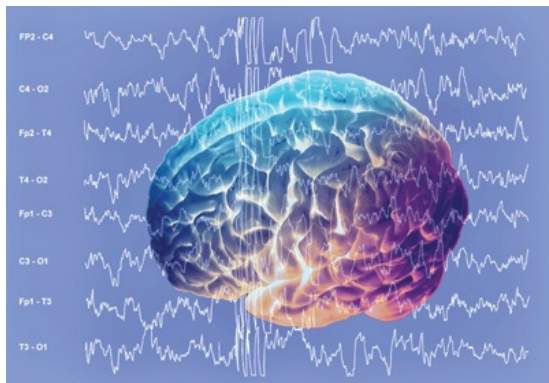


Estadística y proyecciones al azar.



Ricardo Fraiman
Leonardo Moreno
Sebastián Vallejo

UDELAR

SIESTA 2019
Facultad de Ciencias
Económicas y de
Administración

Esquema:

La dimensionalidad de los datos en estadística no paramétrica

Caracterización a través de las proyecciones uno-direccionales.

Proyecciones al azar.

Proyecciones al azar en Independencia y Simetría.

Un test de simetría.

Un test de independencia.

Simulaciones.

Aplicación.

La dimensionalidad de los datos

Según expone Donoho en [Donoho et al., 2000], una preocupación actual en la comunidad es la extensión de métodos clásicos del análisis estadístico cuando la dimensión del espacio es elevada, poder sortear las limitaciones subyacentes es una tarea desafiante.

La dimensionalidad de los datos en estadística no paramétrica

- Estimación de densidades (método del núcleo)

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{k=1}^n K\left(\frac{x - X_i}{h}\right)$$

Si $h \rightarrow 0$ y $nh^d \rightarrow \infty$ entonces $\hat{f}(x) \xrightarrow{P} f(x)$

- Puntos en una Bola:

$$V_d = \frac{\pi^{d/2} R^d}{\Gamma\left(\frac{d}{2} + 1\right)}$$

- Con $R = 1/2$ y $d = 2$ $V_2 = 0.785$
- Con $R = 1/2$ y $d = 5$ $V_5 = 0.164$
- Con $R = 1/2$ y $d = 100$ $V_{100} = 1.8 * 10^{-70}$

Algunos caminos

Uno de los mecanismos de atacar el problema es determinar una transformación del espacio original a uno de menor dimensión que conserve (en la mayor medida posible) la información relevante para el problema en cuestión. Algunos ejemplos

- PCA
- Proyecciones unidireccionales “pursuit”
- ISOMAP (manifold learning)
- Proyecciones al azar (Se proyectan los datos sobre un subespacio elegido al azar)

Proyecciones al azar

Teorema de Lema de Johnson-Lindenstrauss (A la Gupta-Das Gupta)

Para todo $0 < \epsilon < 1$ y para todo natural n , sea k un natural positivo tal que

$$k \geq \frac{24 \log(n)}{3\epsilon^2 - 2\epsilon^3}$$

Entonces para cualquier conjunto de n puntos V en \mathbb{R}^d , existe una función $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ que cumple que para todo $u, v \in V$

$$P\left((1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2\right) \geq 1 - \frac{2}{n^2}$$

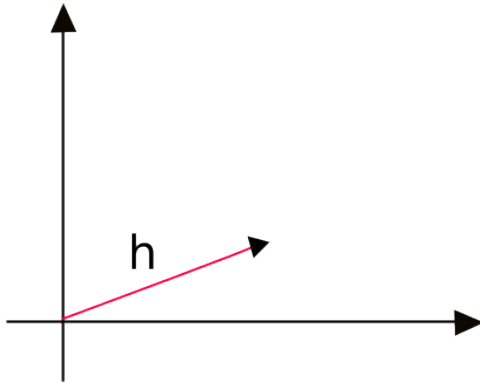
Medida inducida

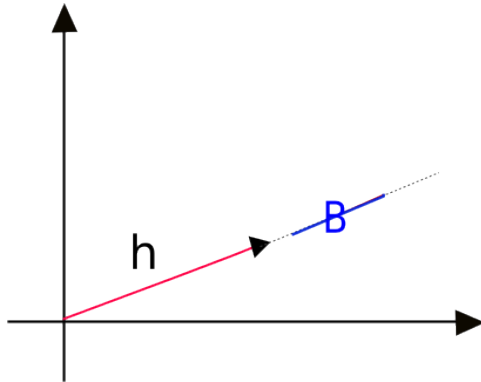
Anotamos $\pi_{\mathbf{h}}$ a la proyección ortogonal de \mathbb{R}^d en el subespacio generado por el vector \mathbf{h} de norma 1, y B un boreliano de este subespacio, entonces la medida inducida en el subespacio es,

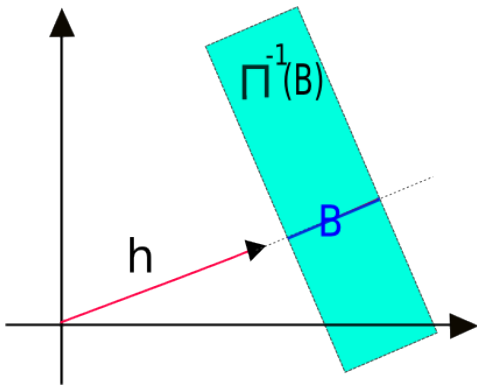
$$P_{\langle \mathbf{h} \rangle}(B) = \mathbb{P} [\pi_{\mathbf{h}}^{-1}(B)]$$

En dimensión infinita,
 $f \in E^*$,

$$P_f(B) = \mathbb{P} [f^{-1}(B)] \tag{1}$$







Teorema de Cramer-Wold

Anotando $\mathcal{E}(P, Q) = \{\mathbf{h} \in \mathbb{R}^d / P_{\langle \mathbf{h} \rangle} = Q_{\langle \mathbf{h} \rangle}\}$, se puede reescribir el teorema de Cramer-Wold,

$$\mathcal{E}(P, Q) = \mathbb{R}^d \Leftrightarrow P = Q. \quad (2)$$

Si \mathbf{X} e \mathbf{Y} son elementos aleatorios en un espacio de Banach separable E si se anota E^* al dual y $\mathcal{E}(\mathbf{X}, \mathbf{Y}) = \{f \in E^* / f(\mathbf{X}) \stackrel{d}{=} f(\mathbf{Y})\}$, se puede enunciar una versión funcional del teorema de Cramér-Wold.

$$\mathcal{E}(\mathbf{X}, \mathbf{Y}) = E^* \Leftrightarrow \mathbf{X} \text{ e } \mathbf{Y} \text{ tienen igual distribución.} \quad (3)$$

Caracterización a través de las proyecciones uno-direccionales

Caracterización de la simetría central

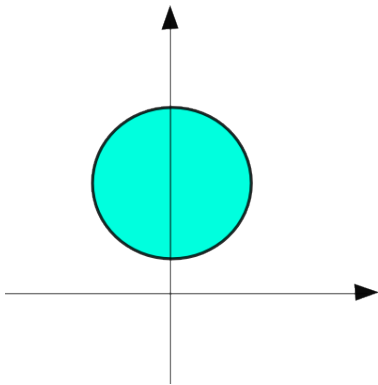
Sea \mathbf{X} un elemento aleatorio definido sobre un espacio de Banach E . \mathbf{X} es simétrico centralmente si y sólo si $f(\mathbf{X})$ y $-f(\mathbf{X})$ son variables aleatorias con la misma distribución para cualquier $f \in E^*$

Caracterización de la independencia

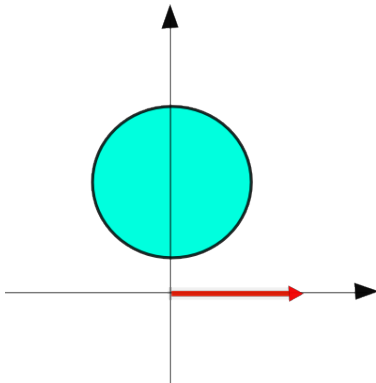
Sea E un espacio de Banach separable. Dos elementos aleatorios \mathbf{X} e \mathbf{Y} en E son independientes si y solo si $f(\mathbf{X})$ y $g(\mathbf{Y})$ son independientes para toda $f, g \in E^*$.

Ambos teoremas pueden ser expresados en un espacio de Hilbert \mathcal{H} , incluyendo el caso de dimensión finita, a partir de la representación de Riesz.

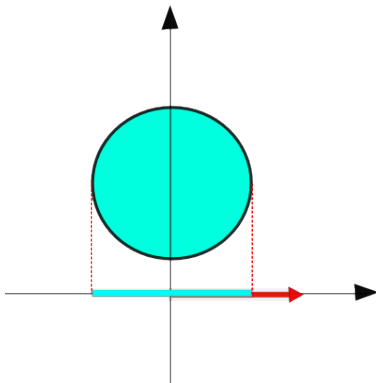
En Simetría ¿Alcanza con una dirección?



En Simetría ¿Alcanza con una dirección?



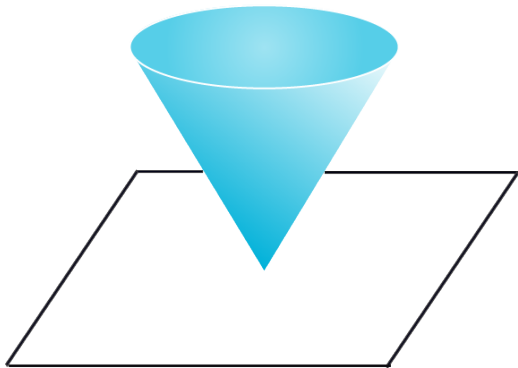
En Simetría ¿Alcanza con una dirección?



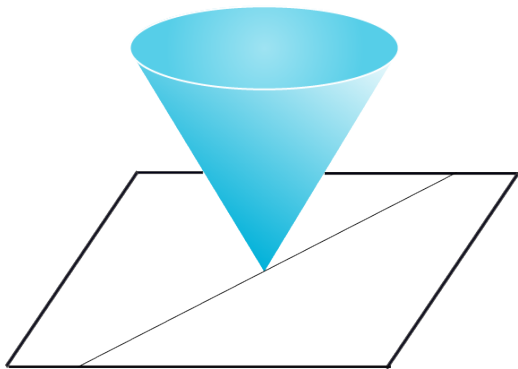
¿Y con infinitas direcciones?



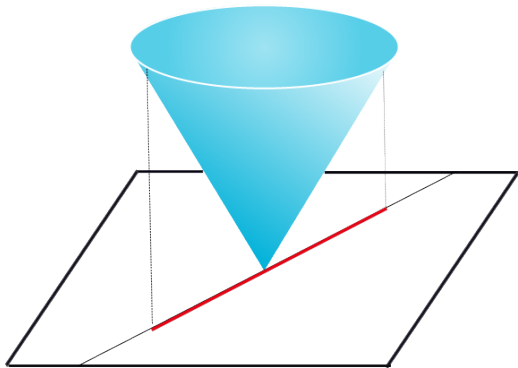
¿Y con infinitas direcciones?



¿Y con infinitas direcciones?



¿Y con infinitas direcciones?



En independencia ¿Alcanza con infinitos pares de direcciones?

Dados $U = V = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu, I)$. $\mu \in \mathbb{R}^2$ y I la matriz identidad de 2×2

Sean $a, b \in \mathbb{R}^2$,

$\langle a, U \rangle$ y $\langle b, V \rangle$ son independientes $\Leftrightarrow a \perp b$

Encontramos infinitos pares de direcciones a y b donde las proyecciones de los vectores son independientes.

$$\mathcal{E}(U, V) = \left\{ \left(a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right) \in \mathbb{R}^2 \times \mathbb{R}^2 / a_1 b_1 + a_2 b_2 = 0 \right\}.$$

En independencia ¿Alcanza con infinitos pares de direcciones?

Dados $U = V = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu, I)$. $\mu \in \mathbb{R}^2$ y I la matriz identidad de 2×2

Sean $a, b \in \mathbb{R}^2$,

$\langle a, U \rangle$ y $\langle b, V \rangle$ son independientes $\Leftrightarrow a \perp b$

Encontramos infinitos pares de direcciones a y b donde las proyecciones de los vectores son independientes.

$$\mathcal{E}(U, V) = \left\{ \left(a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right) \in \mathbb{R}^2 \times \mathbb{R}^2 / a_1 b_1 + a_2 b_2 = 0 \right\}.$$

Entonces....¿Cuántas necesito?

SÓLO UNA SI ES ELEGIDA AL AZAR

Proyecciones al azar

[Cuesta-Albertos et al., 2007]

Sean P y Q medidas de Borel en \mathbb{R}^d donde $d \geq 2$. Si se cumple que,

- P tiene momentos absolutos finitos y verifican Carleman.
- $\mathcal{E}(P, Q)$ no está contenido en alguna hipersuperficie proyectiva en \mathbb{R}^d .

Entonces $P = Q$

En particular, si el conjunto $\mathcal{E}(P, Q)$ tiene H -medida positiva en \mathbb{R}^d , siendo H una medida absolutamente continua respecto de la medida de Lebesgue, entonces dicho conjunto no está contenido en alguna hipersuperficie proyectiva.

[Cuevas and Fraiman, 2009] generalizan el enunciado anterior para elementos aleatorios definidos en espacios de Banach separables E .

[Cuevas and Fraiman, 2009]

Sea μ una medida de Radon gaussiana no degenerada en E^* . Sean Q y M dos medidas de probabilidad en E tal que,

- los momentos absolutos de M son finitos y verifica Carleman.
- el conjunto $\mathcal{E}(M, Q) = \{h \in E^* / Q_h = M_h\}$ tiene μ -medida positiva.

Entonces $Q = M$.

Proyecciones al azar en Independencia y Simetría

Teorema 1

Sea E un espacio de Banach separable y μ una medida de Radon gaussiana no degenerada en E^* . Sea \mathbf{X} un elemento aleatorio en E tal que,

- sus momentos absolutos son finitos y verifican Carleman,
- el conjunto
 $\mathcal{E}(\mathbf{X}, \mathbf{Y}) = \{f \in E^* / f(\mathbf{X}) \text{ es una variable aleatoria simétrica}\}$
tiene μ -medida positiva.

Entonces \mathbf{X} es un elemento aleatorio simétrico centralmente.

Proyecciones al azar en Independencia y Simetría

Teorema 2

Dado $(\Omega, \mathcal{B}, \mathbb{P})$ un espacio de probabilidad. Sean \mathbf{X} e \mathbf{Y} dos elementos aleatorios de E , espacio de Banach separable y μ una medida de Radon gaussiana no degenerada en $(E \times E)^*$. Suponiendo que los momentos absolutos de \mathbf{X} e \mathbf{Y} son finitos y se verifica que la serie,

$$\sum_{n \geq 1} \min \left\{ m_X^{-1/n}(n), m_Y^{-1/n}(n) \right\} \quad \text{Diverge.}$$

Si el conjunto,

$$\mathcal{E}(X, Y) = \{h \in (E \times E)^* / h(X, 0) \text{ y } h(0, Y) \text{ son v.a independientes}\}$$

tiene μ -medida positiva, entonces X e Y son independientes.

Un test de simetría

Sean $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$ un conjunto de elementos aleatorios independientes e idénticamente distribuidos en un espacio de Banach separable E , cuya distribución está determinada por sus momentos. Se quiere realizar una prueba de simetría central en E , o sea

$$\begin{aligned} H_0) \mathbf{X} \text{ y } -\mathbf{X} \text{ tienen igual distribución} \\ H_1) \mathbf{X} \text{ y } -\mathbf{X} \text{ no tienen igual distribución} \end{aligned} \tag{4}$$

Metodología

- Se sortea al azar $\mathbf{h} \in E^*$ con una μ -medida gaussiana. En el caso finito dimensional se sortea una dirección \mathbf{h} con una medida de probabilidad H en \mathbb{R}^d (H absolutamente continua respecto a la medida de Lebesgue) una dirección \mathbf{h} . En ambos casos tomamos $\|\mathbf{h}\| = 1$.
- Fijada \mathbf{h} , se considera la muestra de variables aleatorias independientes e idénticamente distribuidas inducida por \mathbf{h} , $\{\mathbf{h}(\mathbf{X}_1), \mathbf{h}(\mathbf{X}_2), \dots, \mathbf{h}(\mathbf{X}_n)\}$. En dimensión finita se proyecta ortogonalmente la muestra i.i.d $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$ sobre el espacio unidimensional generado por \mathbf{h} , obteniendo así una muestra de variables aleatorias en \mathbb{R} . Se anota $\mathbf{h}(\mathbf{X}) = \langle \mathbf{X}, \mathbf{h} \rangle$.

- Se realiza sobre estos datos proyectados con cierto nivel de significación un test de simetría en \mathbb{R} del tipo Kolmogorov-Smirnov. Si llamamos F^h a la distribución acumulada de $\mathbf{h}(\mathbf{X}_1)$, la prueba en \mathbb{R} a realizar es,

$$H_0) F^h(x) + F^h(-x) - 1 = 0 \quad \forall x \in \mathbb{R} \quad H_1) |F^h(x) + F^h(-x) - 1| > 0 \quad (5)$$

Si se denota F_n^h a la distribución empírica de los datos $\{\mathbf{h}(\mathbf{X}_1), \mathbf{h}(\mathbf{X}_2), \dots, \mathbf{h}(\mathbf{X}_n)\}$ el estadístico propuesto es

$$D^h(n) = \sup_{x \geq 0} |F_n^h(x) + F_n^h(-x^-) - 1|, \quad (6)$$

A valores “grandes” del estadístico se rechaza H_0 de las hipótesis originales. Nos referiremos a este test como por RPK_1 . Tiene distribución libre y es consistente.

Un test de independencia

Queremos implementar un test para contrastar,

H_0) \mathbf{X} e \mathbf{Y} son independientes

H_1) \mathbf{X} e \mathbf{Y} no son independientes,

siendo $\mathbf{X}, \mathbf{Y} : \Omega \rightarrow E$ elementos aleatorios, con E un espacio de Banach separable.

Metodología

Sea X e Y elementos aleatorios en E , un espacio de Banach separable, dada la muestra i.i.d $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Los pasos para la construcción del estadístico son los siguientes,

- Sorteamos $h \in (E \times E)^*$ con una medida gaussiana en $(E \times E)^*$
- Siendo $f(x) = h(x, 0)$ y $g(y) = h(0, y)$ construimos la muestra i.i.d $\{(f(X_1), g(Y_1)), \dots, (f(X_n), g(Y_n))\}$

- Consideramos el estadístico introducido en [Medovikov, 2013], que es simplemente el estadístico del tipo Crámer Von-Mises para cópula introducido por [Deheuvels, 1979], pesado por la función w ,

$$W_n = \int_{[0,1]^2} \mathbb{C}_n^2(u, v) du dv,$$

siendo,

$$\mathbb{C}_n(u, v) = \sqrt{n}[C_n^{f(X), g(Y)}(u, v) - u \cdot v] \sqrt{w(u, v)} \quad \text{tal que} \quad (u, v) \in [0, 1]^2,$$

y $C_n^{f(X), g(Y)}$ la cópula empírica,

- Se podría cambiar acá el estadístico de [Medovikov, 2013] e implementar el test de [García and González-López, 2014]
- Nuestro test hereda las propiedades del test bidimensional.**

Simulaciones: Normal trasladada, $d=2$ y $n=100$

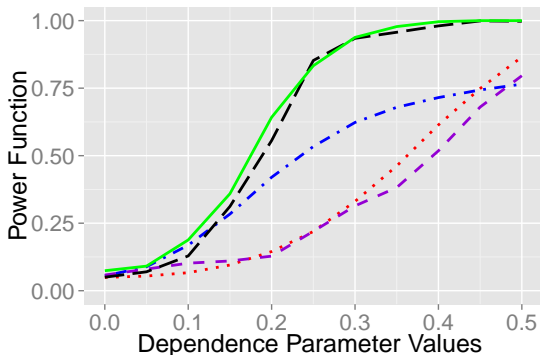


Figure: (— —) potencia del test $RPKW_1$, (—) potencia del test RPK_{10} , (- · -) potencia del test RPK_1 , (· · ·) potencia del test de Marden y (- -) potencia del test de Ley.

Simulaciones : Distribución de Azzalini, $d=2$ y $n=100$

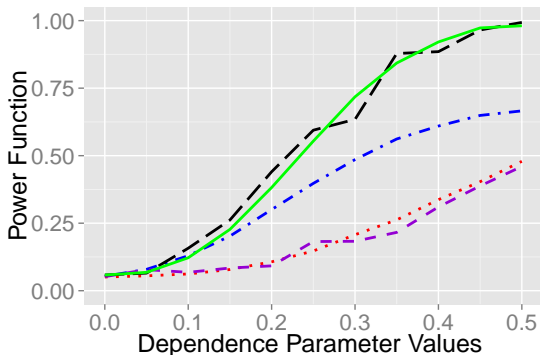


Figure: (— —) potencia del test $RPKW_1$, (—) potencia del test RPK_{10} , (- · -) potencia del test RPK_1 , (· · ·) potencia del test de Marden y (- -) potencia del test de Ley.

Simulaciones: Normal trasladada, $d=3$ y $n=100$

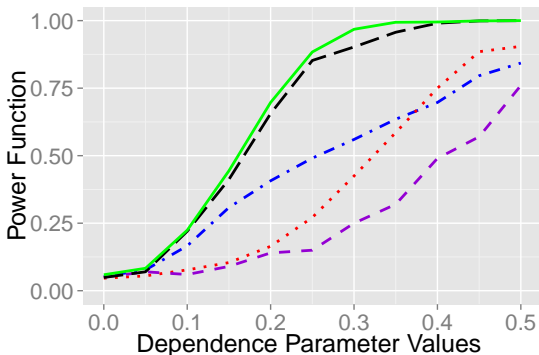


Figure: (— —) potencia del test $RPKW_1$, (—) potencia del test RPK_{10} , (- · -) potencia del test RPK_1 , (· · ·) potencia del test de Marden y (- -) potencia del test de Ley.

Simulaciones: Distribución de Azzalini, $d=3$ y $n=100$

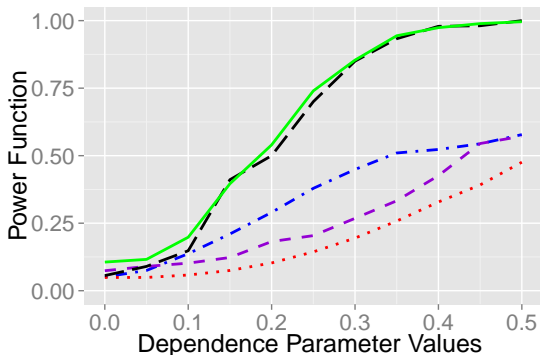


Figure: (— —) potencia del test $RPKW_1$, (—) potencia del test RPK_{10} , (- · -) potencia del test RPK_1 , (· · ·) potencia del test de Marden y (- -) potencia del test de Ley.

Simulaciones: Normal trasladada, $d=50$ y $n=100$

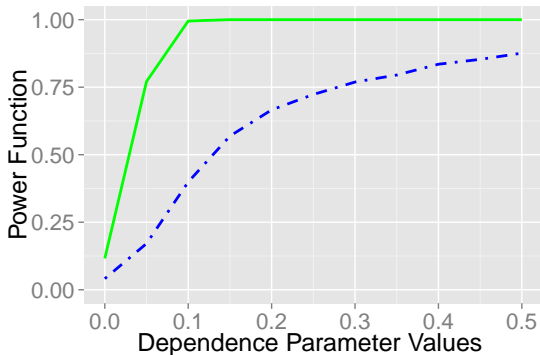


Figure: (—) potencia del test RPK_{10} y (— · —) potencia del test RPK_1 .

Simulaciones: Distribución de Azzalini, $d=50$ y $n=100$

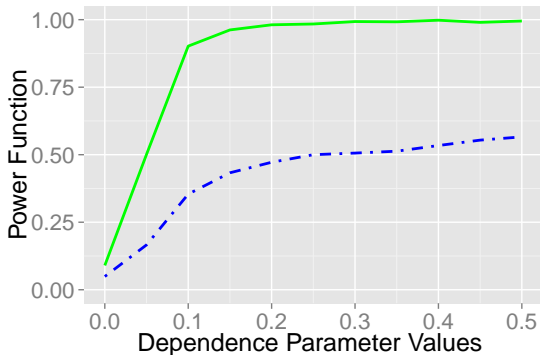


Figure: (—) potencia del test RPK_{10} y (-·-) potencia del test RPK_1 .

Simulaciones: $X(t) = W(t) + mt$ y $W(t)$ un M.B.

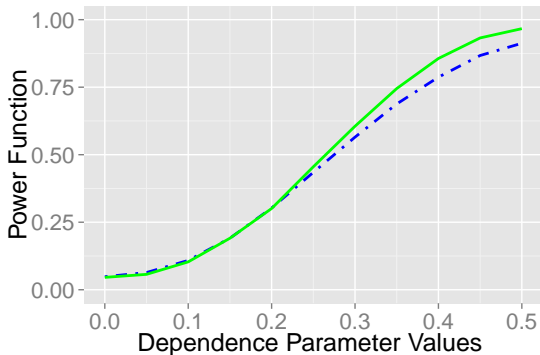


Figure: (—) potencia del test RPK_{10} y (---) potencia del test RPK_1 .

Simulaciones: Los vectores coinciden en un número de coordenadas, $d=50$, $n=20$

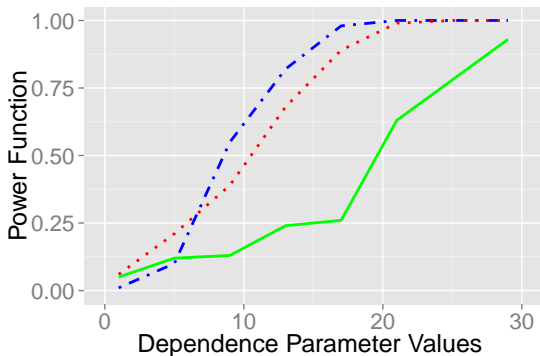


Figure: (— · —) la potencia del test $RPK.GE_{50}$, (····) la potencia del test $RPK.GA_{50}$ y (——) la potencia del test $DIST.COV$

Simulaciones: Los vectores coinciden a partir de un determinado umbral, $d=50$, $n=20$

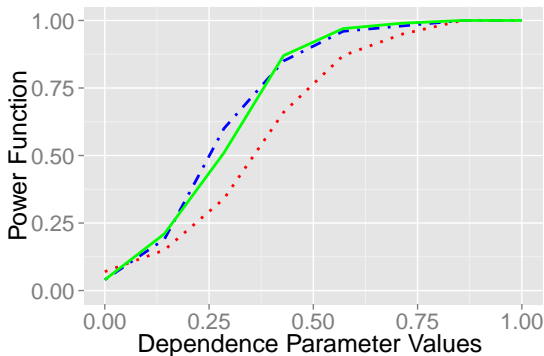


Figure: (— · —) la potencia del test $RPK.GE_{50}$, (····) la potencia del test $RPK.GA_{50}$ y (—) la potencia del test $DIST.COV$

Simulaciones:

$$X = (1 - \epsilon)U + \epsilon Z \text{ y } Y = (1 - \epsilon)V + \epsilon Z$$

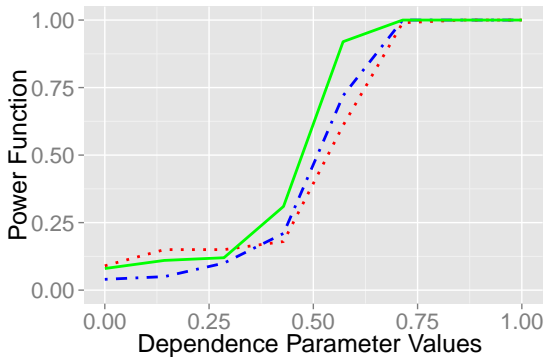
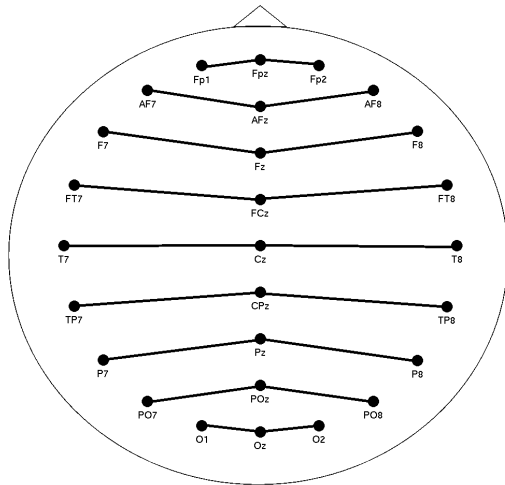
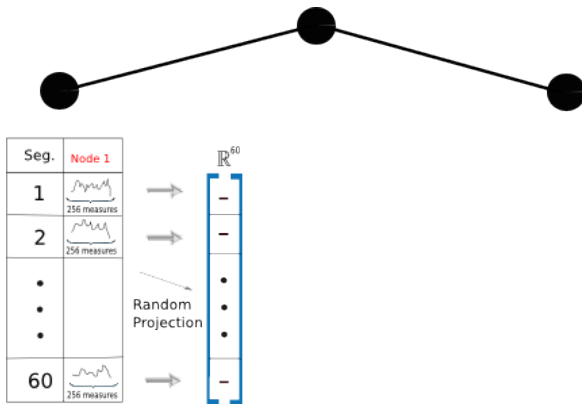


Figure: (— · —) la potencia del test $RPK.GE_{50}$, (· · ·) la potencia del test $RPK.GA_{50}$ and (—) la potencia del test $DIST.COV$

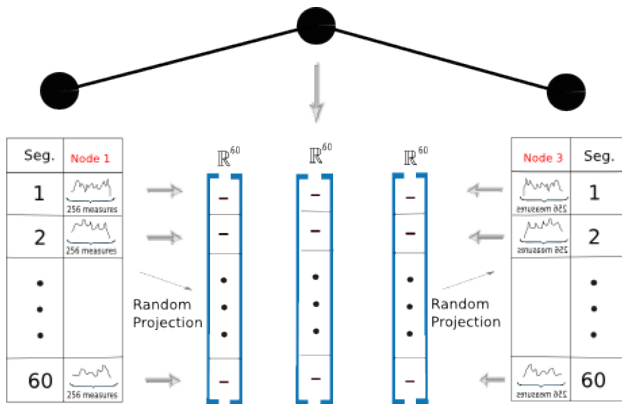
Aplicación



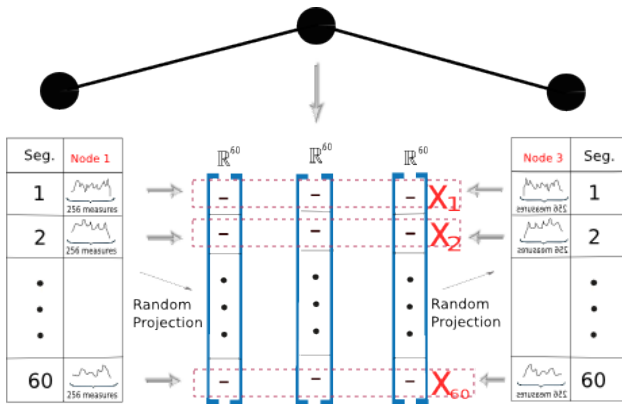
Aplicación



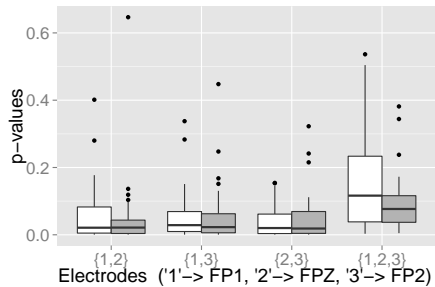
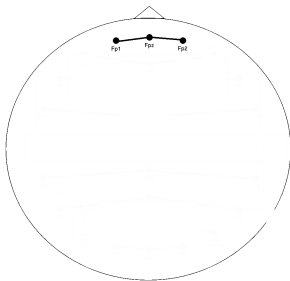
Aplicación



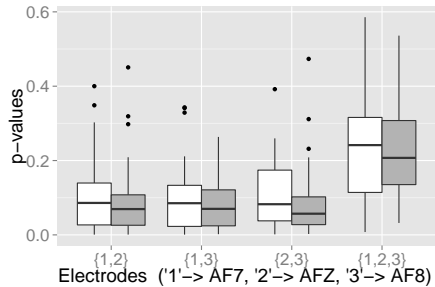
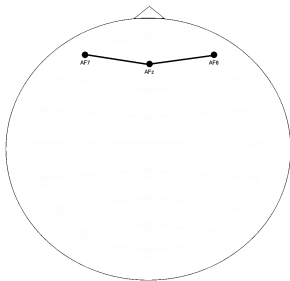
Aplicación



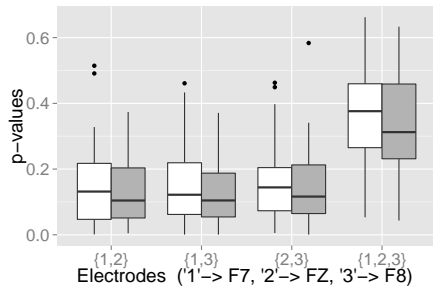
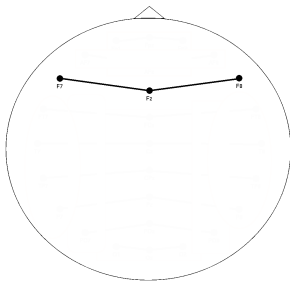
Aplicación



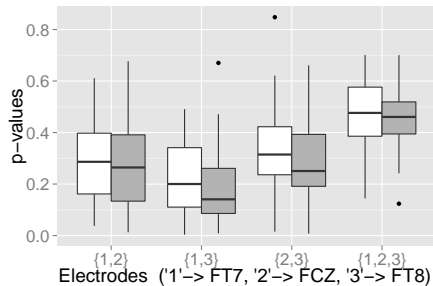
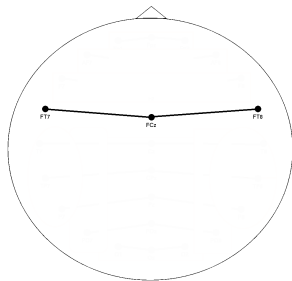
Aplicación



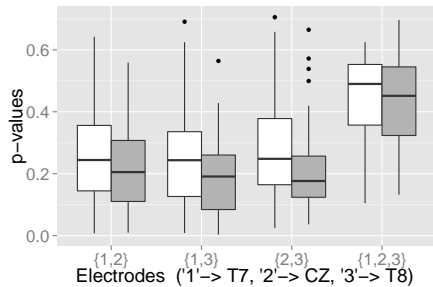
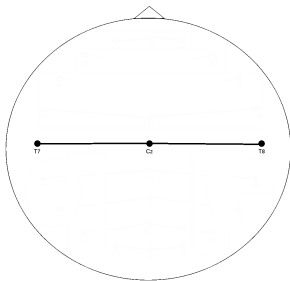
Aplicación



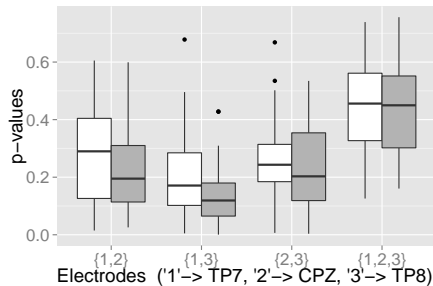
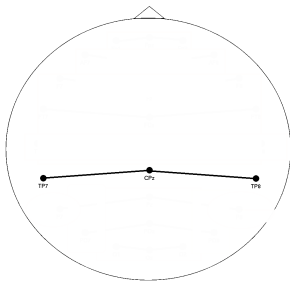
Aplicación



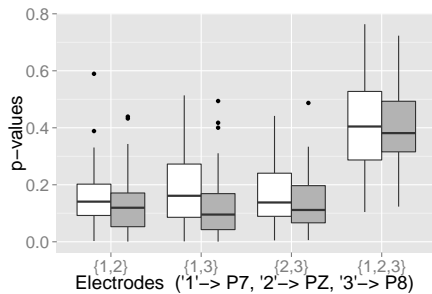
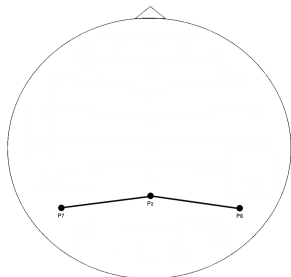
Aplicación



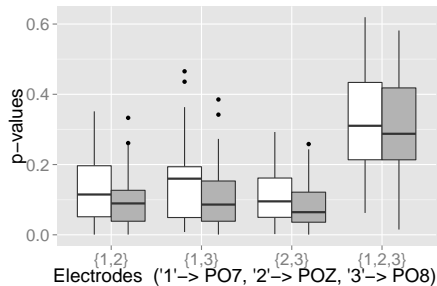
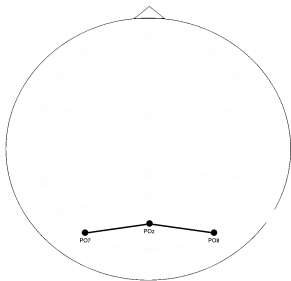
Aplicación



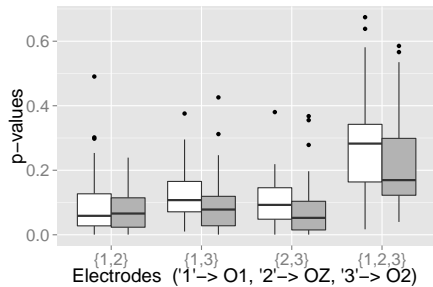
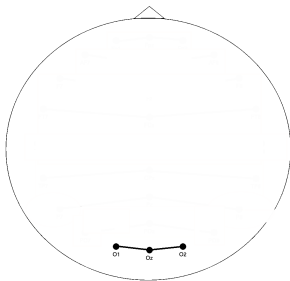
Aplicación



Aplicación





Aplicación






Gracias !



Bibliografía (1)

-  Cuesta-Albertos, J. A., Fraiman, R., and Ransford, T. (2007).
A sharp form of the Cramer–Wold theorem.
Journal of Theoretical Probability, 20:201–209.
-  Cuevas, A. and Fraiman, R. (2009).
On depth measures and dual statistics. A methodology for dealing
with general data.
Journal of Multivariate Analysis, 100:753–766.
-  Deheuvels, P. (1979).
La fonction de dépendance empirique et ses propriétés: Un test
non paramétrique d'indépendance.
Académie Royale de Belgique, 65:274–292.

Bibliografía (2)

-  Donoho, D. L. et al. (2000).
High-dimensional data analysis: The curses and blessings of dimensionality.
AMS math challenges lecture, 1(32):375.
-  García, J. E. and González-López, V. (2014).
Independence tests for continuous random variables based on the longest increasing subsequence.
Journal of Multivariate Analysis, 127:126–146.
-  Medovikov, I. (2013).
A test for independence of random vectors based on weighted empirical copula process.
to be submitted.