1}^{N-1} \left( G_1(Z_{h+1}) - G_1(Z_h) \right) \int_0^{+\infty} \left[ \sum_{i=1}^{h} 1 \{ W_{i-1} < s \} - \frac{h}{N} \sum_{j=1}^{N} 1 \{ W_{j-1} < s \} \right] g_2(s)ds
\]

\[
= \frac{1}{N} \sum_{h=1}^{N-1} \left( G_1(Z_{h+1}) - G_1(Z_h) \right) \int_0^{+\infty} \left[ \sum_{j=1}^{h} 1 \{ W_{j-1} < s \} - \frac{h}{N} \sum_{j=1}^{N} 1 \{ W_{j-1} < s \} \right] g_2(s)ds
\]

\[
= \frac{1}{N} \sum_{h=1}^{N-1} \left( G_1(Z_{h+1}) - G_1(Z_h) \right) \left( G_2(W'_{h+1}) - G_2(W'_h) \right) \left( c(h, j) - \frac{jh}{N} \right),
\]
where \(c(h, j) = \sum_{i=1}^{h} I\{W_i < W_{j+1}^r\}\) is the number of elements of the vector \((W_1, W_2, ..., W_h)\) that are less than \(W_{j+1}^r\) for \(h, j = 1, 2, 3, ..., N - 1\). Thus,
\[
T_n^{(1)} = \sqrt{n - 1} \sum_{h, j=1}^{N-1} \left( G_1(Z_{h+1}) - G_1(Z_h) \right) \left( G_2(W_{j+1}^r) - G_2(W_j^r) \right) c(h, j) - \frac{jhN}{N}.
\]

**Proof of Proposition 3.2.** In accordance with Steps 1 and 2, we put \(N = n(n - 1)\) and re-order \(\{d(X_i, X_j)\}_{i \neq j}\) as \(Z_1, Z_2, ..., Z_N\) such that \(Z_1 < Z_2 < ... < Z_N\) and \(\{d(Y_i, Y_j)\}_{i \neq j}\) as \(W_1, W_2, ..., W_N\) maintaining the same indexing as \(Z\)'s (that is, if \(d(X_i, X_j) = Z_h\), then \(d(Y_i, Y_j) = W_h\)). Observe that, to compute \(T_n(\infty)^{(r,s)}(r,s)\) for all \(r, s > 0\) it is enough to compute \(T_n(\infty)^{(r,s)}(Z, W^r)\) for every \(i, j = 1, 2, ..., N\). Then, the result follows immediately from Steps 4 and 5. \(\square\)

### 10. R-Code

The R-code that we give in this section calculates the \(p\)-value of the independence test based on recurrence rates using \(T_n(2)\) as a statistic test and euclidean distance for \(S_X\) and \(S_Y\), that is \(T_n^{(2,2)}\).

**Step 1.**

1. Input \(k_x\) and \(k_y\) (length of each \(X\) and \(Y\), respectively).
2. Input \(n\) (sample size).
3. Input the data set in the following way:
   - \(x = \text{matrix}($data_x$, n, k_x)\)
   - \(y = \text{matrix}($data_y$, n, k_y)\)
4. Input \(m\) (number of replications to calculate the \(p\)-value).

**Step 2.**

\[
N = n(n - 1)/2
\]

\[
\text{An} = \text{rep}(NA, N)\]
\[
\text{Bn} = \text{rep}(NA, N)\]
\[
\text{Cn} = \text{rep}(NA, N)\]
\[
Z = \text{array}((\text{dist}(x)))\]
\[
Z = \text{pnorm}(Z, \text{mean}(Z), \text{sd}(Z))\]
\[
\text{Zord} = \text{sort}(Z)\]
\[
T = \text{array}((\text{dist}(y)))\]
\[
T = \text{pnorm}(T, \text{mean}(T), \text{sd}(T))\]
\[
\text{Tord} = \text{sort}(T)\]

\[
\text{IRX2} = 1 - (1/(N*N)) * \text{sum}(2*\text{seq}(1, N-1) * \text{Zord})\]
\[
\text{IRY2} = 1 - (1/(N*N)) * \text{sum}(2*\text{seq}(1, N-1) * \text{Tord})\]

\[
\text{for}(i \in 1:N)\{
\text{An}[i] = \text{mean}((1 - (1/2) * (\text{abs}(Z[i] - Z) + Z[i] + Z)) * (1 - (1/2) * (\text{abs}(T[i] - T) + T[i] + T)))
\text{Bn}[i] = \text{mean}((1 - (1/2) * (\text{abs}(T[i] - T) + T[i] + T))
\text{Cn}[i] = \text{mean}((1 - (1/2) * (\text{abs}(Z[i] - Z) + Z[i] + Z))
\}\]
\[
\text{IRXY2} = \text{mean}(\text{An})
\text{IRXRYRXRY} = \text{mean}(\text{Bn} * \text{Cn})
tobs = n*(IRXY2 + IRX2*IRY2 - 2*IRXRYRXRY) # observed value of test statistic

In order to obtain the p-value we perform a permutation method.

t = rep(NA, m)
for (j in 1:m)
{
  x = x[sample(sequence(n))] # permutation of x's
  Z = array(dist(x))
  Z = pnorm(Z, mean(Z), sd(Z))
  Zord = sort(Z)
  IRX2 = 1 - (1/(N*N)) * sum((2*seq(1,N)-1)*Zord)
  for (i in 1:N)
  {
    An[i] = mean((1-(1/2)*abs(Z[i]-Z)+Z[i]+Z)*(1-(1/2)*(abs(T[i]-T)+T[i]+T)))
    Bn[i] = mean((1-(1/2)*abs(T[i]-T)+T[i]+T))
    Cn[i] = mean((1-(1/2)*abs(Z[i]-Z)+Z[i]+Z)))
  }
  IRXY2 = mean(An)
  IRXRYRXRY = mean(Bn*Cn)
  t[j] = n*(IRXY2 + IRX2*IRY2 - 2*IRXRYRXRY) # observed value of test statistic
  # for the j-th permutation.
  print(j)
}
pvalue = mean(t > $tobs)
pvalue

Remark 1. It is possible to run the same code to calculate $T_n^{(2,1)}$ and $T_n^{(2,\infty)}$ by changing in the two places where ir appears dist(x) for dist(x,method="manhattan") or dist(x,method="maximum") respectively and analogously in the place where it appears dist(y).

Acknowledgements

We would like to express our gratitude to Leonardo Moreno for his explanations about the independence test based on random projections and for giving us the R code to use it, and to Gabriel Cazes for giving us the data sets and the map of Figure 5.

References

Arratia, A., A. Cabaña, and E. Cabaña. 2016. A construction of continuous time ARMA models by iterations of Ornstein-Uhlenbeck process. SORT 40 (2):267–302.

Bilodeau, M., and A. Getsop. 2017. Test of mutual or serial independence of random vectors with applications. Journal of Machine Learning 18:1–40.

Cazes Boezio, G., and S. Ortelli. 2018. Minimum-cost numerical prediction system for wind power in Uruguay, with an assessment of the diurnal and seasonal cycles of its quality. Ciencia e Natura 40:206–10.

Feldhoff, J. H., R. V. Donner, J. F. Donges, N. Marwan, and J. Kurths. 2012. Geometric detection of coupling directions by means of inter-system recurrence networks. Physics Letters A 376 (46):3504–13. doi:10.1016/j.physleta.2012.10.008.
