

**UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA**

**DEPARTAMENTO DE MATEMÁTICA**

**Análisis de Imágenes transformadas 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 correlaciones 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 ambos solo se propone una transformación que lleve las  $d$  dimensiones a 1.

## Capítulo 2

# Análisis Estadístico y Procesamiento de Imágenes

### 2.1. Definiciones previas

asdf



## Capítulo 3

# *Maximal Information Coefficient*

La Coeficiente de Información máxima (conocido como MIC por sus siglas en inglés), es una medida estadística propuesta por Reshef et al. en su paper "Detecting Novel Associations in Large Data Sets" [7]. 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 análisis 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  $\hat{\phi}$  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  $\hat{\phi}$  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 [7]. 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) [7]. 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 robusta para encontrar relaciones no lineales entre variables, para nuestro caso en particular, entre imágenes. Como 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 sección discutiremos 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. Además introduciremos el concepto de Equitabilidad de forma más formal y veremos como el  $MIC$  cumple con esta propiedad, para posteriormente estudiar la equitabilidad de los coeficientes que introduciremos en el capítulo 4.

### 3.1. 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" [7]. 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.

En la siguiente sección estudiaremos las definiciones que nos entrega cada coeficiente.

## 3.2. 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:

$$I(X; Y) = \int_Y \int_X P_{(X,Y)}(x, y) \log \left( \frac{P_{(X,Y)}(x, y)}{P_X(x)P_Y(y)} \right) dx dy,$$

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

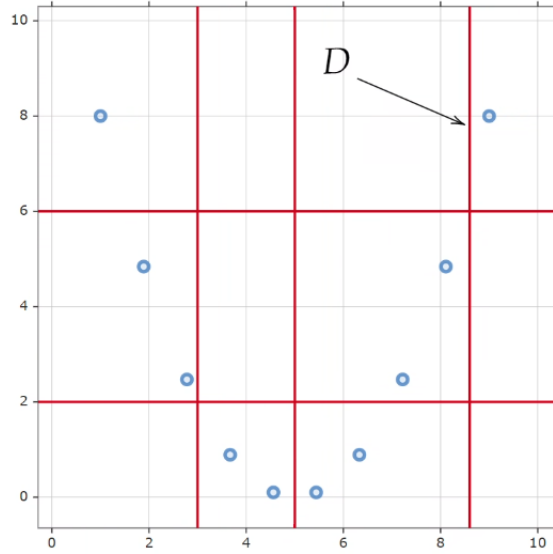


Figura 3.1: Malla  $G$  de  $4 \times 3$  sobre el conjunto de pares ordenados  $D$

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:

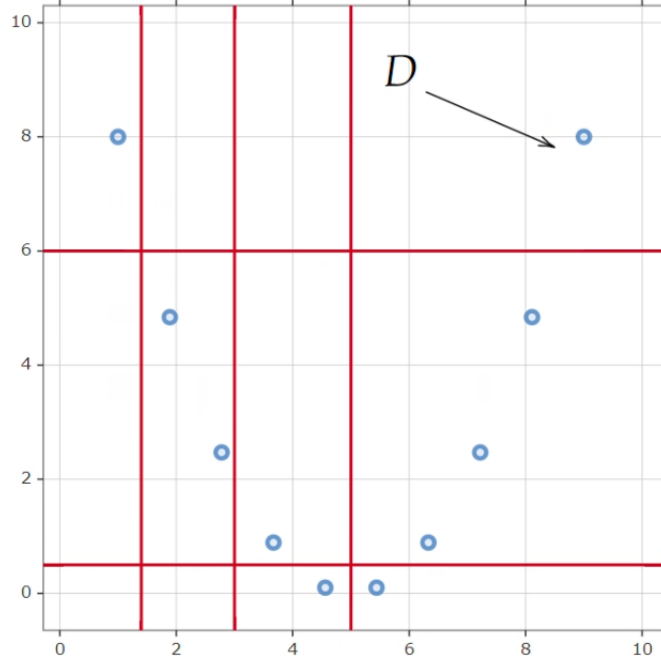


Figura 3.2: Otra malla  $G$  de  $4 \times 3$  sobre el conjunto de pares ordenados  $D$ .

Aquí podemos ver que la función de densidad que nos entrega esta 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  $D \in \mathbb{R}^2$  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 característica del conjunto  $D$ .

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

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

**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  $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$  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 funcionó en práctica en el artículo complementario de Reshef et al. (2011) [7], discuteremos la selección de este parámetro más adelante en 3.8

### 3.3. Formas prácticas de calcular el *MIC*. $MIC_*$ , $TIC_e$ y $MIC_e$

En el artículo "Measuring Dependence Powerfully and Equitably" de Reshef et al. [8], 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.

#### 3.3.1. 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:

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

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 [10, 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

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

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  $k, \ell > 1$ .

Es fácil ver lo siguiente:

**Proposición 3.1.** Sea  $(X, Y)$  un 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.

### 3.3.2. 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. [7]. Recordemos la Definiciones del  $MIC$  y la matriz característica de muestra. Notemos que para evitar confusión denotaremos como  $MIC$  al estadístico  $MIC$  y como  $MIC_*$  al coeficiente de información máxima poblacional.

**Definición 3.7.** (Reshef et al., 2011 [7]) Sea  $D \subset \mathbb{R}^2$  un conjunto de pares ordenados. La matriz característica de muestra  $\widehat{M}(D)$  de  $D$  se define por

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

**Definición 3.8.** (Reshef et al., 2011 [7]) Sea  $D \subset \mathbb{R}^2$  un conjunto de  $n$  pares ordenados, y sea  $B : \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$ . 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) [7] se sugirió que  $B(n)$  se elija como  $n^\alpha$  para alguna constante  $\alpha$  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 [8], 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  $m_\infty$  es el espacio de matrices infinitas equipadas con la norma supremo, y dada una matriz  $A$ , la proyección  $r_i$  anula todas las entradas  $A_{k,\ell}$  para las cuales  $k\ell > i$ .)

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

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

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

Dado que el supremo de una matriz es una función uniformemente continua en  $m_\infty$  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.

### 3.3.3. 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  $M_{k,\ell} = I^*((X, Y), k, \ell) / \log k \leq I^*((X, Y), k, \ell + 1) / \log k = M_{k,\ell+1}$ .  $\square$

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  $M_{k,\uparrow} = \lim_{\ell \rightarrow \infty} M_{k,\ell}$  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

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

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  $(k, \ell)$  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,  $MIC_* \leq \sup \{M_{\uparrow,\ell}\} \cup \{M_{k,\uparrow}\} = \sup \partial M$ .  $\square$

Por otro lado, el Corolario muestra que cada elemento de  $\partial M$  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  $\sup M = MIC_*$ .

### 3.3.4. 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) [7] 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) [7].

En Reshef (2016) [8], esta idea es formalizada a través de un objeto llamado la matriz equicaracterística, la cual es denominada  $[M]$ . La diferencia entre  $[M]$  y la matriz característica  $M$  es la siguiente: mientras que la entrada  $k, \ell$ -th de  $M$  se calcula a partir de la información mutua máxima alcanzable utilizando cualquier cuadrícula de  $k$ -por- $\ell$ , la entrada  $k, \ell$ -th de  $[M]$  se calcula a partir de la información mutua máxima alcanzable utilizando cualquier cuadrícula de  $k$ -por- $\ell$  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  $\sup[M] = \sup M = MIC_*$ . 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  $MIC_*$ .

## 3.4. 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.

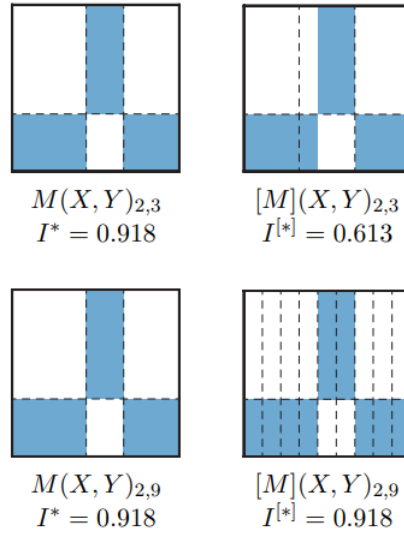


Figura 3.3: Un esquema que ilustra la diferencia entre la matriz característica  $M$  y la matriz equicaracterística  $[M]$ . (Arriba) Cuando se restringe a 2 filas y 3 columnas, la matriz característica  $M$  se calcula a partir de la rejilla óptima de 2 por 3. En contraste, la matriz equicaracterística  $[M]$  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  $Y$ .

Definir

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

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

Definir  $I^\square((X, Y), k, \ell)$  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

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

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  $\partial M$  produce el siguiente corolario como una simple consecuencia.

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

### 3.5. 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

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



**Definición 3.13.** Sea  $D \subset \mathbb{R}^2$  un conjunto de  $n$  pares ordenados, y sea  $B : \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$ . 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  $MIC_e$  es un estimador consistente de  $MIC_*$  mediante argumentos similares a los que aplicamos en el caso de  $MIC$ . (Ver Apéndice G. Reshef 2016) Específicamente, mostramos el siguiente teorema, un análogo del Teorema 6.

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

$$(f \circ r_{B(n)}) (\widehat{[M]}(D_n)) \rightarrow f([M](X, Y)),$$

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

*Demostración.* Apéndice A. Reshef (2016) [8] □

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

### 3.6. Equitabilidad del $MIC_e$

### 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 maximal ( $MIC$ ), 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)[7] 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  $TIC_e$  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  $TIC_e$  realizado en el paper de Reshef (2016)[8] en un conjunto de relaciones elegidas en Simon y Tibshirani (2012)[9], 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 Consistencia de $TIC$

Comenzamos definiendo dos versiones del coeficiente. En la definición

## 3.8. Eligiendo $B(n)$

Recordemos que, como fue propuesto en Reshef 2011 [7], utilizamos funciones de la forma  $B(n) = n^\alpha$ . Valores grandes de  $\alpha$  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  $\alpha$  llevan a un aumento en el error esperado en regímenes de alta señal ( $R^2$  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  $\alpha$  son más adecuados para detectar señales más débiles, mientras que valores más altos de  $\alpha$  son más adecuados para distinguir entre señales más fuertes. Esto concuerda con los re-

sultados observados en nuestro trabajo complementario (Reshef et al., 2015a), que muestran que valores bajos de  $\alpha$  hacen que MICE proporcione pruebas de independencia con mejor potencia, mientras que valores altos de  $\alpha$  hacen que MICE tenga una mejor equidad.

### 3.8.1. Ejemplos

Ya con la función bien definida, veamos unos ejemplos del coeficiente, primero para algunos datos, y luego entre imágenes. Comenzemos por algunos ejemplos usando

el formato de esta sección es temporal

### 3.8.2. Ejemplos

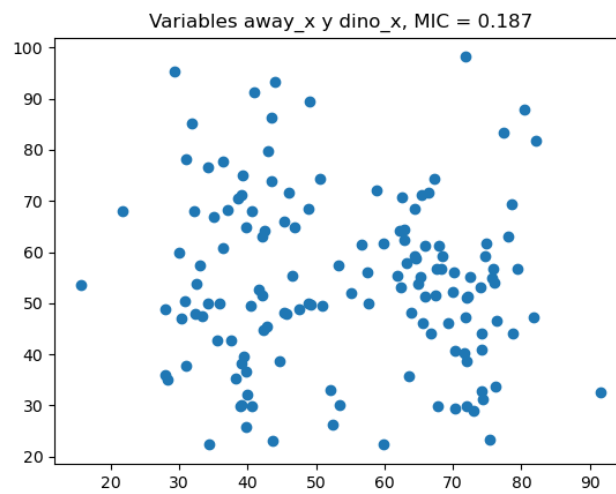


Figura 3.4: MIC = 0.187

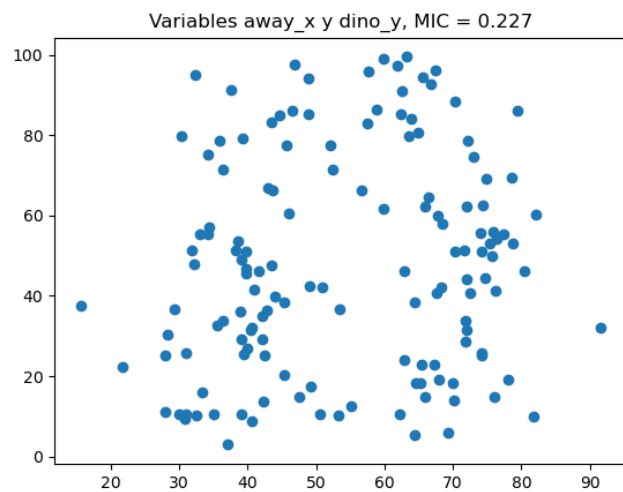


Figura 3.5: MIC = 0.227

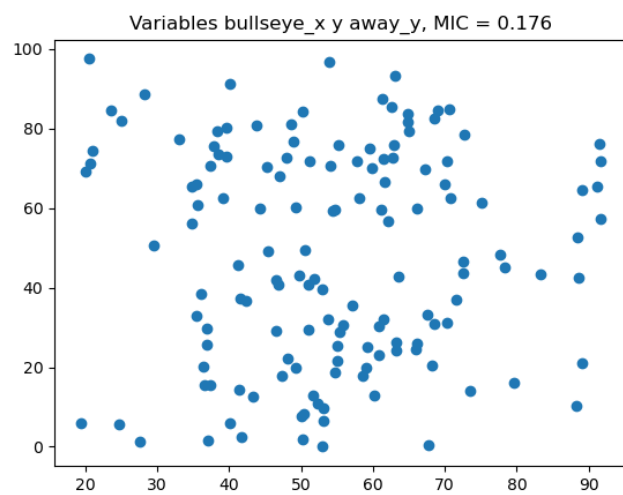


Figura 3.6: MIC = 0.176

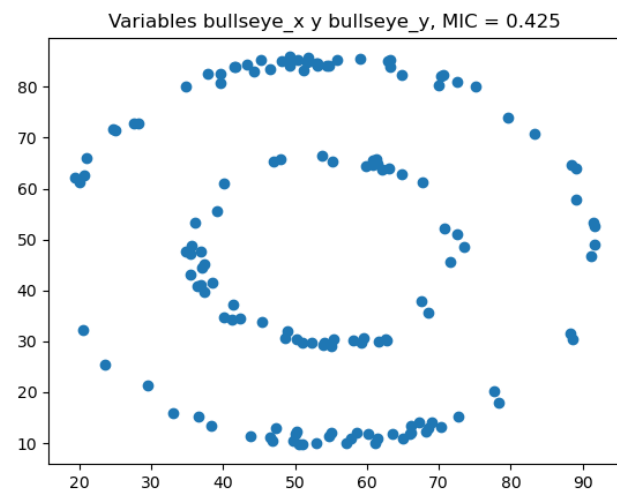


Figura 3.7: MIC = 0.425

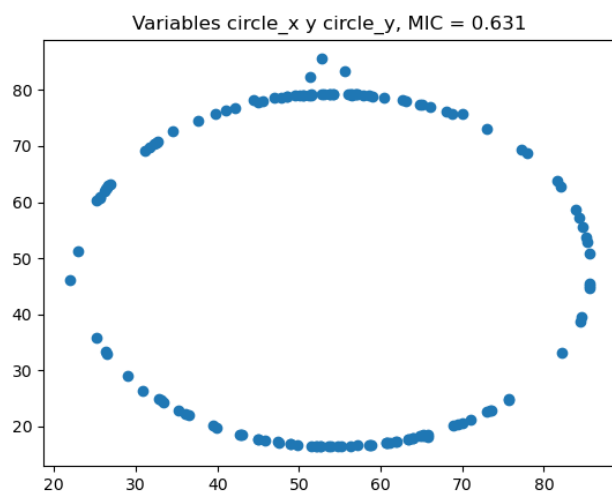


Figura 3.8: MIC = 0.631

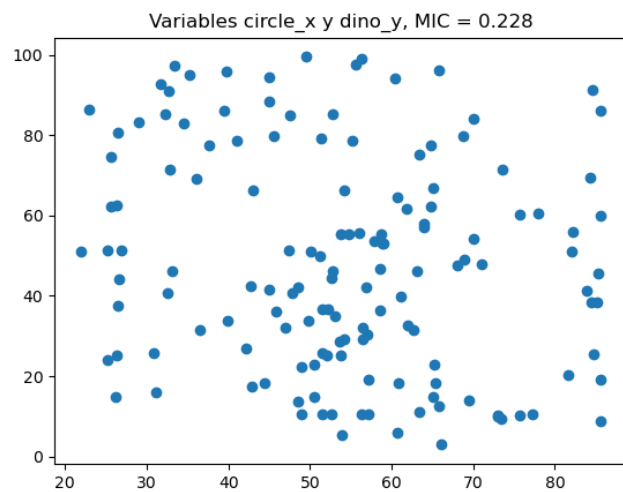


Figura 3.9: MIC = 0.228

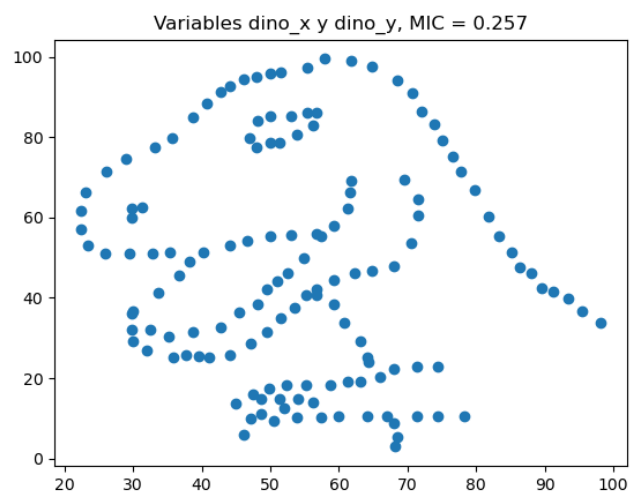


Figura 3.10: MIC = 0.257

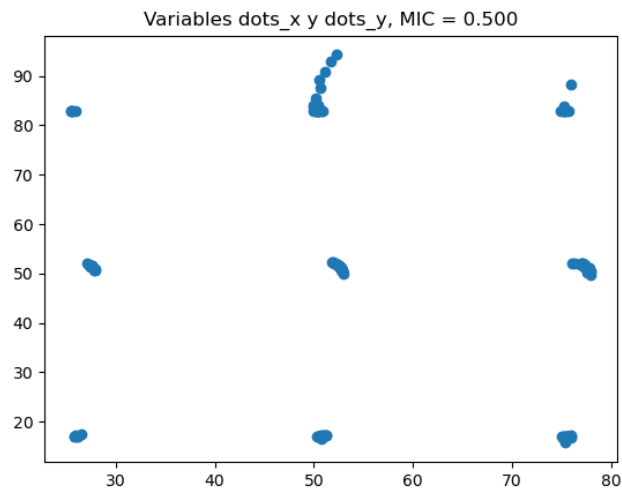


Figura 3.11: MIC = 0.500

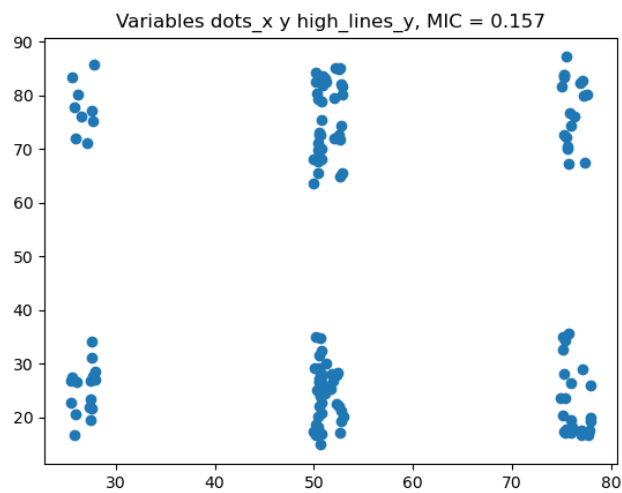


Figura 3.12: MIC = 0.157



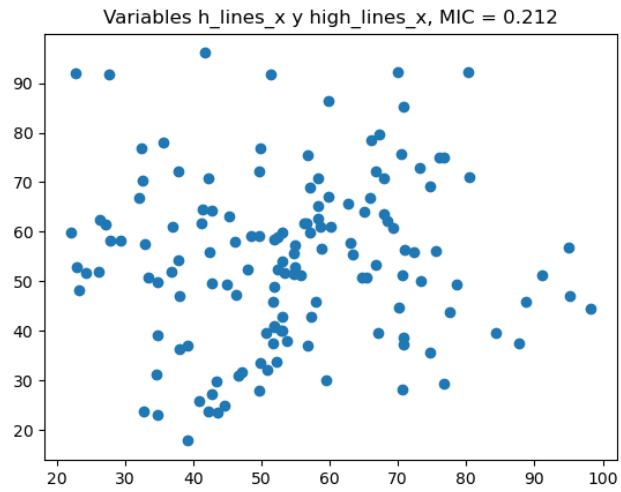


Figura 3.13: MIC = 0.212



## Capítulo 4

# Otros Métodos de Comparación de Imágenes

### 4.1. Correlación local

#### 4.1.1. Discucion sobre el coef.

donde se publico, como su ocupa, despicion en palabras

La correlación local, también 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.

#### 4.1.2. 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.**

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

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  $i = 1, \dots, N$  las observaciones en el conjunto de datos. Sea  $|z_i - z_j|$  la distancia euclidiana. Definimos  $\hat{I}(r) = \frac{1}{N^2} \sum_{i,j=1}^N I(|z_i - z_j| < r)$ . Las distancias obsevasdas son además linealmente transformadas para que se encuentren entres 0 y 1 antes de calcular  $\hat{I}$ .

Notemos que  $\hat{I}$  tiene las propiedades de una función de distribución acumulativa. Es no decreciente entre 0 y 1 y continua por la derecha. La función  $\hat{I}(r)$  describe el patrón global de distancias entre vecinos.

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

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

Donde  $\Delta \hat{I}(r)$  denota un cambio en  $\hat{I}(r)$ . La densidad de vecinos es evaluada en radio discreto  $r$ , con  $r = 0, 1/m, 2/m, \dots, 1$  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  $\hat{D}(r)$  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  $\hat{D}_0(r)$  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  $r$ :

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

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 último, definimos el coeficiente 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  $\hat{D}$  y  $\hat{D}_0$ . En otras palabras, definimos el estadístico  $M$  como la desviación máxima entre dos densidades vecinas subyacentes.

## 4.2. Correlación de Pearson

### 4.2.1. Discusión sobre el coef.

donde se publica, como su ocupa, descripción en palabras

### 4.3. Definiciones

El coef. se define como:

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

Para una muestra de tamaño  $N$ , tenemos:

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

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

# La transformación de Box y Cox

La transformación de Box y Cox, conocida como BoxCox, es una técnica de transformación no lineal que fue propuesta por George Box y David Cox en 1964, cuenta la historia que el Profesor Box 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 [5]. Esta transformación estadística paramétrica fue presentada en su artículo titulado "An Analysis of Transformations" [2]. Como canal de preprocesamiento, la transformación de BoxCox se utiliza comúnmente para convertir un conjunto de datos en una distribución que se asemeja a la normal, lo que a su vez facilita un análisis estadístico posterior.

La transformación de BoxCox pertenece a 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  $\lambda$ . En el contexto de la transformación de BoxCox, el objetivo es encontrar el valor de  $\lambda$  que proporciona el mejor ajuste a una distribución normal. Box y Cox desarrollaron este método con la intención de crear una técnica de transformación flexible que pudiera adaptarse a diversas distribuciones de datos.

Para un  $\lambda \in \mathbb{R}$  dado, la transformación de BoxCox se define como:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \log y & (\lambda = 0) \end{cases} \quad (5.1)$$

$\forall y \in \mathbb{R}_{>0}$ . En la práctica los valores de  $\lambda$  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].

Aunque también existe una versión para datos no positivos dada por:

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

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, como por ejemplo la normalización.

Como mencionamos anteriormente, el objetivo de la transformación es encontrar el valor de  $\lambda$  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:

$$\mathcal{L}(\lambda) \equiv -\frac{n}{2} \log \left[ \frac{1}{n} \sum_{j=1}^n \left( x_j^\lambda - \overline{x^\lambda} \right)^2 \right] + (\lambda - 1) \sum_{j=1}^n \log x_j \quad (5.2)$$

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  $\lambda$ . El valor de  $\lambda$  que maximiza esta verosimilitud es el que se selecciona para la transformación.

La transformación de BoxCox 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 [4] 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  $\lambda$  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.

## 5.1. BoxCox sobre imagenes

En su artículo del 2020 [3], Abbas Cheddad resalta una notable brecha en la aplicación de la transformación de BoxCox 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[6]. 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 transformación es iterativo, en el cual se ha de buscar un parametro  $\lambda$ , 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 comprimido 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:

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

Que correspondo al canal de escala de grises como está definido por el espacio de color  $YC_bC_r$  lo calcula. En base a esto definimos el histograma como

$$\chi(i) = \sum_{i=0}^{255} \mathcal{F}'$$
 , si tiene nivel de gris  $i$

Ahora, denotemos por  $\hat{\lambda}_\chi$  al parametro de la transformación BoxCox 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 demostrado util en problemas de clasificación.

Ahora definimos  $\mathcal{F}'(u, v)^{\lambda_\chi}$  como los datos siendo aplicada la transformación BoxCox definida en (5.1), y por ultimo vamos a definir BoxCox para imagenes o BCI como:



$$BCI = \frac{(\mathcal{F}''(u, v) - \min(\mathcal{F}''(u, v)))}{(\max(\mathcal{F}''(u, v)) - \min(\mathcal{F}''(u, v)))}$$

con  $\mathcal{F}''(u, v) = \mathcal{F}'(u, v)^{\hat{\lambda}_x}$

Nota. mantuve la notación del paper pero creo que esta necesita una revision.  
 Ahora veamos algunos ejemplos:



## Capítulo 6

# Comprando una Imagen con su transformada



## 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] 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.
- [5] 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](https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Introductory_Statistics_(Lane)/16%3A_Transformations/16.04%3A_Box-Cox_Transformations).
- [6] 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.
- [7] David N Reshef et al. «Detecting novel associations in large data sets». En: *science* 334.6062 (2011), págs. 1518-1524.
- [8] Yakir A Reshef et al. «Measuring dependence powerfully and equitably». En: *The Journal of Machine Learning Research* 17.1 (2016), págs. 7406-7468.
- [9] Noah Simon y Robert Tibshirani. «Comment on “Detecting novel associations in large data sets”». En: (2012).
- [10] Joy Thomas. Thomas Cover. *Elements of Information Theory*. New York: John Wiley & Sons, Inc, 2006. Cap. 8.