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

AnÃ¡lisis Estadistico y Procesamiento de Imagenes

2.1. Definiciones previas

Capítulo 3

Coficientes para la comparación entre dos imagenes

En este capítulo discutiremos distintos métodos para realizar comparaciones entre imagenes. Notemos que dado que nuestro objetivo final es comprar una imagen con su transformada de Box y Cox,

3.1. Correlación de Pearson

3.1.1. Discucion sobre el coef.

donde se publico, como su ocupa, despcion en palabras

3.2. Definiciones

El coef. se define como:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.1}$$

Para una muestra de tamaño N, tenemos:

$$r = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2}\sqrt{\sum_i^n (y_i - \bar{y})^2}} \tag{3.2}$$

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

3.3. *Maximal Information Coeficient*

3.3.1. 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"[citar]. 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 su 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 parte 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 coeficiente.

3.3.2. Definiciones

Como mencionamos en la parte anterior, debemos primero encontrar la información mutua entre las variables, para esto definamos:

Definición 3.1 (Información mutua)**.** Para dos variables aleatorias conjuntas X e 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 pokr 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

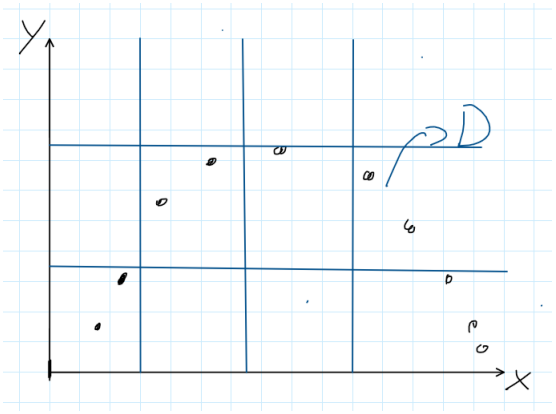

Figura 3.1: Malla G de 4x3 sobre el conjunto de pares ordenados D, figura no final

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

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

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

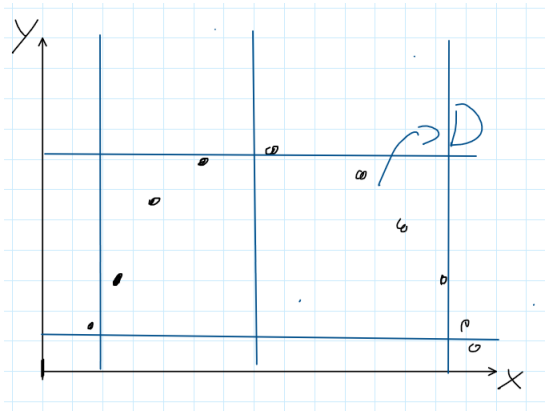

Figura 3.2: Otra malla G de 4x3 sobre el conjunto de pares ordenados D, figura no final

Aquí podemo 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 $I(D|_G)$ denot 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 bivariado 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 menos a $B(n)$ esta dado por:

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

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 $B(n) = n^{0,6}$, que es la función que ocupan en el paper citado al principio del la sección

3.3.3. 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

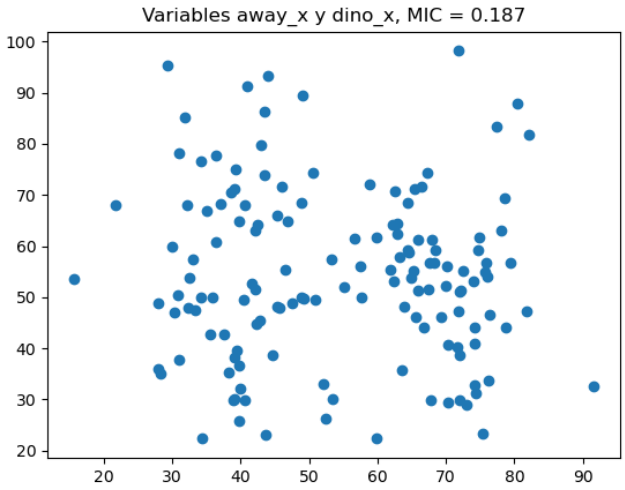

Figura 3.3: MIC = 0.187

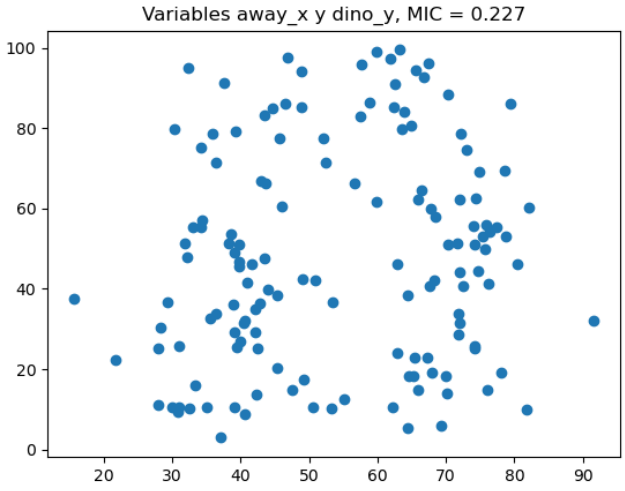

Figura 3.4: MIC = 0.227

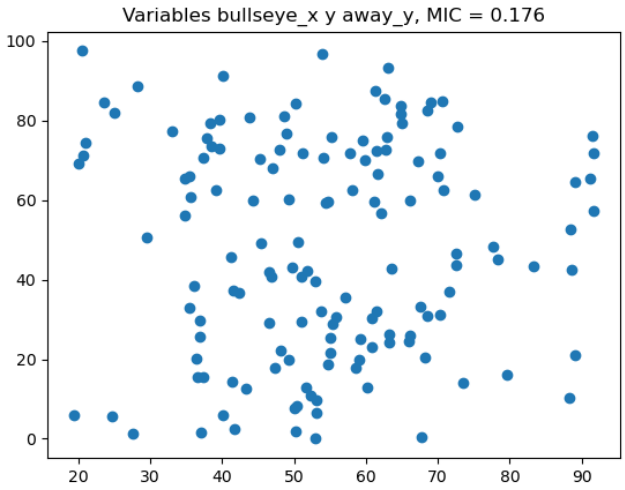

Figura 3.5: MIC = 0.176

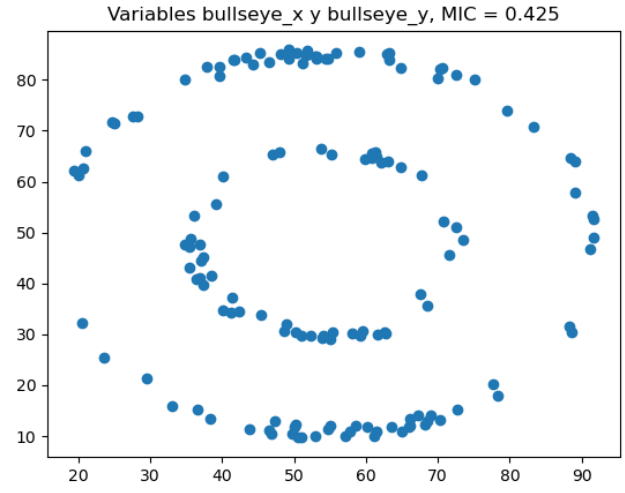

Figura 3.6: MIC = 0.425

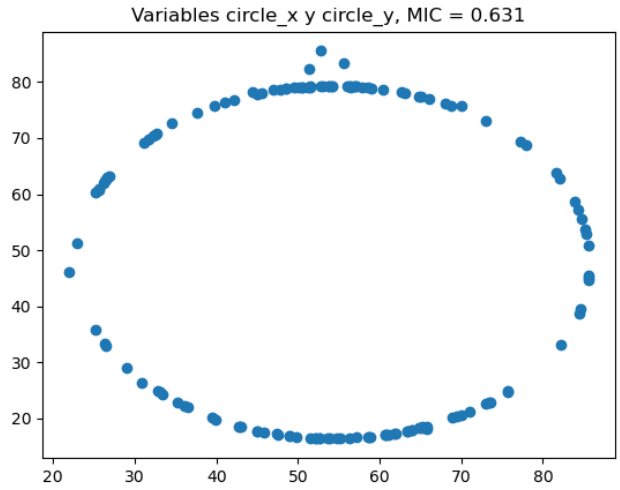

Figura 3.7: MIC = 0.631

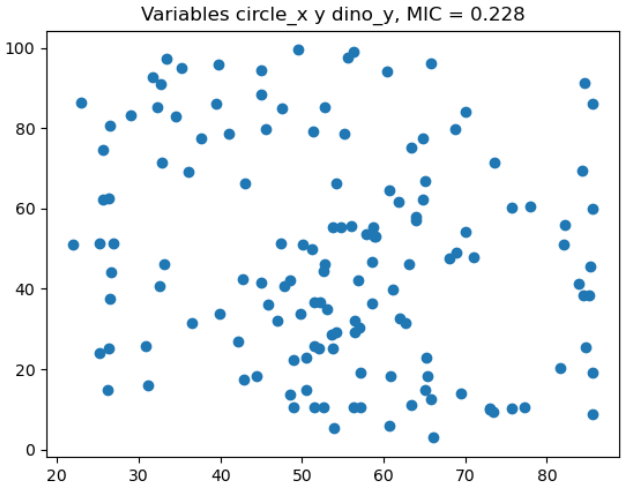

Figura 3.8: MIC = 0.228

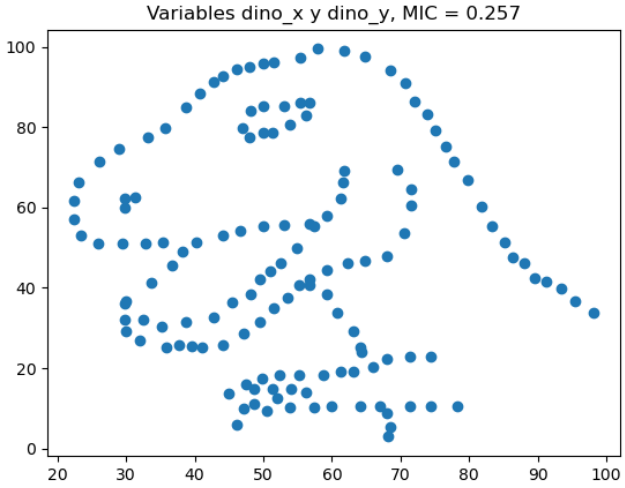

Figura 3.9: MIC = 0.257

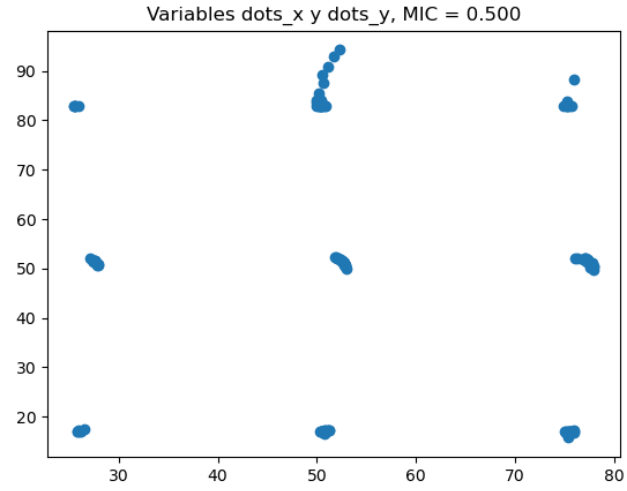

Figura 3.10: MIC = 0.500

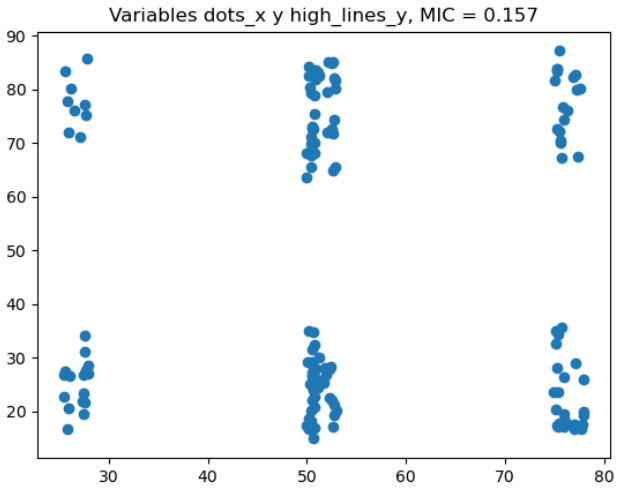

Figura 3.11: MIC = 0.157

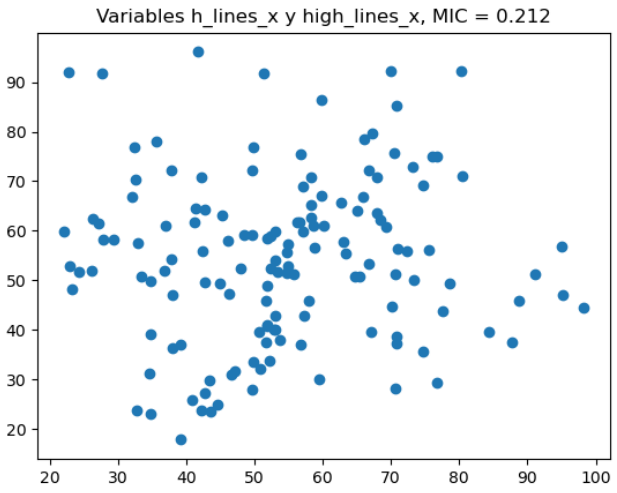

Figura 3.12: MIC = 0.212

3.3.4. Hablar sobre el algoritmo y el papaer del 2016, que es TICe y MICe

3.3.5. Hablar sobre el algoritmo y el papaer del 2016, que es TICe y MICe

3.4. Correlación local

3.4.1. 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.

3.4.2. Definiciones

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

Definición 3.5.

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

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 observaciones en el conjunto de datos. Sea $|z_i - z_j|$ la distancia euclidiana. Definimos $\hat{I}(r) = \frac{1}{N^2} \sum_{i,j=1}^{N} I\left(|z_i - z_j| < r\right)$. 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.

Capítulo 4

La transformación de Box y Cox

La tranformación de Box y Cox, o simplemente BoxCox es una transformación no lineal propuesta por George Box y David Cox en el año 1964 en su paper *An Analysis of Transformations*. Es una transformación estadística paramétrica no lineal que es a menudo utilizada como un canal de preprocesamiento para convertir datos a una distribución normali. El mÃ©todo es parte de las técnicas de tranformación de potencia, o*power transformation*, cuyo objetivo es encuentrar el parámetro λ que maximiza la siguiente verosimilutud.

$$\mathcal{L}(\lambda) \equiv -\frac{n}{2}\log\left[\frac{1}{n}\sum_{j=1}^{n}\left(x_j^{\lambda} - \overline{x^{\lambda}}\right)^2\right] + (\lambda - 1)\sum_{j=1}^{n}\log x_j \tag{4.1}$$

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

Dada la selección de $/lambda$, la transformación está dada por:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & (\lambda \neq 0) \\ \log y & (\lambda = 0) \end{cases} \tag{4.2}$$

$\forall y \in \mathrm{R}_{>0}$.Aunque también existe una versión para datos no positivos dada por

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

Pero esta no es muy usada, dado que antes de aplicar la transformación suele haber un paso de preprocesamiento de los datos que la deja por sobre 0.

Nota: el siguiente parrafo mantiene las citas del paper On Box-Cox Transformation for Image Normality and Pattern Classification

El objetivo de la transformación es asegurar que se logren los supuestos para modelos lineales de modo que técnicas de análisis de varianza estándar puedan aplicarse a los datos. La transformación no cambia el ordenamiento de los datos según Bicego y Baldó [3].

Obviamente, no todos los datos se pueden transformar de esta forma para producir normalidad, sin embargo, Draper y Cox [24] argumentan que incluso en los casos en que ninguna transformación de potencia podría llevar la distribución exactamente a la normalidad, las estimaciones habituales de λ pueden ayudar a regularizar los datos y, finalmente, conducir a una distribución que satisfaga ciertas características como la simetría o la homocedasticidad. Esta última es especialmente útil en el reconocimiento de patrones y aprendizaje automÃ¡tico (p. ej., análisis discriminante lineal de Fisher)

4.1. BoxCox sobre imagenes

Como mencionamos en la introcucción, no hay una gran cantidad de estudios que aborden la transformación BoxCox en conjunto con imÃ¡genes digitales. (ejemplos)

Notemos también que la aplicación de la tranformación es un proceso iterativo en el cual se ha de buscar un parametro λ, esto hace que aplicar la transformación en grandes bancos de imagenes sea demoroso. Una alternativa propuesta por A. Cheddad (On Box-Cox Transformation for Image Normality and Pattern Classification) 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 correspondo 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 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 desmostrado util en problemas de clasificación.

Ahora definimos $\mathcal{F}'(u,v)^{\lambda_\chi}$ como los datos siendo aplicada la transformación BoxCox definida en (4.2), y por ultimo vamos a definir BoxCox para imagenes o BCI como:

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

Nota. mantuve la notación del paper pero creo que esta necesita una revision.

Ahora veamos algunos ejemplos:

Capítulo 5

Comprarando una Imgaen con su tranformada

Capítulo 6

Consideraciones Finales

Apéndice A

Demostraciones

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

Bibliografía

Banerjee, A.N., and Magnus, J.R. (1999). The sensitivity of OLS when the variance matrix is (partially) unknown. *Journal of Econometrics* **92**, 295-323.