UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE MATEMÁTICA

Análisis de Imagenes trnasformadas con Box Cox

Memoria de Título presentada por

Fabián Castellano Núñez

como requisito parcial para optar al título de

Ingeniero Civil Matemático

Profesor Guía

Dr. Ronny Vallejos S.

Martes XX de Diciembre, 2023.

Capítulo 1

Introducción

A lo largo de la literatura se suele aplicar la transformación a vectores unidimensionales, y no ha sido extendida a matrices d-dimensionales en las que existe corrlaciones de adyacencia, excepto en el trabajo de Bicego y Baldo *(Properties of the Box-Cox Transformation for Pattern Classification)*, y en *(MR Image Segmentation Using a Power Transformation Approach)*, en amdos solo se propone una transformación que lleve las d dimensiones a 1.

Capítulo 2

Análisis Estadistico y Procesamiento de Imagenes

2.1. Introducción

2.2. Definiciones previas

asdf

Capítulo 3

Maximal Information Coeficient

3.1. Introducción

El Coeficiente de Información máxima (conocido como MIC por sus siglas en inglés), es una medida estadística propuesta por Reshef et al. en su trabajo "Detecting Novel Associations in Large Data Sets"[11]. Este coeficiente fue creado en el contexto de la ciencia generadora de hipótesis, en la cual los conjuntos de datos se utilizan para ayudar a los investigadores a formular nuevas hipótesis en lugar de probar las existentes. En este enfoque, se utilizan medidas de dependencia, que son estadísticas empleadas para evaluar pares de variables candidatas. Estos avances en el campo de analisis de datos nos han entregado muchas herramientas, tanto para la comparación de datos en si mismos, junto con formas de evaluar estas medidas en si mismas.

Sea $\hat{\varphi}$ una medida de dependencia, una forma de medir la utilidad de esta es la *potencia contra la independencia*, i.e., la capacidad de prueba de independencia basada en $\hat{\varphi}$ para detectar varios tipos de relaciones no triviales. Este es un objetivo importante para conjuntos de datos que tienen muy pocas relaciones no triviales, o solo relaciones muy débiles que son difíciles de detectar. Sin embargo, a menudo el número de relaciones declaradas estadísticamente significativas por una medida de dependencia supera con creces el número de relaciones que luego se pueden explorar más a fondo.

Para abordar este problema, se introdujo un segundo método de evaluación de una medida de dependencia llamado equitabilidad [11]. Las estadísticas equitativas asignan puntuaciones similares a relaciones igualmente fuertes, independientemente de su tipo. El objetivo es definir medidas de dependencia que logren una buena equitabilidad con respecto a medidas relevantes de la fuerza de la relación.

La idea de equitabilidad ha motivado el desarrollo de varias medidas de dependencia, con diferentes formalizaciones y enfoques en aspectos específicos de la fuerza de la relación. El desafío radica en definir medidas de dependencia que logren una buena equitabilidad con respecto a medidas importantes de la fuerza de la relación, como se ve en el artículo complementario de Reshef et al (2011) [11]. Esta línea de investigación tiene como objetivo proporcionar un enfoque más poderoso y equitativo para medir la dependencia, lo que permite una identificación y priorización más precisas de las relaciones en conjuntos de datos complejos.

En este contexto, el coeficiente de información máxima (MIC) nos entrega una medida rebusta para encontrar relaciónes no lineales entre variables, para nuestro caso en particular, entre imágenes. Cómo veremos en la sección 5, la transformación de Box-Cox es no lineal, por lo que el coeficiente nos ayudará a cuantificar la relación entre las imágenes transformadas y las originales, y podemos aprovechar la equitabilidad de este para comparar diferentes versiones de la transformación.

En esta secci/'on discuteremos la definición del MIC, algunas de sus propiedades y un par de caracterizaciones que nos ayudarán a llegar al MIC_* y posteriormente al MIC_e, ambos siendo estimadores de MIC que nos harán posible calcular este valor.

3.2. Sobre el coeficiente

El coeficiente de información máxima (Maximal Information Coefficient o MIC) es una medida estadística propuesta por Reshef et al. en su paper "Detecting Novel Associations in Large Data Sets"[11]. Este coeficiente mide la correlación entre dos variables en un conjunto de datos y se basa en la idea de que una relación fuerte entre dos variables debería ser capaz de predecir una variable a partir de la otra de manera precisa.

En este paper, Reshef et al. presentan un enfoque innovador para detectar asociaciones nobles en grandes conjuntos de datos, en lugar de buscar correlaciones fuertes entre dos variables, el coeficiente MIC permite detectar relaciones débiles pero aún importantes que pueden no ser evidentes al simplemente mirar los datos. Esto es posible gracias a que el coeficiente MIC es capaz de capturar no solo la fuerza de la correlación entre dos variables, sino también su precisión.

Para calcular el coeficiente, se comienza de la idea de que la información mutua entre dos variables es una medida de la precisión con la que se puede predecir una variable a partir de la otra. Por lo tanto, el coeficiente se calcula como la información mutua máxima posible entre dos variables, dado un conjunto de datos. Esto se hace a través de un procedimiento iterativo en el que se prueban diferentes particiones de los datos en conjuntos de entrenamiento y prueba, y se selecciona aquella que maximiza la información mutua.

El la siguiente sección estudiaremos las definiciones que nos entrega cada coefi-

ciente.

3.3. Definiciones

Como mencionamos en la parte anterior, debemos primero encontrar la información mutua entre las variables.

Definición 3.1 (Información mutua)**.** Para un vector aleatorio bivariado (X, Y), se define la información mutua como:

$$\mathrm{I}(X;Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} P_{(X,Y)}(x,y) \log \left(\frac{P_{(X,Y)}(x,y)}{P_X(x) P_Y(y)} \right) dxdy,$$

donde $P_{(X,Y)}$ es la función de densidad de probabilidad conjunta y P_X, P_Y, las distribuciones marginales de X e Y respectivamente.

Luego, sea D un conjunto finito de pares ordenados, podemos particionar los valores de la primera coordenada en x contenedores, y los valores de la segunda en y de estos. Dado una malla G, sea $D|_G$ la distribución inducida por los puntos de D en las celdas de G, i.e., la distribución en las celdas de G obtenida al dejar que la función de densidad de probabilidad en cada celda sea la fracción de puntos de D que caen en esa celda. Veamos un ejemplo

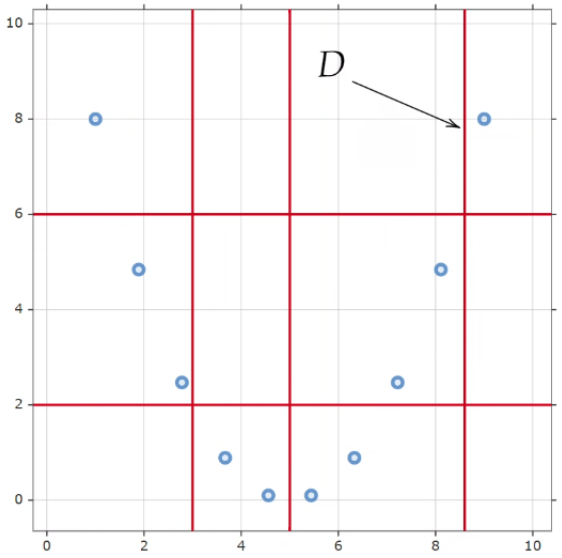

Figura 3.1: Malla G de 4x3 sobre el conjunto de pares ordenados D

Para la Figura 3.1, la función de densidad quedaría de la forma:

$$f_{D|_G}(i,j) = \begin{cases} \frac{1}{10} & \text{si } (i,j) \in \{(1,3),(4,1)\} \\ \frac{2}{10}, & \text{si } (i,j) \in \{(1,2),(2,1),(3,1),(3,2)\} \\ 0, & \text{Otro caso.} \end{cases}$$

Notemos que para un D fijo, aunque fijemos el grosor de la malla, la distribución de esta puede variar dependiendo de donde hagamos los cortes, por ejemplo:

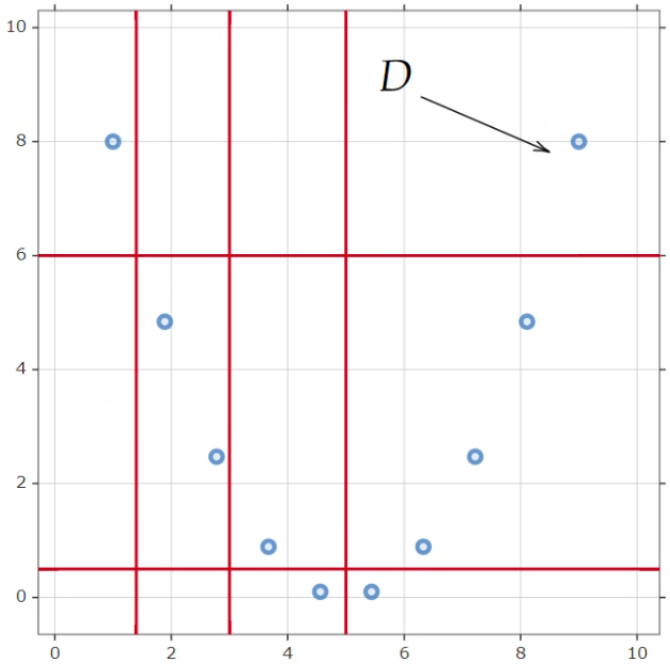

Figura 3.2: Otra malla G de 4x3 sobre el conjunto de pares ordenados D.

Aquí podemos ver que la función de densidad que nos entrega está malla es distinta a la definida para la Figura 3.2 . Este es un hecho que explotamos en la siguiente definición:

Definición 3.2. Para un conjunto finito $D \in \mathrm{R}^2$ y enteros positivos i, j, definimos:

$$I^*(D,i,j) = \max I(D|_G),$$

donde el máximo es sobre todas las mallas G con i columnas y j filas, con $I(D|_G)$ denota la información mutua de $D|_G$.

Ya teniendo este valor procedemos a definir la matriz caracteristica del conjunto D.

Definición 3.3. La matriz caracteristica $M(D)$ de un conjunto de pared ordenados D es una matriz infinita con entradas:

$$M(D)_{x,y} = \frac{I^*(D,x,y)}{\log \text{mín}\{x,y\}}.$$

Definición 3.4. El coeficiente de información máxima o *MIC* de un conjunto bivariado D de tamaño n y una malla de tamaño menor a $B(n)$ esta dado por:

$$\text{MIC}(D) = \max_{xy<B(n)} \{M(D)_{x,y}\},$$

donde $\omega(1) < B(n) \leq O\left(n^{1-\varepsilon}\right)$ para algún $0 < \varepsilon < 1$.

Observación 3.1. A menos que se especifique de otra forma, al momento de trabajar con esta medida usaremos $B(n) = n^{0,6}$, función de la cuál se encontró que funconá en práctica en el artículo complementario de Reshef et al. (2011) [11], discuteremos la selección de este parámetro más adelante en 3.6.2

3.4. Formas prácticas de calcular el MIC. MIC_*, TIC_e y MIC_e

æ En el artículo "Measuring Dependence Powerfully and Equitably"de Reshef et al. [12], los autores presentan y caracterizan teóricamente dos nuevas medidas de dependencia: MIC_* y MIC_e. MIC_* es una medida de dependencia poblacional, y el artículo presenta tres formas de ver esta cantidad. Los autores demuestran que MIC_* es el valor poblacional del coeficiente de información máxima (MIC), una suavización mínima de la información mutua y el supremo de una secuencia infinita. Estas caracterizaciones simplifican el cálculo y fortalecen los resultados teóricos.

Además, los autores desarrollan algoritmos eficientes para aproximar MIC_* en la práctica y estimarlo de manera consistente a partir de una muestra finita. Introducen MIC_e, un estimador consistente de MIC_*, que es computable de manera eficiente y más rápido en la práctica que el algoritmo heurístico para calcular MIC. A través de simulaciones, demuestran que MIC_e tiene mejores propiedades de sesgo/varianza y supera a los métodos existentes en términos de equitabilidad con respecto a R2 en un amplio conjunto de relaciones funcionales ruidosas.

3.4.1. Definiciones y propiedades de MIC_*

En esta sección, abordaremos las definiciones esenciales para el cálculo del MIC_e. El coeficiente máximo de información poblacional puede expresarse de diversas maneras equivalentes, como veremos más adelante. Sin embargo, comenzaremos con la definición más sencilla.

Definición 3.5. Sea (X, Y) un vector aleatorio bivariado. El coeficiente de información máxima poblacional (MIC_*) de (X, Y) se define como:

$$MIC_*(X, Y) = \sup_G \frac{I\left((X, Y)|_G\right)}{\log \|G\|},$$

donde $||G||$ denota el mínimo entre el número de filas y el número de columnas de la malla G.

Ya que $I(X, Y) = sup_G I((X, Y)|G)$ (Cover y Thomas, 2006 [14, Cap. 8]), esto puede interpretarse como una versión regularizada de la información mutua que sanciona las rejillas complejas y garantiza que el resultado esté dentro del rango entre cero y uno.

Previo a continuar, introducimos una definición equivalente y sencilla de MIC_* que resulta útil para los resultados en esta sección. Esta definición considera a MIC_* como el supremo de una matriz denominada matriz característica poblacional, que se define a continuación.

Definición 3.6. Sea (X, Y) una pareja de variables aleatorias conjuntamente distribuidas. Sea

$$I^*((X, Y), k, \ell) = \max_{G \in G(k,\ell)} I\left((X, Y)|_G\right),$$

la matriz característica poblacional de (X,Y), denotada por M(X,Y), se define como

$$M(X, Y)_{k,\ell} = \frac{I^*((X, Y), k, \ell)}{\log \min\{k, \ell\}}.$$

para k, $l > 1$.

Es fácil ver lo siguiente:

Proposición 3.1. Sea (X, Y) una vector aleatorio bidimensional de variables aleatorias conjuntamente distribuidas. Tenemos

$$MIC_*(X, Y) = \sup M(X, Y),$$

donde $M(X, Y)$ es la matriz característica poblacional de (X, Y).

La matriz característica poblacional recibe este nombre porque, al igual que el MIC_*, el supremo de esta matriz, captura una noción de la intensidad de la relación, y otras propiedades de esta matriz se relacionan con diferentes características de las relaciones. Por ejemplo, más adelante en este documento presentamos una propiedad adicional de la matriz característica, el coeficiente de información total, que es útil para comprobar la presencia o ausencia de una relación en lugar de cuantificar la intensidad de la relación.

3.4.2. El MIC_* es el valor poblacional del MIC

Con el MIC_* definido, presentamos nuestra primera caracterización alternativa de este, como el límite de muestra grande del estadístico MIC introducido en Reshef et al. [11]. Recordemos la Definiciones del MIC y la matriz característica de muestra. Notemos que para evistar confución denotaremos como MIC al estadístico MIC y como MIC_* al coeficiente de información máxima poblacional.

Definición 3.7. (Reshef et al., 2011 [11]) Sea $D \subset \mathbb{R}^2$ un conjunto de pares ordenados. La matriz característica de muestra $\widehat{M}(D)$ de D se define por

$$\widehat{M}(D)_{k,\ell} = \frac{I^*(D,k,\ell)}{\log \text{mín}\{k,\ell\}}.$$

Definición 3.8. (Reshef et al., 2011 [11]) Sea $D \subset \mathbb{R}^2$ un conjunto de n pares ordenados, y sea $B : \mathbb{Z}^+ \to \mathbb{Z}^+$. Definimos

$$MIC_B(D) = \max_{k\ell \leq B(n)} \widehat{M}(D)_{k,\ell},$$

donde la función $B(n)$ es especificada por el usuario. En el paper Reshef et al. (2011) [11] se sugirió que $B(n)$ se elija como n^α para alguna constante α en el rango de 0.5 a 0.8. (Los estadísticos que presentaremos más adelante tendrán un parámetro análogo; véase la Sección 4.4.1.)

El siguiente resultado, demostrado en el paper de [12], sobre la convergencia de funciones de la matriz característica de muestra a sus contrapartes poblacionales, una consecuencia de lo cual es la convergencia de MIC a MIC_*. (En la declaración del teorema a continuación, recordemos que m_∞ es el espacio de matrices infinitas equipadas con la norma supremo, y dada una matriz A, la proyección ri anula todas las entradas $A_{k,\ell}$ para las cuales $k\ell > i$.)

Teorema 3.1. *Sea $f : m^\infty \to \mathbb{R}$ uniformemente continua, y suponga que $f \circ r_i \to f$ puntualmente. Entonces, para cada variable aleatoria (X,Y), tenemos*

$$\left(f \circ r_{B(n)}\right)\left(\widehat{M}\left(D_n\right)\right) \to f(M(X,Y)),$$

en probabilidad donde D_n es una muestra de tamaño n de la distribución de (X,Y), siempre que $\omega(1) < B(n) \leq O\left(n^{1-\varepsilon}\right)$ para algún $\varepsilon > 0$.

Dado que el supremo de una matriz es una función uniformemente continua en m_∞ y se puede realizar como el límite de máximos de segmentos cada vez más grandes de la matriz, este teorema genera nuestra afirmación sobre MIC_* como corolario.

Corolario 3.2. *MIC es un estimador consistente de MIC_* siempre que $\omega(1) < B(n) \leq O(n^{1-\epsilon})$ para algún $\epsilon > 0$*

Con esto podemos trabajar, bajo ciertas condiciones, con el MIC_* como reemplazo del MIC. Pero, ¿Cuál es la ventaja de trabajar con este nuevo estimador?

En pocas palabras, es más fácil de estimar, y esto lo veremos en la en una sección más adelante. Antes de esto debemos revisar una caracterización del MIC_* que nos permitirá contruir un estimador de este.

3.4.3. El MIC_* es el supremo de la matriz característica de muestra

Ahora mostramos la una vista alternativa de MIC_*: que puede definirse de manera equivalente como el supremo sobre un límite de la matriz característica en lugar de como un supremo sobre todas las entradas de la matriz. Esta caracterización de MIC_* servirá como base tanto para nuestro enfoque de aproximación de $MIC_*(X, Y)$ como para el nuevo estimador de MIC_* que presentamos más adelante en este artículo.

Comenzamos definiendo lo que entendemos por límite de la matriz característica. Nuestra definición se basa en la siguiente observación.

Proposición 3.2. Sea M una matriz característica poblacional. Entonces, para $\ell \geq k$, $M_{k,\ell} \leq M_{k,\ell+1}$.

Demostración. Sea (X, Y) la variable aleatoria en cuestión. Siempre podemos dejar una fila/columna vacía, sabemos que $I^*((X, Y), k, \ell) \leq I^*((X, Y), k, \ell + 1)$. Y dado que $\ell, \ell + 1 \geq k$, sabemos que $M_{k,\ell} = I^*((X, Y), k, \ell)/\log k \leq I^*((X, Y), k, \ell + 1)/\log k = M_{k,\ell+1}$. □

Dado que las entradas de la matriz característica están acotadas, el teorema de convergencia monótona nos da el siguiente corolario. En el corolario y en adelante, dejamos $M_{k,\uparrow} = \lim_{\ell\to\infty} M_{k,\ell}$ y definimos $M_{\uparrow,\ell}$ de manera similar.

Corolario 3.3. *Sea M una matriz característica poblacional. Entonces, $M_{k,\uparrow}$ existe, es finito e igual a $\sup_{\ell\geq k} M_{k,\ell}$. Lo mismo es válido para $M_{\uparrow,\ell}$.*

El corolario anterior nos permite definir el límite de la matriz característica.

Definición 3.9. Sea M una matriz característica poblacional. El límite de M es el conjunto

$$\partial M = \{M_{k,\uparrow} : 1 < k < \infty\} \bigcup \{M_{\uparrow,\ell} : 1 < \ell < \infty\}.$$

El teorema siguiente da una relación entre el límite de la matriz característica y MIC_*.

Teorema 3.4. *Sea (X, Y) un vector aleatorio bivariado. Tenemos*

$$MIC_*(X, Y) = \sup \partial M(X, Y),$$

donde $M(X, Y)$ es la matriz característica poblacional de (X, Y).

Demostración. El siguiente argumento muestra que cada entrada de M es, como máximo, $\sup \partial M$: fije un par (k, ℓ) y observe que, o bien $k \leq \ell$, en cuyo caso $M_{k,\ell} \leq M_{k,\uparrow}$, o bien $\ell \leq k$, en cuyo caso $M_{k,\ell} \leq M_{\uparrow,\ell}$.
Por lo tanto, $\text{MIC}_* \leq \sup \{M_{\uparrow,\ell}\} \cup \{M_{k,\uparrow}\} = \sup \partial M$. □

Por otro lado, el Corolario muestra que cada elemento de ∂M es un supremo sobre algunos elementos de M. Por lo tanto, $\sup \partial M$, al ser un supremo sobre supremos de elementos de M, no puede exceder $\sup M = \text{MIC}_*$.

3.4.4. Estimando el MIC_* con MIC_e

Como hemos revisado, MIC_* es el valor poblacional del estadístico MIC introducido en Reshef et al. (2011). Sin embargo, aunque es consistente, el estadístico MIC no se conoce por ser eficientemente computable y en Reshef et al. (2011) [11] se calculó en su lugar un algoritmo heurístico de aproximación llamado Approx-MIC. En esta sección, revisaremos un estimador de MIC_* que es tanto consistente como eficientemente computable. El nuevo estimador, llamado MIC_e, tiene una mejor complejidad de tiempo de ejecución incluso que el algoritmo heurístico Approx-MIC y es órdenes de magnitud más rápido en la práctica.

El estimador MIC_e se basa en la caracterización alternativa de MIC_* probada en la sección anterior, si MIC_* puede considerarse como el supremo del límite de la matriz característica en lugar de la matriz completa, entonces solo el límite de la matriz debe estimarse con precisión para estimar MIC_*. Esto tiene la ventaja de que, mientras que calcular entradas individuales de la matriz característica de muestra implica encontrar rejillas óptimas (bidimensionales), estimar las entradas del límite nos requiere solo encontrar particiones óptimas (unidimensionales). Si bien el primer problema es computacionalmente difícil, el segundo puede resolverse utilizando el algoritmo de programación dinámica de Reshef et al.(2011)[11].

En Reshef (2016) [12], esta idea es formalizada a travez de un objeto llamado la matriz equicaracterística, la cual es denominadad $[M]$. La diferencia entre $[M]$ y la matriz característica M es la siguiente: mientras que la entrada k, ℓ-th de M se calcula a partir de la información mutua máxima alcanzable utilizando cualquier cuadrícula de k-por- ℓ, la entrada k, ℓ-th de $[M]$ se calcula a partir de l información mutua máxima alcanzable utilizando cualquier cuadrícula de k-por-ℓ que equiparticiona la dimensión con más filas/columnas. 3.3 (Ver Figura 1.) A pesar de esta diferencia, a medida que la equipartición en cuestión se vuelve más y más fina, se vuelve indistinguible de una partición óptima del mismo tamaño Esta intuición se puede formalizar para mostrar que el límite de $[M]$ es igual al límite de M, y por lo tanto que $\sup[M] = \sup M = \text{MIC}*$. Entonces, se deducirá que estimar $[M]$ y tomar el supremo, como lo hicimos con M en el caso de MIC, proporciona una estimación consistente de MIC$*$.

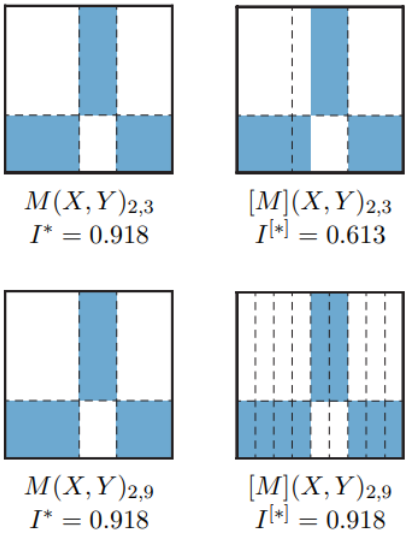

Figura 3.3: Un esquema que ilustra la diferencia entre la matriz característica M y la matriz equicaracterística $[M]$. (Arriba) Cuando se restringe a 2 filas y 3 columnas, la matriz característica M se calcula a partir de la rejilla óptima de 2 por 3. En contraste, la matriz equicaracterística $[M]$ aún optimiza la partición más pequeña de tamaño 2 pero está restringida a tener la partición más grande como una equipartición de tamaño 3. Esto resulta en una información mutua más baja de 0.613. (Abajo) Cuando se permiten 9 columnas en lugar de 3, la rejilla encontrada por la matriz característica no cambia, ya que la rejilla con 3 columnas ya era óptima. Sin embargo, ahora la matriz equicaracterística utiliza una equipartición en columnas de tamaño 9, cuya resolución es capaz de capturar completamente la dependencia entre X e Y.

3.5. La matriz equicaracterística

Ahora definimos la matriz equicaracterística y mostramos que su supremo es efectivamente MIC^*. Para hacerlo, primero definimos una versión de I^* que equiparticiona la dimensión con más filas/columnas. Observe que en la definición, los corchetes se utilizan para indicar la presencia de una equipartición.

Definición 3.10. Sea (X,Y) variables aleatorias conjuntamente distribuidas. Definir

$$I^*((X,Y),k,[\ell]) = \max_{G\in G(k,[\ell])} I\left((X,Y)|_G\right),$$

donde $G(k,[\ell])$ es el conjunto de rejillas de k por $[\ell]$ cuya partición del eje y es una equipartición de tamaño ℓ. Definir $I^*((X,Y),[k],\ell)$ análogamente.

Definir $I^{[]}((X,Y),k,\ell)$ igual a $I^(X,Y),k,[\ell])$ si $k<\ell$ y $I^(X,Y),[k],\ell)$ en caso contrario.

Ahora definimos la matriz equicaracterística en términos de $I^{[*]}$. En la definición a continuación, continuamos nuestra convención de usar corchetes para denotar la presencia de equiparticiones.

Definición 3.11. Sea (X, Y) variables aleatorias conjuntamente distribuidas. La matriz equicaracterística de población de (X, Y), denotada por $[M](X, Y)$, se define por

$$[M](X,Y)_{k,\ell} = \frac{I^{[*]}((X,Y),k,\ell)}{\log \text{mín}\{k,\ell\}},$$

para $k, \ell > 1$.

La frontera de la matriz equicaracterística se puede definir mediante un límite de la misma manera que la matriz característica. Luego tenemos el siguiente teorema.

Teorema 3.5. *Sea (X, Y) variables aleatorias conjuntamente distribuidas. Entonces $\partial[M] = \partial M$.*

Demostración. Apéndice F de Reshef 2016 □

Dado que cada entrada de la matriz equicaracterística está dominada por alguna entrada en su frontera, la equivalencia de $\partial[M]$ y ∂M produce el siguiente corolario como una simple consecuencia.

Corolario 3.6. *Sea (X, Y) variables aleatorias conjuntamente distribuidas. Entonces* $\sup[M](X, Y) = MIC_*(X, Y)$.

3.6. El estimador MIC_e

Con la matriz equicaracterística definida, podemos ahora definir nuestro nuevo estimador MIC_e en términos de la matriz equicaracterística de muestra, de manera análoga a cómo definimos MIC con la matriz característica de muestra.

Definición 3.12. Sea $D \subset \mathbb{R}^2$ un conjunto de pares ordenados. La matriz equicaracterística de muestra $\widehat{[M]}(D)$ de D se define como

$$\widehat{[M]}(D)_{k,\ell} = \frac{I^{[*]}(D,k,\ell)}{\log \text{mín}\{k,\ell\}}.$$

Definición 3.13. Sea $D \subset \mathbb{R}^2$ un conjunto de n pares ordenados, y sea $B : \mathbb{Z}^+ \to \mathbb{Z}^+$. Definimos

$$MIC_{e,B}(D) = \max_{k\ell \leq B(n)} \widehat{[M]}(D)_{k,\ell}.$$

Con la equivalencia establecida entre la frontera de la matriz característica y el de la matriz equicaracterística, es fácil demostrar que MICe es un estimador

consistente de MIC$*$ mediante argumentos similares a los que aplicamos en el caso de MIC. (Ver Apéndice G. Reshef (2016)[12]) Específicamente, mostramos el siguiente teorema, un análogo del Teorema 6.

Teorema 3.7. *Sea $f : m^{\infty} \to \mathbb{R}$ uniformemente continua, y suponga que $f \circ r_i \to f$ puntualmente. Entonces para cada variable aleatoria (X, Y), tenemos:*

$$\left(f \circ r_{B(n)}\right)\left(\widehat{M}\right]\left(D_n\right)\Big) \to f([M](X, Y)),$$

en probabilidad donde D_n es una muestra de tamaño n de la distribución de (X, Y), siempre que $\omega(1) < B(n) \leq O\left(n^{1-\varepsilon}\right)$ para algún $\varepsilon > 0$.

Demostración. Apéndice A. Reshef (2016) [12] □

Al establecer $f([M]) = \sup[M]$, obtenemos como corolario la consistencia de MIC_e.

Corolario 3.8. *MIC_B es un estimador consistente de MIC_* siempre que $\omega(1) < B(n) \leq O\left(n^{1-\varepsilon}\right)$ para algún $\varepsilon > 0$.*

3.6.1. Computando MIC_e

Tanto el MIC como el MIC_e son estimadores consistentes de MIC_*. La diferencia entre ellos radica en que, mientras que el MIC actualmente solo se puede calcular de manera eficiente a través de una aproximación heurística, el MIC_e se puede calcular de manera exacta y muy eficiente mediante un enfoque similar al utilizado para aproximar MIC_* que involucra la subrutina *OptimizeXAxis*. Ahora repasaremos los detalles de este enfoque.

Recordemos que, dada una partición fija del eje x Q en ℓ columnas, un conjunto de n puntos de datos, una partición "maestra" del eje y Π, y un número k, la subrutina *OptimizeXAxis* encuentra, para cada $2 \leq i \leq k$, una partición del eje y $P_i \subset \Pi$de tamaño como máximo i que maximiza la información mutua inducida por la cuadrícula (P_i, Q). El algoritmo realiza esto en un tiempo de $O(|\Pi|^2 k\ell)$. Para obtener más detalles sobre *OptimizeXAxis*, consulte la Sección 3.5 de Reshef el al. (2016) [12]

En el par de teoremas a continuación, se muestran dos formas en que *OptimizeXAxis* se puede utilizar para calcular eficientemente el MIC_e.

Teorema 3.9. *Existe un algoritmo EQUICHAR que, dada una muestra D de tamaño n y algún $B \in \mathbb{Z}^+$, computa la porción $r_{B(n)}(\widehat{[M]}(D))$ de la matriz equicaracteristica poblacional en tiempo $O\left(n^2 B^2\right)$, que es equivalente a $O\left(n^{4-2\varepsilon}\right)$ para $B(n) = O\left(n^{1-\varepsilon}\right)$ con $\varepsilon > 0$.*

Demostración. Describimos el algoritmo y simultáneamente acotamos su tiempo de ejecución. Lo hacemos únicamente para las entradas k, ℓ-ésimas de $\widehat{[M]}(D)$

que satisfacen $k \leq \ell, k\ell \leq B$. Esto es suficiente, ya que por simetría, calcular el resto de las entradas requeridas como máximo duplica el tiempo de ejecución.

Para calcular $\widehat{[M]}(D)_{k,\ell}$ con $k \leq \ell$, debemos fijar una partición equitativa en ℓ columnas en el eje x y luego encontrar la partición óptima del eje y de tamaño como máximo k. Si configuramos la partición maestra Π del algoritmo *OptimizeXAxis* como una partición equitativa en filas de tamaño n, entonces realiza precisamente la optimización requerida. Además, para un ℓ fijo, puede llevar a cabo la optimización simultáneamente para todos los pares $(2,\ell),\ldots,(B/\ell,\ell)$ en tiempo $O\left(|\Pi|^2(B/\ell)\ell\right) = O\left(n^2B\right)$. Para un ℓ fijo, este conjunto contiene todos los pares (k,ℓ) que satisfacen $k \leq \ell, k\ell \leq B$. Por lo tanto, para calcular todas las entradas requeridas de $\widehat{[M]}(D)$, solo necesitamos aplicar este algoritmo para cada $\ell = 2,\ldots,B/2$. Hacerlo resulta en un tiempo de ejecución de $O\left(n^2B^2\right)$. □

El algoritmo mencionado anteriormente, aunque es de tiempo polinómico, no es lo suficientemente eficiente para su uso en la práctica. Sin embargo, una modificación simple resuelve este problema sin afectar la consistencia de las estimaciones resultantes. La modificación se basa en el hecho de que OptimizeXAxis puede usar particiones maestras Π además de la partición equitativa de tamaño n que utilizamos anteriormente. Específicamente, configurar Π en el algoritmo anterior como una partición equitativa en ck "grupos", donde k es el tamaño de la partición óptima más grande que se está buscando, acelera significativamente el cálculo. Esta modificación proporciona una estadística ligeramente diferente, pero que tiene todas las propiedades teóricas de MICe, es decir, una estimación consistente de MIC$*$ y un cálculo exacto eficiente. Estas propiedades se formalizan en el siguiente teorema

Teorema 3.10. *Sea (X,Y) un vertor aleatorio bivariado, y sea D_n y sea una muestra de tamaño n de la distribucion (X,Y). Para cada $c \geq 1$, existe una matriz $\{\widehat{M}\}^c(D_n)$ tal que:*

1. *La función*
$$\widetilde{MIC}_{e,B}(\cdot) = \max k\ell \leq B(n)\{\widehat{M}\}^c(\cdot)k,\ell,$$
es un estimador concistente de MIC_* dado $\omega(1) < B(n) \leq O\left(n^{1-\varepsilon}\right)$ para algún $\varepsilon > 0$. 2. Existe un algoritmo EQUICHARCLUMP para comparar $r_B\left(\{\widehat{M}\}^c(D_n)\right)$ en tiempo $O\left(n + B^{5/2}\right)$, que equivale a $O\left(n + n^{5(1-\varepsilon)/2}\right)$ cuando $B(n) = O\left(n^{1-\varepsilon}\right)$.

Demostración. Apéndice H. Reshef (2016) [12] □

Para un análisis del efecto del parámetro c en el teorema anterior en los resultados del algoritmo EQUICHARCLUMP, consulte el Apéndice H.3 de Reshef el al. (2016) [12]. Estableciendo $\varepsilon = 0{,}6$ en el teorema anterior, obtenemos el siguiente corolario.

Corolario 3.11. *MIC_* puede estimarse de manera consistente en tiempo lineal.*

Por supuesto, en tamaños de muestra pequeños, establecer $\varepsilon = 0{,}6$ sería indeseable. Sin embargo, nuestro el artículo complementario (Reshef et al.,(2015a) [4]) demuestra empíricamente que en tamaños de muestra grandes, esta estrategia funciona muy bien en relaciones típicas.

Cabe destacar que el algoritmo EQUICHARCLUMP dado anteriormente es asintóticamente más rápido incluso que el algoritmo heurístico APPROX-MIC utilizado para calcular MIC en la práctica, que se ejecuta en tiempo $O\left(B(n)^4\right)$. Como se demostró en el artículo complementario (Reshef et al., 2015a [12]), esta diferencia se traduce en una diferencia sustancial en los tiempos de ejecución para un rendimiento similar en una gama de tamaños de muestra realistas, que va desde una aceleración de 30 veces en $n = 500$ hasta más de 350 veces en $n = 10,000$.

3.6.2. Eligiendo $B(n)$

Recordemos que, como fue propuesto en Reshef el al. (2011) [11], utilizamos funciones de la forma $B(n) = n^\alpha$. Valores grandes de α conducen a un aumento en el error esperado en regímenes de baja señal (R^2 bajos) debido tanto a un sesgo positivo en esos regímenes como a un aumento general en la varianza que afecta predominantemente a esos regímenes. Por otro lado, valores pequeños de α llevan a un aumento en el error esperado en regímenes de alta señal (R2 altos) al generar un sesgo negativo en esos regímenes y desplazar la varianza del estimador hacia esos regímenes. En otras palabras, valores más bajos de α son más adecuados para detectar señales más débiles, mientras que valores más altos de α son más adecuados para distinguir entre señales más fuertes. Esto concuerda con los resultados observados en el trabajo complementario (Reshef et al., (2015a) [4]), que muestran que valores bajos de α hacen que MICe proporcione pruebas de independencia con mejor potencia, mientras que valores altos de α hacen que MICe tenga una mejor equidad.

3.7. *Total Information coefficient (TIC)*

Hasta ahora hemos presentado resultados sobre estimadores del coeficiente de información maximal de la población, una cantidad para la cual la equitabilidad es la principal motivación. Ahora introducimos y analizamos una nueva medida de dependencia, el coeficiente de información total (TIC). A diferencia del coeficiente de información maxima(MIC), el coeficiente de información total no está diseñado para la equitabilidad, sino más bien como una estadística de prueba para probar una hipótesis nula de independencia.

Comenzamos dando alguna intuición. Recordemos que el coeficiente de información maximal es el supremo de la matriz característica. Si bien estimar el

supremo de esta matriz tiene muchas ventajas, esta estimación implica tomar un máximo sobre muchas estimaciones de las entradas individuales de la matriz característica. Dado que los máximos de conjuntos de variables aleatorias tienden a aumentar a medida que crece el número de variables, se puede imaginar que este procedimiento puede llevar a un sesgo positivo indeseable en el caso de la independencia estadística, cuando la matriz característica de la población es igual a 0. Esto podría ser perjudicial para las pruebas de independencia, donde la distribución muestral de una estadística bajo una hipótesis nula de independencia es crucial.

La intuición detrás del coeficiente de información total es que si en cambio consideramos una propiedad más estable, como la suma de las entradas en la matriz característica, podríamos esperar obtener una estadística con un sesgo más pequeño en el caso de independencia y, por lo tanto, una mejor potencia. En resumen, si nuestro único objetivo es distinguir cualquier dependencia de ruido completo, entonces ignorar toda la matriz característica de la muestra, excepto su valor máximo, puede descartar una señal útil, y el coeficiente de información total evita esto al sumar todas las entradas.

Cabe destacar que en Reshef et al. (2011)[11] se sugiere que otras propiedades de la matriz característica pueden permitirnos medir otros aspectos de una relación dada además de su fuerza, y se definieron varias de estas propiedades. El coeficiente de información total encaja dentro de este marco conceptual. En la siguiente sección definimos el coeficiente de información total en el caso de la matriz característica (TIC) y la matriz equicaracterística (TIC_e). Luego demostramos que tanto TIC como TICe producen pruebas de independencia que son consistentes frente a todas las alternativas dependientes. (Al igual que en el caso de MIC y MIC_e, TIC_e es más fácil de calcular que TIC). Finalmente, revisaremos un estudio de simulación sobre la potencia de las pruebas de independencia basadas en TICe realizado en el paper de Reshef (2016)[12] en un conjunto de relaciones elegidas en Simon y Tibshirani (2012)[13], mostrando que TIC_e supera a otras medidas comunes de dependencia en muchas de las relaciones y se ajusta estrechamente a su rendimiento en el resto.

3.7.1. Definición y Consistenc/'ia de TIC

Comenzamos definiendo dos versiones del coeficiente. En la siguiente definición notemos qué $\widehat{M}$ denota la matriz caracteristica poblacional y $\widehat{[M]}$ denota la matriz equicaracteristica de poblacional.

Definición 3.14. Sea $D \subset \mathbb{R}^2$ un conjunto de n pares ordenados, y sea $B : \mathbb{Z}^+ \to \mathbb{Z}^+$. Definimos:

$$TIC_B(D) = \sum_{k\ell \leq B(n)} \widehat{M}(D)_{k,\ell},$$

y

$$TIC_{e,B}(D) = \sum_{k\ell \leq B(n)} \widehat{[M]}(D)_{k,\ell}.$$

Ahora nos interesa mostrar que estos estadísticos nos entregan test de independencia concistente, para esto debemos detenernos y analizar el comportamiento de las cantidades poblacionales análogas.

Definición 3.15. Para una matriz A y un número positivo B, la B-parcial suma de A, denotada por $S_B(A)$, es:

$$S_B(A) = \sum_{k\ell \leq B} A_{k,\ell}.$$

Cuando A es una matriz (equi)caracteristica, $S_B(A)$ es la suma sobre todas las entradas correpondientes a mallas con al menos B celdas totales. Por tanto, si $\widehat{M}(D)$ es una matriz equicaracteristica poblacional de $D, S_B(\widehat{M}(D)) = \mathrm{TIC}_B(D)$, y lo mismo se mantiene cierto para $S_B(\widehat{[M]}(D))$ and $\mathrm{TIC}_{e,B}(D)$.

Es claro que si X e Y son variables aleatorias estadísticamente independientes, se tiene que ambas matrices caracteristicas $M(X,Y)$ y la matriz equicaracteristica $[M](X,Y)$ son identicamente 0 , tal que $S_B(M(X,Y)) = S_B([M](X,Y)) = 0$ para todo B. Sin embargo, también nos interesa como estas cantidades se comportan cuando las variables X e Y son dependietes. Las siguientes proposiciones nos ayudan a entender esto. La primera nos muestra una cota inferior para los valores de las entradas de ambas $M(X,Y)$ y $[M](X,Y)$. La segunda nos muestra una caracterización asintótica de como crecen $S_B(M)$ y $S_B([M])$ como funciones de B. Estas dos proposiciones son el corazón técnico de por qué el coeficiente de información total produce un test de independencia consistente.

Proposición 3.3. Sea (X,Y) un vector aleatorio bivariado. Si X e Y son estadísticamente independientes, entonces $M(X,Y) \equiv [M](X,Y) \equiv 0$. Si no, existe algún $a > 0$ y algún entero $\ell_0 \geq 2$ tal que:

$$M(X,Y)_{k,\ell}, [M](X,Y)_{k,\ell} \geq \frac{a}{\log \min\{k,\ell\}},$$

ya sea para todo $k \geq \ell \geq \ell_0$, o para todo $\ell \geq k \geq \ell_0$.

Demostración. Apéndice K.1. Reshef (2016) [12] □

Proposición 3.4. Sean (X,Y) un vector aleatorio bivariado. Si X e Y independites, se tiene que $S_B(M(X,Y)) = S_B([M](X,Y)) = 0$ para todo $B > 0$. Si no, tenemos que $S_B(M(X,Y))$ y $S_B([M](X,Y))$ son ambos $\Omega(B \log\log B)$.

Demostración. Apéndice K.2. Reshef (2016) [12] □

Con las proposiciones presentadas, y siguiendo la misma lógica de los argumentos de convergencia presentados anteriormente, podemos mostrar el resultado principal de esta sección, que las estadísticas TIC y TIC_e producen tests de independencia consistentes.

Teorema 3.12. *Los estadisticos TIC_B y $TIC_{e,B}$ proporcionan pruebas coherentes de cola derecha de independencia, siempre que $\omega(1) < B(n) \leq O\left(n^{1-\varepsilon}\right)$ para algún $\varepsilon > 0$.*

Demostración. Apéndice K.3. Reshef (2016) [12] □

En la práctica, usualemente se utiliza el algoritmo EQUICHARCLUMP [12, Sección 4.3] para computar la matriz equicaracterística, de donde calculamos TIC_e. Este algoritmo no computa la matriz equicaracteristica exactamente. Pero, como es el caso con MIC_e, el uso del algoritmo no afecta las propiedades teóricas de la estadística. Esta demostración se puede encontrar en Reshef el al. (2016) [12, Apéndice H]

3.7.2. Prueba de Poder de independencia basadas en TIC_e

Ya sabemos que ambos, TIC y TIC_e, son consitentes, ahora nos interesa realizar una evalución empírica de la prueba de poder de independencia basado en TIC_e siendo computado usando al algoritmo *EQUICHARCLUMP*.

En Reshef el at. (2016)[12], para evaluar el poder que tiene esta prueba, se realiza el analisis utilizado por Simon y Tibshirani (2012) [13]. Este corresponde a un conjunto de relaciones definido por:

$$\mathcal{Q} = \left\{(X, f(X) + \varepsilon') : X \sim \text{ Unif}, f \in F, \varepsilon' \sim \mathcal{N}\left(0, \sigma^2\right), \sigma \in \mathbb{R}_{\geq 0}\right\},$$

donde F es el conjunto de relaciones definido en Simon y Tibshirani (2012) [13]. Estas corresponden a relacion: Lineal, Cuadrática, Cúbica, Seno (Perido $\frac{1}{8}$), Seno (Periodo $\frac{1}{2}$), $X^{\frac{1}{4}}$, Circular (tratado como dos semicirculos), y Función escalera. Detalles de la métodología que fue utilizada pueden ser encontrados en la seccón 5.2 de Reshef el at. (2016)[12].

Los resultados del análisis se presentan en la Figura 3.4. En primer lugar, la figura muestra que TICe se compara de manera bastante favorable con la correlación de distancia, un método considerado tener un alto poder (Simon y Tibshirani, (2012)[13]). Específicamente, TIC_e supera de manera uniforme a la correlación de distancia en 5 de los 8 tipos de relaciones examinados y se desempeña de manera comparable en los otros tres tipos de relaciones. Cabe destacar que la correlación de distancia tiene muchas ventajas sobre TIC_e, incluyendo el hecho de que se generaliza fácilmente a relaciones de mayor dimensionalidad y viene con un marco teórico elegante y completo.

El análisis también muestra que TICe supera en gran medida al coeficiente de información maximal (MIC) original, y también supera a MIC_e, respaldando la intuición de que la suma realizada por el primero puede, de hecho, conducir a ganancias sustanciales en poder contra la independencia en comparación con la maximización realizada por el último.

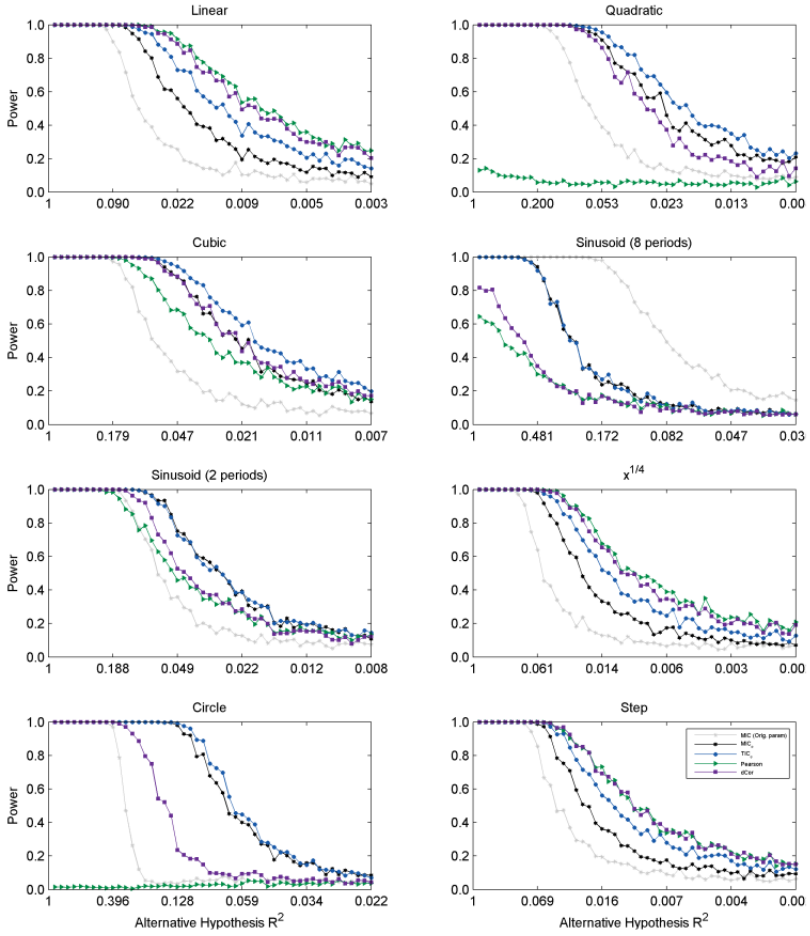

Figura 3.4: Comparación del poder de prueba de independencia basado en TIC_e (azul) con MIC con parámetros predeterminados (gris), MIC_e con los mismos parámetros que TIC_e (negro), correlación de distancia (púrpura) y el coeficiente de correlación de Pearson (verde) en varios tipos de relaciones de hipótesis alternativas elegidos por Simon y Tibshirani (2012 [13]).

3.7.3. Ejemplos

Ahora, con una definición clara de como calcular los coeficientes, podemos procesder a un ejemplo de su uso. Para esto usaremos "The Datasaurus Dozon", un conjutno de datos propuesto por Justin Matejka y George Fitzmaurice (2017) [10]. Este conjunto de datos nos presenta con 13 realciones, visibles en la Figura 3.5, las cuales todas poseen los mismos valores para estadísticos descriptivos comunos (Promedio Marginal, Desviacón Estandar Marginal, y Correlación de Pearson), pero son claramente visualmente distinos.

Figura 3.5: The Datasaurus Dozon", a pesar de compartir los mismos estadísticos descriptivos, son visualmente distintos.

De particular interes para nosotros es el valor de la correlación de Pearson de, esta es básicamente cero, lo que nos "deberíaïndicar que las relaciones son independientes. Sin embargo, como podemos ver en la Figura 3.5, esto no es cierto. Es por esto que estos datos nos son de gran ayuda.

Para este ejemplo, tomamos los vectores marginales de cada una de las relaciones, y los colocamos todos en un solo conjunto de datos. Luego de esto utilizamos TIC_e para encontrar los pares que sean más independientes, y finalmente usamos MIC_e para cuantificar la relacion entre estos pares. Los resultados de este analisis se pueden ver en la Figura

3.7.4. Ejemplos

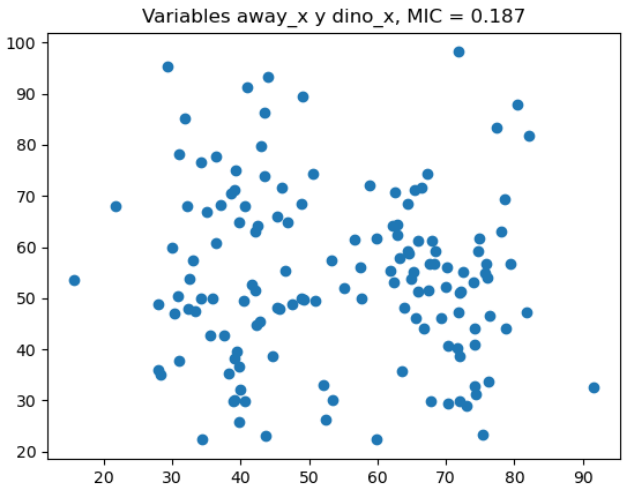

Figura 3.6: MIC = 0.187

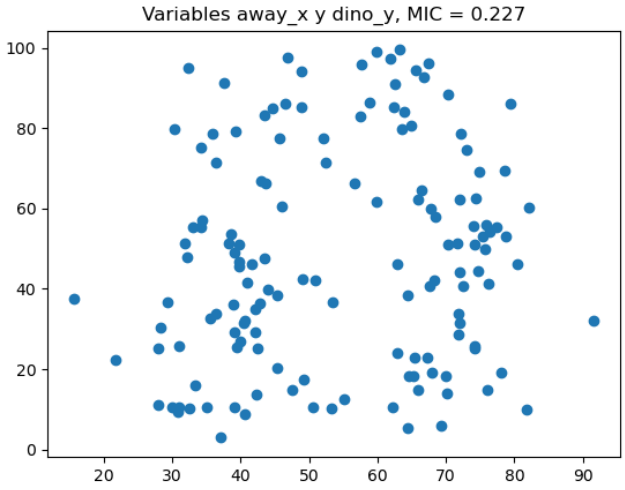

Figura 3.7: MIC = 0.227

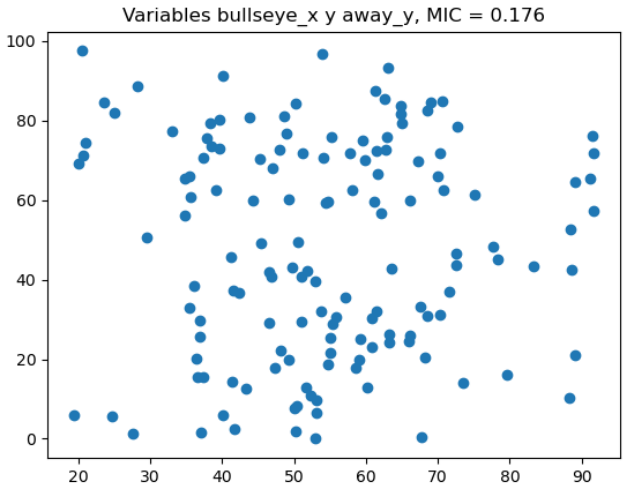

Figura 3.8: MIC = 0.176

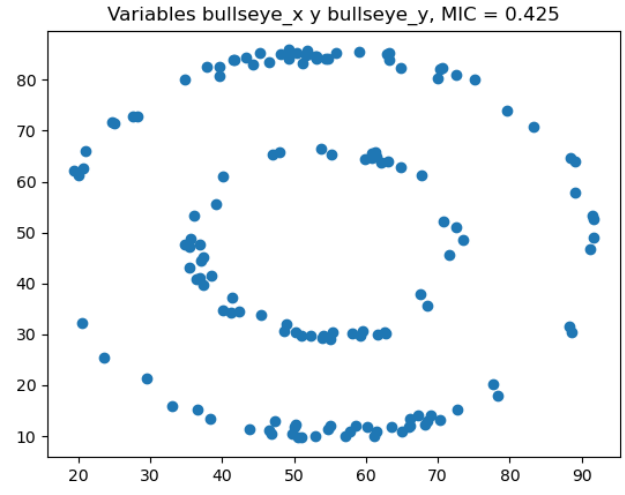

Figura 3.9: MIC = 0.425

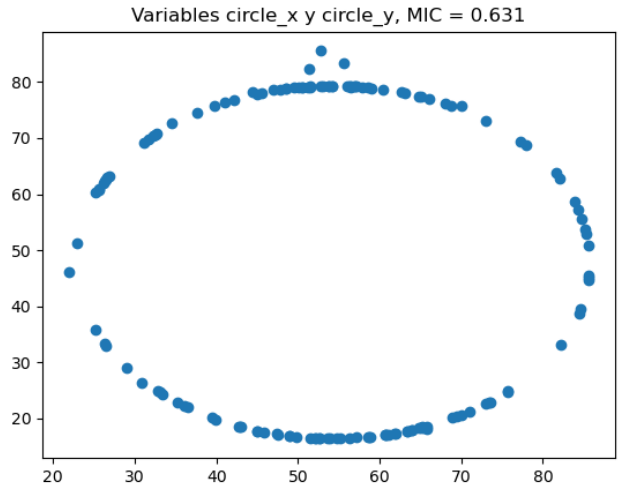

Figura 3.10: MIC = 0.631

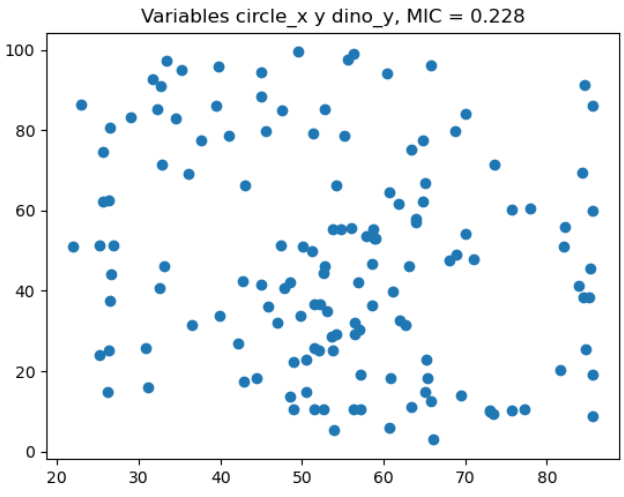

Figura 3.11: MIC = 0.228

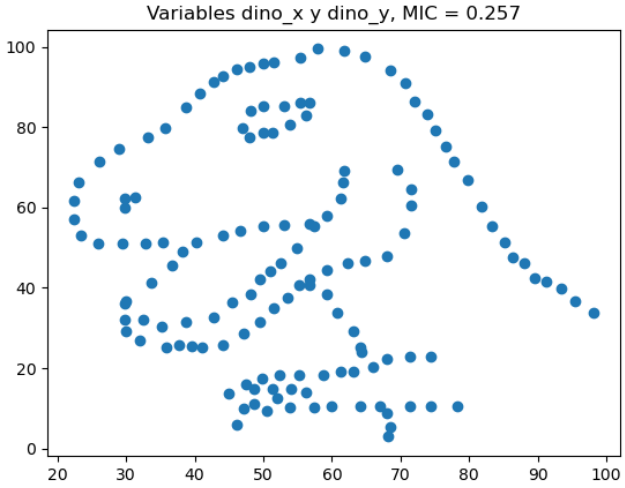

Figura 3.12: MIC = 0.257

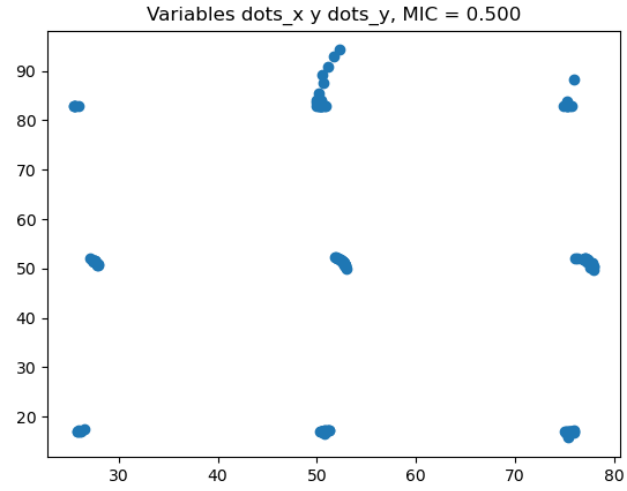

Figura 3.13: MIC = 0.500

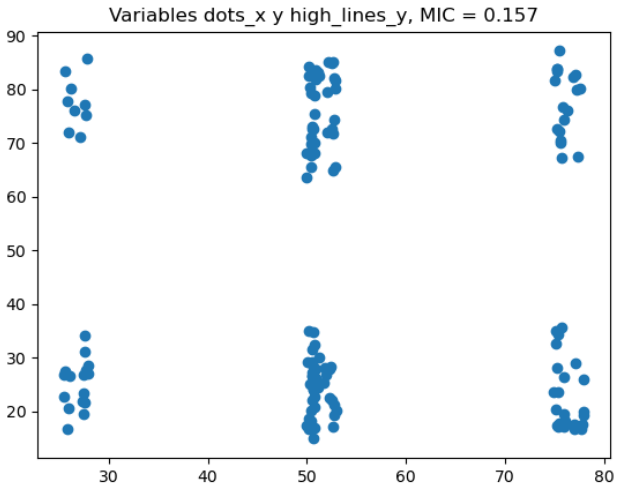

Figura 3.14: MIC = 0.157

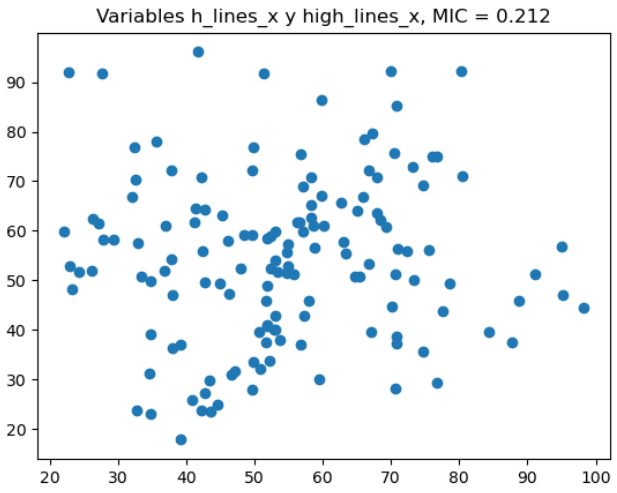

Figura 3.15: MIC = 0.212

Capítulo 4

Otros Métodos de Comparación de Imágenes

4.1. Introducción

4.2. Correlación local

4.2.1. Discucion sobre el coef.

donde se publico, como su ocupa, despcion en palabras

La correlación local, tambíen conodica como coeficiente no paramétrico de Chen, o coeficiente de Chen. Este, sin realizar supuestos sobre distribuciones, detecta relaciones no lineales al invenstigar un montón de correalciones locales.

4.2.2. Definiciones

La definición del método está basada en en el concepto de integrales de correlación, las cuales se definen de la siguiente forma:

Definición 4.1.

$$I(r) = \lim_{N\to\infty} \left\{ \frac{1}{N^2} \sum_{i,j=1}^{N} I\left(|z_i - z_j| < r\right) \right\}$$

La integral de correlación cuantifica el el número promedio de vecinos dentro de un radio r. Notemos que esta definición sigue teniendo sentido cuando los datos no son series de tiempo.

Para desarrollar una medida de asociación entre vectores, x e y, modificamos la definición de $I(r)$ como sigue. Sean $z_i = (x_i, y_i)$ con $i = 1, \ldots, N$ las observa-

ciones en el conjunto de datos. Sea $|z_i - z_j|$ la distancia euclidiana. Definimos $\hat{I}(r) = \frac{1}{N^2} \sum_{i,j=1}^{N} I\left(|z_i - z_j| < r\right)$. Las distancias obsevadas son además linealmente transformadas para que se encuentren entres 0 y 1 antes de calcular $\hat{I}$. Notemos que $\hat{I}$a tiene las propiedades de una función de distribución acumulativa. Es no decreciente entre 0 y 1 y continua por la derecha. La función $\hat{I}(r)$ descrive el patrón global de distancias entre vecinos.

Nuestro interés principal es la definición de una metrica para cuantificar la asosiación no lineal estudiando patrones locale. Dado esto, definimos la densidad de vecinos D de forma similar a la derivada de $\hat{I}$:

$$\hat{D}(r) = \frac{\triangle\ \hat{I}(r)}{\triangle\ r}$$

Donde $\triangle\ \hat{I}(r)$ denota un cambio en $\hat{I}(r)$. La densidad de vecinos es evaluada en radio distreto r, con $r = 0, 1/m, 2/m, \ldots, 1$ y m es un grosor de malla arbitrario. Una función de suavizado automático usando validación cruzada es usada para elegir un óptimo el tamaño m (Vilela et al. 2007) y se aplica para suavizar $D(r)$. En el paper, el tamaño predeterminado m se establece como N, el número de observaciones y en este trabajo usaremos el mismo m. El estadístico $\hat{D}$ es una aproximación discreta de $d\hat{I}(r)/dr$, la cual tiene las propiedades formales de una probabilidad función de densidad. Por lo tanto, con un ligero abuso de terminología nos referimos a $\widehat{D}(r)$ como una distribución.

En base a esto definimos la correlación local. Intuitivamente, las distancias entre los puntos de datos entre dos variables correlacionadas diferirían de las distancias entre dos variables no correlacionadas. Sea $\widehat{D_0}(r)$ la estimación de una distribución nula, que se compone de dos vectores sin asociación. Definimos la correlación local ($\ell(r)$) como la desviación de D de la de la distribución nula a una distancia vecina dada r:

$$\ell(r) = \widehat{D}(r) - \widehat{D}_0(r)$$

Este enfoque no asume ninguna distribución paramétrica. La flexibilidad de este método facilita el cambio de la distribución nula a cualquier distribución de interés.

Por ultimo, definimos el coficiente como de correlación local máxima, o coeficiente de Chen como:

$$M = \max_r\{|\ell(r)|\}$$

La interpretación de $\ell(r)$ como la diferencia de dos distribuciones implica que M puede interpretarse como la distancia bajo la norma del supremo entre $\widehat{D}$ y $\widehat{D_0}$. En otras palabras, definimos el estadístico M como la desviación máxima entre dos densidades vecinas subyacentes.

4.3. *Distance Covariance & Distance Correlation*

4.4. Correlación de Pearson

4.4.1. Discucion sobre el coef.

donde se publico, como su ocupa, despcion en palabras

4.5. Definiciones

El coef. se define como:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{4.1}$$

Para una muestra de tamaño N, tenemos:

$$r = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2}\sqrt{\sum_i^n (y_i - \bar{y})^2}} \tag{4.2}$$

Con x_i, y_i elementos de la muestra y $\bar{x}, \bar{y}$ sus respectivos promedios.

Hablar de The Ineffectiveness of the Correlation Coefficient for Image Comparisons

Capítulo 5

Equitabilidad

5.1. Introducción

La equitabilidad ha sido descrita de manera informal por Reshef et al. como la capacidad de una estadística para .ªsignar puntuaciones similares a relaciones igualmente ruidosas de diferentes tipos"[11]. Aunque útil, esta definición informal es imprecisa en el sentido de que no especifica lo que se entiende por ruidoso.º "similar", y no especifica para qué relaciones debe cumplirse la propiedad mencionada.

En su trabajo posterior "Equitability, interval estimation, and statistical power", Reshef et al. (2015) [15], se formalizó la idea de Equitabilidad a travez del concepto del intervalo interpretativo, que funciona como estimación en intervalos de la cantidad de ruido presente en una relación de tipo desconocida.

En el contexto de este trabajo, nos interesa que las medidas de dependencia sean equitativas puesto que esto nos asegura que la medida de dependencia no está sesgada hacia un tipo de relación en particular, y por lo tanto, nos será útil para comprar comparar las distintas versión de la transformacíon Box-Cox que definiremos en la sección 6.

En esta sección se presentará la definición formal de equitabilidad, y de equitabilidad con respecto a R^2, recrearemos el experimento realizado por Reshef et al. (2016) [12] sobre la equitabilidad del MIC_e, y junto con esto se estudiará la equitabilidad de las otras medidas propuestas en la Sección 4.

5.1.1. Definiciones.

5.1.2. Equitabilidad del MIC_e

Como se mencionó previamente, una de las principales motivaciones para la introducción de MIC fue la equidad, es decir, hasta qué punto una medida de

dependencia captura útilmente alguna noción de la fuerza de una relación en un conjunto de relaciones estándar. En esto contexto en Reshef et al. (2016) [12] se realizõ un análisis empírico de la equidad de MIC_e con respecto a R2 y su desempeño fue comparado con la correlación de distancia (Székely et al., (2007)[6]; Székely and Rizzo, (2009)[7]), la estimación de la información mutua (Kraskov et al., 2004) y la estimación de la correlación máxima (Breiman and Friedman, 1985).

Se comenzó evaluando la equidad en el conjunto de relaciones Q definido anteriormente, un conjunto que ha sido analizado en otros trabajos previos (Reshef et al., 2011, 2015a; Kinney and Atwal, 2014). Los resultados, mostrados en la Figura 5.1, confirman la superior equidad del estimador MIC_e en este conjunto de relaciones.

Para evaluar la equidad de manera más objetiva sin depender de un conjunto de funciones curado manualmente, se analizaron 160 funciones aleatorias extraídas de una distribución de proceso Gaussiano con un kernel de función radial con una de ocho posibles anchuras en el conjunto $\{0{,}01, 0{,}025, 0{,}05, 0{,}1, 0{,}2, 0{,}25, 0{,}5, 1\}$ para representar una variedad de complejidades de relaciones posibles. Los resultados, mostrados en la Figura 5.2, muestran que MIC_e supera a los métodos existentes en términos de equidad con respecto a R^2 en estas funciones también. También se examinó el efecto de las relaciones atípicas en los resultados al muestrear repetidamente subconjuntos aleatorios de 20 funciones de este gran conjunto de relaciones y medir la equidad de cada método en promedio sobre los subconjuntos; los resultados fueron similares.

Una característica del desempeño de MIC_e en estas relaciones elegidas al azar, como se muestra en la Figura 5.2, es que parece ser mínimamente sensible a la anchura del proceso Gaussiano del cual se extrae una relación dada. Esto contrasta, por ejemplo, con la estimación de la información mutua, que muestra una sensibilidad pronunciada a este parámetro que le impide ser altamente equitativa cuando hay relaciones con diferentes anchuras en el mismo conjunto de datos.

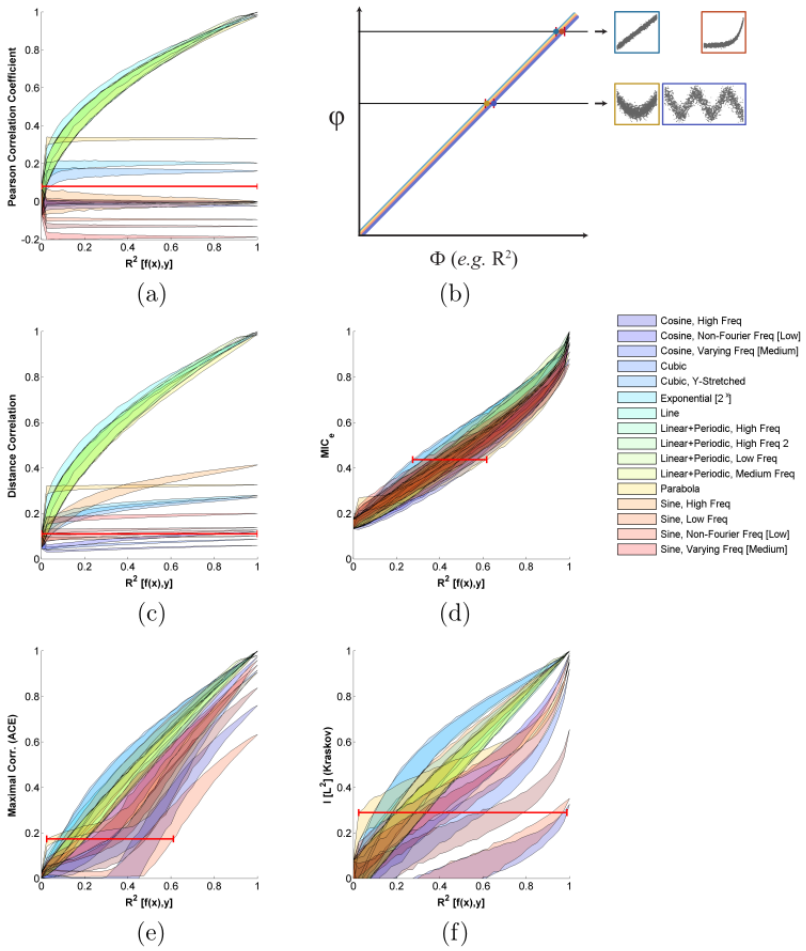

Figura 5.1: Equitabilidad con respecto a R^2 en un conjunto de relaciones funcionales ruidosas de (a) el coeficiente de correlaciõn de Pearson, (b) una medida hipotẽtica de dependencia φ con equitabilidad perfecta, (c) la correlaciõn de distancia, (d) MIC_e, (e) estimaciõn de correlaciõn mãxima y (f) estimaciõn de informaciõn mutua. Para cada relaciõn, una regiõn sombreada denota los valores estimados en el percentil 5 y 95 de la distribuciõn muestral de la estadĩstica en cuestiõn en esa relaciõn en cada valor de R^2. El grãfico resultante muestra quẽ valores de R^2 corresponden a un valor dado de cada estadĩstica. El intervalo rojo en cada grãfico indica el rango mãs amplio de valores de R^2 que corresponden a un valor de la estadĩstica; cuanto mãs estrecho sea el intervalo rojo, mayor serã la equitabilidad. Un intervalo rojo con ancho 0, como en (b), significa que la estadĩstica refleja solo R^2 sin depender del tipo de relaciõn, como se demuestra en los pares de miniaturas de relaciones de diferentes tipos con valores idẽnticos de R^2.

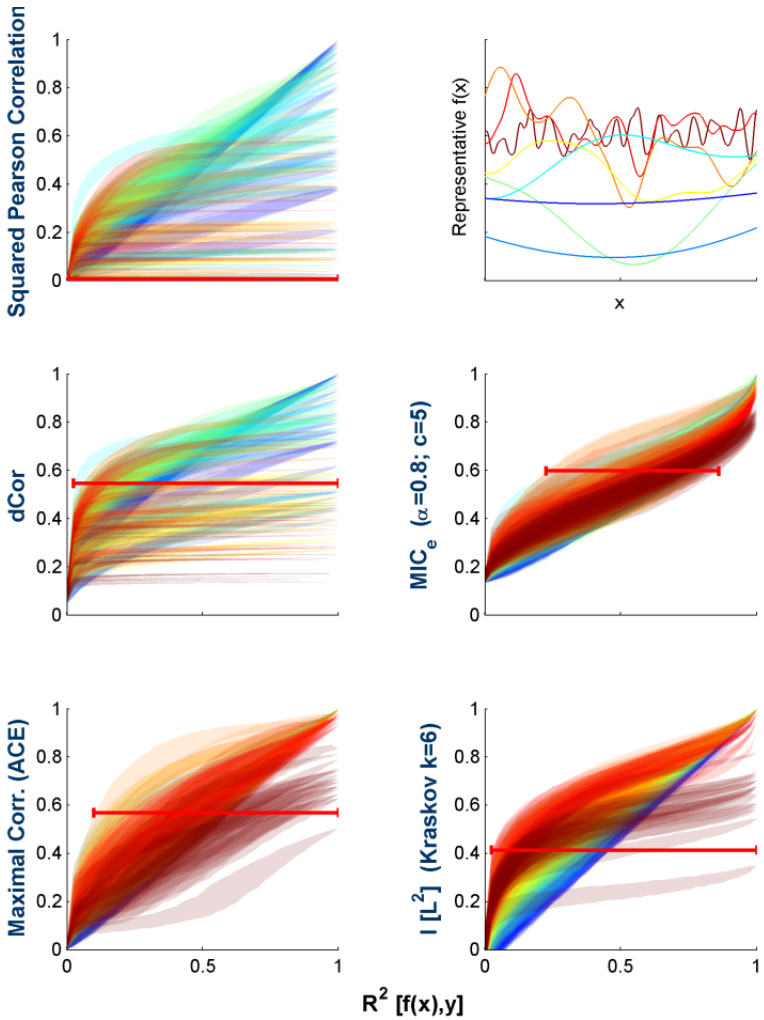

Figura 5.2: Equitabilidad de los mẽtodos examinados en funciones extraĩdas al azar de una distribuciõn de proceso gaussiano. Cada mẽtodo se evalũa como se muestra en la Figura 5.1, con un intervalo rojo que indica el rango mãs amplio de valores de R^2 correspondiente a cualquier valor de la estadĩstica; cuanto mãs estrecho sea el intervalo rojo, mayor serã la equitabilidad. Cada regiõn sombreada corresponde a una relaciõn, y las regiones estãn coloreadas segũn el ancho de banda del proceso gaussiano del que se muestrearon. Las relaciones de muestra para cada ancho de banda se muestran en la esquina superior derecha con colores correspondientes.

Capítulo 6

La transformación de Box y Cox

6.1. Introducción

La transformación de Box y Cox, conocida como Box-Cox, es una técnica de transformación no lineal que fue propuesta por George Box y David Cox en 1964 en su trabajo *"An Analysis of Transformations"*[2]. Cuenta la historia que el Profesor cox estaba visitando al doctor Box en Wisconsin, y decidieron que deberían escribir un artículo juntos dada la similitud de sus nombres, y que ambos eran británicos [8]. Muchos importantes resultados y técnicas en el análisis estadístico de datos toman el supuesto de que los datos poseen una distribucón normal, en los casos cuando este supuesto no se sostiene, una de las alternativas es transformar los datos para que se acerquen a una distribución normal. En este contexto la transformación de Box-Cox fue propuesta para convertir un conjunto de datos en una distribución que se asemeja a la normal, dejando una distribución con menos sesgo que es un poco más simétrica, esto suele ser determinado en base a un test de máxima verosimilitud, más adeltante discutiremos el motivo de esto. La transformación de Box-Cox pertenece una familia de técnicas conocidas como transformaciones de potencia. Estas transformaciones buscan modificar los datos de entrada elevándolos a una potencia determinada, identificada por el parámetro λ.

Box y Cox desarrollaron este método con la intención de crear una técnica de transformación flexible que pudiera adaptarse a diversas distribuciones de datos, esto permite adaptar el coeficiente para funcionar en distintos contextos, y de nuestro interes particular es en el contexto de imágenes. En general la transformación solo es utilizada sobre vectones 1-dimensional. En 2020 Abbas Cheddad público *"On Box-Cox Transformation for Image Normality and Pattern Classification"*[3], donde se discutío el coeficiente como un paso de preprocesamiento de imágenes, tanto para mejoramiento visual, como para mejorar el desempeño

de algoritmos de clasificación. En este trabajo se propuso una nueva forma de aplicar la transformación, que consiste en utilizar el histograma de la imagen como proxy comprimido de la matriz de datos, y así poder aplicar la transformación de forma rápida.

En este capítulo vamos a discutir la transformación de Box-Cox, presentaremos su definición, y discutiremos el motivo de su uso. Luego vamos a discutir el trabajo de Cheddad[3], y como este puede ser aplicado sobre imagenes. Finalmente discutiremos alternativas para calcular λ sobre imágenes.

6.2. Definiciones

Para un $\lambda \in \mathrm{R}$ dado, la transformación de Box-Cox se define como:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & (\lambda \neq 0) \\ \log y & (\lambda = 0) \end{cases} \tag{6.1}$$

$\forall y \in \mathrm{R}_{>0}$. En la práctica los valores de λ se restringen a un intervalo, normalmente $[-2, 2]$ o $[-5, 5]$, notemos además que en la practica se toma la segunda forma cuando $|\lambda| < 0{,}01$[3].

Además existe una versión para datos no positivos dada por:

$$y^{(\lambda)} = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & (\lambda_1 \neq 0)\,, \\ \log\left(y+\lambda_2\right) & (\lambda_1 = 0)\,. \end{cases}$$

Esta versión es menos utilizada en la práctica, dado que se suelen realizar otros pasos de preprocesamiento que dejan los datos entre 0 y 1.

Cabe notar que Bicego y Baldo (2016)[1] demostraron que, dado un vector $\mathbf{y} = (y_1, \ldots, y_n) \in \mathrm{R}^n_{>0}$, la transformación no cambia dado el orden de los elementos en el vector, por lo tanto al momento de aplicar la transformación sobre una imagen, o en general una matriz d-dimensional, se puede aplicar la cualquier ordenamiento sobre los datos, y luego aplicar la transformación sobre el vector unidimensional resultante. Dado esto también cabe notar que al ser agnóstica con respecto al orden de los datos, la transformación no altera la relación espacial entre los datos.

Este no es el caso para la definición de lambda, que es lo que discutiremos en la siguiente sección.

6.3. Elguiendo λ

Como mencionamos anteriormente, el objetivo de la transformación es encontrar el valor de λ que proporciona el mejor ajuste a una distribución normal. Para esto, Box y Cox proponen un criterio de máxima verosimilitud, el cual se define como:

$$\mathcal{L}(\lambda) \equiv -\frac{n}{2}\log\left[\frac{1}{n}\sum_{j=1}^{n}\left(x_j^{\lambda}-\overline{x^{\lambda}}\right)^2\right]+(\lambda-1)\sum_{j=1}^{n}\log x_j \tag{6.2}$$

donde $\overline{x^{\lambda}}$ es el promedio muestral del vector transformado.

La verosimilitud juega un papel crucial en el proceso de transformación de Box-Cox. En términos simples, la verosimilitud se refiere a la probabilidad de que un conjunto de datos observados se derive de una distribución estadística particular. En este caso, la verosimilitud se utiliza para medir qué tan bien una distribución normal se ajusta a los datos transformados para diferentes valores de λ. El valor de λ que maximiza esta verosimilitud es el que se selecciona para la transformación.

La transformación de Box-Cox persigue un objetivo esencial en el análisis estadístico: garantizar el cumplimiento de los supuestos necesarios para la aplicación de modelos lineales. Esta garantía posibilita el uso de técnicas de análisis de varianza estándar en los datos transformados. En este sentido, Bicego y Baldó [1] resaltan que esta transformación no altera el ordenamiento de los datos, manteniendo intacta la relación inherente entre ellos.

Es importante aclarar, sin embargo, que no todos los conjuntos de datos pueden ser transformados de tal manera que resulten en una distribución normal perfecta. A pesar de esta limitación, Draper y Cox [5]argumentan que la transformación de potencia puede ser efectiva en muchos casos. Incluso en situaciones donde la transformación no logra una normalidad exacta, las estimaciones habituales del parámetro λ pueden desempeñar un papel vital en la regularización de los datos.

Este proceso de regularización conduce a una distribución que cumple con ciertos criterios deseables, como la simetría o la homocedasticidad. Esta última característica, que se refiere a la constancia de la varianza a lo largo del conjunto de datos, es especialmente útil en campos como el reconocimiento de patrones y el aprendizaje automático. Por ejemplo, en el análisis discriminante lineal de Fisher, la homocedasticidad facilita la diferenciación entre diferentes clases de datos, potenciando la eficacia de este tipo de técnicas de aprendizaje automático.

6.4. Box-Cox sobre imagenes

En su artículo del 2020 [3], Abbas Cheddad resalta una notable brecha en la aplicación de la transformación de Box-Cox a imágenes digitales. Según Cheddad, existe una carencia significativa de estudios en este ámbito, destacando el trabajo de JD Lee en 2009 como una excepción[9]. En el estudio de Lee, se presentó un método de segmentación para imágenes de resonancia magnética cerebral a través de una técnica de transformación de distribución. En este enfoque, la transformación de Box-Cox se aplicó a las imágenes de resonancia magnética cerebral para normalizar la distribución de intensidad de los píxeles.

Es relevante señalar que, en este estudio, las imágenes se trataron como un vector de datos en lugar de una matriz, lo que implica un enfoque unidimensional en la manipulación y análisis de la imagen.

Cabe destacar que el proceso de aplicar la tranformación es iterativo, en el cual se ha de buscar un parametro λ, esto hace que aplicar esta la transformación en grandes bancos de imagenes sea demoroso. Una alternativa propuesta por A. Cheddad [3] es utilizar el histograma como proxy comprido de la matriz de datos, dado que este refleja la probabilidad estimada de que un pixel esa de un tono en particular. En lo que continua de la sección discutiremos este método.

Dada una imagen en el espacio de color RGB, definimos:

$$\mathcal{F}(u,v) = \{R(u,v), G(u,v), B(u,v)\}$$

donde (u,v) son las coordenadas en el espacio de pixeles que cumplen $u = 1,\dots U$, $v = 1,\dots V$ y (U,V) son las dos dimensiones de la foto. Notemos que cada elemento de la imagen es vector de tres dimensiones con los canales rojo, verde, y azul, pero en la literatura se suele trabajar con imagenes en escala de grises, para esto definimos:

$$\mathcal{F}' = (0{,}299\mathrm{R} + 0{,}587\mathrm{G} + 0{,}114\ \mathrm{B})$$

Que corresponde al canal de escala de grises como está definido por el espacio de color $\mathrm{YC_bC_r}$ lo calcula. En base a esto definimos la función de probabilidad de imágen, i.e., el histograma como:

$$\chi(i) = \sum_{i=0}^{255} \mathcal{F}'_i,$$

donde i es el nivel de gris.

Ahora, denotemos por $\hat{\lambda}_\chi$ al parametro de la transformación Box-Cox seleccionado usando el histograma, y de forma analoga definamos $\lambda_{\mathcal{F}'}$ al seleccionado usando los datos completos. Fue obserado por Cheddad que estos no coinciden (de hecho la correlación entre estos es $r^2 = -0{,}3022$) pero aun así este calculo se ha desmostrado util en problemas de clasificación.

Ahora definimos $\mathcal{F}'(u,v)^{\lambda_\chi}$ como los datos siendo aplicada la transformación Box-Cox definida en (6.1), y por ultimo vamos a definir Box-Cox para imagenes o BCI como:

$$BCI = \frac{(\mathcal{F}''(u,v) - \text{mín}\,(\mathcal{F}''(u,v)))}{\left(\text{máx}\left(\mathcal{F}''^{(u,v)}\right) - \text{mín}\left(\mathcal{F}''^{(u,v)}\right)\right)}$$
$$\text{con } \mathcal{F}''(u,v) = \mathcal{F}'(u,v)^{\hat{\lambda}_\chi}$$

Notemos que este ultimo paso se realiza para que los datos esten entre 0 y 1, y así poder ser representados en una imagen. En la Figura se puede ver un ejemplo de la transformación aplicada sobre una imagen, con ambas versiones de λ.

6.5. Propuestas de λ para imagenes.

Capítulo 7

Comprarando una Imgaen con su tranformada

Capítulo 8

Consideraciones Finales

Apéndice A

Demostraciones

En este apéndice se derivan algunos resultados preliminares asociados con la matrices de derivadas

Bibliografía

[1] Manuele Bicego y Sisto Baldo. «Properties of the Box–Cox transformation for pattern classification». En: *Neurocomputing* 218 (2016), págs. 390-400.

[2] George EP Box y David R Cox. «An analysis of transformations». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2 (1964), págs. 1518-1524.

[3] Abbas Cheddad. «On box-cox transformation for image normality and pattern classification». En: *IEEE Access* 8 (2020), págs. 154975-154983.

[4] Pardis C. Sabeti David N. Reshef Yakir A. Reshef y Michael Mitzenmacher. «An empirical study of leading measures of dependence.» En: *arXiv preprint* (2015).

[5] Norman R Draper y David R Cox. «On distributions and their transformation to normality». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 31.3 (1969), págs. 472-476.

[6] et al. Gábor J. Székely Maria L. Rizzo. «Measuring and testing dependence by correlation of distances.» En: *The Annals of Statistics* (6).35 (2012), págs. 2769-2794.

[7] Maria L. Rizzo. Gábor J. Székely. «Brownian distance covariance». En: *MThe Annals of Applied Statisticsl.* (4).3 (2012), págs. 1236-1265.

[8] David M Lane. *Introduction to Statistics*. Online Statistics Education: An Interactive Multimedia Course of Study, 2003. Cap. 16: Transformations. URL: https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Introductory_Statistics_(Lane)/16%3A_Transformations/16.04%3A_Box-Cox_Transformations.

[9] Juin-Der Lee et al. «MR image segmentation using a power transformation approach». En: *IEEE transactions on medical imaging* 28.6 (2009), págs. 894-905.

[10] Justin Matejka y George Fitzmaurice. «OSame Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.» En: (2017).

[11] David N Reshef et al. «Detecting novel associations in large data sets». En: *science* 334.6062 (2011), págs. 1518-1524.

[12] Yakir A Reshef et al. «Measuring dependence powerfully and equitably». En: *The Journal of Machine Learning Research* 17.1 (2016), págs. 7406-7468.

[13] Noah Simon y Robert Tibshirani. «Comment on "Detecting novel associations in large data sets".» En: (2012). URL: https://arxiv.org/pdf/1401.7645.pdf.

[14] Joy Thomas. Thomas Cover. *Elements of Information Theory.* New York: John Wiley & Sons, Inc, 2006. Cap. 8.

[15] Pardis C. Sabeti Yakir A. Reshef David N. Reshef y Michael Mitzenmacher. «Equitability, interval estimation, and statistical power.» En: *arXiv preprint* (2015).