

Test de independencia basado en porcentajes de recurrencias

Juan Kalemkerian.

`jkalem@cmat.edu.uy`, `jkalem@fing.edu.uy`.

Facultad de Ciencias (Centro de Matemática),
Facultad de Ingeniería (IMERL).
Universidad de la República.
Trabajo en colaboración con Diego Fernández.

Seminario SIESTA, Montevideo, Uruguay. 9 de abril de
2019.

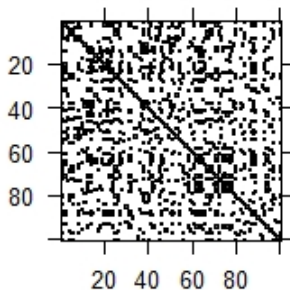
Contenidos

- 1 Recurrence Plots
- 2 Algunas fechas
 - Test de independencia
 - Recurrence plots
- 3 Planteo del test
- 4 Resultados teóricos
 - Distribución asintótica del test propuesto
 - Consistencia bajo un espectro amplio de alternativas
 - Alternativas contiguas
- 5 Implementación del test
 - Cálculo del estadístico
 - Selección de la función de pesos
- 6 Performance del test
 - X e Y variables aleatorias
 - X e Y vectores aleatorios
 - X e Y series de tiempo
- 7 A futuro

- Dada una serie de tiempo X_1, X_2, \dots, X_n (mediciones de alguna variable a tiempos $t = 1, 2, \dots, n$) y dado $r > 0$ construimos una matriz $R(r)$ de dimensiones $n \times n$ donde $R_{ij}(r) = 1$ si $d(X_i, X_j) < r$ y $R_{ij}(r) = 0$ en caso contrario. $d(x, y)$ representa la distancia entre los puntos x e y .
- Si graficamos los valores de esta matriz en un cuadrado $[1, n] \times [1, n]$, poniendo puntos donde van los unos y dejando en blanco donde van los ceros, estamos ante un caso particular de los llamados gráficos de recurrencia (recurrence plots).
- Si las mediciones X_i son vectores, es posible definir de igual manera el gráfico de recurrencia para un valor de r .
- Los gráficos de recurrencia son una importante herramienta visual para detectar patrones, periodicidades e incluso pueden sugerir la modelación de la serie de tiempo mediante modelos determinísticos o probabilísticos.

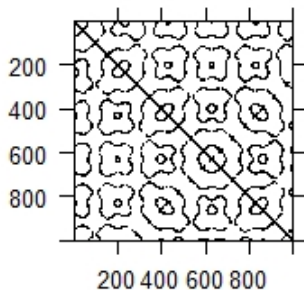
- Es crucial la elección del parámetro r para el cual hay varios criterios que se encuentran en la literatura.
- Estos gráficos fueron introducidos por Eckmann, Kamphorst & Ruelle (Europhys. Lett, 1987), mostrando que es una herramienta gráfica muy importante para el abordaje de la dinámica de una serie en alta dimensión.
- Líneas verticales, horizontales y paralelas a la diagonal (junto con sus longitudes) son de fácil interpretación.
- Vemos ahora algunos gráficos de recurrencia según la serie de tiempo considerada (con el r considerado en cada caso).

Ruido blanco $N(0,1)$ $r=0.2$



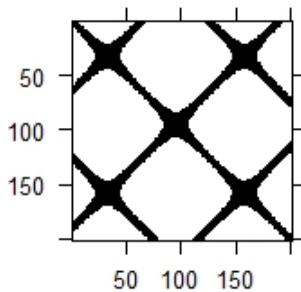
Dimensions: 100 x 100

$$x = \sin(t)\sin(5t)$$
$$r = 0.01 \text{sd}(x)$$



Dimensions: 1001 x 1001

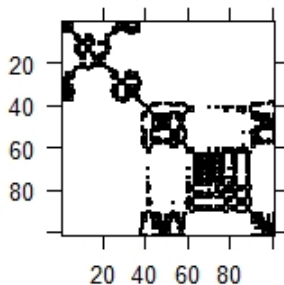
$$x = \sin(t)$$
$$r = 0.2 \text{sd}(x)$$



Dimensions: 201 x 201

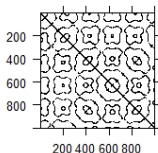
Movimiento browniano

$r=0.2sd(x)$



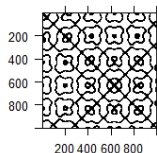
Dimensions: 101 x 101

$$x = \sin(t)\sin(5t)$$
$$r = 0.01\text{sd}(x)$$



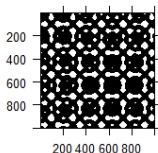
Dimensions: 1001 x 1001

$$x = \sin(t)\sin(5t)$$
$$r = 0.2\text{sd}(x)$$

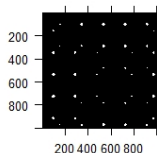


Dimensions: 1001 x 1001

$$x = \sin(t)\sin(5t)$$
$$r = 1.3\text{sd}(x)$$



$$x = \sin(t)\sin(5t)$$
$$r = 3\text{sd}(x)$$



- Elegir un valor de r demasiado grande, deja el cuadrado prácticamente pintado todo de negro.
- Elegir un valor de r demasiado pequeño, oculta las posibles periodicidades o patrones que pueda tener la serie de tiempo.
- El r adecuado depende de cada serie en particular.

Contenidos

- 1 Recurrence Plots
- 2 Algunas fechas
 - Test de independencia
 - Recurrence plots
- 3 Planteo del test
- 4 Resultados teóricos
 - Distribución asintótica del test propuesto
 - Consistencia bajo un espectro amplio de alternativas
 - Alternativas contiguas
- 5 Implementación del test
 - Cálculo del estadístico
 - Selección de la función de pesos
- 6 Performance del test
 - X e Y variables aleatorias
 - X e Y vectores aleatorios
 - X e Y series de tiempo
- 7 A futuro

- Correlación de Pearson es introducida por Galton (“Co-relations and their measurement, chiefly from anthropometric data”, Proceedings of the Royal Society of London, 1888) y su formulación actual es de Pearson (misma revista, 1896).
- Test de rangos de Spearman (The American Journal of Phsichology, 1904).
- Test de rangos de Kendall (Biometrika, 1938).
- Test de Wilks (Econometrica, Journal of the Econometric Society, 1935) se aborda por primera vez la independencia entre vectores aleatorios.
- Hoeffding (Annals of Mat. Stat, 1948).
- Genest & Remilliard (Test, 2004) utilizan cópulas, para vectores aleatorios continuos.
- Beran, Bilodeau & Lafaye de Micheaux (JMVA, 2007), plantean un test universalmente consistente, para vectores aleatorios, a partir de las distribuciones multidimensionales empíricas.

- Criterio de independencia de Hilbert-Schmidt, planteado en Greton, Bousquet, Smola & Schölkof (Advances in Neural information processign systems, 2007), universalmente consistente.
- Székely, Rizzo & Bakirov (The Annals of statistics, 2007), plantean la distance covariance, universalmente consistente.
- Heller, Heller & Gorfine (Biometrika, 2012) universalmente consistente.
- Eckmann, Kamphorst & Ruelle (Europhys. Lett, 1987) introducen los gráficos de recurrencia, mostrando que es una herramienta gráfica muy importante para el abordaje de la dinámica de una serie en alta dimensión.

Dada $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ muestra i.i.d de (X, Y) donde $X \in S_X, Y \in S_Y$ (espacios métricos cualesquiera) y fijados $r, s > 0$.

Definimos el porcentaje de recurrencias de X e Y como

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

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

y el porcentaje conjunto (X, Y) como

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

Definimos también

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

$$p_X(r) := P(d(X_1, X_2) < r) \text{ y } p_Y(s) := P(d(Y_1, Y_2) < s).$$

Observar que para cualesquiera $r, s > 0$,

$$RR_n^X(r) \xrightarrow{a.s.} p_X(r), \quad RR_n^Y(s) \xrightarrow{a.s.} p_Y(s) \text{ and } RR_n^{X,Y}(r,s) \xrightarrow{a.s.} p_{X,Y}(r,s). \quad (1)$$

Queremos testear $H_0 : X$ e Y son independientes, contra $H_1 :$
No H_0 .

Si H_0 es cierto, entonces $p_{X,Y}(r,s) = p_X(r)p_Y(s)$ para todos $r, s > 0$, y esperamos que si n es grande, entonces $RR_n^{X,Y}(r,s) \cong RR_n^X(r)RR_n^Y(s)$ para todos $r, s > 0$. Proponemos trabajar con el proceso $\{E_n(r,s)\}_{r,s>0}$ donde

$$E_n(r,s) := \sqrt{n} \left(RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s) \right). \quad (2)$$

Luego, es natural rechazar H_0 cuando $T_n > c$ siendo

$$T_n := n \int_0^{+\infty} \int_0^{+\infty} \left(RR_n^{X,Y}(r,s) - RR_n^X(r)RR_n^Y(s) \right)^2 dG(r,s) \quad (3)$$

para ciertas constante c y G alguna medida. Le llamaremos ϕ y φ a las funciones de distribución y densidad de una variable $N(0, 1)$ y definimos los conjuntos $I_m^n = \{(i_1, \dots, i_m) : i_j \neq i_k \text{ para } j \neq k, i_j \in \{1, \dots, n\} \text{ para } j = 1, \dots, m\}$.

Contenidos

- 1 Recurrence Plots
- 2 Algunas fechas
 - Test de independencia
 - Recurrence plots
- 3 Planteo del test
- 4 Resultados teóricos**
 - Distribución asintótica del test propuesto
 - Consistencia bajo un espectro amplio de alternativas
 - Alternativas contiguas
- 5 Implementación del test
 - Cálculo del estadístico
 - Selección de la función de pesos
- 6 Performance del test
 - X e Y variables aleatorias
 - X e Y vectores aleatorios
 - X e Y series de tiempo
- 7 A futuro

Distribución asintótica del test estadístico

Primero obtenemos las covarianzas asintóticas

Lemma

Dados $r, r', s, s' > 0$, y $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ i.i.d en $\mathbb{R}^p \times \mathbb{R}^q$ siendo X e Y independentientes, entonces

$$\lim_{n \rightarrow +\infty} \text{COV} (E_n(r, s), E_n(r', s')) =$$

$$4 \left(p_X^{(3)}(r \wedge r') - p_X(r)p_X(r') \right) \left(p_Y^{(3)}(s \wedge s') - p_Y(s)p_Y(s') \right). \quad (4)$$

El próximo lema reduce la convergencia del proceso $\{E_n(r, s)\}_{r, s > 0}$ a la convergencia de un U -process que lo aproxima que llamamos $\{E'_n(r, s)\}_{r, s > 0}$ definido como

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

$$\sum_{(i,j,k,h) \in I_4^n} \left(\mathbf{1}_{\{d(X_i, X_j) < r, d(Y_i, Y_j) < s\}} - \mathbf{1}_{\{d(X_i, X_j) < r, d(Y_h, Y_k) < s\}} \right). \quad (5)$$

Lemma

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

donde

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

La convergencia del proceso $\{E_n(r, s) - \mathbb{E}(E_n(r, s))\}_{r,s>0}$ a un proceso gaussiano centrado con covarianzas definidas en (4),

la obtenemos utilizando el teorema 4.10 obtenido por Arcones & Giné (1993):

Dado (S, \mathcal{S}, P) espacio de probabilidad, $X_i : S \rightarrow S$ i.i.d. sucesión tal que $\mathcal{L}(X_i) = P$. Dado m , consideramos \mathcal{F} una clase de funciones medibles en S^m , se define el U -proceso basado en P e indexado por \mathcal{F} como

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

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

Dado $\varepsilon > 0$, supongamos que existen $\mathcal{L} = \{l_1, l_2, \dots, l_v\}$, $\mathcal{U} = \{u_1, u_2, \dots, u_v\}$ tales que $\mathcal{L}, \mathcal{U} \subset L^2$ y para $f \in \mathcal{F}$,

existen $l_f \in \mathcal{L}$ y $u_f \in \mathcal{U}$ donde $l_f \leq f \leq u_f$ a.s. y $\mathbb{E}(u_f - l_f)^2 < \varepsilon^2$. (6)

$$N_{[\cdot]}^{(2)}(\varepsilon, F, P^m) = \min \{v : (6) \text{ se cumple}\}. \quad (7)$$

Theorem (Theorem 4.10. Arcones & Giné (1993))

Si

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

entonces

$$\mathcal{L}(\sqrt{n}(U_m^n - P^m)f) \xrightarrow{w} \mathcal{L}(mG_P \circ P^{m-1}f) \text{ in } l^\infty(\mathcal{F}) \quad (8)$$

siendo G_P el puente Browniano asociado a P .

A partir de este teorema obtenemos el resultado asintótico del proceso $\{E_n(r, s)\}$.

Theorem

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ i.i.d en $S_X \times S_Y$ entonces

$$\{E_n(r, s) - \mathbb{E}(E_n(r, s))\}_{r,s>0} \xrightarrow{w} \{E(r, s)\}_{r,s>0} \quad (9)$$

donde $\{E(r, s)\}_{r,s>0}$ es un proceso Gaussiano centrado cuya función de covarianzas viene dada por (4).

Corollary

Dada $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ muestra i.i.d de (X, Y) siendo $X \in S_X, Y \in S_Y$ y fijados $r, s > 0$, entonces

$$\sqrt{n}(E_n(r, s) - \mathbb{E}(E_n(r, s))) \xrightarrow{d} N(0, \sigma_{X,Y}^2(r, s)).$$

donde

$$\sigma_{X,Y}^2(r, s) = 4 \left(p_X^{(3)}(r) - p_X^2(r) \right) \left(p_Y^{(3)}(s) - p_Y^2(s) \right). \quad (10)$$

- Observar que la distribución asintótica del test estadístico es válida tanto bajo H_0 como bajo H_1 (en el caso en que H_0 es cierto se tiene $\mathbb{E}(E_n(r, s)) = 0$ para todos $r, s > 0$).
- Observar que la distribución asintótica depende de la distribución original de los X e Y , pero son fáciles de estimar (para cada r, s) mediante

$$\widehat{p}_X(r) = RR_n^X(r),$$

$$\widehat{p}_X^{(3)}(r) = \frac{1}{n(n-1)(n-2)} \sum_{(i,j,k) \in I_3^n} \mathbf{1}_{\{d(X_i, X_j) < r, d(X_i, X_k) < r\}}$$

y análogamente se estiman $p_Y(s)$ y $p_Y^{(3)}(s)$. Luego

$\widehat{\sigma}_{X,Y}^2(r, s) = 4 \left(\widehat{p}_X^{(3)}(r) - \widehat{p}_X^2(r) \right) \left(\widehat{p}_Y^{(3)}(s) - \widehat{p}_Y^2(s) \right)$ es un estimador consistente de $\sigma_{X,Y}^2(r, s)$.

Corollary

Bajo H_0 cierto, se cumple que

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

para todos $r, s > 0$.

Observar que para I y J conjuntos finitos, se cumple que

$$n \sum_{(r,s) \in I \times J} \left(RR_n^{X,Y}(r,s) - RR_n^X(r) RR_n^Y(s) \right)^2 \xrightarrow{d} \sum_{i=1}^h \lambda_i Z_i^2$$

siendo h el cardinal de $I \times J$, Z_1, Z_2, \dots, Z_h normales típicas y λ_i constantes positivas. Luego, se puede testear la hipótesis

mediante $T'_n := n \sum_{(r,s) \in I \times J} \left(RR_n^{X,Y}(r,s) - RR_n^X(r) RR_n^Y(s) \right)^2$

Corollary

Dada $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n) \in \mathbb{R}^2$ muestra i.i.d de (X, Y) donde las marginales X, Y son $N(0, 1)$. Entonces $\sqrt{n} \left(RR_n^{X,Y}(r, s) - RR_n^X(r) RR_n^Y(s) \right) \xrightarrow{d} N \left(0, \sigma_{X,Y}^2(r, s) \right)$ donde

$$\sigma_{X,Y}^2(r, s) =$$

$$4 \left(\int_{-\infty}^{+\infty} (\phi(x+r) - \phi(x-r))^2 \varphi(x) dx - \left(2\phi(r/\sqrt{2}) - 1 \right)^2 \right) \times$$

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

(11)

En el caso en que tomemos $s = r$ se tiene

$$4 \left(\int_{-\infty}^{+\infty} (\phi(x+r) - \phi(x-r))^2 \varphi(x) dx - \left(2\phi\left(r/\sqrt{2}\right) - 1 \right)^2 \right)^2. \quad (12)$$

Consistencia bajo un espectro amplio de alternativas

El próximo teorema nos muestra que si H_1 es cierto, el test propuesto es consistente bajo un amplio espectro de alternativas.

Theorem

Si $dG(r, s) = g(r, s)drds$ y $g(r, s) > 0$ para todos $r, s > 0$, y $d(X_1, X_2)$, $d(Y_1, Y_2)$ son continuas y no independientes, entonces $T_n \xrightarrow{P} +\infty$ as $n \rightarrow +\infty$.

Consistencia en el caso normal multivariado

El próximo teorema nos muestra que en el caso de la normal, si H_1 es cierto, entonces existen $r > 0$ y $s > 0$ tales que $E_n(r, s)$ es no acotado en probabilidad, lo que permitirá obtener un resultado de consistencia del test propuesto.

Theorem

Dada $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ muestra i.i.d de $(X, Y) \sim N(0, \Sigma)$ en $\mathbb{R}^p \times \mathbb{R}^q$. Si X e Y no son independientes. Entonces, existen $r > 0$ y $s > 0$ tales que

$$\sqrt{n} \left(RR_n^{X,Y}(r, s) - RR_n^X(r) RR_n^Y(s) \right) \xrightarrow{P} +\infty.$$

Corollary

Si $(X, Y) \sim N(0, \Sigma)$, X e Y no son independientes, entonces $T_n \xrightarrow{P} +\infty$ cuando $n \rightarrow +\infty$.

Alternativas contiguas

Dada $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ i.i.d. en $\mathbb{R}^p \times \mathbb{R}^q$,
consideramos

$$H_0 : f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{for all } (x, y)$$

(i.e. X Y independientes), vs

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

donde $f_{X,Y}^{(n)}(x, y) = c_n(\delta) f_X(x) f_Y(y) \left(1 + \frac{\delta}{2\sqrt{n}} k_n(x, y)\right)^2$, $\delta > 0$,

$c_n(\delta)$ es una constante que permite que $f_{X,Y}^{(n)}(x, y)$ sea
densidad, y las funciones k_n verifican las condiciones (i) y (ii)
dadas abajo:

Definimos $L_0^2 = L^2(dF_0)$ para $dF_0(x, y) = f_X(x)f_Y(y)dxdy$, la distribution (X, Y) bajo H_0 , análogamente definimos L_0^1 .

(i) Existe una función $K \in L_0^1$ tal que $k_n \leq K$ para todo n

(ii) Existe $k \in L_0^2$ tal que $k_n \xrightarrow{L_0^2} k$, $\|k\| = 1$.

- Se puede probar que las condiciones (i) y (ii) implican contigüidad (Cabaña [6]). En la siguiente proposición utilizamos la notación $\mathbb{E}^{(n)}(T)$ y $P^{(n)}((X, Y) \in A)$ para el valor esperado de T , y para la probabilidad del conjunto $\{(X, Y) \in A\}$ bajo H_n respectivamente.

Theorem

Bajo H_n

$\mathbb{E}^{(n)}(E_n(r, s)) \rightarrow \delta\mu(r, s)$ cuando $n \rightarrow +\infty$ para todos $r, s > 0$.

donde

$$\mu(r, s) = \iiint\limits_{A_{r,s}} g(x_1, x_2, y_1, y_2) dx_1 dx_2 dy_1 dy_2, \quad (13)$$

$$g(x_1, x_2, y_1, y_2) = (k(x_1, y_1) + k(x_2, y_2)) f_X(x_1) f_Y(y_1) f_X(x_2) f_Y(y_2) \quad (14)$$

y $A_{r,s} :=$

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

Con un poco más de trabajo, es posible usar el Tercer Lema de Le Cam (Le Cam & Yang, [13] junto a las condiciones presentadas en Oosterhoff & Van Zwet, [15]) para obtener que

bajo H_n , se cumple que

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

donde $\{E(r, s)\}_{r, s > 0}$ es el proceso límite bajo H_0 y $\mu(r, s)$ es la definida en el teorema anterior. Luego, bajo H_n cierto, se cumple que

$$T_n \xrightarrow{w} \int_0^{+\infty} \int_0^{+\infty} (E(r, s) + \delta\mu(r, s))^2 dG(r, s). \quad (15)$$

Contenidos

- 1 Recurrence Plots
- 2 Algunas fechas
 - Test de independencia
 - Recurrence plots
- 3 Planteo del test
- 4 Resultados teóricos
 - Distribución asintótica del test propuesto
 - Consistencia bajo un espectro amplio de alternativas
 - Alternativas contiguas
- 5 Implementación del test**
 - Cálculo del estadístico
 - Selección de la función de pesos
- 6 Performance del test
 - X e Y variables aleatorias
 - X e Y vectores aleatorios
 - X e Y series de tiempo
- 7 A futuro

- Si X e Y son reales con distribución continua, aplicamos el test a $X' = \phi^{-1}(F_X(X))$ e $Y' = \phi^{-1}(F_Y(Y))$, donde F_X y F_Y las distribuciones de X e Y respectivamente. Aplicar el test a X' e Y' en lugar de aplicarlo a X e Y tiene la ventaja de dejar a las variables en la misma escala, además de que “nos acerca” a las hipótesis del teorema anteriormente mencionado. Además, nos permite calcular el valor crítico para un nivel α dado.
- Si X e Y son vectores aleatorios, no conocemos la distribución del estadístico bajo H_0 , pero mediante un argumento de permutación podemos estimar el p-valor de la prueba.
- Si X o Y están en dimensión infinita, podemos usar el mismo argumento de permutación.

Cálculo del estadístico

- Para simplificar le llamamos $N = n(n - 1)$. Elegimos una función de pesos de la forma $dG(r, s) = g_1(r)g_2(s)drds$ donde g_1, g_2 son funciones de densidad cuyas respectivas distribuciones les llamamos G_1 y G_2 respectivamente.
 - El cálculo del estadístico puede ser realizado en los siguientes pasos.
- 1 Ordeno en forma creciente los valores $d(X_i, X_j)$ con $i, j \in \{1, 2, 3, \dots, n\}$ tales que $i \neq j$ los que reindizamos como Z_1, Z_2, \dots, Z_N . Análogamente le llamaremos T_1, T_2, \dots, T_N a los valores $d(Y_i, Y_j)$ utilizando la misma reindización. Le llamaremos a su vez $Z_1^*, Z_2^*, \dots, Z_N^*$ a los estadísticos de orden y análogamente $T_1^*, T_2^*, \dots, T_N^*$.
 - 2 Calculamos $A_n =$

$$\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (1 - G_1(\max\{Z_i, Z_j\})) (1 - G_2(\max\{T_i, T_j\})).$$

3 Calculamos $B_n =$

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

4 Calculamos $C_n =$

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

5 Finalmente el estadístico queda

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

Selección de la función de pesos

- El investigador de acuerdo a cada caso en particular a detectar independencia, puede seleccionar la función de pesos que le parezca adecuada.
- Una idea que se puede utilizar de manera universal es definir g_1 como la densidad de una normal cuyas media y varianza sean el promedio y la varianza de las Z_1, Z_2, \dots, Z_N . Análogamente g_2 con T_1, T_2, \dots, T_N .
- Es posible también evitarse la elección de la función de pesos y utilizar un estadístico del tipo Kolmogorov-Smirnov entre los porcentajes de recurrencias.

Contenidos

- 1 Recurrence Plots
- 2 Algunas fechas
 - Test de independencia
 - Recurrence plots
- 3 Planteo del test
- 4 Resultados teóricos
 - Distribución asintótica del test propuesto
 - Consistencia bajo un espectro amplio de alternativas
 - Alternativas contiguas
- 5 Implementación del test
 - Cálculo del estadístico
 - Selección de la función de pesos
- 6 **Performance del test**
 - X e Y variables aleatorias
 - X e Y vectores aleatorios
 - X e Y series de tiempo
- 7 A futuro

- En los siguientes resultados mostramos las potencias del test propuesto y lo comparamos con otros existentes ampliamente citados en la literatura reciente.
- Todas las potencias son calculadas al nivel de significación del 5% para tamaños de muestra de $n = 30$, $n = 50$ y $n = 80$.
- Los valores críticos del test propuesto fueron calculados mediante la simulación con $m = 50.000$ réplicas, mientras que las potencias fueron calculadas mediante $m = 10.000$ réplicas.
- En todos los casos a todas las coordenadas que integran tanto el vector X como el Y se les aplica la transformación $\phi^{-1} \circ F_X$.
- Compararemos las potencias con respecto a los recientes test propuestos HHG (Heller, Heller & Gorfine, 2012), DCOV (Székely, Rizzo & Bakirov, 2007), HSIC (Greton, Bousquet, Smola & Schölkopf, 2005) junto con los clásicos test de correlación de Pearson, Spearman y Kendall

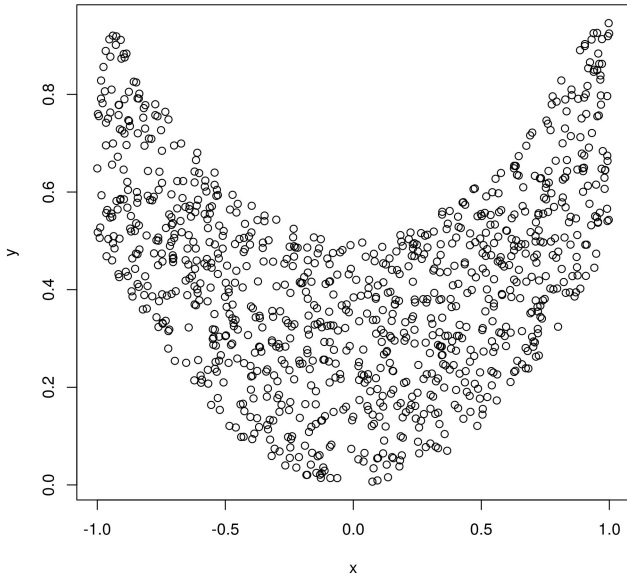
(PSK).

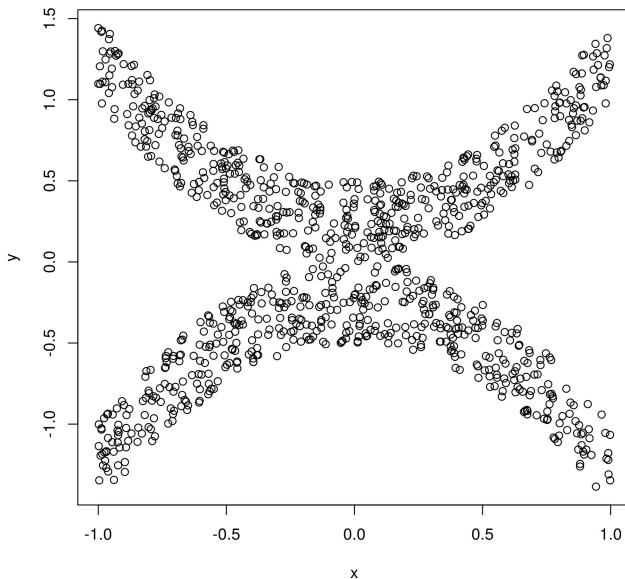
- En las siguientes tablas, en cada columna irán las potencias de estos tests considerados, en la columna llamada PSK se toma la más alta de las potencias entre Pearson, Spearman y Kendall.
- En las siguientes columnas se consideran las potencias del test de recurrencia, para distintas funciones de pesos G . En la última columna (g_1, g_2) utilizamos la función de pesos sugerida en el trabajo.
- En todos los casos se considera $G(r, s) = g(r)g(s)$ siendo g la densidad de una $N(\mu, \sigma^2)$ para distintos valores de μ, σ^2 .

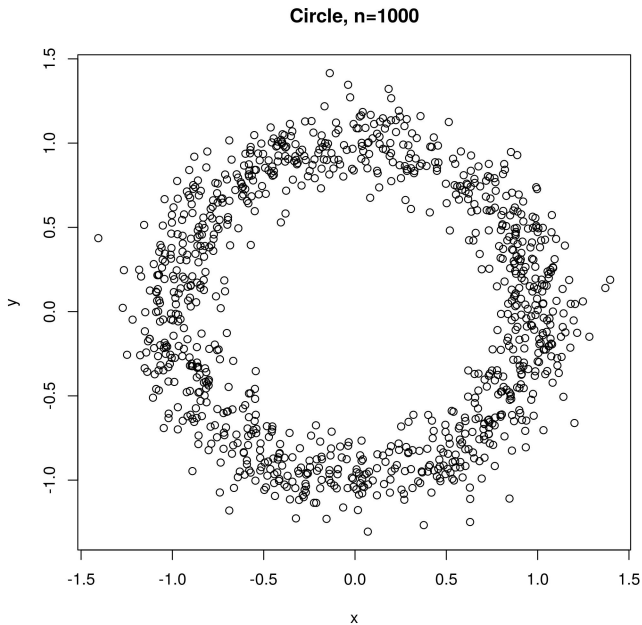
Primer caso. X e Y variables aleatorias

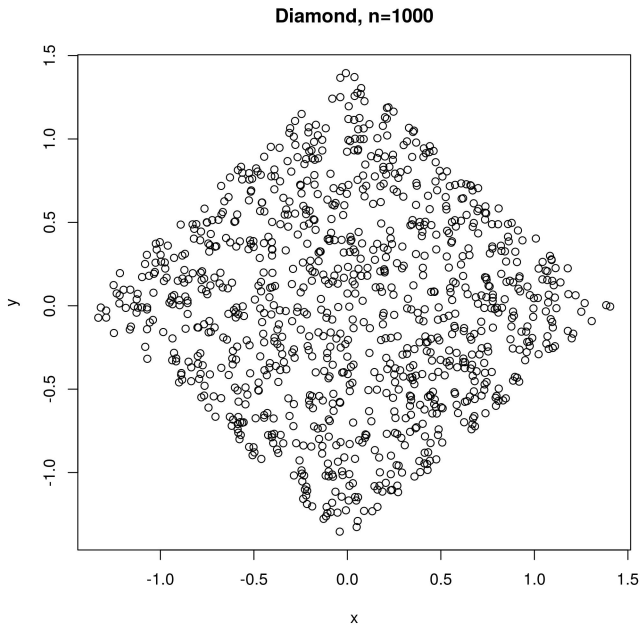
Consideramos las alternativas incluídas en la Tabla 3 del trabajo de Heller et al:

- Parabola: $X \sim U(-1, 1)$, $Y = (X^2 + U(0, 1)) / 2$.
- Two parabolas: $X \sim U(-1, 1)$, $Y = (X^2 + U(0, 1)) / 2$ con probabilidad $1/2$, $Y = -(X^2 + U(0, 1)) / 2$ con probabilidad $1/2$.
- Circle: $U \sim U(-1, 1)$, $X = \sin(\pi U) + N(0, 1) / 8$,
 $Y = \cos(\pi U) + N(0, 1) / 8$.
- Diamond: $U_1, U_2 \sim U(-1, 1)$ independientes,
 $X = \sin(\theta) U_1 + \cos(\theta) U_2$, $Y = -\sin(\theta) U_1 + \cos(\theta) U_2$
for $\theta = \pi/4$.

Parabola, $n=1000$ 

Two parabolas, $n=1000$ 





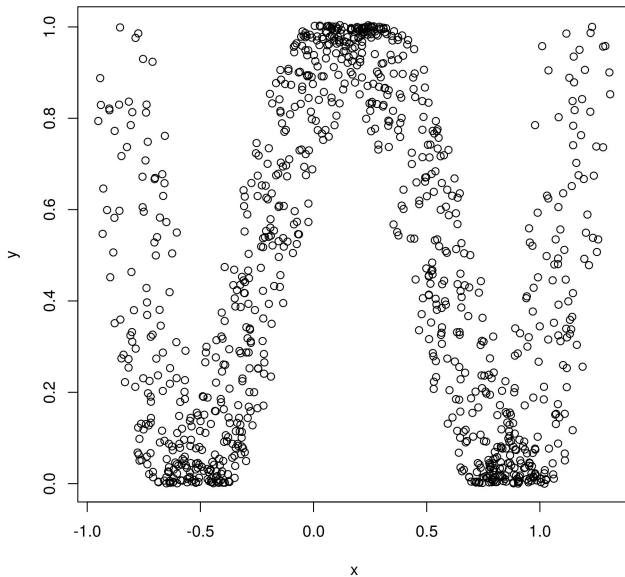
W-shape, n=1000

Tabla 1. Potencias, caso $n = 30$.

Test	HHG	DCOV	HSIC	PSK	N(1,1)	N(0,1)	N(1,4)	g_1, g_2
Parabola	0.791	0.522	0.733	0.103	0.824	0.831	0.814	0.817
2 parabolas	0.962	0.204	0.849	0.194	1.000	1.000	1.000	1.000
Circle	0.646	0.051	0.488	0.096	0.923	0.716	0.947	0.823
Diamond	0.283	0.030	0.262	0.016	0.422	0.139	0.477	0.395
W-shape	0.908	0.569	0.856	0.179	0.788	0.887	0.782	0.874

Tabla 2. Potencias, caso $n = 50$.

Test	HHG	DCOV	HSIC	PSK	N(1,1)	N(0,1)	N(1,4)	g_1, g_2
Parabola	0.983	0.854	0.957	0.114	0.979	0.983	1.000	0.975
2 parabolas	1.000	0.354	0.997	0.198	1.000	1.000	1.000	1.000
Circle	0.985	0.075	0.914	0.008	0.999	0.997	1.000	0.995
Diamond	0.664	0.048	0.545	0.013	0.836	0.630	0.884	0.761
W-shape	0.999	0.935	0.988	0.077	0.989	0.998	0.987	0.979

Tabla 3. Potencias, caso $n = 80$.

Test	HHG	DCOV	HSIC	PSK	N(1,1)	N(0,1)	N(1,4)	g_1, g_2
Parabola	1.000	0.994	1.000	0.105	1.000	1.000	1.000	1.000
2 parabolas	1.000	0.700	1.000	0.201	1.000	1.000	1.000	1.000
Circle	1.000	0.196	0.999	0.004	0.999	1.000	1.000	1.000
Diamond	0.948	0.096	0.853	0.003	0.836	0.953	1.000	0.999
W-shape	1.000	0.999	1.000	0.085	0.988	1.000	1.000	1.000

Segundo caso. X e Y vectores aleatorios

Consideramos las dos últimas alternativas de la Tabla 3, y alguna de la Tabla 4 de Heller et al.

- Logarithmic: $X, Y \in \mathbb{R}^5$ donde las X_i son $N(0, 1)$ independientes, $Y_i = \log(X_i^2)$ para $i = 1, 2, 3, 4, 5$.
- Epsilon: $X, Y, \varepsilon \in \mathbb{R}^5$ donde las X_i, ε_i son $N(0, 1)$ independientes, $Y_i = \varepsilon_i X_i$ para $i = 1, 2, 3, 4, 5$.
- Quadratic: $X, Y, \varepsilon \in \mathbb{R}^5$ donde X_i, ε_i son independientes, $X_i \sim N(0, 1)$, $\varepsilon_i \sim N(0, 3)$, $Y_i = X_i + 4X_i^2 + \varepsilon_i$ $i = 1, 2$, $Y_i = \varepsilon_i$ para todo $i = 3, 4, 5$.

Agregamos una alternativa extraída de la monografía de Boglioni.

- 2D-pairwise independent: $X, Z_0, Y_1 \sim N(0, 1)$ independientes, $Y = (Y_1, Y_2)$ donde $Y_2 = |Z_0| \operatorname{sign}(XY_1)$.

Tabla 4. Potencias, caso $n = 30$.

Test	HHG	DCOV	HSIC	N(1,1)	N(1,4)	N(0,4)	N(2,4)	g_1, g_2
Log	0.594	0.154	0.610	0.710	0.759	0.321	0.885	0.813
Epsilon	0.784	0.226	0.484	0.470	0.576	0.194	0.749	0.858
Quadratic	0.687	0.302	0.530	0.197	0.155	0.170	0.147	0.144
2D-indep	0.161	0.175	0.403	0.177	0.264	0.106	0.263	0.112

Tabla 5. Potencias, caso $n = 50$.

Test	HHG	DCOV	HSIC	N(1,1)	N(1,4)	N(0,4)	N(2,4)	g_1, g_2
Log	0.936	0.386	0.958	0.998	0.999	1.000	1.000	0.995
Epsilon	0.969	0.298	0.689	0.895	0.967	0.968	0.999	0.984
Quadratic	0.934	0.485	0.904	0.362	0.293	0.315	0.733	0.236
2D-indep	0.27	0.359	0.798	0.281	0.219	0.261	0.198	0.172

Tabla 6. Potencias, caso $n = 80$.

Test	HHG	DCOV	HSIC	N(1,1)	N(1,4)	N(0,4)	N(2,4)	g_1, g_2
Log	1.000	0.793	1.000	1.000	1.000	1.000	1.000	1.000
Epsilon	0.999	0.382	0.896	0.998	1.000	1.000	1.000	1.000
Quadratic	0.996	0.725	0.971	0.595	0.545	0.535	0.480	0.416
2D-indep	0.544	0.751	0.993	0.489	0.348	0.466	0.263	0.284

Tercer caso. X e Y series de tiempo (a tiempo discreto y continuo)

En las siguientes tablas, ε y ε' representan un ruido blanco con varianza 1 independiente de los restantes procesos generados.

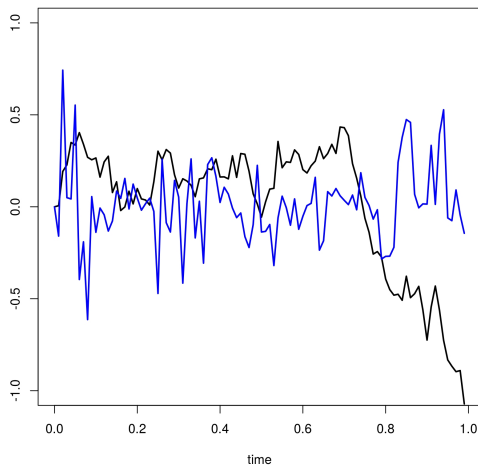
En la Tabla 7, Z representa un ruido blanco con varianza 0.4 independiente de todos los demás procesos considerados. En

la Tabla 7, tenemos X e Y series de tiempo de parámetro discreto. El caso llamado ARMA(2, 1), los parámetros son $\theta = (0.2, 0.5)$ y $\phi = 0.2$. En la Tabla 8 y la 9, tenemos series de tiempo de parámetro continuo.

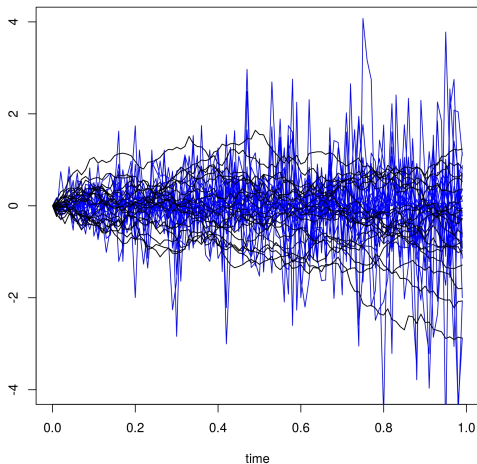
Tabla7. Potencias para distintos n , Alternativa serie de tiempo de parámetro discreto.

n	X	$Y = X^2 + 3\varepsilon$	$Y = \sqrt{ X } + Z$	$Y = \varepsilon X$
30	AR(0, 1)	0.350	0.214	0.772
50	AR(0, 1)	0.592	0.402	0.962
100	AR(0, 1)	0.999	0.698	1.000
30	AR(0, 9)	1.000	0.903	1.000
50	AR(0, 9)	1.000	0.998	1.000
100	AR(0, 9)	1.000	1.000	1.000
30	ARMA(2, 1)	0.817	0.323	0.925
50	ARMA(2, 1)	0.986	0.566	0.996
100	ARMA(2, 1)	1.000	0.921	1.000

En la figura tenemos una realización de un fBm (X) para $H = 0.7$ (en negro) y en azul $Y = \varepsilon X$. A simple vista es muy difícil detectar la dependencia de ambas series. El test planteado aquí detecta la dependencia entre X e Y con probabilidad 1, (al nivel del 5%) para un tamaño de muestra de $n = 20$, mientras que si $n = 10$ detecta la dependencia con probabilidad del 85%.



20 observaciones de $X = mBf$ con $H = 0.7$ (en negro) y sus correspondientes $Y = \varepsilon X$ en azul.



- En la Tabla 8. Bm significa un movimiento browniano estándar, mientras que fBm representa un movimiento browniano fraccional con parámetro de Hurst $H = 0.7$.

Tabla 8. Potencias para distintos n . Series de tiempo a parámetro continuo.

n	X	$Y = X^2 + 3\varepsilon$	$Y = \sqrt{ X } + \varepsilon$	$Y = \varepsilon X + 3\varepsilon'$
30	Bm	0.770	0.519	0.402
50	Bm	0.924	0.752	0.656
80	Bm	0.994	0.923	0.839
30	fBm	0.732	0.550	0.366
50	fBm	0.883	0.805	0.586
80	fBm	0.987	0.93	0.804

- En la Tabla 9, FOU representa un proceso de Ornstein-Uhlenbeck fraccionario, es decir que $Y_t = \sigma \int_{-\infty}^t e^{-\lambda(t-s)} dX_s$ donde $\lambda = 2.8$, $\sigma = 1$.
- FOU(2) representa un proceso de Ornstein-Uhlenbeck fraccionario iterado de orden 2, con parámetros $\lambda_1 = 2.8$ y $\lambda_2 = 0.3$, es decir $Y_t = \frac{\lambda_1}{\lambda_1 - \lambda_2} \sigma \int_{-\infty}^t e^{-\lambda_1(t-s)} dX_s + \frac{\lambda_2}{\lambda_2 - \lambda_1} \sigma \int_{-\infty}^t e^{-\lambda_2(t-s)} dX_s$.
Bm y fBm son como en la Tabla 8.

Tabla 9. Potencias para distintos n cuando la dependencia es del tipo FOU o FOU(2).

n	X	$Y = \text{FOU}$	$Y = \text{FOU}(2)$
60	Bm	0.13	0.07
150	Bm	0.35	0.31
300	Bm	0.62	0.67
60	fBm	0.12	0.15
150	fBm	0.27	0.25
300	fBm	0.76	0.65

Contenidos

1 Recurrence Plots

2 Algunas fechas

- Test de independencia
- Recurrence plots

3 Planteo del test

4 Resultados teóricos

- Distribución asintótica del test propuesto
- Consistencia bajo un espectro amplio de alternativas
- Alternativas contiguas

5 Implementación del test

- Cálculo del estadístico
- Selección de la función de pesos

6 Performance del test

- X e Y variables aleatorias
- X e Y vectores aleatorios
- X e Y series de tiempo

7 A futuro

Contenidos

1 Recurrence Plots

2 Algunas fechas

- Test de independencia
- Recurrence plots

3 Planteo del test

4 Resultados teóricos

- Distribución asintótica del test propuesto
- Consistencia bajo un espectro amplio de alternativas
- Alternativas contiguas

5 Implementación del test

- Cálculo del estadístico
- Selección de la función de pesos

6 Performance del test

- X e Y variables aleatorias
- X e Y vectores aleatorios
- X e Y series de tiempo

7 A futuro



Arcones, M. A. and Giné, E. (1993). Limit Theorems for U-Processes. The Annals of Probability Vol 21-3, 1494-1542.



Arratia, A., Cabaña, A. & Cabaña, E., (2016). A construction of Continuous time ARMA models by iterations of Ornstein-Uhlenbeck process, *SORT* Vol 40 (2) 267-302.



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



Blomqvist, N. (1950). On a measure of dependence between two random variables. The Annals of Mathematical Statistics 593-600.



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

<https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/187>

glioni_Beaulieu_Guillaume_2016_memoire.pdf?sequence=2.



Cabaña, E. M. (1997). Contiguidad, pruebas de ajuste y Procesos Empí ricos Transformados. Décima escuela venezolana de Matemáticas.




Eckmann, J. P, Oliffson Kamphorst, S, Ruelle, D. (1987).
Recurrence plots of dynamical systems. Europhys. Lett. 4,
973-977.



Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*. 45(273-279):135–145.



Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidts norms. International Conference on algorithmic learning theory. 63-77. Springer.

-  Heller, R., Heller, Y., and Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika*. 100(2):503–510.
-  Kalemkerian, J. (2017). Fractional Iterated Ornstein-Uhlenbeck Processes. arXiv preprint arXiv:1709.07143v1[math.ST].
-  Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.
-  Le Cam, L. & Yang, G. L. (1990). *Asymptotics in Statistics. Some Basic Concepts*. Springer, New York.
-  Marwan, N. (2008). A Historical Review of Recurrence Plots. *European Physical Journal—Special Topics*, 164, 3-12.
-  Oosterhoof, J. & Van Zwet, W. R. (1979). A note on contiguity and Hellinger distance. *Contributions to*

Statistics. Jaroslav Hájek Memorial Volume (J. Jorecová, ed) Reidel Dordrecht. 157-166



Pearson, K. (1898). 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.



Spearman, C. (1904). The proof and measurement of association between two things. The American journal of psychology. 15(1):72–101.



Székel, 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.



Székel, G. J., Rizzo, M. L., et al. (2009). Brownian distance covariance. The annals of applied statistics, 3(4):1236–1265.

