UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Test de independencia

Basado en análisis de recurrencia

Diego Gabriel Fernández Raíz

Programa de Posgrado en Ingeniería Matemática
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
Julio de 2021

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Test de independencia

Basado en análisis de recurrencia

Diego Gabriel Fernández Raíz

Tesis de Maestría presentada al Programa de Posgrado en Ingeniería Matemática, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magister en Ingeniería Matemática.

Director de tesis:
Dr. Prof. Juan Kalemkerian

Codirector:
Dr. Prof. Roberto Markarian

Director académico:
Dr. Prof. Juan Kalemkerian

Montevideo – Uruguay
Julio de 2021

Fernández Raíz, Diego Gabriel

Test de independencia / Diego Gabriel Fernández Raíz. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2021.

xi, 78 p. 29, 7cm.

Director de tesis:

Juan Kalemkerian

Codirector:

Roberto Markarian

Director académico:

Juan Kalemkerian

Tesis de Maestría – Universidad de la República, Programa de Ingeniería Matemática, 2021.

Referencias bibliográficas: p. 53 – 56.

1. Test de Independencia, 2. Tasa de recurrencia, 3. Series de tiempo, 4. Gráficos de recurrencia. I. Kalemkerian, Juan *et al.* II. Universidad de la República, Programa de Posgrado en Ingeniería Matemática. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Dr. Prof. Enrique Cabaña

Dr. Prof. José León

Dr. Prof. Alejandro Cholaquidis

Dr. Prof. Martín Puchet

Montevideo – Uruguay
Julio de 2021

A Rodrigo y Matilda.

Agradecimientos

En primer lugar quisera agradecer el valioso apoyo de mis tutores sin los cuales no hubiera sido posible realizar este trabajo.

En segundo lugar queremos dar nuestro agradecimiento a José Rafael León, Ricardo Fraiman, Ernesto Mordecki y Jorge Graneri por sus comentarios que fueron de gran utilidad en la elaboración de este trabajo. También agradecemos a Leonardo Moreno por explicarnos el test de independencia basado en proyecciones aleatorias y darnos el código en R para aplicarlo, y a Gabriel Cazes por darnos el conjunto de datos y el mapa de la Figura 4.2.

Por último y tal vez lo más importante agradecer a mi familia por el apoyo brindado.

(Epígrafe:) *El azar no existe;*
Dios no juega a los dados.

Albert Einstein

RESUMEN

Dada una muestra $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ i.i.d. de (X, Y), cuando tenemos un test de hipótesis de la forma: $H_0 : X$ e Y son independientes estamos ante los llamados test de independencia. En esta tesis se presenta una nueva prueba de independencia entre dos elementos aleatorios X e Y que toman valores en espacios métricos. La prueba se basa en porcentajes de recurrencias obtenidos a partir de la distancia entre puntos de cada muestra. Se obtiene la distribución asintótica del estadístico bajo la hipótesis nula y se demuestra que la misma tiene un sesgo bajo alternativas contiguas. Se prueba también la consistencia de la prueba para una amplia clase de alternativas, que incluyen el caso particular en el que (X, Y) siguen una distribución normal multivariada. La performance de la prueba, medida a través de la comparación de la potencia respecto de varias alternativas muestra muy buenos resultados, mostrando una mejora con respecto a otras pruebas en muchos casos para diferentes dimensiones. Finalmente se aplica el test a datos reales de tipo metereológico y económico. Como se verá, se detecta muy claramente la dependencia entre todas las series consideradas.

Palabras claves:
Test de Independencia, Tasa de recurrencia, Series de tiempo, Gráficos de recurrencia.

ABSTRACT

When we have a hypothesis test of the form: $H_0 : X$ and Y are independent, we are faced with the so-called independence test. This thesis presents a new test of independence between two random elements found in metric spaces. The test is based on recurrence percentages obtained from the distance between points of each sample. The asymptotic distribution of the statistic is obtained and it is show that the distribution of the test statistic under contiguous alternatives has a bias. The consistency of the test is also tested for a wide class of alternatives, including the particular case in which (X, Y) follows a multivariate normal distribution. The performance of the test, measured through the comparison of the power with respect to several alternatives, shows very good results, showing an improvement with respect to other tests in many cases for different dimensions. Finally, the test is applied to real meteorological and economic data. As seen, is detected the dependence between all the series considered.

Keywords:
Independence tests, Recurrence rates, Time series, Recurrence Plot.

Tabla de contenidos

Capítulo 1

Introducción y motivaciones

1.1. Interés y participación en la Tesis

Mi interés en la realización del presente trabajo de tesis para la maestría de Ingeniería Matemática tiene como antecedentes mi investigación en la tesis de la maestría de Economía de la Facultad de Ciencias Económicas y Administración bajo la dirección del Dr. Roberto Markarian y el Dr. Martin Puchet.

En dicha tesis de economía me interesó profundizar en las herramientas de análisis de recurrencia para el estudio de la dinámica de índices de precios bursátiles. Se trató de un trabajo de aplicación de métodos numéricos computacionales donde no se desarrolla una teoría ni económica ni matemática, pero se innovó en la aplicación de dichas técnicas en la Facultad de Economía.

Para dicho trabajo conté con la orientación del Dr. Markarian y el Dr. Kalemkerian fue parte del tribunal de la tesis.

Mi interés en profundizar en la aplicación de estas técnicas matemáticas en el análisis económico me llevó a realizar la maestría de Ingeniería Matemática bajo la orientación del Dr. Kalemkerian y del Dr. Markarian como cotutores.

Mi primer interés era desarrollar una prueba de hipótesis que permita determinar cuando un sistema dinámico "causa" o "conduce". Si bien el estudio de causalidad lo entendemos posible, en la tesis de Ingeniería matemática se estudia el concepto de independencia.

Con la ayuda fundamental del Dr. Kalemkerian se ha logrado desarrollar fundamentos teóricos matemáticos a las técnicas numéricas del análisis de recurrencia y ha sido posible desarrollar una prueba de Hipótesis de Independencia.

Junto al Dr. Kalemkerian se han realizado tres trabajos cuya sistematiza-

ción se plasman en esta tesis. En dichos trabajos y en la tesis mi participación ha sido primeramente en el planteo de la temática y propuesta de análisis en base al cálculo de probabilidad condicionada de recurrencia para el estudio o determinación de si un sistema A es independiente o no de un sistema B.

La programación en R y el desarrollo teórico matemático de las técnicas de recurrencia son de autoría del Dr. Kalemkerian. Además de redactar conjuntamente los tres trabajos indicados, me puse al día en el uso de estas técnicas y en la obtención de varios resultados por medio de cálculos en el software R. Mi aporte principal en esta parte del trabajo fue la de aplicación mediante su instrumentación. En ese sentido trabajé en el terreno explorativo de la performance del test, calculando algunos de los valores de las tablas que incluyen potencias bajo ciertas alternativas, lo que redundó en que se hicieron muchos otros cálculos de potencias que no aparecen en las publicaciones ni en esta tesis pero que fueron parte fundamental en el proceso de elaboración de las tablas finales.

1.2. Introducción

Detectar la dependencia entre datos es una tarea fundamental en el análisis científico de muchos sistemas complejos e irregulares cualquiera sea su índole: económico, físico, etc. Dada una muestra $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ i.i.d. de (X, Y), donde $X \in S_X$ e $Y \in S_Y$, siendo S_X y S_Y espacios métricos cualesquiera. Cuando tenemos un test de hipótesis de la forma: H_0 : X e Y son independientes estamos ante los llamados test de independencia. Los test de independencia han sido desarrollados en primer lugar para el caso $S_X = S_Y = \mathbb{R}$ en los pioneros trabajos de Galton primero [10] y luego Pearson [24] (el famoso test de correlación, muy utilizado hasta nuestros días).

Existe una amplia colección de test no paramétricos [7]. Uno de los más destacados fue desarrollado por Wald y Wolfowitz [29] y se basa en la ocurrencia repetida del mismo valor o categoría de una variable como el signo. Más recientemente, Heller et al. [14] proponen un test que está basado en las distancias entre los elementos que componen la muestra de X y las de los correspondientes elementos que componen la muestra de Y.

Estas ideas de distancias entre los elementos de una muestra se introducen en la metodología de análisis de recurrencia que es la base de esta tesis. El análisis de recurrencia se puede realizar para las muestras X e Y de forma

separada (análisis univariado) o para las dos muestras en conjunto (análisis bivariado).

1.3. Análisis de recurrencia univariado

1.3.1. Gráfico de Recurrencia

Eckman et al. (1987) [8] incorporaron estas ideas en una herramienta cualitativa denominada gráfico de recurrencia *RP*. La intención original fue proporcionar una herramienta que permitiera brindar información sobre sistemas dinámicos de alta dimensión, cuyas trayectorias en espacios de fase son muy difíciles de visualizar. Un RP permite investigar la trayectoria de un espacio de fase m-dimensional a través de una representación de sus recurrencias en dos dimensiones.

El RP representa los tiempos en que los estados x_i en un espacio de fase recurren. Se estudian las trayectorias en el espacio de fases o en una adecuada reconstrucción de la dinámica subyacente a dichos espacios [22] proporcionando pistas importantes sobre las características de los sistemas en que se desarrollan.

La reconstrucción del espacio de fase es el punto inicial para la construcción del gráfico de recurrencia.

La observación de un proceso real por lo general no brinda todas las variables de estado posibles. A veces no se conocen todas las variables de estado o no todas pueden ser medidas. Muy a menudo solo está disponible una observación *u(t)*. Dado que las mediciones resultan en series de tiempo discretas, las observaciones serán escritas u_i, donde $t = i\ \Delta t$. Variables con subíndices serán medidas en tiempo discreto (Por ejemplo, x_i, $R_{i,j}$, mientras que entre paréntesis t denota variables en tiempo continuo (Por ejemplo, *x(t)*, $R(t_1, t_2)$).

La unión entre los elementos del sistema implica que cada componente individual contiene información esencial sobre la dinámica de todo el sistema. Por lo tanto una trayectoria del espacio de estado equivalente, que preserva las estructuras topológicas del espacio de fases original puede ser reconstruida utilizando solo una observación o serie de tiempo, respectivamente. Un método utilizado frecuentemente para la reconstrucción de tal trayectoria $\hat{x}(t)$ es el método de los retardos: $\hat{x}_i = (u_i, u_{i+\tau}, ..., u_{i+(m-1)\tau})^T$, donde m es la dimensión de incrustación y τ es el retardo temporal. La preservación de las estructuras

topológicas de la trayectoria original están garantizadas si $m \geq 2d + 1$, donde d es la dimensión del atractor [22].

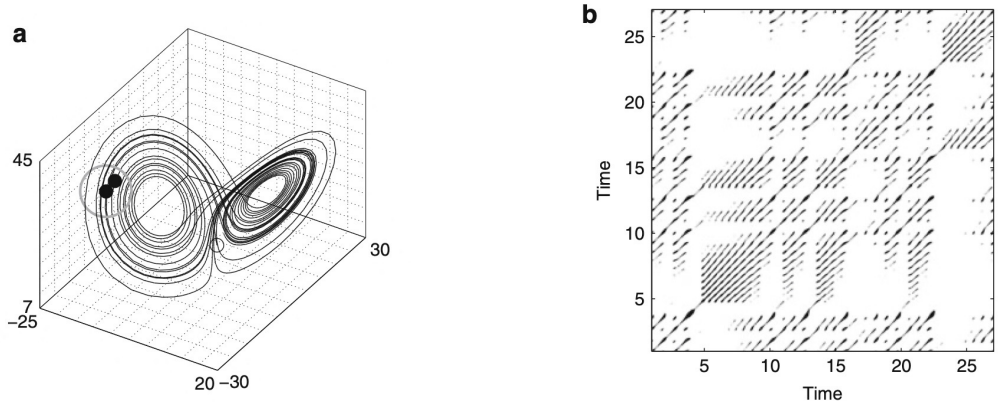

Figura 1.1: (a) Segmento de trayectoria espacio de fases sistema Lorenz (b) Gráfico de recurrencia de (a)

Por ejemplo podemos considerar un segmento de la trayectoria en el espacio de fases del sistema dinámico de Lorenz cuya representación gráfica se presenta en la Figura 1.1 y cuyas ecuaciones mostramos a continuación:

$$\begin{aligned} \dot{x} &= -\sigma\,(x - y)\,, \\ \dot{y} &= -xz + rz - y, \\ \dot{z} &= xy - bz. \end{aligned}$$

Tomando los valores de los parámetros $r = 28$, $\sigma = 10$ y $b = 8/3$ obtenemos la representación de la Figura 1.1 parte a). La parte b) de esta figura es el RP correspondiente a este sistema calculado para un radio $\epsilon = 5$. Un punto de la trayectoria en j que cae en la cercanía (círculo gris en parte a) de un punto dado i es considerado un punto recurrente (punto negro en la trayectoria de la parte a). Esto se marca en la Figura con un punto negro en el RP en la ubicación (i, j). Un punto que no esté en la cercanía (círculo pequeño en parte a) produce un punto blanco en el RP.

1.3.2. Medidas de recurrencias

A comienzos de 1990s Zbilut y Webber introducen definiciones y procedimientos para cuantificar estructuras en RP [32]. Ellos definen un conjunto de variables de recurrencia que constituyen medidas de complejidad basadas en estructuras de líneas diagonales en RP y acuñaron el nombre de análisis de cuantificación de recurrencias *(recurrence quantification analysis, RQA)*.

En el libro [30] se hace un compendio de la teoría y aplicaciones del RQA. El primer trabajo, *Mathematical and Computational Foundations of Recurrence Quantifications*, escrito por los responsables de la recopilación incluye una puesta al día de esas técnicas. En esta introducción seguimos las líneas de esa exposición.

La recurrencia de un estado en el momento i en un momento diferente j se representa dentro de una matriz cuadrada bidimensional $\mathbf{R}$ con puntos, donde ambos ejes son ejes de tiempo [22]:

$$R_{i,j}^{m,\varepsilon_i} = \mathcal{H}\left(\varepsilon_i - \|x_i - x_j\|\right), \quad x_i \in \mathbb{R}^m, \quad i,j = 1...n, \tag{1.1}$$

donde n es el número de estados considerados x_i; ε_i es un umbral de distancia, $\|\cdot\|$ una norma, y $\mathcal{H}(\cdot)$ la función de Heaviside.

La función anterior define por lo tanto una matriz simétrica formada por unos y ceros de acuerdo a si se supera o no el umbral de distancia definido. Lo importante de esta matriz es que tiene su equivalente en el gráfico donde cada valor uno de la matriz le va a corresponder un punto en el RP como por ejemplo observamos en la Figura 1.1 (b). Es a partir de los diferentes patrones de puntos es que se logra caracterizar las trayectorias temporales.

Dado que $R_{i,i} = 1 \quad (i = 1...n)$ por definición, el RP tiene una línea diagonal principal negra, llamada la *línea de identidad*, con un ángulo $\pi/4$. Debe notarse que un punto de recurrencia aislado (i,j) no contiene ninguna información sobre los estados actuales en momento i y j. Sin embargo, del total de puntos recurrentes es posible reconstruir la trayectoria del espacio de fase (ver página 7, [30]).

En la práctica no es útil y casi es imposible encontrar recurrencias completas. Por lo tanto, una recurrencia es definida cuando un estado x_j está lo suficientemente cercano a x_i. Esto significa que aquellos estados x_j que caen en un entorno m-dimensional de tamaño ε_i centrado en x_i son recurrentes.

Estos x_j se denominan *puntos recurrentes*. En la Ec. (1.1), esto es expresado por medio de la función de Heaviside y su argumento ε_i.

En la definición inicial de los RP, el entorno es una bola, es decir, se utiliza la norma L_2 y el radio es elegido de forma de contener un número fijo de estados cercanos x_j [22]. Con este entorno, el radio ε_i cambia para cada $x_i \quad (i = 1...n)$ y puede ocurrir que $R_{i,j} \neq R_{j,i}$ porque el entorno de x_i no tiene que ser similar al del x_j. Esta propiedad origina un RP asimétrico, pero todas las columnas del RP tienen la misma densidad de recurrencias. Sin embargo, el entorno más utilizado es con un radio fijo $\varepsilon_i = \varepsilon, \forall i$. Un radio fijo asegura que $R_{i,j} = R_{j,i}$, es decir, un RP simétrico.

El umbral de recurrencia ε es un parámetro crucial en el análisis del RP. A pesar de que varios trabajos han contribuido a esta discusión (ver página 8, [30]), aún se necesita un estudio general y sistemático para la selección del umbral de recurrencia.

La primer variable en RQA es definida en 1994 en el trabajo [31] y es denominada *porcentaje de recurrencia (REC)* o *tasa de recurrencia (RR)*

$$RR(\varepsilon, n) = \frac{1}{n^2 - n} \sum_{i \neq \mathrm{J}=1}^{n} R_{i,j}^{m,\varepsilon}. \tag{1.2}$$

RR calcula el porcentaje de puntos negros sobre el total de puntos en el RP excluyendo la línea diagonal principal (*LOI*, line of identity). Es una medida de la densidad relativa de los puntos recurrentes en una matriz esparsa y está relacionada con la definición de suma de correlación [22]. En el límite de series de tiempo largas

$$P = \lim_{n \to \infty} RR(\varepsilon, n), \tag{1.3}$$

es la probabilidad de encontrar un punto recurrente dentro del RP.

1.4. Análisis de recurrencia bivariado

El análisis de recurrencia bivariado permite el estudio de correlaciones y sincronizaciones entre sistemas dinámicos.

Si nos preguntamos si dos sistemas tienen una estructura de recurrencia similar, es decir, si sus estados se repiten de manera simultánea, utilizaremos el *análisis de recurrencia conjunto* (ver página 40, [30].) Consideramos las

recurrencias de las trayectorias de los dos sistemas en sus respectivos espacios de fases por separado y buscamos los casos en que ambos recurren de manera simultánea, es decir, cuando ocurre una recurrencia conjunta. La matriz de recurrencia conjunta para dos sistemas $\mathbf{x}$ y $\mathbf{y}$ es

$$JR_{i,j}^{x,y}\left(\varepsilon^{x},\varepsilon^{y}\right)=\mathcal{H}\left(\varepsilon^{x}-\|x_{i}-x_{j}\|\right)\mathcal{H}\left(\varepsilon^{y}-\|y_{i}-y_{j}\|\right),\quad i,j=1,...,n. \tag{1.4}$$

En esta aproximación, una recurrencia ocurre si un punto x_j en la primer trayectoria retorna al entorno de un punto anterior x_i, y simultáneamente el punto y_j en la segunda trayectoria retorna a un entorno de un punto anterior y_i.

El análisis de recurrencia conjunto puede ser utilizado para estimar probabilidades conjuntas y condicionales [22]. Supongamos que tengo dos sistemas. Si dos vectores del espacio de fases del segundo sistema en i y j son cercanos (puntos negros) y si dos vectores del espacio de fase del primer sistema en los mismos i y j son también cercanos (puntos negros), tenemos un punto negro en el JRP en la ubicación (i, j).

La tasa de recurrencia del JRP se calcula con la siguiente ecuación:

$$RR(\varepsilon^{1},\varepsilon^{2})=\frac{1}{n^{2}}\sum_{i,j=1}^{n}\prod_{k=1}^{2}R_{i,j}^{m,\varepsilon^{k}}. \tag{1.5}$$

1.5. Motivaciones

Dentro de la rica literatura del análisis de recurrencia hay varios autores [34, 33] que analizan la relación de dependencia entre variables pero no se aplica una prueba de hipótesis en ninguno de ellos. Recientemente en [18, 17, 19], se propone un test de independencia basado en análisis de recurrencia y se aplica a datos reales.

En teoría de probabilidades, se dice que dos sucesos aleatorios son independientes entre sí cuando la probabilidad de cada uno de ellos no está influida porque el otro suceso ocurra o no, es decir, cuando ambos sucesos no están relacionados.

Los test de hipótesis presentados en esta tesis fueron diseñados en base al cálculo de probabilidades marginales y conjuntas donde los sucesos refieren a la presencia de recurrencias asintóticas en las trayectorias de los sistemas. Los

cálculos de las probabilidades marginales y conjuntas se basan en las ecuaciones (1.2) y (1.5) respectivamente que son reescritas en el Capítulo 2 para una mejor interpretación en la elaboración del test de hipótesis propuesto.

Siguiendo el estudio realizado en [18], el objetivo de esta tesis es la presentación de un test de independencia basado en porcentaje de recurrencias. Se estudian sus propiedades matemáticas y se muestra a través de simulaciones su comportamiento mediante un estudio de potencias y se lo compara con otros tests propuestos en la literatura. Por último, siguiendo a [17, 19] se aplica el test propuesto a datos reales.

Para cumplir el objetivo propuesto la tesis se organiza de la siguiente manera. En el Capítulo 2, se muestra la idea en la cual está basado el test y se enuncian las propiedades teóricas que tiene. También se muestra cómo puede ser implementado el test. En el Capítulo 3, se muestra un estudio de simulación sobre el comportamiento del test para distintos escenarios posibles de dependencia entre X e Y y para distintas dimensiones, cubriendo el caso en el cual X e Y son variables aleatorias, vectores alestarios o series de tiempo, tanto a tiempo discreto como continuo. En el Capítulo 4, se aplica el test a datos reales de tipo económico y meteorológico. En el Capítulo 5, se plantean las principales conclusiones. Finalmente en el Apéndice 1 se presenta la demostración de los resultados enunciados en el Capítulo 2 mientras que en el Anexo 1 se brindan los códigos en R utilizados para obtener los resultados a datos reales.

Capítulo 2

Formulación e implementación del Test de Independencia

Siguiendo el estudio realizado en [18], el objetivo de este capítulo es la presentación de un nuevo test para detectar dependencia entre dos elementos aleatorios X e Y, basado en el análisis de recurrencia [30].

Dada una muestra $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ i.i.d. de (X, Y), donde $X \in S_X$ e $Y \in S_Y$, siendo S_X y S_Y espacios métricos cualesquiera. Nos planteamos realizar el test H_0 : X e Y son independientes versus H_1 : X e Y no son independientes. En el test propuesto, X e Y pueden tomar valores en cualquier espacio métrico. Por lo tanto el test puede ser utilizado para analizar si X e Y son independientes en el caso en el cual X e Y son variables aleatorias, vectores aleatorios o series de tiempo.

2.1. Formulación del test y propiedades teóricas

Sea $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ i.i.d. una muestra de (X, Y) donde $X \in S_X$, $Y \in S_Y$, S_X y S_Y son espacios métricos, y dado $r, s > 0$. Para simplificar la notación y sin riesgo de confusión, se utilizará la misma letra d para la misma función de distancia en ambos espacios métricos S_X y S_Y.

Para una mejor interpretación se reescribe la ecuación (1.2) y se define la tasa de recurrencia para las muestras de X e Y como

$$RR_n^X(r) := \frac{1}{n^2-n}\sum_{i\neq j}\mathbf{1}_{\{d(X_i,X_j)<r\}}, \qquad RR_n^Y(s) := \frac{1}{n^2-n}\sum_{i\neq j}\mathbf{1}_{\{d(Y_i,Y_j)<s\}},$$

respectivamente. Estas tasas de recurrencia como se presentó en el Capítulo 1, son una medida de la densidad relativa de puntos recurrentes sobre el total de puntos de los RP formados a partir de las muestras de X e Y.

También se reescribe la ecuación (1.5) definiendo la tasa de recurrencia conjunta para (X, Y) como

$$RR_n^{X,Y}(r,s) := \frac{1}{n^2-n}\sum_{i\neq j}\mathbf{1}_{\{d(X_i,X_j)<r\ ,\ d(Y_i,Y_j)<s\}}.$$

Se define como $p_X(r) := P\left(d\left(X_1, X_2\right) < r\right)$ la probabilidad de que la distancia entre dos elementos cualquiera de la muestra X sea menor que r. De forma similar, se define la probabiliad entre tres puntos como $p_X^{(3)}(r) := P\left(d\left(X_1, X_2\right) < r,\ d\left(X_1, X_3\right) < r\right)$ y de forma análoga p_Y y $p_Y^{(3)}$.

Es necesario definir también

$p_{X,Y}(r,s) := P\left(d\left(X_1, X_2\right) < r,\ d\left(Y_1, Y_2\right) < s\right).$

La ley fuerte de los grandes números para estadísticas-U ([15]) permite afirmar que para cualquier $r, s > 0$, la convergencia es casi segura (a.s.).

$$RR_n^X(r) \overset{a.s.}{\to} p_X(r), \quad RR_n^Y(s) \overset{a.s.}{\to} p_Y(s) \text{ and } RR_n^{X,Y}(r,s) \overset{a.s.}{\to} p_{X,Y}(r,s). \quad (2.1)$$

Queremos testear H_0 : X e Y son independientes, contra H_1 : H_0 no se cumple.

Si H_0 es cierta, entonces $p_{X,Y}(r,s) = p_X(r)p_Y(s)$ para todos $r, s > 0$, por lo que se espera que si n es grande, $RR_n^{X,Y}(r,s) \cong RR_n^X(r)RR_n^Y(s)$ para todos $r, s > 0$. Entonces, se propone construir el test estadístico a partir del proceso-U $\{E_n(r,s)\}_{r,s>0}$ defindo como

$$E_n(r,s) := \sqrt{n}\left(RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s)\right). \quad (2.2)$$

Por lo tanto, es natural rechazar H_0 cuando $T_n^{(2)} > c$ siendo

$$T_n^{(2)} := n\int_0^{+\infty}\int_0^{+\infty}\left(RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s)\right)^2 dG(r,s), \quad (2.3)$$

donde c es una constante y G es una función de distribución fijada de antemano.

Se utiliza la notación ϕ y φ para la función de distribución y densidad de una variable aleatoria $N(0,1)$, respectivamente, y para cada m definimos los conjuntos

$$I_m^n := \{(i_1, ..., i_m) : \ i_j \neq i_k \ j \neq k, \ i_j \in \{1, ..., n\} \ \ j \in \{1, ..., m\}\}.$$

Una ventaja del test es que en lugar de escoger los valores apropiados r y s, se utiliza la información generada por ambas muestras para todos los valores posibles de r y s.

A continuación se formulan los resultados asintóticos del test estadístico cuyas demostraciones se presentan en el Apéndice 1. Primero, se formula un resultado que garantiza la distribución asintótica de $T_n^{(2)}$ bajo H_0. También se presenta un resultado que establece la consistencia del estadístico bajo una amplia clase de alternativas. En segundo lugar, se analiza el sesgo asintótico bajo alternativas contiguas.

2.1.1. Resultados asintóticos bajo H_0 y consistencia

Se comienza con el siguiente lema, en el que se obtiene la fórmula para la función de autocovarianza asintótica del proceso $\{E_n(r,s)\}_{r,s>0}$ bajo H_0.

Lema 1. *Dados $r, r', s, s' > 0$, y $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, i.i.d. en $S_X \times S_Y$, donde X y Y son independientes. Entonces*

$$\begin{aligned} &\lim_{n\to+\infty} Cov\left(E_n(r,s), E_n(r',s')\right) = \\ &= 4\left(p_X^{(3)}(r \wedge r') - p_X(r)p_X(r')\right)\left(p_Y^{(3)}(s \wedge s') - p_Y(s)p_Y(s')\right). \end{aligned} \tag{2.4}$$

El siguiente lema será útil para reducir la convergencia asintótica del proceso $\{E_n(r,s)\}_{r,s>0}$ a la convergencia de un proceso-U que lo aproxima que se llamará $\{E'_n(r,s)\}_{r,s>0}$ y se define como

$$E'_n(r,s) := \frac{\sqrt{n}}{n(n-1)(n-2)(n-3)} \times$$

$$\sum_{(i,j,k,h)\in I_4^n} \left(\mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_j)<s\}} - \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}}\right). \tag{2.5}$$

Lema 2. *Dados* $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ *i.i.d. en* $S_X \times S_Y$*, entonces*

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

donde

$$0 \leq H_n(r,s) \leq \frac{4}{\sqrt{n}} \text{ para todos } r,s>0.$$

El próximo teorema prueba la convergencia del proceso $\{E_n\}$.

Teorema 3. *Dada la muestra* $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ *i.i.d. en* $S_X \times S_Y$*. Si las funciones de distribución de* $d(X_1,X_2)$ *y* $d(Y_1,Y_2)$ *son continuas, entonces*

$$\{E_n(r,s) - E(E_n(r,s))\}_{r,s>0} \xrightarrow{w} \{E(r,s)\}_{r,s>0}, \tag{2.6}$$

donde $\{E(r,s)\}_{r,s>0}$ *es un proceso Gausiano centrado.*

Observación 1. *Se observa que el proceso* $\{E_n(r,s)\}_{r,s>0}$ *se encuentra en* $L^2(dG)$ *(porque* G *es una medida de probabilidad). Por lo tanto, el estadístico* $T_n^{(2)}$ *es igual a* $||\{E_n(r,s)\}_{r,s>0}||$ *entonces dado que el funcional es continuo,* $T_n^{(2)}$ *convergerá a* $||\{E(r,s)\}_{r,s>0}||$*.*

Observación 2. *Dados* $r,s>0$ *y* $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n) \in \mathbb{R}^2$ *muestra i.i.d. de* (X,Y) *donde las marginales* X,Y *son* $N(0,1)$ *independientes. Entonces*

$$\sqrt{n}\left(RR_n^{X,Y}(r,s) - RR_n(r)RR_n^Y(s)\right) \xrightarrow{w} N\left(0,\sigma_{X,Y}^2(r,s)\right),$$

donde

$$\sigma_{X,Y}^2(r,s) = 4\left(\int_{-\infty}^{+\infty}\left(\phi(x+r)-\phi(x-r)\right)^2\varphi(x)\,dx - \left(2\phi\left(r/\sqrt{2}\right)-1\right)^2\right)$$

$$\times \left(\int_{-\infty}^{+\infty} \left(\phi\left(x+s\right) - \phi\left(x-s\right) \right)^2 \varphi\left(x\right) dx - \left(2\phi\left(s/\sqrt{2} \right) - 1 \right)^2 \right). \tag{2.7}$$

El siguiente teorema prueba que si $d(X_1, X_2)$ y $d(Y_1, Y_2)$ son no independientes, entonces el test es consistente.

Teorema 4. *Dada la muestra* $\left(X_1, Y_1\right), \left(X_2, Y_2\right), ..., \left(X_n, Y_n\right)$ *i.i.d. en* $S_X \times S_Y$. *Si* $dG(r,s) = g(r,s)drds$, $g(r,s) > 0$ *para todo* $r, s > 0$, *y* $d\left(X_1, X_2\right)$, $d\left(Y_1, Y_2\right)$ *son variables aleatorias continuas y no independientes, luego* $T_n^{(2)} \xrightarrow{P} +\infty$ *cuando* $n \to +\infty$.

El siguiente corolario surge del Teorema 4.

Corolario 1. *Si* $\left(X, Y\right) \sim N\left(0, \Sigma\right)$, *donde* X *e* Y *son no independientes, y* $dG(r,s) = g(r,s)drds$, $g(r,s) > 0$ *para todos* $r, s > 0$, *entonces* $T_n^{(2)} \xrightarrow{P} +\infty$ *cuando* $n \to +\infty$.

Observación 3. *Si* $\left(X_1, Y_1\right)$, $\left(X_2, Y_2\right)$ *en* $\mathbb{R}^2$ *i.i.d. con densidad conjunta* $f_{X,Y}$ *y distribución conjunta* F *tal ques* $|X_1 - X_2|$ *y* $|Y_1 - Y_2|$ *son independientes, entonces*

$$\alpha\left(r,s\right) := P\left(|X_1 - X_2| \leq r,\ \ |Y_1 - Y_2| \leq s\right) =$$

$$= \iint_{\mathbb{R}^2} f_{X,Y}(x_1, y_1) dx_1 dy_1 \int_{x_1 - r}^{x_1 + r} dx_2 \int_{y_1 - s}^{y_1 + s} f_{X,Y}(x_2, y_2) dy_2$$

$$= \iint_{\mathbb{R}^2} P\left(x_1 - r \leq X_1 \leq x_1 + r,\ \ y_1 - s \leq Y_2 \leq y_1 + s\right) f_{X,Y}(x_1, y_1) dx_1 dy_1$$

$$= E\left(F\left(X + r, Y + s\right) - F\left(X + r, Y - s\right)\right)$$

$$-E\left(F\left(X - r, Y + s\right) + F\left(X - r, Y - s\right)\right).$$

De forma similiar,

$$\beta\left(r,s\right) := P\left(|X_1 - X_2| \leq r\right) P\left(\ \ |Y_1 - Y_2| \leq s\right)$$

$$= E\left(F_X\left(X+r\right) - F_X\left(X-r\right)\right) E\left(F_Y\left(Y+r\right) - F_Y\left(Y-r\right)\right).$$

Luego, $\alpha\left(r,s\right) = \beta\left(r,s\right)$ *para todos* $r,s>0$.

La condición $\alpha\left(r,s\right) = \beta\left(r,s\right)$ *para todo* $r,s>0$ *equivale a la independencia entre las variables* $|X_1 - X_2|$ *e* $|Y_1 - Y_2|$ *siendo* X_1, X_2 *independientes con distribución como la de* X *e* Y_1, Y_2 *independientes con distribución como la de* Y *lo que no asegura que* X *e* Y *sean necesariamente independientes. Las variables* X *e* Y *que verifiquen la condición* $\alpha\left(r,s\right) = \beta\left(r,s\right)$ *para todos* $r,s>0$ *pero que no sean independientes no verificarán las hipótesis del teorema por lo que no se puede garantizar la consistencia del test en estos casos.*

2.1.2. Alternativas contiguas

En esta subsección se analiza el comportamiento del test bajo alternativas contiguas. Más explícitamente, dada la muestra $\left(X_1,Y_1\right), \left(X_2,Y_2\right), ..., \left(X_n,Y_n\right)$ i.i.d. en $\mathbb{R}^p \times \mathbb{R}^q$, consideramos

$$H_0 : f_{X,Y}(x,y) = f_X(x) f_Y(y) \quad \text{para todo } (x,y)\, vs$$

$$H_n : f_{X,Y}(x,y) = f^{(n)}_{X,Y}(x,y) \quad \text{para todo } (x,y)$$

donde $f^{(n)}_{X,Y}(x,y) = c_n\left(\delta\right) f_X(x) f_Y(y) \left(1 + \frac{\delta}{2\sqrt{n}} k_n(x,y)\right)^2$, $\delta > 0$, $c_n\left(\delta\right)$ es una constante tal que $f^{(n)}_{X,Y}(x,y)$ es una densidad, y las funciones k_n verifican las condiciones (i) y (ii) que se dan a continuación:

Se define $L^2_0 = L^2\left(dF_0\right)$ para $dF_0(x,y) = f_X(x) f_Y(y) dxdy$, la función de distribución de (X,Y) bajo H_0, análogamente se define L^1_0;

(i) Existe una función $K \in L^1_0$ tal que $k_n \leq K$ para todo n;
(ii) Existe $k \in L^2_0$ tal que $k_n \xrightarrow{L^2_0} k$, $\|k\| = 1$.

Puede probarse que las condiciones (i) y (ii) implican contiguidad (Cabaña [6]). El coeficiente δ se introduce para permitir la normalización $\|k\| = 1$. La función δk se llama apartamiento asintótico.

Se muestra en las siguientes líneas que bajo H_n, el proceso $\{E_n(r,s)\}_{r,s>0}$ tiene el mismo límite asintótico que bajo H_0 más un sesgo determinístico.

Se utiliza la notación $E^{(n)}(T)$ y $P^{(n)}((X,Y) \in A)$ para el valor esperado de T, y la probabilidad del conjunto $\{(X,Y) \in A\}$ bajo H_n respectivamente. De forma análoga se utiliza $E^{(0)}(T)$ y $P^{(0)}((X,Y) \in A)$ bajo H_0.

Proposición 1. *Bajo* H_n

$$E^{(n)}(E_n(r,s)) \to \delta\mu(r,s) \quad \textit{cuando } n \to +\infty \quad \textit{para todos } r,s>0,$$

donde

$$\mu(r,s) = \iiiint_{A_{r,s}} (k(x_1,y_1)+k(x_2,y_2))\, f_X(x_1)f_Y(y_1)f_X(x_2)f_Y(y_2)dx_1dx_2dy_1dy_2, \tag{2.8}$$

$$A_{r,s} := \{(x_1,y_1,x_2,y_2) \in \mathbb{R}^{2p+2q} : \ d(x_1,x_2) < r, \ d(y_1,y_2) < s\}.$$

Con un poco más de trabajo, utilizando el tercer lema de Le Cam (Le Cam y Yang [21] y Oosterhoff y Van Zwet [23]) es posible probar que bajo H_n,

$$\{E_n(r,s)\}_{r,s>0} \xrightarrow{w} \{E(r,s)+\delta\mu(r,s)\}_{r,s>0},$$

donde $\{E(r,s)\}_{r,s>0}$ es el límite del proceso bajo H_0 y

$$\mu(r,s) = \iiiint_{A_{r,s}} (k(x_1,y_1)+k(x_2,y_2))\, f_X(x_1)f_Y(y_1)f_X(x_2)f_Y(y_2)dx_1dx_2dy_1dy_2.$$

Por lo tanto, bajo H_n tendremos que

$$T_n^{(2)} \xrightarrow{w} \int_0^{+\infty}\int_0^{+\infty} (E(r,s)+\delta\mu(r,s))^2\, dG(r,s).$$

2.2. Implementación del test

2.2.1. X e Y son variables aleatorias

En el caso donde X e Y son variables aleatorias continuas, se observa que decir que X e Y son independientes es equivalente a decir que $X' = \phi^{-1}(F_X(X))$ e $Y' = \phi^{-1}(F_Y(Y))$ son independientes, donde F_X y F_Y son las funciones de distribución de X e Y, respectivamente. Si se aplica el procedimiento del test a X' e Y', luego se tiene la ventaja que ahora las variables están en la misma escala y cada una tiene una distribución normal centrada que se aproxima a las hipótesis de la Observación 2. Otra ventaja adicional es

que bajo H_0 (X' e Y' son independientes y $N(0,1)$), para valores pequeños de n, es posible calcular los valores críticos al 5 % u otro nivel porque se conoce la distribución límite de $T_n^{(2)}$ bajo H_0. Cuando X e Y son vectores aleatorios, se puede aplicar la misma transformación en cada coordenada.

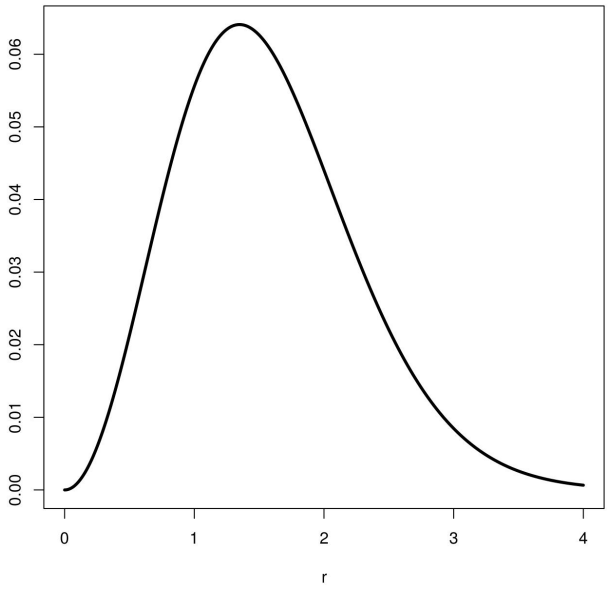

Figura 2.1: $\sigma^2_{X,Y}(r,r)$ (varianza asintótica del proceso $\{E_n(r,r)\}_{r>0}$) en función de r, en el caso X e Y son independientes y $N(0,1)$.

Para dar una idea de la variabilidad del proceso $\{E_n(r,s)\}_{r,s>0}$, en la Figura 2.1 se muestran los valores de $\sigma^2_{X',Y'}(r,r)$ para diferentes valores de r. El máximo es 0.064 y se alcanza cuando $r = 1.35$.

2.2.2. Caso general

En muchas aplicaciones estadísticas, se tiene un tamaño de muestra pequeño. Entonces, puede tomarse una decisión errónea si el investigador utiliza los p-valores (o los valores críticos) obtenidos a través de la distribución asintótica para tomar la decisión en el test de hipótesis. Por lo tanto, cuando se tiene una muestra de tamaño n, es preferible estimar el p-valor (o el valor crítico) estimando la distribución de $T_n^{(2)}$ para dicho valor de n. Además, la distribución asintótica de $T_n^{(2)}$ es difícil de obtener porque se necesita realizar muchas simulaciones de un proceso Gausiano continuo centrado indexado en $D = (0,+\infty) \times (0,+\infty)$ y luego, calcular la integral en D.

Para calcular el p-valor o el valor crítico del test para n fijo podemos proceder como se explica a continuación en las siguientes líneas. Dado n, si H_0 es cierta, no se conoce la distribución de $T_n^{(2)}$, pero dado el valor observado a partir de la muestra que denominamos t_{obs}, se puede generar, mediante un procedimiento de permutación, una muestra grande de $T_n^{(2)}$ con la cual es posible estimar $P\left(T_n^{(2)} \geq t_{obs}\right)$. Dado la muestra $(X_1, Y_1), ..., (X_n, Y_n)$ i.i.d. de (X, Y). Se observa que la distribución de T_n depende de la distribución conjunta de $(X_1, Y_1), ..., (X_n, Y_n)$. Si H_0 es cierta, y si se considera cualquier $\sigma : \{1, ..., n\} \to \{1, ..., n\}$ función de permutación, luego la distribución conjunta de $(X_1, Y_1), ..., (X_n, Y_n)$ y la distribución conjunta de $\left(X_{\sigma(1)}, Y_1\right), ..., \left(X_{\sigma(n)}, Y_n\right)$ son la misma. Consideremos $S(n) = \{\sigma_1, ..., \sigma_{n!}\}$ el conjunto de todas las permutaciones $\sigma : \{1, ..., n\} \to \{1, ..., n\}$. Supongamos que la muestra $(X_1, Y_1), ..., (X_n, Y_n)$ es fija y consideremos Z definida por $Z = T_n\left(\left(X_{\sigma_i(1)}, Y_1\right), ..., \left(X_{\sigma_i(n)}, Y_n\right)\right)$ con probabilidad $1/n!$ para cada $i \in \{1, ..., n!\}$. Si se toma $Z_1, ..., Z_m$, i.i.d., muestra de Z, es posible estimar el valor de $p_n = P(T_n \geq t_{obs})$ simplemente utilizando $\widehat{p}_n^{(m)} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\{Z_i \geq t_{obs}\}}$ para m suficientemente grande. Se definen las variables aleatorias $B_i = \sum_{j=1}^{m} \mathbf{1}_{\left\{Z_j = T_n\left(\left(X_{\sigma_i(1)}, Y_1\right), ..., \left(X_{\sigma_i(n)}, Y_n\right)\right)\right\}}$. para $i \in \{1, ..., n!\}$. Se observa que B_i tiene distribución $\text{Bin}(m, 1/n!)$ para cada $i \in \{1, ..., n!\}$. Luego

$$\widehat{p}_n^{(m)} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}_{\{Z_j \geq t_{obs}\}} = \frac{1}{m}\sum_{i=1}^{n!} B_i \mathbf{1}_{\left\{T_n\left(\left(X_{\sigma_i(1)}, Y_1\right), ..., \left(X_{\sigma_i(n)}, Y_n\right)\right) \geq t_{obs}\right\}}$$

converge cuando $m \to +\infty$ hacia $\frac{1}{n!}\sum_{i=1}^{n!} \mathbf{1}_{\left\{T_n^{(2)}\left(\left(X_{\sigma_i(1)}, Y_1\right), ..., \left(X_{\sigma_i(n)}, Y_n\right)\right) \geq t_{obs}\right\}}$ casi seguramente. Si ahora se considera que $(X_1, Y_1), ..., (X_n, Y_n)$ son elementos aleatorios tales que exista su valor esperado, entonces se obtiene $E\left(\widehat{p}_n^{(m)}\right) \underset{m \to +\infty}{\to} p_n$, luego $\widehat{p}_n^{(m)}$ es un estimador asintóticamente insesgado de p_n.

2.2.3. Un método simple para obtener la función de pesos

La performance del test depende de la elección de la función de pesos. La función de pesos puede elegirse por el investigador en cada caso particular. De acuerdo al Teorema 4, se puede utilizar cualquier función G tal que $dG(r, s) = g(r, s)drds$ donde $g(r, s) > 0$ para cualquier $r, s > 0$. Sería interesante estudiar

algún tipo de optimalidad en la elección de la función G, bajo cierto tipo de alternativas, pero se propone en esta subsección un método simple para escoger la función G. Como se analiza en la próxima sección, esta elección simple de G, tiene muy buena performance bajo las alternativas estudiadas en esta tesis.

Se define $dG(r,s) = g_1(r)g_2(s)drds$, donde g_1 y g_2 son densidades Gausianas. En el caso de g_1 se puede utilizar $\mu_1 = E\left(d(X_1, X_2)\right)$ y $\sigma_1^2 = V\left(d(X_1, X_2)\right)$. Los valores de μ_1 y σ_1 pueden ser fácilmente estimados mediante la muestra $d\left(X_i, X_j\right)$ con $(i,j) \in I_2^n$. Se puede proceder de forma similar para μ_2 y σ_2. De esta forma, damos más peso en las cercanías de la distancia promedio entre dos observaciones independientes X_1 y X_2 para g_1, y análogamente para g_2. Observar que es posible evitar el problema de escoger G, si utilizamos $T_n' = ||\{E_n(r,s)\}_{r,s>0}||_\infty = \sqrt{n}\sup_{r,s>0}\left|RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s)\right|$ para testear independencia porque todos los resultados teóricos obtenidos para $T_n^{(2)}$ son aún válidos para T_n'.

2.2.4. Cálculo del estadístico

En esta subsección se ve como calcular el estadístico $T_n^{(2)}$. Se considera el caso en que $dG(r,s) = g_1(r)g_2(s)drds$, donde g_1 y g_2 son funciones de densidad con G_1 y G_2 como sus funciones de distribución respectivamente.

$$\begin{aligned}
&\int_0^{+\infty}\int_0^{+\infty}\left(RR_n^{X,Y}(r,s) - RR_n^X(r)\,RR_n^Y(s)\right)^2 g_1(r)\,g_2(s)\,drds \\
&\qquad = \int_0^{+\infty}\int_0^{+\infty}\left[RR_n^{X,Y}(r,s)\right]^2 g_1(r)\,g_2(s)\,drds \\
&\qquad + \int_0^{+\infty}\left[RR_n^X(r)\right]^2 g_1(r)\,dr \int_0^{+\infty}\left[RR_n^Y(s)\right]^2 g_2(s)\,ds \\
&-2\int_0^{+\infty}\int_0^{+\infty} RR_n^{X,Y}(r,s)\,RR_n^X(r)\,RR_n^Y(s)\,g_1(r)\,g_2(s)\,drds := A_n + B_n - 2C_n.
\end{aligned} \tag{2.9}$$

Para simplificar la notación y para el resto de esta subsección, se denomina $N = n(n-1)$. También se indexa $d\left(X_i, X_j\right)$ con $(i,j) \in I_2^n$ en la forma $Z_1, Z_2, ..., Z_N$. Análogamente, se definen los $T_1, T_2, ..., T_N$ a los valores $d\left(Y_i, Y_j\right)$ utilizando la misma indexación que la de los $Z's$. Se denomina $Z_1^*, Z_2^*, ..., Z_N^*$ estadísticos de orden de $Z's$, y análogamente $T_1^*, T_2^*, ..., T_N^*$.

Se verá ahora como calcular A_n, B_n y C_n.

$$\int_0^{+\infty} \left[RR_n^X(r)\right]^2 g_1(r)\,dr = \frac{1}{N^2}\sum_{i\neq j}\sum_{k\neq h}\int_0^{+\infty} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(X_h,X_k)<r\}} g_1(r)\,dr$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_0^{+\infty} \mathbf{1}_{\{Z_i<r,\ Z_j<r\}} g_1(r)\,dr$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(1-G_1\left(\text{máx}\left\{Z_i,Z_j\right\}\right)\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(1-G_1\left(\text{máx}\left\{Z_i^*,Z_j^*\right\}\right)\right)$$

$$= 1-\frac{1}{N^2}\sum_{i=1}^{N}\left(2\sum_{j=1}^{i-1}G_1\left(Z_i^*\right)+G_1\left(Z_i^*\right)\right)$$

$$= 1-\frac{1}{N^2}\sum_{i=1}^{N}\left(2\left(i-1\right)G_1\left(Z_i^*\right)+G_1\left(Z_i^*\right)\right) = 1-\frac{1}{N^2}\sum_{i=1}^{N}\left(2i-1\right)G_1\left(Z_i^*\right).$$

Análogamente,

$$\int_0^{+\infty} \left[RR_n^Y(s)\right]^2 g_2(s)\,ds = 1-\frac{1}{N^2}\sum_{i=1}^{N}\left(2i-1\right)G_2\left(T_i^*\right).$$

Luego,

$$A_n = \int_0^{+\infty}\int_0^{+\infty} \left[RR_n^{X,Y}(r,s)\right]^2 g_1(r)\,g_2(s)\,dr ds$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_0^{+\infty} \mathbf{1}_{\{Z_i<r,\ Z_j<r\}} g_1(r)\,dr \int_0^{+\infty} \mathbf{1}_{\{T_i<s,\ T_j<s\}} g_2(s)\,ds$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(1-G_1\left(\text{máx}\left\{Z_i,Z_j\right\}\right)\right)\left(1-G_2\left(\text{máx}\left\{T_i,T_j\right\}\right)\right), \qquad (2.10)$$

$$B_n = \left(1 - \frac{1}{N^2}\sum_{i=1}^{N}(2i-1)\,G_1\left(Z_i^*\right)\right)\left(1 - \frac{1}{N^2}\sum_{i=1}^{N}(2i-1)\,G_2\left(T_i^*\right)\right), \quad (2.11)$$

$$C_n = \int_0^{+\infty}\int_0^{+\infty} RR_n^{X,Y}(r,s)\,RR_n^X(r)\,RR_n^Y(s)\,g_1(r)\,g_2(s)\,drds$$

$$= \frac{1}{N^3}\sum_{i\neq j}\sum_{k\neq h}\sum_{l\neq m}\int_0^{+\infty}\int_0^{+\infty} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_j)<s,\ d(X_h,X_k)<r,\ d(Y_l,Y_m)<s\}} g_1(r)\,g_2(s)\,drds$$

$$= \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(1 - G_1\left(\max\left\{Z_i, Z_j\right\}\right)\right)\left(1 - G_2\left(\max\left\{T_i, T_k\right\}\right)\right). \quad (2.12)$$

Entonces

$$T_n^{(2)} = n(A_n + B_n - 2C_n), \quad (2.13)$$

donde A_n, B_n y C_n están dados por las fórmulas (2.10), (2.11) y (2.12) respectivamente.

2.3. Otros posibles estadísticos para utilizar en el test

Vimos que cuando se testea $H_0 : X$ e Y son independientes, contra $H_1 :$ H_0 no lo son, se propone rechazar H_0 cuando $T_n^{(2)} > c$, donde

$$T_n^{(2)} := n\int_0^{+\infty}\int_0^{+\infty}\left(RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s)\right)^2 dG(r,s), \quad (2.14)$$

donde c es una constante y G es una función de distribución elegida de forma adecuada. Se observa que $T_n^{(2)}$ es un funcional del tipo L^2 Cramér–von Mises aplicado al proceso $\{E_n(r,s)\}_{r,s>0}$ donde

$$E_n(r,s) := \sqrt{n}\left(RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s)\right). \quad (2.15)$$

Los resultados teóricos planteados sobre el proceso definido en (2.15), son válidos para cualquier función de distancia d_X y d_Y, y se mantienen válidos si se considera otros funcionales continuos tales como los del tipo L^1-Cramér–von Mises o del tipo Kolmogorov–Smirnov.

Si $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ son i.i.d. en $S_X \times S_Y$, se compara el estadístico $T_n^{(2)}$, con los estadísticos definidos como

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

y

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

Se observa que en el caso general en que X e Y yacen en espacios métricos (S_X, d_X) y (S_Y, d_Y), los estadísticos $T_n^{(1)}, T_n^{(2)}$ y $T_n^{(\infty)}$ dependen de las funciones de distancia d_X y d_Y. En la sección 3.2 se compara la potencia bajo varias pruebas alternativas basado en $T_n^{(1)}, T_n^{(2)}$ y $T_n^{(\infty)}$ para diferentes funciones de distancia d_X y d_Y.

En el caso en que X e Y son series de tiempo discretas, utilizaremos las distancias clásicas l^1, l^2 y l^∞, esto es, $d_X(x, x') = \sum_{n\geq 1} |x_n - x'_n|$, $d_X(x, x') = \sqrt{\sum_{n\geq 1} (x_n - x'_n)^2}$ y $d_X(x, x') = \sup_{n\geq 1} |x_n - x'_n|$ respectivamente y análogamente para d_Y. De forma análoga, cuando X e Y son series de tiempo continuas, se utiliza las distancias clásicas L^1, L^2, L^∞, esto es , $d_X(x, x') = \int_{-\infty}^{+\infty} |x(t) - x'(t)|\, dt$, $d_X(x, x') = \sqrt{\int_{-\infty}^{+\infty} (x(t) - x'(t))^2\, dt}$ y $d_X(x, x') = \sup_{t\in\mathbb{R}} |x(t) - x'(t)|$ respectivamente. Se utiliza la notación $T_n^{(i,j)}$ donde $i, j = 1, 2, \infty$ para el estadístico $T_n^{(i)}$ donde las funciones de distancias utilizadas son la distancia l^j (o L^j).

En todos los casos se utiliza una función de pesos G tal como $dG(r,s) = g_1(r)g_2(s)drds$ donde g_1 y g_2 son $g_1(z) = \varphi\left(\frac{z-\mu_X}{\sigma_X}\right)$ con φ siendo la función de densidad de una variable aleatoria $N(0,1)$ y $\mu_X = \mathbb{E}(d(X_1, X_2))$, $\sigma_X^2 = \mathbb{V}(d(X_1, X_2))$ siendo X_1, X_2 variables aleatorias independientes con la misma distribución que X. Análogamente, $g_2(t) = \varphi\left(\frac{t-\mu_Y}{\sigma_Y}\right)$. En la práctica μ_X y σ_X son desconocidos, pero pueden ser estimados naturalmente por $\widehat{\mu}_X = \frac{1}{N}\sum_{i\neq j} d(X_i, X_j)$ y $\widehat{\sigma}_X^2 = \frac{1}{N}\sum_{i\neq j} (d(X_i, X_j) - \widehat{\mu}_X)^2$ donde $N = n(n-1)$, y análogamente con $\widehat{\mu}_Y$ y $\widehat{\sigma}_Y^2$.

2.4. Cálculo de los Estadísticos $T_n^{(1)}, T_n^{(2)}$ y $T_n^{(\infty)}$

Dados $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ i.i.d. en $S_X \times S_Y$, y elegidas las funciones de pesos g_1, g_2 a ser utilizadas, los estadísticos $T_n^{(1)}, T_n^{(2)}$ y $T_n^{(\infty)}$ pueden ser calculados en los pasos indicados en las siguientes tres proposiciones.

Proposición 2. *Cálculo de* $T_n^{(2)}$.

Paso 1. Calcular $d(X_i, X_j)$ *y* $d(Y_i, Y_j)$ *para todo* $i, j \in \{1, 2, 3, ..., n\}$ *donde* $i \neq j$ *y poner* $N = n(n-1)$.

Paso 2. Reordenar $\{d(X_i, X_j)\}_{i \neq j}$ *como* $Z_1, Z_2, ..., Z_N$ *tal que* $Z_1 < Z_2 < ... < Z_N$ *y* $\{d(Y_i, Y_j)\}_{i \neq j}$ *como* $T_1, T_2, ..., T_N$ *manteniendo la misma indexación que las de los* $Z's$ *(esto es, si* $d(X_i, X_j) = Z_h$ *entonces* $d(Y_i, Y_j) = T_h$*).*

Paso 3. Calcular los estadísticos de orden para $T's$*, esto es,* $T_1^* < T_2^* < ... < T_N^*$.

Paso 4. Calcular

$$A_n = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(1 - G_1\left(\text{máx}\left\{Z_i, Z_j\right\}\right)\right) \left(1 - G_2\left(\text{máx}\left\{T_i, T_j\right\}\right)\right),$$

$$B_n = \left(1 - \frac{1}{N^2} \sum_{i=1}^{N} (2i-1)\, G_1\left(Z_i\right)\right) \left(1 - \frac{1}{N^2} \sum_{i=1}^{N} (2i-1)\, G_2\left(T_i^*\right)\right),$$

$$C_n = \frac{1}{N^3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \left(1 - G_1\left(\text{máx}\left\{Z_i, Z_j\right\}\right)\right) \left(1 - G_2\left(\text{máx}\left\{T_i, T_k\right\}\right)\right).$$

Paso 5. Calcular

$$T_n^{(2)} = n(A_n + B_n - 2C_n).$$

Proposición 3. *Cálculo de* $T_n^{(1)}$.

Paso 1. Calcular $d(X_i, X_j)$ *y* $d(Y_i, Y_j)$ *para todo* $i, j \in \{1, 2, 3, ..., n\}$ *donde* $i \neq j$ *y poner* $N = n(n-1)$.

Paso 2. Reordenar $\{d(X_i, X_j)\}_{i \neq j}$ *como* $Z_1, Z_2, ..., Z_N$ *tal que* $Z_1 < Z_2 < ... < Z_N$ *y* $\{d(Y_i, Y_j)\}_{i \neq j}$ *como* $T_1, T_2, ..., T_N$ *manteniendo la misma indexación que las de los* $Z's$ *(esto es, si* $d(X_i, X_j) = Z_h$ *entonces* $d(Y_i, Y_j) = T_h$*).*

Paso 3. Calcular los estadísticos de orden para $T's$*, esto es,* $T_1^* < T_2^* < ... < T_N^*$.

Paso 4. Para cada $h, j \in \{1, 2, 3, ..., N-1\}$ *calcular* $c(h, j) = \sum_{i=1}^{h} \mathbf{1}_{\{T_i < T_{j+1}^*\}}$*, esto es, el número de elementos del vector* $(T_1, T_2, ..., T_h)$ *que son menos que* T_{j+1}^* *para* $h, j = 1, 2, 3, ..., N-1$.

Paso 5. Calcular

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

Proposición 4. *Cálculo de* $T_n^{(\infty)}$.

Paso 1. Calcular $d(X_i, X_j)$ *y* $d(Y_i, Y_j)$ *para todo* $i, j \in \{1, 2, 3, ..., n\}$ *donde* $i \neq j$ *y poner* $N = n(n-1)$.

Paso 2. Reordenar $\{d(X_i, X_j)\}_{i \neq j}$ *como* $Z_1, Z_2, ..., Z_N$ *tal que* $Z_1 < Z_2 < ... < Z_N$ *y* $\{d(Y_i, Y_j)\}_{i \neq j}$ *como* $T_1, T_2, ..., T_N$ *manteniendo la misma indexación que las de los* $Z's$ *(esto es, si* $d(X_i, X_j) = Z_h$ *entonces* $d(Y_i, Y_j) = T_h$*).*

Paso 3. Calcular los estadísticos de orden para $T's$*, esto es,* $T_1^* < T_2^* < ... < T_N^*$.

Paso 4. Calcular la matrix $(N-1) \times (N-1)$ C *tal que*

$$C_{ij} = \left| \sum_{k=1}^{N} \mathbf{1}_{\{Z_k \leq Z_i,\ T_k \leq T_j^*\}} - \frac{ij}{N} \right|.$$

Paso 5. Calcular

$$T_n^{(\infty)} = \frac{\sqrt{n}}{N} \max_{i,j} C_{ij}.$$

Capítulo 3

Estudio de la perfomance del test mediante simulaciones

Siguiendo el estudio realizado en [18, 19], en este capítulo se realiza un estudio sobre la performance del test en diferentes aspectos por medio de simulaciones.

En primer lugar, en la sección 3.1 se analiza la performance del test frente a otras alternativas. En segundo lugar, en la sección 3.2 se analiza la performance de los test estadísticos $T_n^{(1)}, T_n^{(2)}$ y $T_n^{(\infty)}$ para diferentes funciones de distancia. Finalmente, en la sección 3.3 se muestra, utilizando una comparación de potencias, que el test de independencia de tasas de recurrencias en dimensión alta supera el rendimiento de otros test competidores en casi todos los casos, y se estudia la incidencia de las funciones de distancia consideradas (d_X y d_Y) en la performance del test. Como es esperable, se muestra que el test estadístico en alta dimensión tiene alguna sensibilidad respecto a la elección de la función de distancia, d_X o d_Y. Además, se comparan la performance de los test $T_n^{(1)}, T_n^{(2)}$ y $T_n^{(\infty)}$.

3.1. Comparación de la performance del test $T_n^{(2)}$ con respecto a otros test de independencia

En esta sección se compara la performance del test propuesto frente a otros. En las Tablas 1 a 6 se muestra la comparación de la potencia entre

el test para diferentes elecciones de la función G y otros tests, para tamaños de muestra de $n = 30$, $n = 50$ y $n = 80$. Todos los cálculos de potencias que se consideran fueron realizados a un nivel de significación de 5 %. Los cálculos fueron realizados utilizando (2.3) y tomando como función de pesos $dG(r, s) = g_1(r)g_2(s)drds$ donde $g_1 = g_2 = g$ es la función de densidad de una variable aleatoria $N(\mu, \sigma^2)$ para diferentes valores de μ y σ^2, excepto para la última columna, donde se toma las funciones g_1 y g_2 sugeridas en la subsección 2.2.3. Se compara la potencia del test respecto al test propuesto en Heller et al. [14] (que se llamará HHG), el test de distancia de covarianza propuesto en Székely et al. [28] (que se llamará DCOV) y el test propuesto en Gretton et al. [12] (que se llamará HSIC).

3.1.1. X e Y son variables aleatorias

Se consideran los test de Heller et al.'s [14], que se denominan "Parábola", "2 parábolas", "Círculo", "Diamante", "Forma de W" y "4 nubes" definidos en la Tabla 3.1. Se observa que en "4 nubes", H_0 es verdadera, y la potencia en todos los casos debería estar cerca de 0.05. En todos los casos, los valores críticos del test fueron calculados a partir de 50000 replicaciones y la potencia para cada alternativa a partir de 10000 replicaciones. Las columnas 2, 3 y 4 de la Tabla 3.1 brindan la potencia de los tests HHG, DCOV y HSIC. La columna 5 brinda la máxima potencia entre los tests clásicos de correlación: Pearson, Spearman y Kendall, que se denominan PSK. Las columnas 6, 7 y 8 brindan la potencia del test para diferentes funciones $g = g_1 = g_2$ consideradas como función de pesos G. En la columna 9, se utiliza las funciones g_1 y g_2 propuestas en la subsección 2.2.3, análogamente en la Tabla 3.2 y Tabla 3.3. La Figura 3.1 nos brinda $n = 1000$ simulaciones de las alternativas consideradas en esta subsección.

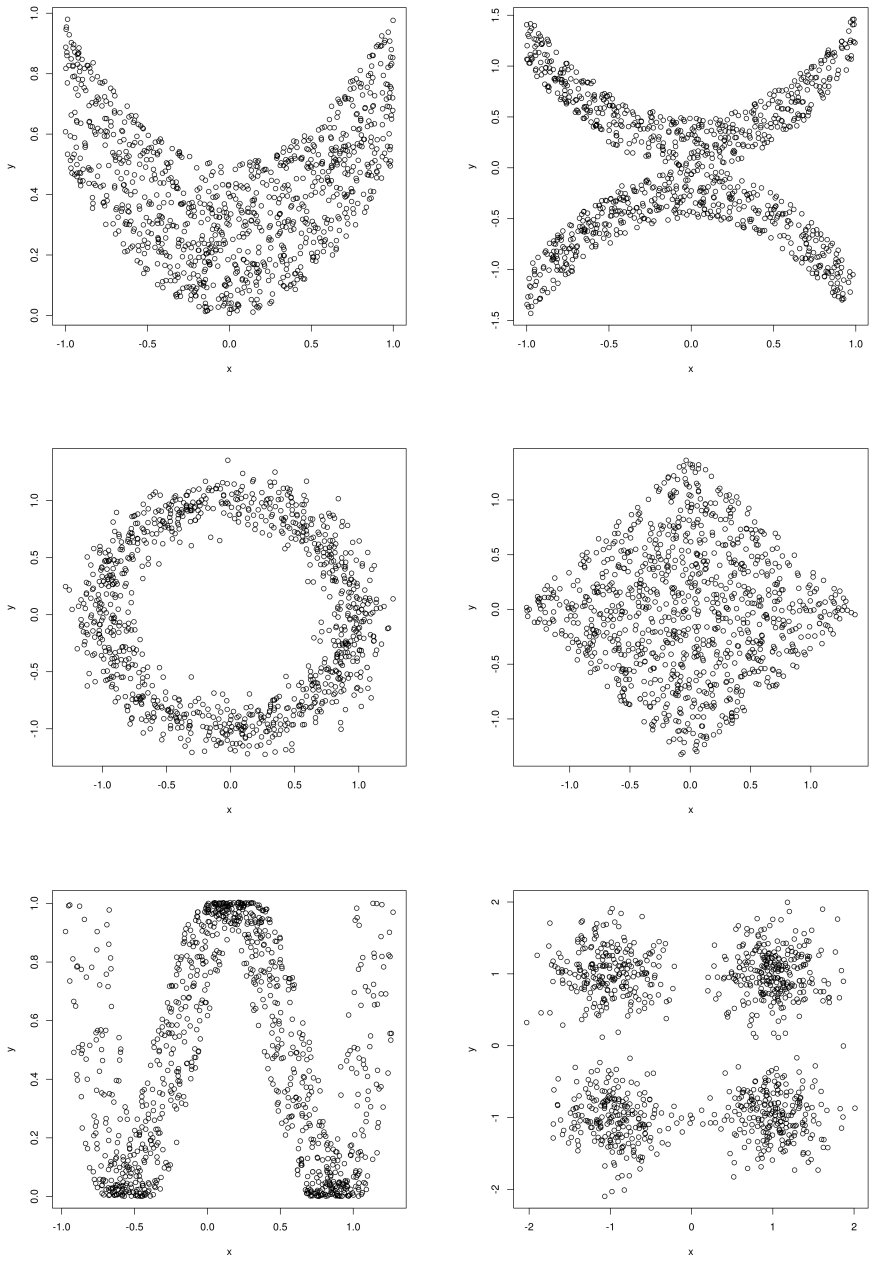

Figura 3.1: $n = 1000$ simulaciones de las alternativas consideradas en esta subsección: Parábola: $X \sim U(-1,1)$, $Y = \left(X^2 + U(0,1)\right)/2$; 2 parábolas: $X \sim U(-1,1)$, $Y = \left(X^2 + U(0,1)/2\right)$ con probabilidad $1/2$ y $Y = -\left(X^2 + U(0,1)/2\right)$ con probabilidad $1/2$; Círculo: $U \sim U(-1,1)$, $X = \sin(\pi U) + N(0,1)/8$, $Y = \cos(\pi U) + N(0,1)/8$; Diamante: $U_1, U_2 \sim U(-1,1)$ independiente, $X = \sin(\theta)\, U_1 + \cos(\theta)\, U_2$, $Y = -\sin(\theta)\, U_1 + \cos(\theta)\, U_2$ para $\theta = \pi/4$; Forma de W: $U \sim U(-1,1)$, $U_1, U_2 \sim U(0,1)$ independiente. $X = U + U_1/3$ y $Y = 4\left(U^2 - 1/2\right)^2 + U_2/n$; 4 nubes: $X = 1 + Z_1/3$ con probabilidad $1/2$, $X = -1 + Z_2/3$ con probabilidad $1/2$ e $Y = 1 + Z_3/3$ con probabilidad $1/2$, $Y = -1 + Z_4/3$ con probabilidad $1/2$, donde $Z_1, Z_2, Z_3, Z_4 \sim N(0,1)$ son independientes.

Tabla 3.1: Comparación de probabilidades de rechazo empíricas para los diferentes tests para tamaños de muestra $n = 30$. Parábola: $X \sim U(-1,1)$, $Y = \left(X^2 + U(0,1)\right)/2$; 2 parábolas: $X \sim U(-1,1)$, $Y = \left(X^2 + U(0,1)/2\right)$ con probabilidad $1/2$ y $Y = -\left(X^2 + U(0,1)/2\right)$ con probabilidad $1/2$; Círculo: $U \sim U(-1,1)$, $X = \sin(\pi U) + N(0,1)/8$, $Y = \cos(\pi U) + N(0,1)/8$; Diamante: $U_1, U_2 \sim U(-1,1)$ independientes, $X = \sin(\theta)U_1 + \cos(\theta)U_2$, $Y = -\sin(\theta)U_1 + \cos(\theta)U_2$ para $\theta = \pi/4$; Forma de W: $U \sim U(-1,1)$, $U_1, U_2 \sim U(0,1)$ independientes. $X = U + U_1/3$ e $Y = 4\left(U^2 - 1/2\right)^2 + U_2/n$; 4 nubes: $X = 1 + Z_1/3$ con probabilidad $1/2$, $X = -1 + Z_2/3$ con probabilidad $1/2$ e $Y = 1 + Z_3/3$ con probabiliad $1/2$, $Y = -1 + Z_4/3$ con probabilidad $1/2$, donde $Z_1, Z_2, Z_3, Z_4 \sim N(0,1)$ son independientes.

Test	HHG	DCOV	HSIC	PSK	N(1,1)	N(0,1)	N(1,4)	g_1, g_2
Parábola	0.79	0.52	0.73	0.10	0.82	**0.83**	0.81	0.81
2 parábolas	0.96	0.20	0.85	0.19	**1.00**	**1.00**	**1.00**	**1.00**
Círculo	0.65	0.05	0.49	0.10	0.92	0.72	**0.95**	0.82
Diamante	0.28	0.03	0.26	0.02	0.42	0.14	**0.48**	0.39
Forma de W	**0.91**	0.57	0.86	0.18	0.79	0.89	0.78	0.87
4 nubes	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

3.1.2. X e Y son vectores aleatorios

En el test, la distancia considerada para el cálculo de las medidas de recurrencias es la distancia euclideana. Teniendo en cuenta que la distancia euclideana aumenta con la dimensión, se agregan en la columna 7 y 8 las densidades de $N(0,4)$ y $N(2,4)$. En esta subsección, se considera las dos últimas alternativas en Tabla 3, y en la Tabla 4 de Heller et al. [14], que se denominan "Log", "Epsilon" y "Cuadrático" y que están definidas en la Tabla 3.4. También se agregan las alternativas consideradas en Boglioni [5] , que se denominan "pares-2D" y son definidas en la Tabla 3.4. En todos los casos, los valores críticos del test son calculados a partir de 50000 replicaciones y la potencia para cada alternativa a partir 10000 replicaciones.

De forma de obtener una idea del tamaño del test para vectores aleatorios, se simulan $X, Y \in \mathbb{R}^5$ independientes con distribución $N(0, I)$. Las probabilidades de rechazo empíricas del test son 0.051, 0.048 y 0.052 para tamaños de muestra de $30, 50$ y 80, respectivamente.

Tabla 3.2: Comparación de probabilidades de rechazo empíricas para los diferentes tests para tamaños de muestra $n = 50$. Parábola: $X \sim U(-1,1)$, $Y = \left(X^2 + U(0,1)\right)/2$; 2 parábolas: $X \sim U(-1,1)$, $Y = \left(X^2 + U(0,1)/2\right)$ con probabilidad $1/2$ y $Y = -\left(X^2 + U(0,1)/2\right)$ con probabilidad $1/2$; Círculo: $U \sim U(-1,1)$, $X = \sin(\pi U) + N(0,1)/8$, $Y = \cos(\pi U) + N(0,1)/8$; Diamante: $U_1, U_2 \sim U(-1,1)$ independientes, $X = \sin(\theta)U_1 + \cos(\theta)U_2$, $Y = -\sin(\theta)U_1 + \cos(\theta)U_2$ para $\theta = \pi/4$; Forma de W: $U \sim U(-1,1)$, $U_1, U_2 \sim U(0,1)$ independientes. $X = U + U_1/3$ y $Y = 4\left(U^2 - 1/2\right)^2 + U_2/n$; 4 nubes: $X = 1 + Z_1/3$ con probabilidad $1/2$, $X = -1 + Z_2/3$ con probabilidad $1/2$ y $Y = 1 + Z_3/3$ con probabilidad $1/2$, $Y = -1 + Z_4/3$ con probabilidad $1/2$, donde $Z_1, Z_2, Z_3, Z_4 \sim N(0,1)$ son independientes.

Test	HHG	DCOV	HSIC	PSK	N(1,1)	N(0,1)	N(1,4)	g_1, g_2
Parábola	0.98	0.85	0.96	0.11	0.98	0.98	**1.00**	0.98
2 parábolas	**1.00**	0.35	0.99	0.20	**1.00**	**1.00**	**1.00**	**1.00**
Círculo	0.98	0.07	0.91	0.01	0.99	0.99	**1.00**	0.99
Diamante	0.66	0.05	0.54	0.01	0.84	0.63	**0.88**	0.76
Forma de W	**1.00**	0.93	0.99	0.078	0.99	**1.00**	0.99	0.98
4 nubes	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

3.1.3. X e Y son series de tiempo

En esta subsección, se considera el caso en que X e Y son series de tiempo. En todos los casos X e Y son series de tiempo de longitud 100 y la potencia (debido al costo computacional) es calculada por el método de permutación propuesto en la subsección 2.2.2 para $m = 1.000$ replicaciones (Tabla 3.7 y Tabla 3.8) y $m = 100$ replicaciones (Tabla 3.9). Se utilizan g_1 y g_2 propuestas en subsección 2.2.3. La potencia de las diferentes alternativas y los tamaños muestrales en el caso discreto figuran en la Tabla 3.7. El AR(0.1) y AR(0.9) significa que la serie de tiempo X es un AR(1) con parámetros 0.1 y 0.9, respectivamente. El caso llamado ARMA$(2,1)$, es un modelo ARMA$(2,1)$ con parámetros $\phi = (0.2, 0.5)$ y $\theta = 0.2$. En la columna 4 de la Tabla 3.7, Z representa un ruido blanco donde σ es la desviación estándar de $\sqrt{|X|}$. En la Tabla 3.7 y Tabla 3.8, ε y ε' son ruidos blancos independientes con $\sigma = 1$. En la Tabla 3.8 se presenta la potencia para diferentes alternativas y tamaños muestrales en el caso continuo. En esta tabla, Bm significa que X es un movimiento browniano con $\sigma = 1$ observado en $[0,1]$ (en momentos $0, 1/100, 2/100, ..., 99/100$) y fBm es un movimiento Browniano fraccional con parámetro de Hurst $H = 0.7$. Finalmente, la Tabla 3.9 muestra la potencia para los casos en que la dependencia entre X y Y es más difícil de detectar. En estos casos, Y es un proceso de Ornstein-Uhlenbeck fraccional conducido por

Tabla 3.3: Comparación de probabilidades de rechazo empíricas para los diferentes tests para tamaños de muestra $n = 80$. Parábola: $X \sim U(-1,1)$, $Y = \left(X^2 + U(0,1)\right)/2$; 2 parábolas: $X \sim U(-1,1)$, $Y = \left(X^2 + U(0,1)/2\right)$ con probabilidad $1/2$ e $Y = -\left(X^2 + U(0,1)/2\right)$ con probabilidad $1/2$; Círculo: $U \sim U(-1,1)$, $X = \sin(\pi U) + N(0,1)/8$, $Y = \cos(\pi U) + N(0,1)/8$; Diamante: $U_1, U_2 \sim U(-1,1)$ independientes, $X = \sin(\theta)U_1 + \cos(\theta)U_2$, $Y = -\sin(\theta)U_1 + \cos(\theta)U_2$ para $\theta = \pi/4$; Forma de W: $U \sim U(-1,1)$, $U_1, U_2 \sim U(0,1)$ independientes. $X = U + U_1/3$ e $Y = 4\left(U^2 - 1/2\right)^2 + U_2/n$; 4 nubes: $X = 1 + Z_1/3$ con probabilidad $1/2$, $X = -1 + Z_2/3$ con probabilidad $1/2$ e $Y = 1 + Z_3/3$ con probabilidad $1/2$, $Y = -1 + Z_4/3$ con $1/2$, donde $Z_1, Z_2, Z_3, Z_4 \sim N(0,1)$ son independientes.

Test	HHG	DCOV	HSIC	PSK	N(1,1)	N(0,1)	N(1,4)	g_1, g_2
Parábola	**1.00**	0.99	**1.00**	0.10	**1.00**	**1.00**	**1.00**	**1.00**
2 parábolas	**1.00**	0.70	**1.00**	0.20	**1.00**	**1.00**	**1.00**	**1.00**
Círculo	**1.00**	0.20	**1.00**	0.01	**1.00**	**1.00**	**1.00**	**1.00**
Diamante	0.95	0.10	0.85	0.00	0.84	0.95	**1.00**	**1.00**
Forma de W	**1.00**	**1.00**	**1.00**	0.08	0.99	**1.00**	**1.00**	**1.00**
4 nubes	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Tabla 3.4: Comparación de la potencia para diferentes tests para tamaños de muestra de $n = 30$. Log: $X, Y \in \mathbb{R}^5$ donde $X_i \sim N(0,1)$ son independientes, $Y_i = \log\left(X_i^2\right)$ para $i \in \{1,2,3,4,5\}$; Epsilon: $X, Y, \varepsilon \in \mathbb{R}^5$ donde $X_i, \varepsilon_i \sim N(0,1)$ son independientes, $Y_i = \varepsilon_i X_i$ para $i \in \{1,2,3,4,5\}$; Cuadrático: $X, Y, \varepsilon \in \mathbb{R}^5$ donde X_i, ε_i son independientes, $X_i \sim N(0,1)$, $\varepsilon_i \sim N(0,3)$, $Y_i = X_i + 4X_i^2 + \varepsilon_i$ $i \in \{1,2\}$, $Y_i = \varepsilon_i$ para todo $i \in \{3,4,5\}$; pares-2D: $X, Z_0, Y_1 \sim N(0,1)$ independientes, $Y = (Y_1, Y_2)$ donde $Y_2 = |Z_0|\, sign(XY_1)$.

Test	HHG	DCOV	HSIC	N(1,1)	N(1,4)	N(0,4)	N(2,4)	g_1, g_2
Log	0.56	0.15	0.61	0.71	0.76	0.32	**0.88**	0.81
Epsilon	0.78	0.23	0.48	0.47	0.58	0.19	0.75	**0.86**
Cuadrático	**0.69**	0.30	0.53	0.20	0.15	0.17	0.15	0.14
pares-2D	0.16	0.17	**0.40**	0.18	0.26	0.11	0.26	0.11

un fBm (X) para $H = 0.5$ y $H = 0.7$, que se denominan OU y FOU, respectivamente. Una combinación lineal particular de FOU, que se llama $FOU(2)$, y cuya definición y desarrollo teórico se encuentra en [20], es un caso particular de los modelos propuestos en [2]. La Tabla 3.9 considera los parámetros $\sigma = 1, \lambda = 0.3$ (columna 3) y $\sigma = 1, \lambda_1 = 0.3, \lambda_2 = 0.8$ (columna 4). Más explícitamente, $Y_t = \sigma \int_{-\infty}^{t} e^{-\lambda(t-s)} dX_s$ en columna 3 (donde $X = \{X_t\}$ es un fBm), e $Y_t = \dfrac{\lambda_1}{\lambda_1 - \lambda_2}\sigma \int_{-\infty}^{t} e^{-\lambda_1(t-s)} dX_s + \dfrac{\lambda_2}{\lambda_2 - \lambda_1}\sigma \int_{-\infty}^{t} e^{-\lambda_2(t-s)} dX_s$ en columa 4 (donde $X = \{X_t\}$ es un fBm). Para tener una idea sobre el tamaño del test, en la columna 5 Y es un Bm independiente de X.

Tabla 3.5: Comparación de la potencia para diferentes tests para tamaños de muestra de $n = 50$. Log: $X, Y \in \mathbb{R}^5$ donde $X_i \sim N(0,1)$ son independientes, $Y_i = \log\left(X_i^2\right)$ para $i \in \{1,2,3,4,5\}$; Epsilon: $X, Y, \varepsilon \in \mathbb{R}^5$ donde $X_i, \varepsilon_i \sim N(0,1)$ son independientes, $Y_i = \varepsilon_i X_i$ para $i \in \{1,2,3,4,5\}$; Cuadrático: $X, Y, \varepsilon \in \mathbb{R}^5$ donde X_i, ε_i son independientes, $X_i \sim N(0,1)$, $\varepsilon_i \sim N(0,3)$, $Y_i = X_i + 4X_i^2 + \varepsilon_i$ $i \in \{1,2\}$, $Y_i = \varepsilon_i$ para todo $i \in \{3,4,5\}$; pares-2D: $X, Z_0, Y_1 \sim N(0,1)$ independientes, $Y = (Y_1, Y_2)$ donde $Y_2 = |Z_0|\, sign(XY_1)$.

Test	HHG	DCOV	HSIC	N(1,1)	N(1,4)	N(0,4)	N(2,4)	g_1, g_2
Log	0.94	0.39	0.96	**1.00**	**1.00**	**1.00**	**1.00**	0.99
Epsilon	0.97	0.30	0.69	0.89	0.97	0.97	**1.00**	0.98
Cuadrático	**0.93**	0.48	0.90	0.36	0.29	0.31	0.73	0.24
pares-2D	0.27	0.36	**0.80**	0.28	0.22	0.26	0.20	0.17

Tabla 3.6: Comparación potencia para diferentes test para tamaño muestra de $n = 80$. Log: $X, Y \in \mathbb{R}^5$ donde $X_i \sim N(0,1)$ son independientes, $Y_i = \log\left(X_i^2\right)$ para $i \in \{1,2,3,4,5\}$; Epsilon: $X, Y, \varepsilon \in \mathbb{R}^5$ donde $X_i, \varepsilon_i \sim N(0,1)$ son independientes, $Y_i = \varepsilon_i X_i$ para $i \in \{1,2,3,4,5\}$; Cuadrático: $X, Y, \varepsilon \in \mathbb{R}^5$ donde X_i, ε_i son independientes, $X_i \sim N(0,1)$, $\varepsilon_i \sim N(0,3)$, $Y_i = X_i + 4X_i^2 + \varepsilon_i$ $i \in \{1,2\}$, $Y_i = \varepsilon_i$ para todo $i \in \{3,4,5\}$; pares-2D: $X, Z_0, Y_1 \sim N(0,1)$ independiente, $Y = (Y_1, Y_2)$ donde $Y_2 = |Z_0|\, sign(XY_1)$.

Test	HHG	DCOV	HSIC	N(1,1)	N(1,4)	N(0,4)	N(2,4)	g_1, g_2
Log	**1.00**	0.79	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Epsilon	**1.00**	0.38	0.90	0.99	**1.00**	**1.00**	**1.00**	**1.00**
Cuadrático	**0.99**	0.72	0.97	0.59	0.54	0.53	0.48	0.42
pares-2D	0.54	0.75	**0.99**	0.49	0.35	0.47	0.26	0.28

3.2. Análisis de la potencia en alta dimensión

Cuando X e Y se encuentran en espacios de dimensión alta, es interesante analizar la performance de los test estadísticos $T_n^{(1)}, T_n^{(2)}$ y $T_n^{(\infty)}$ para diferentes funciones de distancia d_X y d_Y. En esta sección se compara la potencia de los 9 test estadísticos $T_n^{(i,j)}$ para $i, j = 1, 2, \infty$ en los casos en que X e Y son series de tiempo discretas y continuas bajo varias alternativas. En todos los casos se utiliza la misma función de distancia para X e Y, esto es, si X e Y son series de tiempo discreta, entonces se utiliza l^j para ambas X e Y para $j = 1, 2, \infty$, y de forma análoga en el caso en que X e Y son series de tiempo continuas. En todos los casos, X e Y son series de tiempo de longitud 100 y la potencia (debido al costo computacional) son calculadas al nivel del 5 % a partir de 500 replicaciones. Cada p-valor es calculado por el método de permutación, que se

Tabla 3.7: Probabilidades empíricas de rechazo para el caso de series de tiempo discretas y diferentes tamaños de muestra. Los parámetros en el caso ARMA$(2,1)$ son $\phi = (0.2, 0.5)$ y $\theta = 0.2$. Z representa un ruido blanco donde σ es la desviación estándar de $\sqrt{|X|}$. ε representa un ruido blanco con $\sigma = 1$ independiente de X.

n	X	$Y = X^2 + 3\varepsilon$	$Y = \sqrt{\lvert X\rvert} + Z$	$Y = \varepsilon X$	$Y = \varepsilon$
30	AR$(0,1)$	0.35	0.21	0.77	0.05
50	AR$(0,1)$	0.59	0.40	0.96	0.05
100	AR$(0,1)$	1.00	0.70	1.00	0.05
30	AR$(0,9)$	1.00	0.90	1.00	0.03
50	AR$(0,9)$	1.00	1.00	1.00	0.05
100	AR$(0,9)$	1.00	1.00	1.00	0.04
30	ARMA$(2,1)$	0.82	0.32	0.92	0.06
50	ARMA$(2,1)$	0.99	0.57	1.00	0.05
100	ARMA$(2,1)$	1.00	0.92	1.00	0.05

Tabla 3.8: Probabilidades empíricas de rechazo para el caso de series de tiempo continuas y diferentes tamaños de muestra. Bm y fBm representan un movimiento Browniano y un movimiento Browniano fraccional con $H = 0.7$ respectivamente. ε y ε' son ruidos blancos independientes con $\sigma = 1$.

n	X	$Y = X^2 + 3\varepsilon$	$Y = \sqrt{\lvert X\rvert} + \varepsilon$	$Y = \varepsilon X + 3\varepsilon'$	$Y = \varepsilon$
30	Bm	0.770	0.519	0.402	0.060
50	Bm	0.924	0.752	0.656	0.052
80	Bm	0.994	0.923	0.839	0.040
30	fBm	0.732	0.550	0.366	0.039
50	fBm	0.883	0.805	0.586	0.040
80	fBm	0.987	0.930	0.804	0.051

sugiere en el Capítulo 2, para 100 replicaciones.

3.2.1. El caso discreto

Se analizan dos escenarios para X: uno de ellos es cuando X es AR(1) donde $\phi = 0.1$, que se denomina simplemente AR(0.1) y el segundo caso, es cuando X es ARMA$(2,1)$ con parámetros $\phi = (0.2, 0.5)$ y $\theta = 0.2$. En ambos casos se consideran tres posibles Y: $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + \sigma\varepsilon$ donde σ^2 significa la varianza de $\sqrt{|X|}$ e $Y_3 = \varepsilon X$. En todos los casos, ε es un ruido blanco Gaussiano $(N(0,1))$ independiente de X. En la Tabla 3.10 y Tabla 3.11 se muestra la potencia para $n = 30$ y $n = 50$, respectivamente, en el caso en que X es un proceso AR(0.1) para los 9 tests considerados. De forma similar la Tabla 3.12 y Tabla 3.13 muestran la potencia para el caso en que X es un proceso

Tabla 3.9: Potencia donde la dependencia es entre un movimiento Browniano fraccional y sus FOU y FOU(2) asociados, para los casos $H = 0.5$ (Bm), $H = 0.7$ (fBm) y probabilidades empíricas de rechazo donde X e Y son independientes.

n	X	Y =FOU	Y =FOU(2)	$Y = Bm$
30	Bm	0.775	0.183	0.053
50	Bm	0.906	0.541	0.046
80	Bm	0.986	0.880	0.056
30	fBm	0.380	0.106	0.045
50	fBm	0.516	0.282	0.039
80	fBm	0.707	0.542	0.042

ARMA$(2, 1)$. Las Tablas 3.10–3.13 no muestran diferencias importantes entre la utilización de $T_n^{(2)}$, $T_n^{(1)}$ o $T_n^{(\infty)}$. En la Figura 3.2 se muestra la potencia como una función del tamaño muestral, donde los estadísticos considerados son $T_n^{(2)}$, esto es $T_n^{(2,1)}$, $T_n^{(2,2)}$ y $T_n^{(2,\infty)}$. El comportamiento de $T_n^{(1)}$ y $T_n^{(2)}$ es similar. La Figura 3.2 sugiere que la potencia crece al cambiar la función de distancia desde d_∞ (distancia l^∞) hacia d_1 (distancia l^1). También para la alternativa $Y = Y_2$, el estadístico basado en la distancia l^∞ tiene dificultades en la detección de las dependencias entre X e Y (que crece muy lentamente al crecer n), mientras que para $n = 60$ la potencia del test basado en las distancias l^1 o l^2 es próximo a la unidad.

Tabla 3.10: Comparación de potencias, al nivel de 5 %, para los diferentes tests, donde X es AR(0.1) y $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + \sigma\varepsilon$, $Y_3 = \varepsilon X$ para tamaño muestral de $n = 30$.

$n = 30$	$T_n^{(1,1)}$	$T_n^{(1,2)}$	$T_n^{(1,\infty)}$	$T_n^{(2,1)}$	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$	$T_n^{(\infty,1)}$	$T_n^{(\infty,2)}$	$T_n^{(\infty,\infty)}$
$Y = Y_1$	0.39	**0.40**	0.29	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	0.11	0.69	0.19	0.05
$Y = Y_3$	0.91	0.79	0.28	0.87	0.77	0.34	**0.92**	0.77	0.27

Tabla 3.11: Comparación de potencias, al nivel de 5 %, para los diferentes tests, donde X es AR(0.1) y $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + \sigma\varepsilon$, $Y_3 = \varepsilon X$ para tamaño muestral de $n = 50$.

$n = 50$	$T_n^{(1,1)}$	$T_n^{(1,2)}$	$T_n^{(1,\infty)}$	$T_n^{(2,1)}$	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$	$T_n^{(\infty,1)}$	$T_n^{(\infty,2)}$	$T_n^{(\infty,\infty)}$
$Y = Y_1$	0.90	**0.92**	0.74	0.54	0.59	0.47	0.49	0.57	0.34
$Y = Y_2$	0.34	0.50	0.38	**0.97**	0.81	0.16	0.92	0.89	0.79
$Y = Y_3$	**1.00**	0.94	0.64	**1.00**	0.94	0.61	0.99	0.92	0.57

Tabla 3.12: Comparación de potencias, al nivel de 5 %, para los diferentes test, donde X es ARMA$(2,1)$ e $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + \sigma\varepsilon$, $Y_3 = \varepsilon X$ para tamaño muestral de $n = 30$.

$n = 30$	$T_n^{(1,1)}$	$T_n^{(1,2)}$	$T_n^{(1,\infty)}$	$T_n^{(2,1)}$	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$	$T_n^{(\infty,1)}$	$T_n^{(\infty,2)}$	$T_n^{(\infty,\infty)}$
$Y = Y_1$	**0.85**	0.82	0.53	0.83	0.78	0.57	0.77	**0.85**	0.47
$Y = Y_2$	0.56	0.27	0.08	**0.84**	0.78	0.34	0.49	0.34	0.08
$Y = Y_3$	**1.00**	**1.00**	0.93	0.99	0.93	0.50	0.96	0.88	0.23

Tabla 3.13: Comparación de potencias, al nivel de 5 %, para los diferentes test, donde X es ARMA$(2,1)$ y $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + \sigma\varepsilon$, $Y_3 = \varepsilon X$ para tamaño muestral de $n = 50$.

$n = 50$	$T_n^{(1,1)}$	$T_n^{(1,2)}$	$T_n^{(1,\infty)}$	$T_n^{(2,1)}$	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$	$T_n^{(\infty,1)}$	$T_n^{(\infty,2)}$	$T_n^{(\infty,\infty)}$
$Y = Y_1$	**0.98**	**0.98**	0.82	0.96	0.97	0.82	0.96	0.96	0.82
$Y = Y_2$	0.76	0.48	0.04	**0.98**	**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**	0.75

3.2.2. El caso continuo

En esta subsección, se considera que X es un movimiento Browniano fraccional con $\sigma = 1$ observado en $[0, 1]$ (en los instantes $0, 1/100, 2/100, ..., 99/100$) para $H = 0.5$ (movimiento Browniano estándar) y $H = 0.7$. Se consideran 7 casos de dependencia entre X e Y. Los primeros tres son para el caso en que X es un movimiento Browniano estándar (Bm) y la dependencia es definida por medio de $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + \varepsilon$, $Y_3 = \varepsilon X + 3\varepsilon'$ donde ε y ε' son ruidos blancos Gaussianos con $\sigma = 1$ tal que X, ε y ε' son independientes. En las últimas 4 alternativas, se explora la potencia cuando Y es un funcional lineal de X. Más explícitamente, se considera el caso en que Y es un proceso fraccional de Ornstein–Uhlenbeck conducido por un movimiento Browniano (X) para $H = 0.5$ (Bm) y un movimiento Browniano fraccional para $H = 0.7$ (fBm), que se llaman los procesos OU y FOU, respectivamente. Una combinación lineal de FOU, que se llama $FOU(2)$, y cuya definición, desarrollo teórico y simulaciones se encuentran en [20] y [16], es un caso particular de los modelos propuestos en [2]. Más explícitamente, el proceso FOU es definido por $Y_t = \sigma \int_{-\infty}^{t} e^{-\lambda(t-s)} dX_s$ (donde $X = \{X_t\}$ es un fBm), y el proceso $FOU(2)$ es definido por $Y_t = \dfrac{\lambda_1}{\lambda_1 - \lambda_2}\sigma \int_{-\infty}^{t} e^{-\lambda_1(t-s)} dX_s + \dfrac{\lambda_2}{\lambda_2 - \lambda_1}\sigma \int_{-\infty}^{t} e^{-\lambda_2(t-s)} dX_s$ (donde $X = \{X_t\}$ es un fBm). Cuando $H = 0.5$, se llamará simplemente el proceso OU y $OU(2)$, como se define en [2]. En las Tablas 3.14 y 3.15 se muestra la potencia para $n = 30$ y $n = 50$, respectivamente, para las 7 alternativas.

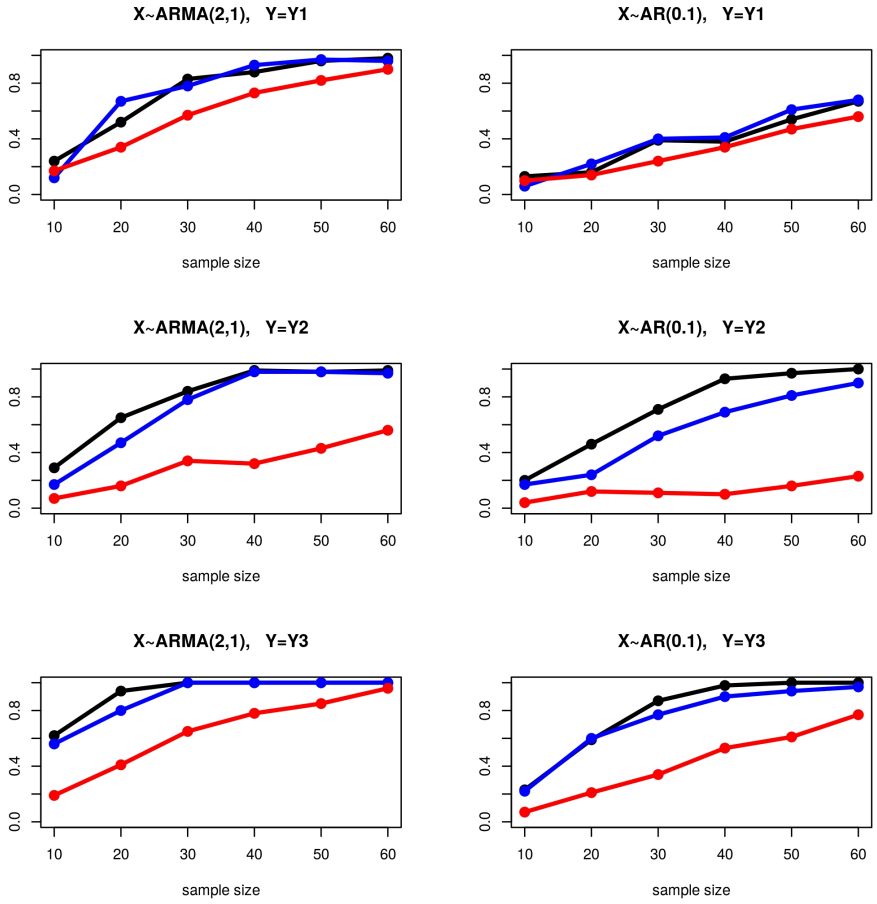

Figura 3.2: Potencia al nivel de 5 % bajo diferentes alternativas para el estadístico $T_n^{(2)}$ utilizando la distancia de Manhattan ($T_n^{(2,1)}$ en negro), la distancia Euclideana ($T_n^{(2,2)}$ en azul) y la distancia del máximo ($T_n^{(2,\infty)}$ en rojo). Y_1, Y_2 y Y_3 están definidas en las Tablas 1–4.

En estas tablas, Y_4 significa un proceso OU conducido por X con parámetros $\sigma = 1, \lambda = 0.3$. Similarmente Y_5 es un proceso FOU con parámetros $\sigma = 1, H = 0.7, \lambda = 0.3$, $Y_6 \sim OU(2)$ con parámetros $\sigma = 1, \lambda_1 = 0.3, \lambda_2 = 0.8$ e $Y_7 \sim FOU(2)$ con parámetros $\sigma = 1, H = 0.7, \lambda_1 = 0.3, \lambda_2 = 0.8$. No se presenta la performance del test para otras elecciones de parámetros, porque el comportamiento es similar. Como se espera, para valores de σ mayores que 1, la dependencia entre X e Y es más difícil de detectar, y es necesario incrementar el tamaño muestral. Lo mismo sucede si se toma λ_1 próximo a λ_2 en $OU(2)$ y $FOU(2)$. Las Tablas 3.14 y 3.15 muestran, como el caso discreto, que no hay diferencias sustanciales entre la performance de los tres estadísticos ($T_n^{(1)}, T_n^{(2)}$

o $T_n^{(\infty)}$). Con respecto a qué distancia entre los elementos de X e Y es más apropiada, la Tabla 3.14 y la Tabla 3.15 muestran que la distancia L^∞ tiene una pobre performance bajo las alternativas Y_1, Y_2 y Y_3, pero mejor bajo las alternativas Y_4, Y_5, Y_6 y Y_7. La performance de las distancias L^1 y L^2 es similar en las 7 alternativas consideradas. La Figura 3.3 expande la información dada en la Tabla 3.14 y Tabla 3.15 para los casos Y_1, Y_2 y Y_3 porque muestra la potencia de los estadísticos $T_n^{(i,j)}$ para $i, j = 1, 2, \infty$ para tamaños muestrales de $n = 10$ a $n = 50$. La Figura 3.3 muestra de forma clara que la distancia L^∞ tiene una performance más pobre que las distancias L^1 y L^2. La Figura 3.4 muestra la potencia como una función del tamaño muestral para el estadístico $T_n^{(2)}$ en los casos Y_4, Y_5, Y_6 e Y_7. El comportamiento de los estadísticos $T_n^{(1)}$ y $T_n^{(2)}$ es similar. Contrariamente a lo que sucede en los casos Y_1, Y_2 y Y_3, la distancia L^∞ tiene una mejor performance que las distancias L^1 y L^2 . La Figura 3.4 también muestra que la performance del estadístico $T_n^{(2)}$ aumenta al cambiar de utilizar la distancia L^1 a utilizar la distancia L^∞. Por otro lado, la Figura 3.4 muestra que la potencia en el caso de la alternativa OU es mayor que para la alternativa $OU(2)$ (y lo mismo para FOU versus $FOU(2)$), lo que es razonable, porque la dependencia entre X e Y es más simple en el caso OU (FOU) que en el caso $OU(2)$ ($FOU(2)$). Además, la potencia en el caso OU ($OU(2)$) es mayor que en el caso FOU ($FOU(2)$), que es lo esperado porque cuando $H = 0.7$, el movimiento Browniano fraccional es una serie de memoria larga, y por lo tanto es razonable que la dependencia entre X e Y sea más difícil de detectar.

Para concluir esta sección, se observa que el test de independencia basado en las tasas de recurrencia tiene una potencia que aumenta al aumentar n para los 9 estadísticos considerados, $T_n^{(i,j)}$ para $i, j = 1, 2, \infty$ (como se espera de acuerdo a la teoría desarrollada en el Capítulo 2) en todas las alternativas consideradas para ambos casos discretos y continuos. En la mayoría de los casos, el test tiene una potencia cercana a la unidad para tamaños de muestra moderadamente pequeños. Tomando en cuenta lo que se observa en esta sección, se puede decir que no hay preferencia en utilizar el test basado en $T_n^{(1)}, T_n^{(2)}$ o $T_n^{(\infty)}$, pero en los tres casos, la performance es mejor en general al cambiar la función de distancia desde la distancia L^1 (l^1) hacia la distancia L^∞ (l^∞) en algunos casos, y en la dirección opuesta en otros casos. Por lo tanto, se puede sugerir que se utilice el test estadístico utilizando la distancia L^1 (l^1) o L^2 (l^2) y la distancia L^∞ (l^∞) para cubrir ambas posibilidades.

Tabla 3.14: Comparación de potencias, al nivel de 5 %, de los diferentes tests, donde $X \sim Bm$ en alternativas Y_1, Y_2, Y_3, Y_4, Y_6 e $X \sim fBm$ con $H = 0.7$ en alternativas Y_5, Y_7 donde $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + \varepsilon$, $Y_3 = \varepsilon X + 3\varepsilon'$, $Y_4 = OU, Y_5 = FOU, Y_6 = OU(2)$, e $Y_7 = FOU(2)$ para tamaño muestral de $n = 30$.

$n = 30$	$T_n^{(1,1)}$	$T_n^{(1,2)}$	$T_n^{(1,\infty)}$	$T_n^{(2,1)}$	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$	$T_n^{(\infty,1)}$	$T_n^{(\infty,2)}$	$T_n^{(\infty,\infty)}$
$Y = Y_1$	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	0.31	0.38	**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

Tabla 3.15: Comparación de potencias, al nivel de 5 %, de los diferentes tests, donde $X \sim Bm$ en alternativas Y_1, Y_2, Y_3, Y_4, Y_6 e $X \sim fBm$ con $H = 0.7$ en alternativas Y_5, Y_7 donde $Y_1 = X^2 + 3\varepsilon$, $Y_2 = \sqrt{|X|} + \varepsilon$, $Y_3 = \varepsilon X + 3\varepsilon'$, $Y_4 = OU, Y_5 = FOU, Y_6 = OU(2)$, e $Y_7 = FOU(2)$ para tamaño muestral de $n = 50$.

$n = 50$	$T_n^{(1,1)}$	$T_n^{(1,2)}$	$T_n^{(1,\infty)}$	$T_n^{(2,1)}$	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$	$T_n^{(\infty,1)}$	$T_n^{(\infty,2)}$	$T_n^{(\infty,\infty)}$
$Y = Y_1$	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

3.3. Comparación con otros tests en alta dimensión

En la sección 3.1 se muestra la muy buena performance del test de tasas de recurrencia para variables y vectores aleatorios. En esta sección, se compara el test cuando X e Y están en espacios de alta dimensión. De acuerdo a lo visto en la sección anterior, se ha considerado el test utilizando el estadístico $T_n^{(2,2)}$ y $T_n^{(2,\infty)}$. Se consideran tres competidores: el bien conocido test de distancia de covarianza propuesto en [28] y adaptado para tener una mejor performance en dimensión alta en [27], el criterio de información de Hilbert–Schmidt propuesto en [13], y el propuesto más recientemente en [9] basado en proyecciones aleatorias. Básicamente, este último test está basado en la idea de escoger K pares de direcciones aleatorias, y observar que si X e Y son independientes, entonces las proyecciones de X e Y en cada uno de los K pares de direcciones

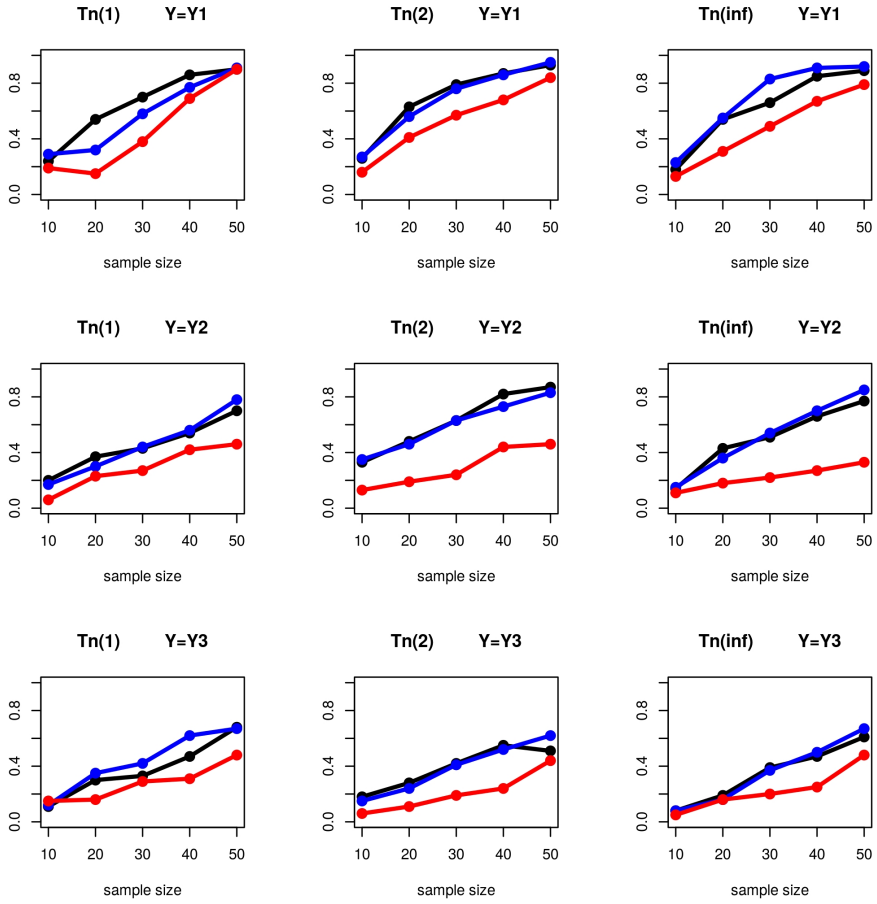

Figura 3.3: Comparación de potencias, al nivel del 5 %, donde X es un movimiento Browniano estándar, bajo varias alternativas para los estadísticos $T_n^{(1)}, T_n^{(2)}$ y $T_n^{(\infty)}$ utilizando la distancia de Manhattan ($T_n^{(i,1)}$ en negro), la distancia Euclideana ($T_n^{(i,2)}$ en azul) la distancia del máximo ($T_n^{(i,\infty)}$ en rojo) para $i = 1, 2, \infty$. Y_1, Y_2 e Y_3 son definidos como en las Tablas 3.14 y 3.15.

son independientes. Este test es universalmente consistente. Para aplicar este test, es necesario escoger previamente el número de pares de proyecciones (K), y luego se aplican K test de hipótesis de independencia. Si al menos uno de estos test rechazan la hipótesis de independencia, entoces H_0 es rechazada. Para trabajar al nivel del 5 %, se propone en [9] utilizar una corrección de Bonferroni, esto es, para calcular la proporción de p-values menores que $0.05/K$ para aplicar cada uno de los K tests unidimensionales. Se llamará el test de RPK.

En la Tabla 3.16 se reporta una comparación de potencias al nivel del 5 %, cuando X es una realización de una serie de tiempo discreta de largo 100 en

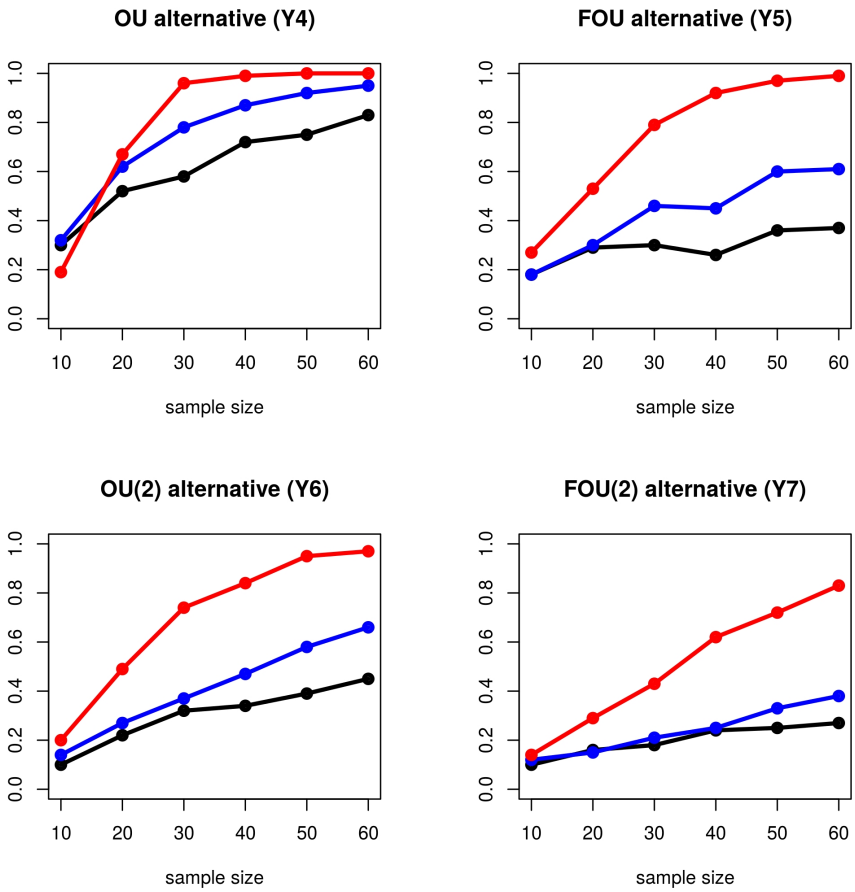

Figura 3.4: Potencia al nivel de 5 %, bajo varias alternativas para el estadístico $T_n^{(2)}$ utilizando la distancia de Manhattan ($T_n^{(2,1)}$ en negro), la distancia Euclideana ($T_n^{(2,2)}$ en azul) y la distancia del máximo ($T_n^{(2,\infty)}$ en rojo).

tres escenarios posibles, donde hay tres alternativas para Y en cada escenario.

La performance de RPK es muy malo en estos casos, y la potencia utilizando la corrección de Bonferroni es 0.

Por esta razón se presenta en la Tabla 3.16 la potencia del test RPK utilizando $0.05/4$ en lugar de $0.05/K$ para $K = 100$ proyecciones aleatorias. La Tabla 3.16 muestra que el test propuesto basado $T_n^{(2,2)}$ supera a las otros test en los 9 casos considerados. La Tabla 3.17 muestra la comparación al nivel del 5 %, de la potencia en 12 escenarios en los cuales X e Y son realizaciones de una serie de tiempo continua vista en 100 puntos equidistantes en $[0, 1]$. En esta tabla, se considera el test RPK test para $K = 5$ proyecciones aleatorias y se utiliza la corrección de Bonferroni. Se escogen $K = 5$ proyecciones porque este

es el valor de K para el cual la potencia del test de RPK alcanza el máximo. La Tabla 3.17 muestra que el test propuesto basado en $T_n^{(2,2)}$ o $T_n^{(2,\infty)}$ supera a los otros competidores en 6 escenarios, el test HSIC, DCOV y RPK tienen la mejor peformance en casos cada uno.

Tabla 3.16: Comparación al nivel del 5 % de las potencias de los 4 test de independencia considerados en el caso de series de tiempo discretas y diferentes tamaños muestrales. Los parámetros en el caso ARMA$(2,1)$ son $\phi = (0.2, 0.5)$ y $\theta = 0.2$. El parámetro σ en $Y = \sqrt{|X|} + \sigma\varepsilon$ denota la desviación estandard de $\sqrt{|X|}$. ε denota un ruido blanco con $\sigma = 1$, independiente de X.

$X \sim$ ARMA(2,1)	n	RPK	HSIC	DCOV	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$
$Y = X^2 + 3\varepsilon$	30	0.533	0.324	0.282	**0.785**	0.566
	50	0.570	0.381	0.313	**0.975**	0.825
	100	0.660	0.537	0.377	**1.000**	0.994
$Y = \sqrt{\lvert X\rvert} + \sigma\varepsilon$	30	0.549	0.294	0.240	**0.779**	0.341
	50	0.592	0.364	0.270	**0.976**	0.427
	100	0.702	0.572	0.373	**0.921**	0.860
$Y = \varepsilon X$	30	0.466	0.501	0.467	**0.925**	0.498
	50	0.468	0.583	0.535	**0.996**	0.778
	100	0.473	0.674	0.567	**1.000**	0.984
$X \sim$ AR(0.1)	n	RPK	HSIC	DCOV	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$
$Y = X^2 + 3\varepsilon$	30	**0.484**	0.134	0.114	0.398	0.236
	50	0.487	0.132	0.134	**0.592**	0.465
	100	0.950	0.162	0.131	**0.999**	0.834
$Y = \sqrt{\lvert X\rvert} + \sigma\varepsilon$	30	0.518	0.207	0.182	**0.523**	0.114
	50	0.508	0.222	0.184	**0.810**	0.157
	100	0.509	0.273	0.217	**0.698**	0.504
$Y = \varepsilon X$	30	0.474	0.349	0.331	**0.772**	0.345
	50	0.486	0.380	0.354	**0.945**	0.613
	100	0.507	0.404	0.372	**1.000**	0.938
$X \sim$ AR(0.9)	n	RPK	HSIC	DCOV	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$
$Y = X^2 + 3\varepsilon$	30	0.886	0.997	0.949	**1.000**	0.992
	50	0.978	**1.000**	0.993	**1.000**	**1.000**
	100	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
$Y = \sqrt{\lvert X\rvert} + \sigma\varepsilon$	30	0.785	0.887	0.649	**0.903**	0.778
	50	0.924	0.992	0.838	**0.998**	0.978
	100	**1.000**	**1.000**	0.993	**1.000**	**1.000**
$Y = \varepsilon X$	30	0.560	0.933	0.870	**1.000**	0.948
	50	0.562	0.983	0.945	**1.000**	**1.000**
	100	0.570	**1.000**	0.985	**1.000**	**1.000**

Tabla 3.17: Comparación al nivel del 5 % de las potencias de los 4 test de independencia considerados en el caso de series de tiempo continua y diferentes tamaños muestrales. Bm y fBm denotan un movimiento Browniano y movimiento Browniano fraccional con $H = 0.7$. ε y ε' son ruidos blancos independientes con $\sigma = 1$ (e independentes de X).

$X \sim$ Bm	n	RPK	HSIC	DCOV	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$
$Y = X^2 + 3\varepsilon$	30	0.767	**0.815**	0.596	0.757	0.570
	50	0.951	**0.977**	0.829	0.947	0.836
	80	0.999	**1.000**	0.977	0.994	0.968
$Y = \sqrt{\lvert X\rvert} + \varepsilon$	30	**0.954**	0.906	0.550	0.519	0.224
	50	**0.998**	0.996	0.862	0.802	0.416
	80	**1.000**	**1.000**	0.992	0.923	0.834
$Y = \varepsilon X + 3\varepsilon'$	30	0.054	0.099	0.118	**0.421**	0.185
	50	0.050	0.119	0.125	**0.619**	0.437
	80	0.056	0.128	0.126	**0.839**	0.648
$Y \sim$ OU	30	0.765	0.770	0.826	0.651	**0.956**
	50	0.927	0.977	0.988	0.906	**1.000**
	80	0.992	**1.000**	**1.000**	0.986	**1.000**
$Y \sim$ OU(2)	30	0.574	**0.965**	0.961	0.374	0.744
	50	0.790	**1.000**	**1.000**	0.584	0.947
	80	0.938	**1.000**	**1.000**	0.880	0.998
$X \sim$ fBm	n	RPK	HSIC	DCOV	$T_n^{(2,2)}$	$T_n^{(2,\infty)}$
$Y = X^2 + 3\varepsilon$	30	**0.742**	0.728	0.544	0.732	0.546
	50	0.962	0.946	0.814	0.883	0.758
	80	0.997	0.998	0.970	0.987	0.922
$Y = \sqrt{\lvert X\rvert} + \varepsilon$	30	**0.962**	0.925	0.579	0.580	0.266
	50	**1.000**	0.999	0.902	0.830	0.440
	80	**1.000**	**1.000**	**1.000**	0.930	0.680
$Y = \varepsilon X + 3\varepsilon'$	30	0.054	0.109	0.125	**0.366**	0.246
	50	0.053	0.098	0.131	**0.586**	0.404
	80	0.062	0.119	0.132	**0.804**	0.634
$Y \sim$ FOU	30	0.585	0.394	0.509	0.460	**0.806**
	50	0.760	0.705	0.801	0.581	**0.978**
	80	0.913	0.956	0.984	0.707	**1.000**
$Y \sim$ FOU(2)	30	0.443	0.847	**0.909**	0.206	0.426
	50	0.665	0.987	**0.996**	0.326	0.722
	80	0.820	**1.000**	**1.000**	0.542	0.928
$X \sim$ OU(λ_1)	30	0.509	0.445	0.462	0.304	**0.672**
$Y \sim$ OU(λ_2)	50	0.717	0.777	0.769	0.448	**0.888**
	80	0.889	0.978	0.965	0.582	**0.980**
$X \sim$ FOU(λ_1)	30	0.305	0.180	0.175	0.106	**0.332**
$Y \sim$ FOU(λ_2)	50	0.475	0.299	0.313	0.204	**0.524**
	80	0.644	0.596	0.574	0.210	**0.738**

Capítulo 4

Aplicación del Test de Independencia a datos reales

Siguiendo el estudio relalizado en [17, 19], en este Capítulo se realizan aplicaciones del test de independencia a datos reales. En la primer sección se analizan datos meteorológicos y en la segunda sección datos económicos.

4.1. Datos meteorológicos

4.1.1. Temperatura, humedad, viento y evaporación

En esta subsección se consideran los datos meteorológicos presentados en la Tabla 7.2 de [25]. Los datos son 46 observaciones agrupadas en 11 variables definidas como: Y_1 =“temperatura máxima diaria del aire”, Y_2 =“temperatura mínima diaria del aire”, Y_3 =“área integrada bajo la curva de temperatura del aire diaria”, Y_4 =“temperatura máxima diaria del suelo”, Y_5 =“temperatura mínima diaria del suelo”, Y_6 =“área integrada bajo la curva de temperatura diaria del suelo”, Y_7 =“humedad relativa diaria máxima”, Y_8 =“humedad relativa mínima diaria”, Y_9 =“área integrada bajo curva de humedad diaria”, Y_{10} =“viento total (en millas por día)” e Y_{11} =“evaporación”. Se consideran los vectores $Z_1 = (Y_1, Y_2, Y_3)$, $Z_2 = (Y_4, Y_5, Y_6)$, $Z_3 = (Y_7, Y_8, Y_9)$ y las variables $Z_4 = Y_{10}$ y $Z_5 = Y_{11}$. Teniendo en cuenta lo visto en el Capítulo 3, que no hay diferencias importantes entre utilizar los test estadísticos $T_n^{(1)}$, $T_n^{(2)}$ o $T_n^{(\infty)}$, se aplica el test de independencia entre parejas de Z's utilizando el test estadístico $T_n^{(2,2)}$. En la Tabla 4.1 se muestran los p-valores del test en cada caso. En la Figura 4.1 se muestra el dependograma de orden 2 del test de

independencia mutua de las Z's, esto es, los valores críticos al 5 % y 10 % y el valor del estadístico. Las aproximaciones de los p-valores y valores críticos fueron calculados a partir de $m = 1000$ replicaciones por medio del método de permutación planteado en el Capítulo 2. El test concluye que Z_1, Z_2, Z_3 y Z_5 son independientes de a pares, pero el viento (Z_4) no muestra ninguna dependencia con ninguna de las otras variables. Por otro lado, estas conclusiones son equivalentes a las obtenidas en [3].

Tabla 4.1: p-valores para el test entre parejas de las Z's.

	Z_2	Z_3	Z_4	Z_5
Z_1	0.000	0.000	0.109	0.000
Z_2		0.000	0.394	0.000
Z_3			0.373	0.000
Z_4				0.403

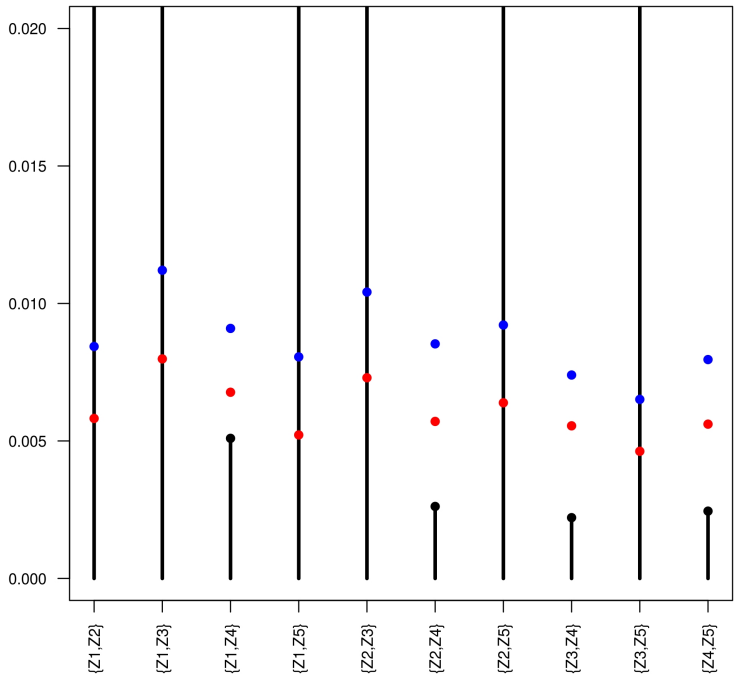

Figura 4.1: Valores críticos al 5 % (azul), 10 % (rojo) y valores observados (negro) para la prueba de independencia por pares entre las variables $Z's$.

Por otra lado, se observa que en los casos en los cuales el test no rechaza

H_0, la diferencia entre lo observado y los valores críticos es muy grande. El valor observado más grande es 0.2887, alcanzado en el caso Z_1, Z_2.

4.1.2. Temperatura, viento del oeste, viento del este

En esta subsección se considera la base de datos formada por la temperatura prevista (T), el viento en dirección oeste (U) y el viento en dirección este (V) a 850 hPa (alrededor de 1200m sobre el nivel del mar) de cada día desde enero de 2012 hasta diciembre de 2012. En total hay 341 pronósticos debido al hecho de que faltan 25 datos. El dominio numérico se muestra en la Figura 4.2 y consiste en un total de $117 \times 75 = 8775$ puntos geográficos.

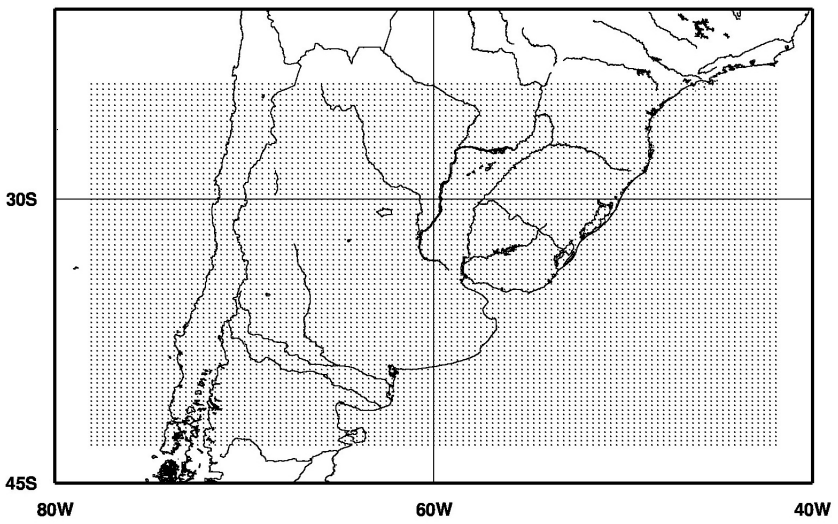

Figura 4.2: $117 \times 75 = 8775$ puntos geográficos donde se realizan los pronósticos diarios.

El horizonte temporal de los pronósticos es de 24 horas, y son para las 0:00 GMT hora de cada día. Las simulaciones numéricas fueron obtenidas utilizando el modelo regional WRF planteado en [26], y las condiciones de frontera inicial y lateral fueron obtenidas del Sistema de Pronóstico Global NCEP, como en [4]. Si se considera $(U_1, V_1, T_1), (U_2, V_2, T_2), ..., (U_{341}, V_{341}, T_{341})$ donde $U_i, V_i, T_i \in \mathbb{R}^{8775}$ para todo $i = 1, 2, 3, ..., 341$, los p-valores para el test de independencia entre U y V son iguales a cero, y así sucesivamente para la prueba entre U y T, y V y T. Esto se espera porque para cada punto i,

las variables U_i, V_i y T_i son dependientes de a pares. Ahora se considera (para cada día) cada vector $U \in \mathbb{R}^{8775}$ descompuesto como $U = (U_1, U_2, ..., U_{75})$ donde $U_i \in \mathbb{R}^{117}$. De esta forma, cada U_i representa el pronóstico de los 117 puntos geográficos en latitud i y puede ser visto como una discretización de una curva en latitud i, $(U(i))$. Es decir que, $i = 1$ indica la latitud más sureste dada en la Figura 4.2 y $i = 117$ la latitud más noreste. Se consideran los primeros 30 pronósticos, correspondientes a enero 2012. De esta forma, se obtiene una muestra de 30 curvas para cada latitud i, y se testeará la independencia mutua entre U_i y U_j para $i = 1, 2, 3, ..., 38$ y $j = 76 - i$. Se descompone T y V de forma análoga. Es de esperar que, al menos para valores pequeños de i, las variables U_i y U_j sean independientes, debido a la distancia geográfica, y lo mismo para las variables V y T.

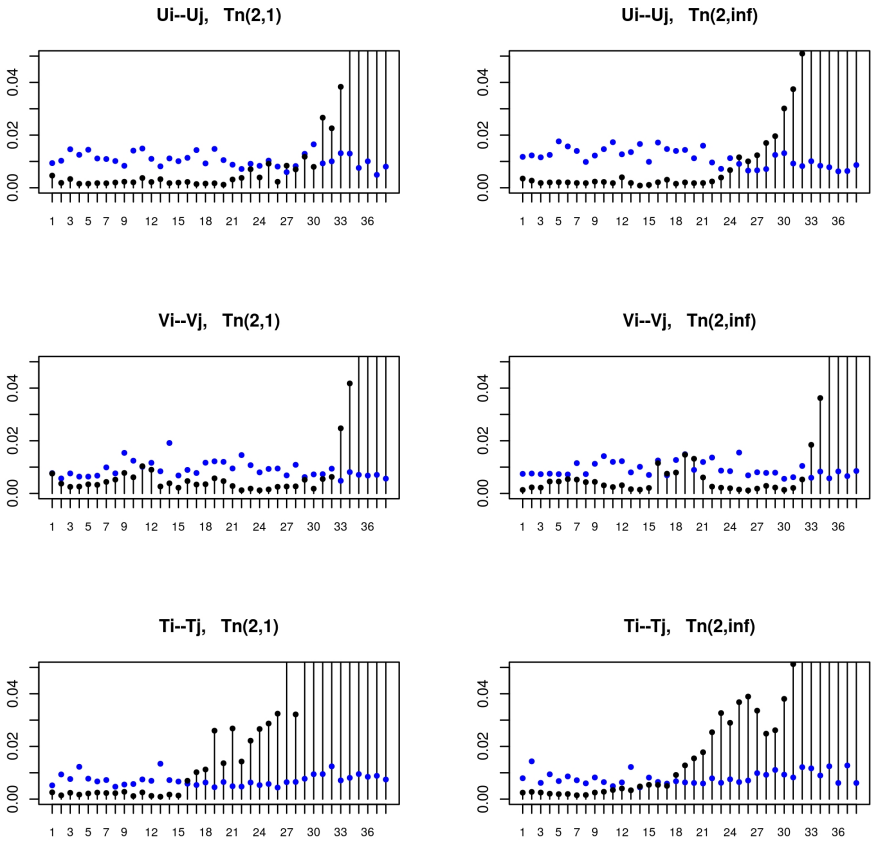

Figura 4.3: Comparación entre dependogramas para $T_n^{(2,1)}$ y $T_n^{(2,\infty)}$ entre U, V y T.

En las Figuras 4.3, 4.4, 4.5 y 4.6 se muestran los dependogramas para la prueba de independencia, utilizando los estadísticos $T_n^{(2,1)}$ y $T_n^{(2,\infty)}$, entre U_i y U_{76-i} para cada $i = 1, 2, ..., 38$ y lo mismo para las variables V, T y las otras combinaciones entre U, V y T. En la Figura 4.3, se muestran resultados similares entre $T_n^{(2,1)}$ y $T_n^{(2,\infty)}$. Sin embargo, en el caso de U_i y U_j, $T_n^{(2,\infty)}$ detecta la dependencia en más casos que $T_n^{(2,1)}$. Ambas pruebas muestran que cuando i y $j = 76 - i$ están cerca, entonces las variables U_i y U_j son dependientes. Lo mismo sucede con V_i, V_j y T_i, T_j. Además, la región geográfica en la cual los vectores son dependientes es más larga para T que para U y V.

La Figura 4.4 muestra que $T_n^{(2,1)}$ funciona mejor que $T_n^{(2,\infty)}$ porque para $i \geq 32$ la prueba basada en $T_n^{(2,1)}$ detecta una dependencia para ambos casos: U_i, V_j y V_i, U_j.

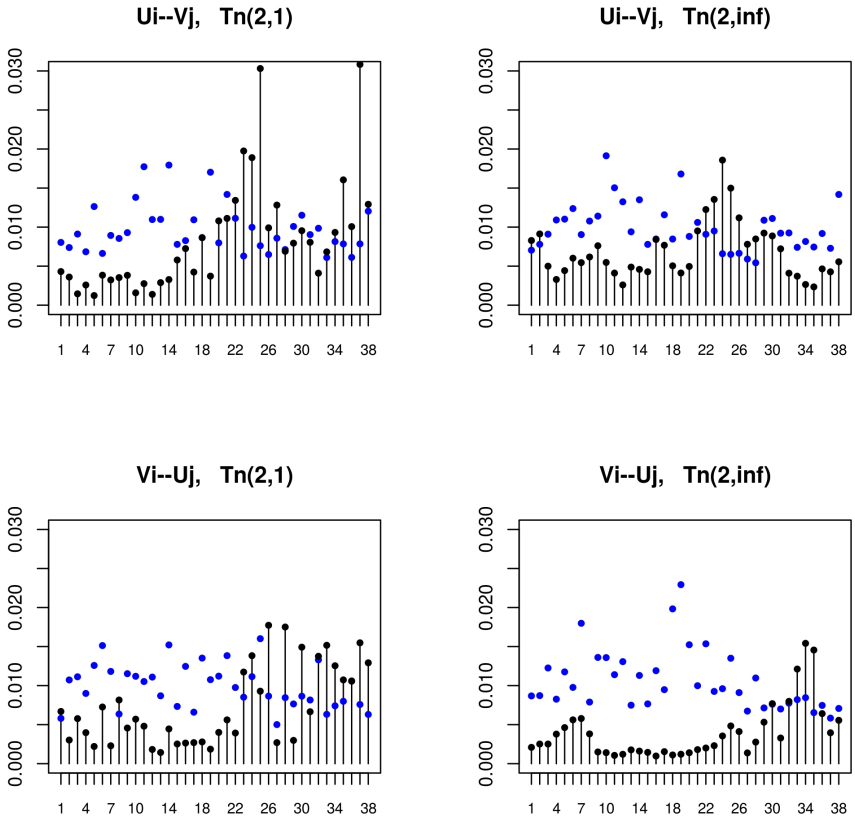

Figura 4.4: Comparación entre dependogramas para $T_n^{(2,1)}$ y $T_n^{(2,\infty)}$ entre U y V.

Las Figuras 4.5 y 4.6 muestran que las pruebas basadas en $T_n^{(2,1)}$ y $T_n^{(2,\infty)}$

se comportan de forma similar. Además, aún estando geográficamente cerca, los vectores U_i y T_j son independientes. Por otro lado, ambas pruebas detectan una dependencia entre T_i y U_j para $i = 21$ a $i = 27$ (Figura 4.5). La Figura 4.6 muestra que en en la mayoría de los casos, T_i y V_j son independientes, mientras que para V_i y T_j la prueba no detecta dependencia salvo en los casos en los cuales i y j están cercanos.

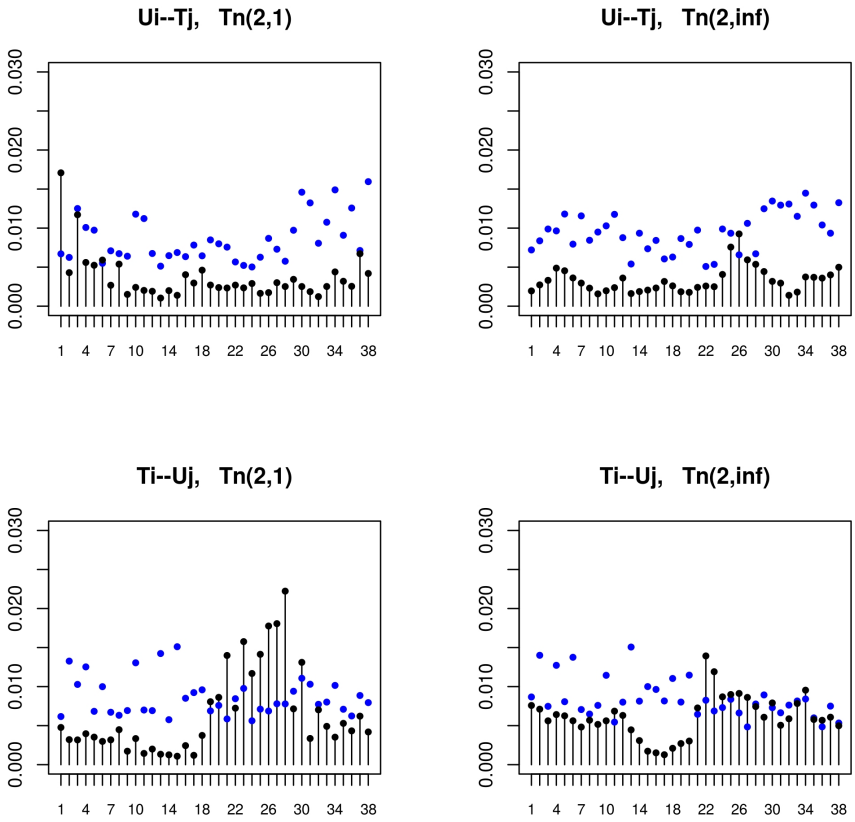

Figura 4.5: Comparación entre dependogramas para $T_n^{(2,1)}$ y $T_n^{(2,\infty)}$ entre U y T.

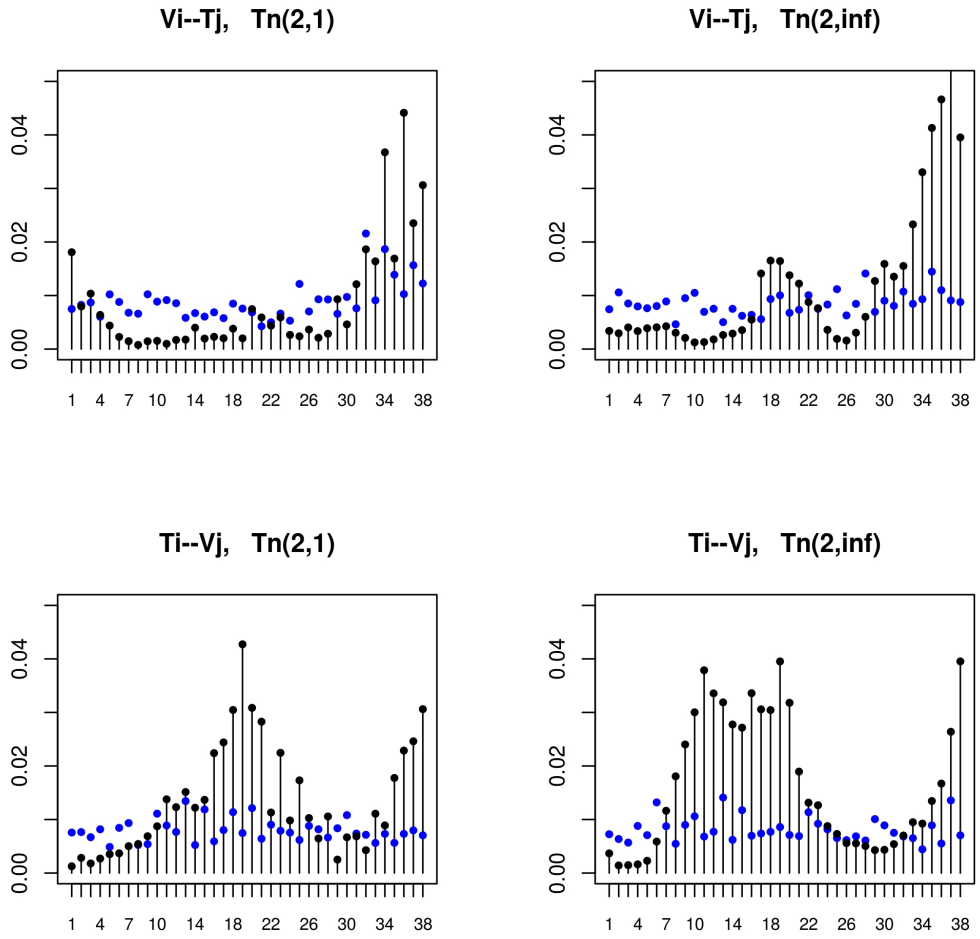

Figura 4.6: Comparación entre dependogramas para $T_n^{(2,1)}$ y $T_n^{(2,\infty)}$ entre V y T.

4.2. Datos Económicos

4.2.1. Tipo de cambio nominal y tasa Libor

En esta subsección se consideran las variables X = "Tipo de cambio nominal del peso contra el dólar", calculada como la cotización interbancaria promedio, e Y = "Tasa Libor a 6 meses", calculada como la cotización al cierre. Se utilizaron los datos diarios de ambas variables desde el 1 de setiembre de 1994 hasta el 31 de julio de 2019, totalizando 6152 datos. En la Figura 4.7 se observan los mismos donde cada punto corresponde a una medición del

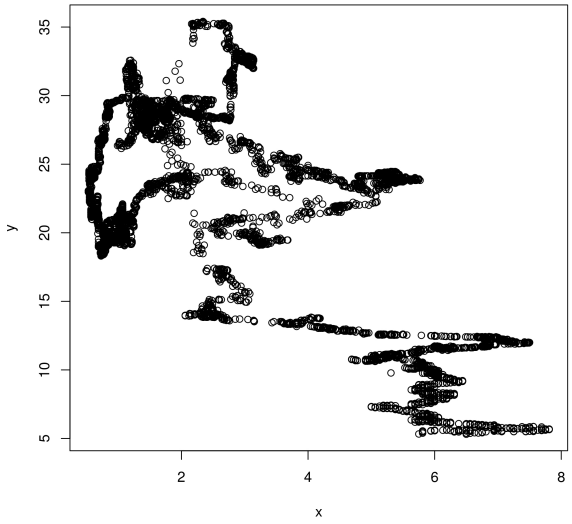

Figura 4.7: Tasa nominal peso contra dólar (x) y tasa Libor (y)

tipo de cambio nominal (y) y la tasa Libor (x) del mismo día. Del gráfico se desprende de manera evidente la dependencia entre los valores de X e Y, por ejemplo se ve que los valores de Y superiores a 30, se encuentran para valores de X menores que 4.

Ahora si definimos $X = (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}, X^{(5)})$ e $Y = (Y^{(1)}, Y^{(2)}, Y^{(3)}, Y^{(4)}, Y^{(5)})$ las series semanales de la tasa de Libor y de la tasa de cambio nominal del peso contra el dólar respectivamente, ¿son independientes? Intuitivamente parece claro que no. En este caso podríamos aplicar el test de independencia para el caso en el cual $X, Y \in \mathbb{R}^5$. Como el test requiere que las muestras semanales sean independientes, aplicamos el test, dejando algunas semanas sin contabilizar. Por ejemplo si dejamos un mes sin observar entre dos observaciones consecutivas tanto de X como de Y el p-valor del test da 0. Lo mismo ocurre (p-valor =0) si consideramos la dependencia o no de las series mensuales, es decir testeamos si $X = (X^{(1)}, X^{(2)}, ..., X^{(25)})$ e $Y = (Y^{(1)}, Y^{(2)}, ..., Y^{(25)})$ son independientes, es decir si testeamos la independencia para series mensuales.

4.2.2. Indicadores de bolsas de valores

En esta subsección se toman en cuenta indicadores de bolsas de valores de Europa (España y Alemania), Estados Unidos, Japón y Sudamérica (Brasil y

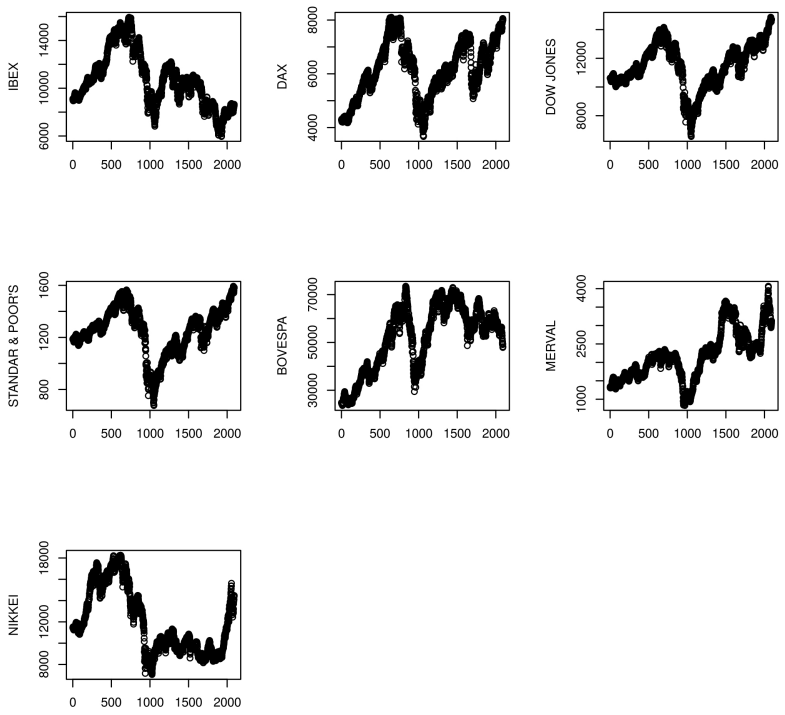

Figura 4.8: Indices de las bolsas de los países considerados

Tabla 4.2: p-valor para el test de independencia entre índices de dos países.

	IBEX 35	DAX 30	DJ	SP	Bovespa	Merval	Nikkei 225
IBEX 35		0.000	0.000	0.000	0.001	0.003	0.000
DAX 30			0.000	0.000	0.000	0.000	0.028
DJ				0.000	0.000	0.012	0.042
SP					0.000	0.012	0.000
Bovespa						0.000	0.000
Merval							0.000

Argentina). Sus respectivos nombres son IBEX (España), DAX (Alemania), Dow Jones y Standard & poor's (Estados Unidos), Nikkei (Japón), Bovespa (Brasil) y Merval (Argentina). Los valores considerados son diarios, desde el 2 de enero de 2005 hasta el 29 de mayo de 2013. En total son 2093 datos diarios. En este rango de tiempo hubo información para todas las series estudiadas, es decir días en los cuales las bolsas de todos los países considerados estaban operando. Es bien conocido que todas estas series están relacionadas entre sí, incluso se puede deducir del gráfico adjunto. Se aplicó el test de independencia a los datos de las series mensuales tomadas de a 25 días. Es decir que en la

aplicación del test consideramos como espacios $S_X = S_Y = \mathbb{R}^{25}$.

Se aplicó el test de independencia a las series de a dos. En la Tabla 4.2, se muestra el p-valor en cada caso. El p-valor fue calculado mediante la simulación de $m = 1000$ permutaciones. Como se ve, el test que planteamos en este trabajo, detecta muy claramente la dependencia entre todas las series consideradas.

Capítulo 5

Consideraciones finales

Detectar la dependencia entre datos es una tarea fundamental en el análisis científico.

En este trabajo se presentó una nueva prueba de independencia entre dos elementos aleatorios que se encuentran en espacios métricos cualesquiera.

La prueba se basa en porcentajes de recurrencias obtenidos a partir de la distancia entre puntos de cada muestra. Se obtuvo la distribución asintótica del estadístico y se demostró que la distribución del límite bajo alternativas contiguas tiene un sesgo.

Se probó también la consistencia de la prueba para una amplia clase de alternativas, que incluyen el caso particular en el que (X, Y) siguen una distribución normal multivariada.

La performance de la prueba, medida a través de la comparación de la potencia respecto de varias alternativas mostró muy buenos resultados, mostrando una mejora con respecto a otras pruebas en muchos casos para diferentes dimensiones.

Se mostró, utilizando una comparación de potencias, que el test de independencia de tasas de recurrencias en altas dimensiones supera el rendimiento de otros test competidores en casi todos los casos, y se estudió la incidencia de las funciones de distancia consideradas (d_X y d_Y) en la performance del test. Como era esperable, se mostró que el test estadístico en alta dimensión presenta sensibilidad respecto a la elección de la función de distancia, d_X o d_Y. Además, se propuso y se comparó el test frente a otros posibles funcionales a ser tenidos en cuenta, tal como un funcional del tipo L^1-Cramér–von Mises y un funcional del tipo Kolmogorov–Smirnov, y se mostró como calcular el

estadístico en cada caso.

Finalmente se aplicó el test a datos reales de tipo metereológico y económico. Como se ve, se detectó muy claramente la dependencia entre todas las series consideradas.

Como trabajo futuro se propone estudiar la adaptación del test para detectar causalidad entre dos o más sistemas dinámicos. Este es un tema que tiene una gran importancia en economía y otras áreas.

Referencias bibliográficas

[1] Arcones, M. A. and Giné, E. (1993). Limit theorems for u-processes. *The Annals of Probability*, 21:1494–1542.

[2] Arratia, A., Cabaña, A., and Cabaña, E. (2016). A construction of continuous time arma models by iterations of ornstein-uhlenbeck processes. *SORT-Statistics and Operations Research Transactions*, 40(2):267–302.

[3] Beran, R., Bilodeau, M., and de Micheaux, P. L. (2007). Nonparametric tests of independence between random vectors. *Journal of Multivariate Analysis*, 98(9):1805–1824.

[4] Boezio, G. C. and Ortelli, S. (2018). Minimum-cost numerical prediction system for wind power in uruguay, with an assessment of the diurnal and seasonal cycles of its quality. *Ciência e Natura*, 40:205–210.

[5] Boglioni, G. (2016). A consistent test of independence between random vectors.

[6] Cabaña, E. (1997). Contigüidad, pruebas de ajuste y procesos empıricos transformados. *Décima escuela venezolana de Matemáticas*.

[7] Dufour, J.-M., Lepage, Y., and Zeidan, H. (1982). Nonparametric testing for time series: a bibliography. *Canadian Journal of Statistics*, 10(1):1–38.

[8] Eckmann, J., Kamphorst, S. O., and Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhysics Ltters*, 5:973–977.

[9] Fraiman, R., Moreno, L., and Vallejo, S. (2017). Some hypothesis tests based on random projection. *Computational Statistics*, 32(3):1165–1189.

[10] Galton, F. (1889). I. co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279):135–145.

[11] Giné, E. and Zinn, J. (1986). Lectures on the central limit theorem for empirical processes. In *Probability and Banach spaces*, pages 50–113. Springer.

[12] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

[13] Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2008). A kernel statistical test of independence, in 'advances in neural information processing systems'.

[14] Heller, R., Heller, Y., and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510.

[15] Hoeffding, W. (1961). The strong law of large numbers for u-statistics. Technical report, North Carolina State University. Dept. of Statistics.

[16] Kalemkerian, J. (2020). Parameter estimation for the discretely observed fractional iterated ornstein–uhlenbeck processes. *arXiv preprint arXiv:2004.10369*.

[17] Kalemkerian, J. and Fernández, D. (2019). Implementación del test de independencia basado en porcentaje de recurrencias para series de tiempo.

[18] Kalemkerian, J. and Fernández, D. (2020a). An independence test based on recurrence rates. *Journal of Multivariate Analysis*, page 104624.

[19] Kalemkerian, J. and Fernández, D. (2020b). An independence test based on recurrence rates: An empirical study and applications to real data. *arXiv preprint arXiv:2009.08883*.

[20] Kalemkerian, J. and León, J. R. (2019). Fractional iterated ornstein-uhlenbeck processes. *ALEA*, 16:1105–1128.

[21] Le Cam, L. and Yang, G. L. (2012). *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media.

[22] Marwan, N., Romano, M. C., Thiel, M., and Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics reports*, 438(5-6):237–329.

[23] Oosterhoff, J. and Van Zwet, W. (1979). A note on contiguity and hellinger distance. contributions to statistics. *Reidel, Dordrecht % Boston, Mass. London*, 157:166.

[24] Pearson, K. (1897). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, 60(359-367):489–498.

[25] Rencher, A. C. (1995). Multivariate analysis of variance. *Methods of multivariate analysis. New York, NY: Wiley*, pages 174–257.

[26] Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X.-Y., Wang, W., and Powers, J. G. (2008). G.: A description of the advanced research wrf version 3. In *NCAR Tech. Note NCAR/TN-475+ STR*. Citeseer.

[27] Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272.

[28] Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

[29] Wald, A. and Wolfowitz, J. (1943). An exact test for randomness in the non-parametric case based on serial correlation. *The Annals of Mathematical Statistics*, 14(4):378–388.

[30] Webber, C. and Marwan, N. (2015). Recurrence quantification analysis. *Theory and Best Practices. Springer*.

[31] Webber Jr, C. L. and Zbilut, J. P. (1994). Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of applied physiology*, 76(2):965–973.

[32] Zbilut, J. P. and Webber Jr, C. L. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics letters A*, 171(3-4):199–203.

[33] Zou, Y., Romano, M. C., Thiel, M., and Kurths, J. (2015). Identifying coupling directions by recurrences. In *Recurrence Quantification Analysis*, pages 65–99. Springer.

[34] Zou, Y., Romano, M. C., Thiel, M., Marwan, N., and Kurths, J. (2011). Inferring indirect coupling by means of recurrences. *International Journal of Bifurcation and Chaos*, 21(04):1099–1111.

APÉNDICES

Apéndice 1

Demostración de los resultados enunciados en el Capítulo 2

Prueba del Lema 1. Se observa que como $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ son i.i.d, entonces

$$P\left(d\left(X_i, X_j\right) < r \ , \ d\left(Y_i, Y_j\right) < s\right) = p_{X,Y}\left(r, s\right)$$

para todos i, j tal que $i \neq j$. Por lo tanto,
$E\left(RR_n^{X,Y}(r,s)\right)$

$$= E\left(\frac{1}{n^2 - n}\sum_{i \neq j} \mathbf{1}_{\{d(X_i,X_j)<r \ , \ d(Y_i,Y_j)<s\}}\right) =$$

$$= \frac{1}{n^2 - n}\sum_{i \neq j} P\left(d\left(X_i, X_j\right) < r \ , \ d\left(Y_i, Y_j\right) < s\right) = p_{X,Y}(r,s).$$

De forma análoga, $E\left(RR_n^X(r)\right) = p_X(r)$ y $E\left(RR_n^Y(s)\right) = p_Y(s)$. Dado que X e Y son independientes, entonces

$$E\left(RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s)\right) = 0.$$

Por lo tanto,

$$Cov\left(E_n(r,s), E_n(r',s')\right) =$$

$$= nE\left[\left(RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s)\right)\left(RR_n^{X,Y}(r',s') - RR_n^X(r')RR_n^Y(s')\right)\right]$$

$$= n[E\left[\left(RR_n^{X,Y}(r,s)\right)\left(RR_n^{X,Y}(r',s')\right)\right] - E\left(RR_n^{X,Y}(r,s)RR_n^X(r')RR_n^Y(s')\right)$$

$$-E\left(RR_n^{X,Y}(r',s')RR_n^X(r)RR_n^Y(s)\right)$$
$$+E\left[RR_n^X(r)RR_n^X(r')\right]E\left[RR_n^Y(s)RR_n^Y(s')\right]]. \tag{1.1}$$

Además,
$E\left(RR_n^X(r)RR_n^X(r')\right)$

$$= E\left(\frac{1}{n^2(n-1)^2}\sum_{i\neq j}\sum_{h\neq k}\mathbf{1}_{\{d(X_i,X_j)<r,\ d(X_h,X_k)<r'\}}\right) = \tag{1.2}$$

$$= \frac{1}{n^2(n-1)^2}\sum_{i\neq j}\sum_{h\neq k}P\left(d\left(X_i,X_j\right)<r,\ d\left(X_h,X_k\right)<r'\right). \tag{1.3}$$

Descomponiendo (1.3) en los términos en que i,j,k,h son diferentes por pares $\{i,j\}=\{h,k\}$ y $\{i,j,h,k\}$ tiene tres elementos, y utilizando que las X_i son i.i.d, se obtiene que (1.3) es igual a

$$\frac{n(n-1)(n-2)(n-3)p_X(r)p_X(r')+2n(n-1)p_X(r)+4n(n-1)(n-2)p_X^{(3)}(r\wedge r')}{n^2(n-1)^2}$$

$$= \frac{n-2}{n(n-1)}\left[(n-3)p_X(r)p_X(r')+4p_X^{(3)}(r\wedge r')\right]+o\left(\frac{1}{n}\right). \tag{1.4}$$

De forma análoga,

$$E\left[\left(RR_n^Y(s)\right)RR_n^Y(s')\right] = \frac{n-2}{n(n-1)}\left((n-3)p_Y(s)p_Y(s')+4p_Y^{(3)}(s\wedge s')\right)+o\left(\frac{1}{n}\right). \tag{1.5}$$

Similarmente, utilizando que los vectores aleatorios (X_i,Y_i) son i.i.d. y también que X e Y son independientes,

$E\left[RR_n^{X,Y}(r,s)RR_n^{X,Y}(r',s')\right]$

$$= \frac{(n-2)(n-3)p_X(r)p_X(r')p_Y(s)p_Y(s')+2p_X(r)p_Y(s)+4(n-2)p_X^{(3)}(r\wedge r')p_Y^{(3)}(s\wedge s')}{n(n-1)}$$

$$= \tfrac{n-2}{n(n-1)}\left[(n-3)p_X(r)p_X(r')p_Y(s)p_Y(s')+4p_X^{(3)}(r\wedge r')p_Y^{(3)}(s\wedge s')\right]+o\left(\tfrac{1}{n}\right). \tag{1.6}$$

Con la misma técnica que en (1.4) y (1.5), se obtiene

$$E\left[RR_n^{X,Y}(r,s)RR_n^X(r')RR_n^Y(s')\right]$$

$$= E\left(\frac{1}{n^3(n-1)^3}\sum_{i\neq j}\sum_{h\neq k}\sum_{l\neq m}\mathbf{1}_{\{d(X_i,X_j)<r, d(Y_i,Y_j)<s,\ d(X_h,X_k)<r', d(Y_l,Y_m)<s'\}}\right)$$

$$= \tfrac{1}{n^3(n-1)^3}\sum_{i\neq j}\sum_{h\neq k}\sum_{l\neq m} P\left(d\left(X_i,X_j\right)<r, d\left(Y_i,Y_j\right)<s,\ d\left(X_h,X_k\right)<r', d\left(Y_l,Y_m\right)<s'\right)$$

$$= \tfrac{1}{n^3(n-1)^3}\sum_{i\neq j}\sum_{h\neq k}\sum_{l\neq m} P\left(d\left(X_i,X_j\right)<r,\ d\left(X_h,X_k\right)<r'\right) P\left(d\left(Y_i,Y_j\right)<s, d\left(Y_l,Y_m\right)<s'\right)$$

$$= \frac{1}{n^3(n-1)^3}\left[n(n-1)(n-2)^2(n-3)^2 p_X(r)p_X(r')p_Y(s)p_Y(s')\right]$$

$$+\frac{1}{n^3(n-1)^3}\left[4n(n-1)(n-2)^2(n-3)p_X(r)p_X(r')p_Y^{(3)}(s\wedge s')\right]$$

$$+\frac{1}{n^3(n-1)^3}[4n(n-1)(n-2)^2(n-3)p_X^{(3)}(r\wedge r')p_Y(s)p_Y(s')$$

$$+8n(n-1)(n-2)p_X(r)p_Y^{(3)}(s\wedge s')]$$

$$+\tfrac{1}{n^3(n-1)^3}\left[8n(n-1)(n-2)p_X^{(3)}(r\wedge r')p_Y(s)+2n(n-1)(n-2)(n-3)p_X(r)p_X(r')p_Y(s)\right]$$

$$+\frac{1}{n^3(n-1)^3}[2n(n-1)(n-2)(n-3)p_X(r)p_Y(s)p_Y(s')$$

$$+16n(n-1)(n-2)^2p_X^{(3)}(r\wedge r')p_Y^{(3)}(s\wedge s')]$$

$$+\frac{1}{n^3(n-1)^3}4n(n-1)p_X(r)p_Y(s).$$

Por lo tanto,

$$E\left[RR_n^{X,Y}(r,s)RR_n^X(r')RR_n^Y(s')\right]$$

$$= \tfrac{1}{n^2(n-1)^2}\left[(n-2)^2(n-3)^2p_X(r)p_X(r')p_Y(s)p_Y(s') + 4(n-2)^2(n-3)p_X(r)p_X(r')p_Y^{(3)}(s)\right]$$

$$+\frac{1}{n^2(n-1)^2}[4(n-2)^2(n-3)p_X^{(3)}(r\wedge r')p_Y(s)p_Y(s')+$$

$$+8(n-2)p_X(r)p_Y^{(3)}(s\wedge s') + 8(n-2)p_X^{(3)}(r)p_Y(s)]$$

$$+\tfrac{1}{n^2(n-1)^2}\left[2(n-2)(n-3)p_X(r)p_X(r')p_Y(s) + 2(n-2)(n-3)p_X(r)p_Y(s)p_Y(s')\right]$$

$$+\frac{1}{n^2(n-1)^2}\left[16(n-2)^2p_X^{(3)}(r\wedge r')p_Y^{(3)}(s\wedge s') + 4p_X(r)p_Y(s)\right]$$

$$= \frac{(n-2)^2(n-3)}{n^2(n-1)^2}[(n-3)p_X(r)p_X(r')p_Y(s)p_Y(s')+$$

$$+4\left(p_X^{(3)}(r\wedge r')p_Y(s)p_Y(s') + p_X(r\wedge r')p_Y^{(3)}(s)\right)] + o(\frac{1}{n}).$$

Sustituyendo (1.4), (1.5) y (1.6) en (1.1), se obtiene que (1.1) es igual a

$$\frac{1}{n^2(n-1)^2}\left[(n-2)(n-3)(4n-6)p_X(r)p_X(r')p_Y(s)p_Y(s')\right]$$

$$+\frac{1}{n^2(n-1)^2}\left[4(n-2)(n^2+3n-8)p_X^{(3)}(r\wedge r')p_Y^{(3)}(s\wedge s')\right]$$

$$+\frac{-4(n-2)^2(n-3)}{n^2(n-1)^2}\left(p_X(r)p_X(r')p_Y^{(3)}(s\wedge s') + p_X^{(3)}(r\wedge r')p_Y^2(s)\right)+o\left(\frac{1}{n}\right).$$

Luego

$$\lim_{n\to+\infty} Cov\left(E_n(r,s),E_n(r',s')\right) = 4\left(p_X^{(3)}(r\wedge r') - p_X(r)p_X(r')\right)\left(p_Y^{(3)}(s\wedge s') - p_Y(s)p_Y(s')\right).$$

□

Prueba del Lema 2. Sea $(S,\mathcal{S},P)$ un espacio de probabilidad, y para todo $i\in\mathbb{N}$, $X_i : S\to S$ una sucesión i.i.d. con ley o distribución de X_i, $\mathcal{L}(X_i) = P$. Dado m, sea $\mathcal{F}$ una clase de funciones medibles en S^m, el U-proceso basado en P e indexado por $\mathcal{F}$ se define como

$$U_m^n(f) = \frac{(n-m)!}{m!}\sum_{(i_1,\ldots,i_m)\in I_m^n} f\left(X_{i_1},X_{i_2},\ldots,X_{i_m}\right),$$

donde $f\in\mathcal{F}$.

Dado $\varepsilon > 0$, se considera el conjunto $A(\varepsilon, \mathcal{F}, P^m)$ de números positivos v que verifican que existe $\mathcal{L} = \{l_1, l_2, ..., l_v\}$, $\mathcal{U} = \{u_1, u_2, ..., u_v\}$ tal que $\mathcal{L}$,$\mathcal{U} \subset L^2$ y para todo $f \in \mathcal{F}$, existe $l_f \in \mathcal{L}$ y $u_f \in \mathcal{U}$ donde $l_f \leq f \leq u_f$ *a.s.* y $E(u_f - l_f)^2 < \varepsilon^2$.

Teorema 5 (Arcones & Giné 1993)**.** *Si* $N_{[\ ]}^{(2)}(\varepsilon, \mathcal{F}, P^m) = \text{mín}\, A(\varepsilon, \mathcal{F}, P^m)$ *y*

$$\int_0^{+\infty} \left(\log N_{[\ ]}^{(2)}(\varepsilon, \mathcal{F}, P^m)\right)^{1/2} d\varepsilon < +\infty, \tag{1.7}$$

luego

$$\mathcal{L}\left(\sqrt{n}\,(U_m^n - P^m)\,f\right) \xrightarrow{w} \mathcal{L}\left(mG_p \circ P^{m-1} f\right) \ \textit{en}\ l^\infty(\mathcal{F}), \tag{1.8}$$

donde G_P *es el puente Browniano asociado con P.*

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

$$= \frac{\sqrt{n}}{n(n-1)} \sum_{(i,j)\in I_2^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_j)<s\}} - \sqrt{n}RR_n^X(r)RR_n^Y(s)$$

$$= \tfrac{\sqrt{n}}{n(n-1)(n-2)(n-3)} \textstyle\sum_{(i,j,h,k)\in I_4^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_j)<s\}} - \sqrt{n}RR_n^X(r)RR_n^Y(s) =$$

$$= E_n'(r,s) - H_n(r,s),$$

donde

$$H_n(r,s) = \sqrt{n}\left(RR_n^X(r)RR_n^Y(s) - \tfrac{1}{n(n-1)(n-2)(n-3)} \textstyle\sum_{(i,j,k,h)\in I_4^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}}\right).$$

Luego, $H_n(r,s)$ es igual a

$$\frac{\sqrt{n}}{n^2(n-1)^2} \sum_{(i,j)\in I_2^n} \mathbf{1}_{\{d(X_i,X_j)<r\}} \sum_{(h,k)\in I_2^n} \mathbf{1}_{\{d(Y_h,Y_k)<s\}}$$

$$-\frac{\sqrt{n}}{n(n-1)(n-2)(n-3)} \sum_{(i,j,k,h)\in I_4^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}}$$

$$= \frac{\sqrt{n}}{n^2(n-1)^2} \frac{1}{n(n-1)} \sum_{(i,j)\in I_2^n} \sum_{(h,k)\in I_2^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}}$$

$$- \frac{\sqrt{n}}{n(n-1)(n-2)(n-3)} \sum_{(i,j,k,h)\in I_4^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}}. \tag{1.9}$$

Ahora, descomponemos

$$\sum_{(i,j)\in I_2^n} \sum_{(h,k)\in I_2^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}} = \sum_{(i,j,k,h)\in I_4^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}}$$

$$+4 \sum_{(i,j,k)\in I_3^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_k)<s\}}$$

$$+2 \sum_{(i,j)\in I_2^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_j)<s\}}$$

y sustituyendo en (1.9) se obtiene que (1.9) es igual a

$$\frac{\sqrt{n}}{n(n-1)} \left(\left(\frac{1}{n(n-1)} - \frac{1}{(n-2)(n-3)} \right) \sum_{(i,j,k,h)\in I_4^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}} \right)$$

$$+\frac{\sqrt{n}}{n^2(n-1)^2} \left(4 \sum_{(i,j,k)\in I_3^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_k)<s\}} + 2 \sum_{(i,j)\in I_2^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_j)<s\}} \right)$$

$$= \frac{\sqrt{n}}{n^2(n-1)^2(n-2)(n-3)} \sum_{(i,j,k,h)\in I_4^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}}$$

$$+\frac{\sqrt{n}}{n^2(n-1)^2} \left(4 \sum_{(i,j,k)\in I_3^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_k)<s\}} + 2 \sum_{(i,j)\in I_2^n} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_j)<s\}} \right). \tag{1.10}$$

Se observa que (1.10) está acotada entre 0 y $\frac{4}{\sqrt{n}}$ ya que

$$\frac{\sqrt{n}}{n^2(n-1)^2} \left(4n(n-1)(n-2) + 2n(n-1)\right) = \frac{1}{\sqrt{n}} \frac{4n-6}{n-1} < \frac{4}{\sqrt{n}}.$$

□

Prueba del Teorema 3. Toda función continua $h : \mathbb{R} \to \mathbb{R}$ con límite finito cuando $x \to \pm\infty$ es uniformemente continua. Por lo tanto dado $\varepsilon > 0$, existe $\delta > 0$ tal que $|F(x) - F(y)| \leq \varepsilon^2/8$ y $|G(x) - G(y)| \leq \varepsilon^2/8$ para todo (x, y) tal que $|x - y| < \delta$, donde F y G son las funciones de distribución de $d(X_1, X_2)$ y

$d(Y_1, Y_2)$ respectivamente. Si H_0 es cierta, se considera para cada $r, s > 0$ las funciones $f_{r,s} : (S_X \times S_Y)^4 \to \mathbb{R}$ definidas por

$$f_{r,s}\left(x, y, x', y', x'', y'', x''', y'''\right) = \mathbf{1}_{\{d(x,x')<r,\ d(y,y')<s\}} - \mathbf{1}_{\{d(x'',x''')<r,\ d(y'',y''')<s\}},$$

donde $x, x', x'', x''' \in S_X$ y $y, y', y'', y''' \in S_Y$ y se considera la familia $\mathcal{F} = \{f_{r,s}\}_{r,s>0}$. Para simplificar la notación, se denomina $z = (x, y, x', y', x'', y'', x''', y''')$ a lo largo de la demostración.

Se observa que

$$E'_n(r, s) = \frac{\sqrt{n}}{n(n-1)(n-2)(n-3)} \sum_{(i,j,k,h)\in I_4^n} f_{r,s}\left(X_i, Y_i, X_j, Y_j, X_h, Y_h, X_k, Y_k\right)$$

y luego el proceso $\{E'_n(r, s)\}_{r,s>0}$ es un $U-$proceso de orden 4.

Para obtener la convergencia débil del proceso $\{E_n(r, s) - E(E_n(r, s))\}_{r,s>0}$ a un proceso Gausiano centrado (con lo cual la distribución asintótica del estadístico $T_n^{(2)}$ definido en (2.3) queda determinada), se utilizará el Teorema 4.10 obtenido por Arcones & Giné [1]:

La convergencia es en el espacio $l^\infty(\mathcal{F})$, es en el sentido de Hoffmann-Jørgensen, ver ([11]).

Para obtener la convergencia, según Teorema de Arcones & Giné, es suficiente probar que

$$\int_0^{+\infty} \left(\log N_{[\,]}^{(2)}\left(\varepsilon, \mathcal{F}, P^4\right)\right)^{1/2} d\varepsilon < +\infty.$$

Si $\varepsilon \geq 2$, se tiene que $-1 \leq f_{r,s}(z) \leq 1$ para todo $z \in (S_X \times S_Y)^4$ y $r, s > 0$. Luego $\mathcal{L} = \{-1\}$, $\mathcal{U} = \{1\}$ satisface el Supuesto 1 definido en Teorema de Arcones & Giné. En consecuencia, $N_{[\,]}^{(2)}(\varepsilon, \mathcal{F}, P^4) = 1$, por lo tanto $\int_0^{+\infty} \left(\log N_{[\,]}^{(2)}(\varepsilon, \mathcal{F}, P^4)\right)^{1/2} d\varepsilon = \int_0^2 \left(\log N_{[\,]}^{(2)}(\varepsilon, \mathcal{F}, P^4)\right)^{1/2} d\varepsilon$.

Si $\varepsilon < 2$, se toma $T > 0$ tales que $\max\{1 - F(T), 1 - G(T)\} < \varepsilon^2/8$, luego se particiona $[0, +\infty)$ en $m+1$ subintervalos de la forma $\left[\frac{iT}{m}, \frac{(i+1)T}{m}\right)$ tal que $\frac{T}{m} < \delta$, donde $\frac{(m+1)T}{m}$ es interpretado como $+\infty$. Se definen las siguientes funciones

$$g_{i,j}(z) = \begin{cases} \mathbf{1}_{\left\{d(x,x')<\frac{iT}{m},\ d(y,y')<\frac{jT}{m}\right\}} & \text{para} \quad i, j \in \{1, ..., m\}, \\ 0 & \text{para} \quad i = 0 \text{ o } j = 0 \end{cases}$$

y

$$h_{i,j}(z) = \begin{cases} \mathbf{1}_{\left\{d(x,x')<\frac{iT}{m},\ d(y'',y''')<\frac{jT}{m}\right\}} & \text{para} \quad i,j \in \{1,...,m\}, \\ \mathbf{1}_{\left\{d(x,x')<\frac{iT}{m}\right\}} & \text{para} \quad i \in \{1,...,m\},\ j = m+1, \\ \mathbf{1}_{\left\{d(y'',y''')<\frac{jT}{m}\right\}} & \text{para} \quad j \in \{1,...,m\},\ i = m+1, \\ 1 & \text{para} \quad i = j = m+1 \end{cases}.$$

Se observa que para cada $r, s > 0$ existen $i, j \in \{0, 1, 2, ..., m\}$ tal que t $\frac{iT}{m} \leq r < \frac{(i+1)T}{m}$ y $\frac{jT}{m} \leq s < \frac{(j+1)T}{m}$.

Luego

$$g_{i,j}(z) - h_{i+1,j+1}(z) \leq f_{r,s}(z) \leq g_{i+1,j+1}(z) - h_{i,j}(z) \text{ para todo } z \in (S_X \times S_Y)^4,$$

Por tanto $\mathcal{L} = \{l_{i,j}\}$ y $\mathcal{U} = \{u_{i,j}\}$ donde $l_{i,j}(z) = g_{i,j}(z) - h_{i+1,j+1}(z)$ y $u_{i,j}(z) = g_{i+1,j+1}(z) - h_{i,j}(z)$ para $i, j \in \{1, ..., m\}$. Además

$$E\left(u_{i,j}(Z) - l_{i,j}(Z)\right)^2 \leq 2\left(E\left(g_{i+1,j+1}(Z) - g_{i,j}(Z)\right)^2 + E\left(h_{i+1,j+1}(Z) - h_{i,j}(Z)\right)^2\right). \tag{1.11}$$

Se definen los conjuntos $A_{i,j} := \left[0, \frac{(i+1)T}{m}\right) \times \left[0, \frac{(j+1)T}{m}\right) - \left[0, \frac{iT}{m}\right) \times \left[0, \frac{jT}{m}\right)$, luego

$$E\left(g_{i+1,j+1}(Z) - g_{i,j}(Z)\right)^2 = E\left(\mathbf{1}_{A_{i,j}}(Z)\right) \leq P\left(\tfrac{iT}{m} \leq d\left(X_1, X_2\right) < \tfrac{(i+1)T}{m}\right) + P\left(\tfrac{jT}{m} \leq d\left(Y_1, Y_2\right) < \tfrac{(j+1)T}{m}\right)$$

$$\leq F\left(\frac{(i+1)T}{m}\right) - F\left(\frac{iT}{m}\right) + G\left(\frac{(j+1)T}{m}\right) - G\left(\frac{jT}{m}\right) \leq \varepsilon^2/4. \tag{1.12}$$

De manera análoga,

$$E\left(h_{i+1,j+1}(Z) - h_{i,j}(Z)\right)^2 \leq \varepsilon^2/4. \tag{1.13}$$

Sustituyendo (1.13) y (1.12) en (1.11) se obtiene que $E\left(u_{i,j}(Z) - l_{i,j}(Z)\right)^2 \leq \varepsilon^2$.

Finalmente, se observa que el cardinal de $\mathcal{L}$ y $\mathcal{U}$ es $(m+1)^2$, luego

$$N_{[\,]}^{(2)}\left(\varepsilon, \mathcal{F}, P^4\right) \leq \frac{cte}{\varepsilon^4}, \text{ por tanto } \int_0^2 \left(\log N_{[\,]}^{(2)}\left(\varepsilon, \mathcal{F}, P^4\right)\right)^{1/2} d\varepsilon < +\infty.$$

□

Prueba del Teorema 4. Se define

$$\mu(r,s) = P\left(d(X_1,X_2) < r,\ d(Y_1,Y_2) < s\right) - P\left(d(X_1,X_2) < r\right) P\left(d(Y_1,Y_2) < s\right)$$

Luego, existen $r_0, s_0 > 0$, tales que $\mu^2(r_0,s_0) > 0$, así existe $\varepsilon > 0$ y $A \subset [0,+\infty)^2$ tal que $(r_0,s_0) \in A$ y $\mu^2(r,s) > \varepsilon$ para todo $(r,s) \in A$. Luego, cuando $n \to +\infty$,

$$n \int_0^{+\infty} \int_0^{+\infty} \mu^2(r,s)\, g(r,s) dr ds \geq n\varepsilon \iint_A g(r,s) dr ds \to +\infty.$$

Ahora, utilizando que $(a+b)^2 \leq 2(a^2+b^2)$ se obtiene que

$$n \int_0^{+\infty} \int_0^{+\infty} \mu^2(r,s)\, g(r,s) dr ds \leq$$

$$\leq 2n \int_0^{+\infty} \int_0^{+\infty} \left(RR_n^{X,Y}(r,s) - RR_n^X(r) RR_n^Y(s) - \mu(r,s)\right)^2 g(r,s) dr ds$$

$$+2n \int_0^{+\infty} \int_0^{+\infty} \left(RR_n^{X,Y}(r,s) - RR_n^X(r) RR_n^Y(s)\right)^2 g(r,s) dr ds.$$

En consecuencia

$T_n^{(2)} = n \int_0^{+\infty} \int_0^{+\infty} \left(RR_n^{X,Y}(r,s) - RR_n^X(r) RR_n^Y(s)\right)^2 g(r,s) dr ds \xrightarrow{P} +\infty$ cuando $n \to +\infty$.

□

Prueba del Corolario 1. Utilizando que todas las normas en $\mathbb{R}^p$ y $\mathbb{R}^q$ son equivalentes, alcanza con dar la prueba en el caso de la norma Euclideana. Se utiliza que si (Z,T) tiene una distribución normal bivariada centrada, entonces $Cov\left(Z^2,T^2\right) = 2\left(Cov(Z,T)\right)^2$.

Se denomina $X = \left(X_{(1)}, X_{(2)}, ..., X_{(p)}\right)$ y $Y = \left(Y_{(1)}, Y_{(2)}, ..., Y_{(q)}\right)$. Luego

$$Cov\left(\|X\|^2, \|Y\|^2\right) = Cov\left(\sum_{i=1}^{p} X_{(i)}^2, \sum_{j=1}^{q} Y_{(j)}^2\right) = 2\sum_{i=1}^{p}\sum_{j=1}^{q} \left(Cov\left(X_{(i)}, Y_{(j)}\right)\right)^2.$$

Si X e Y no son independientes, entonces existen i y j tales que $Cov\left(X_{(i)}, Y_{(j)}\right) \neq 0$, luego $Cov\left(\|X\|^2, \|Y\|^2\right) > 0$, luego $\|X\|^2$ y $\|Y\|^2$ no

son independientes, por lo tanto $\|X\|$ y $\|Y\|$ no son independientes, y entonces existen números positivos r y s tales que $P(\|X\| < r,\ \|Y\| < s) \neq P(\|X\| < r)\, P(\|Y\| < s)$. Si se aplica este argumento para $X_1 - X_2$ y $Y_1 - Y_2$ en lugar de X e Y, entonces se obtiene que

$$P(\|X_1 - X_2\| < r,\ d\|Y_1 - Y_2\| < s) \neq P(\|X_1 - X_2\| < r)\, P(\|Y_1 - Y_2\| < s).$$

Finalmente, el resultado se sigue del Lema 2. □

Prueba de la Proposición 1.

$$E^{(n)}\left(RR_n^{X,Y}(r,s)\right) = E^{(n)}\left(\frac{1}{N}\sum_{(i,j)\in I_2} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_i,Y_j)<s\}}\right) =$$

$$= P^{(n)}\left(d(X_i,X_j) < r,\ d(Y_i,Y_j) < s\right). \tag{1.14}$$

Se define
$A_{r,s} := \{(x_1,y_1,x_2,y_2) \in \mathbb{R}^{2p+2q} :\ d(x_1,x_2) < r,\ d(y_1,y_2) < s\}$, luego (1.14) es igual a

$$c_n^2(\delta) \iiiint_{A_{r,s}} f_{X,Y}^{(n)}(x_1,y_1) f_{X,Y}^{(n)}(x_2,y_2) dx_1 dx_2 dy_1 dy_2 =$$

$$= c_n^2(\delta) \iiiint_{A_{r,s}} f_X(x_1) f_Y(y_1) f_X(x_2) f_Y(y_2)$$

$$\times \left(1 + \tfrac{\delta}{2\sqrt{n}} k_n(x_1,y_1)\right)^2 \left(1 + \tfrac{\delta}{2\sqrt{n}} k_n(x_2,y_2)\right)^2 dx_1 dx_2 dy_1 dy_2 = c_n^2(\delta)\, p_X^{(0)}(r) p_Y^{(0)}(s)$$

$$+ c_n^2 \tfrac{\delta}{\sqrt{n}} \iiiint_{A_{r,s}} \left(k_n(x_1,y_1) + k_n(x_2,y_2)\right) f_X(x_1) f_Y(y_1) f_X(x_2) f_Y(y_2) dx_1 dx_2 dy_1 dy_2 + \varepsilon_n(r,s),$$

donde $|\varepsilon_n(r,s)| \leq \frac{c}{\sqrt{n}}$ para todo $r, s > 0$ y c es una constante. Además,

$$E^{(n)}\left(RR_n^X(r) RR_n^Y(s)\right) = \frac{1}{N^2} E^{(n)}\left(\sum_{(i,j)\in I_2,\ (h,k)\in I_2} \mathbf{1}_{\{d(X_i,X_j)<r,\ d(Y_h,Y_k)<s\}}\right) =$$

$$= c_n^2(\delta) \frac{(n-2)(n-3)}{N} p_X^{(0)}(r) p_Y^{(0)}(s)$$

$$+ \frac{2}{N} P^{(n)}(A_{r,s}) + \frac{4(n-2)}{N} P^{(n)}\left(d(X_1,X_2) < r,\ d(Y_1,Y_3) < s\right).$$

Por lo tanto

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

$$= \sqrt{n}c_n^2\left(\delta\right)\frac{4n-6}{N}p_X^{(0)}(r)p_Y^{(0)}(s) + \delta c_n^2\left(\delta\right)$$

$$\times \iiiint_{A_{r,s}} \left(k_n(x_1,y_1) + k_n(x_2,y_2)\right) f_X(x_1)f_Y(y_1)f_X(x_2)f_Y(y_2)dx_1dx_2dy_1dy_2 + \varepsilon_n\left(r,s\right).$$

Entonces, al tender $n \to +\infty$

$$E^{(n)}\left(E_n(r,s)\right) \to \delta \iiiint_{A_{r,s}} \left(k(x_1,y_1) + k(x_2,y_2)\right) f_X(x_1)f_Y(y_1)f_X(x_2)f_Y(y_2)dx_1dx_2dy_1dy_2.$$

□

Prueba de la Proposición 3.

Reordenamos $d\left(X_i,X_j\right)$ con $(i,j) \in I_2^n$ en la forma $Z_1, Z_2, ..., Z_n$. Asumimos que $Z_1 < Z_2 < ... < Z_n$, y usaremos $T_1, T_2, ..., T_n$ para denotar los valores de $d\left(Y_i,Y_j\right)$ utilizando la misma indexación. También escribiremos $T_1^*, T_2^*, ..., T_n^*$ para los estadísticos de orden $T's$.

$$\int_0^{+\infty}\int_0^{+\infty} \left|RR_n^{X,Y}\left(r,s\right) - RR_n^X\left(r\right)RR_n^Y\left(s\right)\right| g_1\left(r\right)g_2\left(s\right)drds =$$

$$\frac{1}{N}\int_0^{+\infty} g_2(s)ds \times$$

$$\int_0^{+\infty} \left|\sum_{i\neq j} \mathbf{1}_{\{d(X_i,X_j)<r,\, d(Y_i,Y_j)<s\}} - \frac{1}{N}\sum_{i\neq j}\mathbf{1}_{\{d(X_i,X_j)<r,\}}\sum_{h\neq k}\mathbf{1}_{\{d(Y_h,Y_k)<s\}}\right| g_1\left(r\right)dr =$$

$$\frac{1}{N}\int_0^{+\infty} g_2(s)ds \int_0^{+\infty} \left|\sum_{i=1}^N \mathbf{1}_{\{Z_i<r,\, T_i<s\}} - \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\{Z_i<r\}}\sum_{j=1}^N \mathbf{1}_{\{T_j<s\}}\right| g_1\left(r\right)dr. \tag{1.15}$$

Se observa que

$$\int_0^{+\infty} \left|\sum_{i=1}^N \mathbf{1}_{\{Z_i<r,\, T_i<s\}} - \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\{Z_i<r\}}\sum_{j=1}^N \mathbf{1}_{\{T_j<s\}}\right| g_1\left(r\right)dr =$$

$$\sum_{h=1}^{N-1}\int_{Z_h}^{Z_{h+1}}\left|\sum_{i=1}^{h}\mathbf{1}_{\{T_i<s\}}-\frac{h}{N}\sum_{j=1}^{N}\mathbf{1}_{\{T_j<s\}}\right|g_1(r)\,dr=$$

$$\sum_{h=1}^{N-1}\left|\sum_{i=1}^{h}\mathbf{1}_{\{T_i<s\}}-\frac{h}{N}\sum_{j=1}^{N}\mathbf{1}_{\{T_j<s\}}\right|\left(G_1(Z_{h+1})-G_1(Z_h)\right).$$

Luego, (1.15) es igual a

$$\frac{1}{N}\sum_{h=1}^{N-1}\left(G_1(Z_{h+1})-G_1(Z_h)\right)\int_0^{+\infty}\left|\sum_{i=1}^{h}\mathbf{1}_{\{T_i<s\}}-\frac{h}{N}\sum_{j=1}^{N}\mathbf{1}_{\{T_j<s\}}\right|g_2(s)ds=$$

$$\frac{1}{N}\sum_{h=1}^{N-1}\left(G_1(Z_{h+1})-G_1(Z_h)\right)\sum_{j=1}^{N-1}\int_{T_j^*}^{T_{j+1}^*}\left|\sum_{i=1}^{h}\mathbf{1}_{\{T_i<s\}}-\frac{h}{N}\sum_{j=1}^{N}\mathbf{1}_{\{T_j<s\}}\right|g_2(s)ds=$$

$$\frac{1}{N}\sum_{h=1}^{N-1}\left(G_1(Z_{h+1})-G_1(Z_h)\right)\sum_{j=1}^{N-1}\int_{T_j^*}^{T_{j+1}^*}\left|c(h,j)-\frac{jh}{N}\right|g_2(s)ds=$$

$$\frac{1}{N}\sum_{h,j=1}^{N-1}\left(G_1(Z_{h+1})-G_1(Z_h)\right)\left(G_2\left(T_{j+1}^*\right)-G_2\left(T_j^*\right)\right)\left|c(h,j)-\frac{jh}{N}\right|,$$

donde $c(h,j)=\sum_{i=1}^{h}\mathbf{1}_{\left\{T_i<T_{j+1}^*\right\}}$ es el número de elementos del vector $(T_1,T_2,...,T_h)$ menores que T_{j+1}^* para $h,j=1,2,3,...,N-1$. En consecuencia,

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

□

Prueba de la Proposición 4.

De acuerdo a los Pasos 1 y 2, ponemos $N=n(n-1)$ y reordenamos $\{d(X_i,X_j)\}_{i\neq j}$ como $Z_1,Z_2,...,Z_N$ tales que $Z_1<Z_2<...<Z_N$ y $\{d(Y_i,Y_j)\}_{i\neq j}$ como $T_1,T_2,...,T_N$ manteniendo la misma indexación de las $Z's$ (esto es, si $d(X_i,X_j)=Z_h$, entonces $d(Y_i,Y_j)=T_h$). Se observa que para calcular $T_n^{(\infty)}(r,s)$ para todo $r,s>0$ alcanza con calcular $T_n^{(\infty)}(Z_i,T_j^*)$ para todo $i,j=1,2,...,N$. En consecuencia, el resultado se sigue de forma inmediate de los Pasos 4 y 5.

□

ANEXOS

Anexo 1

Código R utilizado para datos reales

En este Anexo se presentan los códigos de los 9 estadísticos ($T_n^{(2)}$, $T_n^{(1)}$ y $T_n^{(\infty)}$ y sus variantes) utilizados en el trabajo para el cálculo de los *pvalor* aplicados a datos reales.

1.1. Estadístico $T_n^{(2)}$

El estadístico de prueba es $T_n^{(2,1)}$ (o sea integral del cuadrado pero usando la distancia L^1). Se parte de $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ muestra bivariada independiente (si las x_i e y_i son series de tiempo tener cuenta que deben ser independientes las x_i entre sí y las y_i entre sí)

1- Ingresar kx y ky (dimensión donde viven los datos por ejemplo si cada x_i es una serie de tiempo, la misma debe ser de longitud igual a kx y anólgamente con ky).

2- Ingresar n (tamaño de la muestra bivariada)

3- Ingresar los datos en una matriz para x y para y de la siguiente forma $x = matrix(datosdex, n, kx)$ $y = matrix(datosdey, n, ky)$ es decir que en la fila i de la matriz x debe ir x_i y anólogo con y.

4- Ingresar m (el *pvalor* se calculará mediante un argumento de permutación, m será la cantidad de permutaciones que consideraremos para hallar el pvalor como el porcentaje de veces en que el test rechazaría H_0 bajo H_0 cierto).

```
n=30
m=100
kx=100
ky=100

N=n*(n-1)/2
 An=rep(NA,N)
 Bn=rep(NA,N)
 Cn=rep(NA,N)

 Z=array(dist(x,method="manhattan"))

 Z=pnorm(Z,mean(Z),sd(Z))
 Zord=sort(Z)
 T=array(dist(y,method="manhattan"))

 T=pnorm(T,mean(T),sd(T))
 Tord=sort(T)

 IRX2=1-(1/N^2)*sum((2*seq(1,N)-1)*Zord)
 IRY2=1-(1/N^2)*sum((2*seq(1,N)-1)*Tord)

 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)

 IRXYRXRY=mean(Bn*Cn)

 tobs=n*(IRXY2+IRX2*IRY2-2*IRXYRXRY)

t=rep(0,m)
```

```
for( j in 1:m)
{
x=x[sample(sequence(n))]
 Z=array(dist(x,method="manhattan"))

  Z=pnorm(Z,mean(Z),sd(Z))
  Zord=sort(Z)

  IRX2=1-(1/N^2)*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)

  IRXYRXRY=mean(Bn*Cn)

  t[j]=n*(IRXY2+IRX2*IRY2-2*IRXYRXRY)
  print(j)

}
pvalor=mean(t>tobs)
pvalor
```

Los códigos para $T_n^{(2,1)}$ y $T_n^{(2,\infty)}$ son el mismo cambiando en el código en los lugares donde dice 'manhattan' por 'euclidean' o 'maximum' respectivamente.

1.2. Estadístico $T_n^{(1)}$

El estadístico de prueba es $T_n^{(1,1)}$ (el planteado en el trabajo del JMVA, o sea integral del cuadrado pero usando la distancia L^1). Se parte de $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ muestra bivariada independiente (si las x_i e y_i son series de tiempo tener cuenta que deben ser independientes las x_i entre

sí y las y_i entre sí)

1- Ingresar kx y ky (dimensión donde viven los datos por ejemplo si cada x_i es una serie de tiempo, la misma debe ser de longitud igual a kx y anólgamente con ky).
2- Ingresar n (tamaño de la muestra bivariada)
3- Ingresar los datos en una matriz para x y para y de la siguiente forma $x = matrix(datosdex, n, kx)$ $y = matrix(datosdey, n, ky)$ es decir que en la fila i de la matriz x debe ir x_i y anólogo con y.
4- Ingresar m (el *pvalor* se calculará mediante un argumento de permutación, m será la cantidad de permutaciones que consideraremos para hallar el pvalor como el porcentaje de veces en que el test rechazaría H_0 bajo H_0 cierto).

```
n=50
m=100
kx=100
ky=100


N=n*(n-1)/2

t=rep(0,m)

jj=seq(1:(N-1))


Z=array(dist(x,method="manhattan"))
Z=pnorm(Z,mean(Z),sd(Z))

debe hacerse antes de ordenar los datos de Z de menor a mayor
T=array(dist(y,method="manhattan"))
T=pnorm(T,mean(T),sd(T))
Z=sort(Z)

dG1=Z[2:length(Z)]-Z[1:length(Z)-1]

T=T[o]
```

```
Tord=sort(T)
dG2=Tord[2:length(Tord)]-Tord[1:length(Tord)-1]

Tordcorrido=matrix(Tord[2:N],1,(N-1))

C=rep(NA,(N-1))
I=rep(NA,(N-1))

for ( h in 1:(N-1))
{
 f=function(x){sum(T[1:h]<x)}
C=apply(Tordcorrido,2, f)
I[h]=sum(abs(C-jj*h/N)*dG2)
}
tobs=sqrt(n)*sum(dG1*I)/N
for (j in 1:m)
{
x=x[sample(sequence(n)),]

Z=array(dist(x,method="manhattan"))
Z=pnorm(Z,mean(Z),sd(Z))
o=order(Z

T=array(dist(y,method="manhattan"))
T=pnorm(T,mean(T),sd(T))
Z=sort(Z)
dG1=Z[2:length(Z)]-Z[1:length(Z)-1]

T=T[o]

Tord=sort(T)
dG2=Tord[2:length(Tord)]-Tord[1:length(Tord)-1]
Tordcorrido=matrix(Tord[2:N],1,(N-1))

C=rep(NA,(N-1))
I=rep(NA,(N-1))
```

```
for ( h in 1:(N-1))
{
 f=function(x){sum(T[1:h]<x)}
C=apply(Tordcorrido,2, f)
I[h]=sum(abs(C-jj*h/N)*dG2)
}
t[j]=sqrt(n)*sum(dG1*I)/N
print(j)
}


pvalor=mean(t>tobs)

pvalor
```

Los códigos para $T_n^{(1,2)}$ y $T_n^{(1,\infty)}$ son el mismo cambiando en el código en los lugares donde dice 'manhattan' por 'euclidean' o 'maximum' respectivamente.

1.3. Estadístico $T_n^{(\infty)}$

El estadístico de prueba es $T_n^{(\infty,1)}$.

Se parte de $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ muestra bivariada independiente (si las x_i e y_i son series de tiempo tener cuenta que deben ser independientes las x_i entre sí y las y_i entre sí)

1- Ingresar kx y ky (dimensión donde viven los datos por ejemplo si cada x_i es una serie de tiempo, la misma debe ser de longitud igual a kx y anólgamente con ky).

2- Ingresar n (tamaño de la muestra bivariada)

3- Ingresar los datos en una matriz para x y para y de la siguiente forma $x = matrix(datosdex, n, kx)$ $y = matrix(datosdey, n, ky)$ es decir que en la fila i de la matriz x debe ir x_i y anólogo con y.

4- Ingresar m (el *pvalor* se calculará mediante un argumento de permutación, m será la cantidad de permutaciones que consideraremos para hallar el pvalor como el porcentaje de veces en que el test rechazaría H_0 bajo H_0 cierto).

```
n=50
 m=100
 kx=100
 ky=100

  N=n*(n-1)/2



 Z=array(dist(x,method="manhattan"))
Zord=sort(Z)

T=array(dist(y,method="manhattan"))
Tord=sort(T)


IndZ=matrix(NA,N,N)
IndT=matrix(NA,N,N)
for (h in 1:N)
{
IndZ[h,]=1*(Z<=Zord[h])
IndT[h,]=1*(T<=Tord[h])
}
C=IndZ%*%t(IndT)/N
ij=seq(1,N)%*%t(seq(1,N))
C=abs(C-ij/(N^2))


tobs=sqrt(n)*max(C)

t=rep(0,m)

for (j in 1:m)
{
x=x[sample(sequence(n)),]
 Z=array(dist(x,method="manhattan"))
```

```
  Zord=sort(Z)
for (h in 1:N)
{
IndZ[h,]=1*(Z<=Zord[h])
}
C=IndZ%*%t(IndT)/N
ij=seq(1,N)%*%t(seq(1,N))
C=abs(C-ij/(N^2))


t[j]=sqrt(n)*max(C)
 print(j)
}
pvalor=mean(t>tobs)


pvalor
```

Los códigos para $T_n^{(\infty,2)}$ y $T_n^{(\infty,\infty)}$ son el mismo cambiando en el código en los lugares donde dice 'manhattan' por 'euclidean' o 'maximum' respectivamente.