UNIVERSIDAD NACIONAL DE COLOMBIA

Comparación de dos metodologías estadísticas para un problema de clasificación binaria de imágenes bidimensionales

Deniz Andrea Sánchez Segura

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2020

Comparación de dos metodologías estadísticas para un problema de clasificación binaria de imágenes bidimensionales

Deniz Andrea Sánchez Segura

Trabajo de grado presentado como requisito parcial para optar al título de:
Magister en Ciencias-Estadística

Director:
Ph.D. en Estadística Rubén Darío Guevara

Codirector:
Ph.D. en Estadística Sergio Alejandro Calderón

Plan de estudio: Profundización

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2020

Dedicado a

Mis padres:

Marleny Segura y Arismendez Sánchez

Agradecimientos

Agradezco profundamente a mis directores Rubén Guevara y Sergio Calderón por su paciencia, cordialidad y apoyo en el desarrollo de este trabajo. A mis padres, hermanos y a Diego por su apoyo incondicional y su amor. De manera especial agradezco al profesor Leonardo Rendón por su amabilidad, sus consejos y su ayuda. Por último, pero no menos importante, a el profesor Luis Hernando Vanegas por su enorme colaboración.

Resumen

Un problema frecuente en el análisis de imágenes médicas es la identificación de regiones de interés en el organismo que son relevantes a ciertas respuestas clínicas. Esto corresponde a un problema de regresión con el uso de las imágenes anatómicas como variables explicativas y el estado del paciente como respuesta, que generalmente es una variable categórica. De esta manera, los coeficientes a estimar corresponden a ciertas señales que se activan en partes específicas del cuerpo y se desea estimar la variable respuesta, que representa el estado de clasificación del individuo. En el presente trabajo, dentro de las metodologías existentes para resolver el problema de categorización asociado a la condición del paciente se hace uso de dos métodos, con el fin de contrastarlos, debido a que no han sido comparados bajo el contexto de clasificación y uso de imágenes. El primer procedimiento que se implementa está basado en la exploración de un modelo matemático-estadístico a partir de arreglos multidimensionales, frecuentemente conocidos como tensores, y se fundamenta en el marco teórico del modelo lineal generalizado de alta dimensionalidad. La segunda metodología se sitúa en el campo del análisis de datos funcionales, más aún, en aquellas funciones que poseen la medida de variación total finita.

Palabras clave: Modelos lineales generalizados, Arreglos multidimensionales, Regresión tensor, Regresión funcional, Variación total, Regresión generalizada escalar-imagen, Clasificación .

Abstract

A common problem that appears in the analysis of medical images is the identification of regions of interest in the body that are relevant to certain clinical responses. This corresponds to a regression problem with the use of anatomical images as independent variables and the patient's state as the response, which is generally a categorical variable. In this way, the coefficients to be estimated correspond to certain signals that are activated in specific parts of the body and it is desired to estimate the response variable, which represents the classification status of the individual. In the present work, within the existing methodologies to solve the categorization problem associated with the patient's condition, two methods will be used, in order to contrast them, since they have not been compared under the context of classification and use of images. The first procedure that is implemented is based on the exploration of a mathematical-statistical model from multidimensional arrangements, often known as tensors. The second methodology is located in the field of functional data analysis, even more, in those functions that have the finite total variation measure.

Keywords: Generalized linear model, Multidimensional array, Tensor regression, Functional regression, Total variation, Generalized scalar-on-image regression, Classification

Índice de figuras

3-1. Tensor 1-dimensional con $x = 1, 2, ..., p_1$ 8
3-2. Tensor 2-dimensional con $x = 1, 2, ..., p_1$ y $y = 1, 2, ..., p_2$. 9
3-3. Tensor 3-dimensional con $x = 1, 2, ..., p_1$, $y = 1, 2, ..., p_2$ y $z = 1, 2, ..., p_3$. . . . 9
3-4. Tensor 4-dimensional con $x = 1, 2, ..., p_1$, $y = 1, 2, ..., p_2$, $z = 1, 2, ..., p_3$ y $t = 1, 2, ..., p_4$. 9
3-5. Ejemplo de representación de una imagen 64×64 como un tensor 2-dimensional, en una escala de grises 10

4-1. Representación de una imagen 64×64 píxeles (derecha) como una superficie continua (izquierda) 19
4-2. Representación de una imagen de tamaño 64×64 píxeles con un cambio de color abrupto (derecha) como una superficie discontinua (izquierda) 19
4-3. Imagen de tamaño 2×2 en la que cada entrada de posición 11, 12, 21 y 22 toma los valores $1 = Azul\ claro$, $2 = Azul\ oscuro$, $3 = Amarillo$ y $4 = Verde$, respectivamente. 22

5-1. Curva ROC, señalando la trayectoria en función de los diferentes valores de α. Tomado de $https : //algolytics.com/wp-content/uploads/2018/05/roc3_en.png$ 31
5-2. Curva ROC, denotando el área bajo la curva (ABC). Detallando que la diagonal de divide el cuadrado de tamaño 1×1 tendra como ABC 0.5. Tomado de $https : //algolytics.com/wp - content/uploads/2018/05/roc5.png$ 32

6-1. Imágenes coeficientes **B**, con una forma específica, con las cuales se realizará el proceso de estimación ($Forma - T$, $Triángulo$, $Círculo$ y $Combinado$). . . 34
6-2. Imagen coeficiente **B**, simulada de la misma manera que las covariables X, con la cual también se realizará el proceso de estimación. ($Simulado$) 34
6-3. $B - coeficientes$, para una de las 100, iteraciones con $n = 200$ ($n_e = 160$). Fijando el rango de la regresión tensor en 2. 39
6-4. $B - coeficientes$ estimados, para una de las 100 iteraciones, con $n = 1250$ ($n_e = 1000$). Fijando el rango de la regresión tensor en 2. 40
6-5. Curvas ROC, para una de las 100 iteraciones, con $n = 1250$ ($n_e = 1000$). Fijando el rango de la regresión tensor en 2. Las etiquetas corresponden a: (T) $Forma - T$, (Tr) $triángulo$, (C) $Círculo$, (Co) $Combinado$ 41

6-6. Imagen coeficiente **B** que será fijada. Es simulada como un producto de dos rejillas con el rango de una función senosoidal, en el intervalo [0,128]. 42
6-7. Superposición de un ejemplo de las imágenes cerebrales de la base de datos de individuos sanos ($\mathbf{X_i}$ en fucsia) con la imagen señal fija (**B** en verde). Ambas imágenes tienen un tamaño de 128×128 píxeles. 43
6-8. Representación de las estimaciones de **B**, de tamaño 128×128 píxeles. **B** original (izquierda), estimación vía regresión tensor de rango 2 (centro) y estimación vía variación total (derecha). Cada fila de imágenes emplea un tamaño de muestra de $n_0 = 2560$, $n_1 = 2560$, $n_2 = 1280$, $n_3 = 640$ y $n_4 = 320$ para las estimaciones respectivamente 44
6-9. Se simulan 8000 variables X_i siguiendo los parámetros de las bases de Haar, luego se estima el modelo para sub muestras de tamaño $n = 500$, $n = 1000$, $n = 5000$ y $n = 8000$ para **B** de la $Forma - T$. En la figura se muestran las estimaciones del coeficiente imagen. 45
6-10. Se simulan 8000 variables X_i siguiendo los parámetros de las bases de Haar, luego se estima el modelo para sub muestras de tamaño $n = 500$, $n = 1000$, $n = 5000$ y $n = 8000$ para **B** de la *Combinado*. En la figura se muestran las estimaciones del coeficiente imagen. 45
6-11. Dos ejemplos de imágenes cerebrales (MRI dos-dimensional) pertenecientes a la base de datos: brain tumor dataset. A la izquierda se presenta una imagen de un paciente sano y a la derecha una imagen de un paciente enfermo. . . . 46
6-12. Curva ROC para la clasificación para tamaño de muestra de $n = 253$, dividido en $n_e = 202$ observaciones de entrenamiento y $n_p = 51$ observaciones de prueba, para cada uno de las dos metodologías regresión tensor (Azul) y variación total (naranja). 47
6-13. Curva ROC para la clasificación para tamaño de muestra de $n = 253$ si se reducen las imágenes a un tamaño de 64×64, dividido en $n_e = 202$ observaciones de entrenamiento y $n_p = 51$ observaciones de prueba, para cada uno de las dos metodologías regresión tensor (Azul) y variación total (naranja). . . . 48

Índice de cuadros

3-1. Cambio del número de parámetros a estimar, sin reducción de dimensionalidad (columna 5) y con reducción de dimensionalidad (columna 6), para un modelo de imágenes bidimensionales ($D = 2$) de tamaño 28×28 píxeles, 64×64 píxeles y 128×128 píxeles. Se fija $p_0 = 3$ y un rango $R = 2$. 13

6-1. Valor del ABC de clasificación y su respectivo variación, con un tamaño de muestra de $n = 200$, dividido en $n_e = 160$ observaciones de entrenamiento y $n_p = 40$ observaciones de prueba, para cada uno de las diferentes formas de $\mathbf{B}$ (el coeficiente imagen) 36

6-2. Tiempo promedio computacional (en segundos), con un tamaño de muestra de $n = 200$, dividido en $n_e = 160$ observaciones de entrenamiento y $n_p = 40$ observaciones de prueba, para cada uno de las diferentes formas de $\mathbf{B}$ (el coeficiente imagen) 36

6-3. Valor del ABC de clasificación y su respectivo variación, con un tamaño de muestra de $n = 625$, dividido en $n_e = 500$ observaciones de entrenamiento y $n_p = 125$ observaciones de prueba, para cada uno de las diferentes formas de $\mathbf{B}$ (el coeficiente imagen) 37

6-4. Tiempo promedio computacional (en segundos), con un tamaño de muestra de $n = 625$, dividido en $n_e = 500$ observaciones de entrenamiento y $n_p = 125$ observaciones de prueba, para cada uno de las diferentes formas de $\mathbf{B}$ (el coeficiente imagen) 37

6-5. Valor del ABC de clasificación y su respectivo variación, con un tamaño de muestra de $n = 1250$, dividido en $n_e = 1000$ observaciones de entrenamiento y $n_p = 250$ observaciones de prueba, para cada uno de las diferentes formas de $\mathbf{B}$ (el coeficiente imagen) 37

6-6. Tiempo promedio computacional (en segundos), con un tamaño de muestra de $n = 1250$, dividido en $n_e = 1000$ observaciones de entrenamiento y $n_p = 250$ observaciones de prueba, para cada uno de las diferentes formas de $\mathbf{B}$ (el coeficiente imagen) 38

6-7. La cantidad de observaciones n se establece como $n_1 = 2560$, $n_2 = 1280$, $n_3 = 640$ y $n_4 = 320$. Para cada uno de estos tamaños de muestra se estima el coeficiente imagen. La tabla muestra el tiempo promedio computacional que requiere cada uno de los dos modelos, en segundos, para la estimación de $\mathbf{B}$, de tamaño 128×128 píxeles, empleando un tamaño n observaciones. 43

6-8. Valor del ABC de clasificación y su respectivo tiempo computacional con tamaño de muestra de $n = 253$, dividido en $n_e = 202$ observaciones de entrenamiento y $n_p = 51$ observaciones de prueba, para cada uno de las dos metodologías regresión tensor (Rango $R = 1$) y variación total. 47

6-9. Valor del ABC de clasificación y su respectivo tiempo computacional si se reducen las imágenes a un tamaño de 64×64, con tamaño de muestra de $n = 253$, dividido en $n_e = 202$ observaciones de entrenamiento y $n_p = 51$ observaciones de prueba, para cada uno de las dos metodologías regresión tensor (Rango $R = 2$) y variación total. 48

Contenido

Agradecimientos vii

Resumen ix

1. Introducción 1

2. Preliminares 3
- 2.1. Familia exponencial de distribuciones . . . 3
- 2.2. Modelo Lineal Generalizado-GLM . . . 4

3. Arreglos multidimensionales tensores 8
- 3.1. Definición, principales operaciones y propiedad de descomposición . . . 8
- 3.2. Modelo de Regresión Tensor . . . 12
- 3.3. Método de estimación . . . 13

4. Funcionales, variación total y espacio de variación acotada 16
- 4.1. Modelo Lineal Escalar-Función de valor real . . . 24
- 4.2. Modelos de regresión lineal generalizado Escalar-Función . . . 27
  - 4.2.1. Modelo de variación acotada . . . 27

5. Prevalencia en los modelos de clasificación: curva ROC 29

6. Resultados 33
- 6.1. Proceso de simulación . . . 33
  - 6.1.1. Resultados . . . 36
- 6.2. Ejercicio preliminar . . . 42
- 6.3. Aplicación a datos reales . . . 46

7. Conclusiones y Recomendaciones 50
- 7.1. Conclusiones . . . 50
- 7.2. Recomendaciones . . . 51

A. Anexo: Total variation Minimization by Augmented Lagrangian Alternating Direction Algorithms-TVAL3 52

B. Anexo: bases de Haar 54

C. Anexo: ejemplo puntual del calculo de las operaciones tensoriales definidas **56**

Bibliografía **58**

1. Introducción

El procesamiento de imágenes, el reconocimiento de patrones y los sistemas de clasificación a través de imágenes se han convertido en aplicaciones bastante estudiadas debido al desarrollo tecnológico. Su investigación ha crecido notablemente en los últimos años en un intento para ofrecer herramientas de sistemas más precisos y seguros. Un problema frecuente, es el análisis de imágenes médicas, que consiste en la identificación de regiones de interés en el cuerpo que son relevantes a ciertas respuestas clínicas. Dicho problema se puede ver desde el punto de vista estadístico como un modelo regresión donde las variables exógenas son las imágenes cerebrales y el estado del paciente es la variable respuesta o endógena, de tal manera que, los coeficientes a estimar correspondan a ciertas señales que se activan en partes específicas del cuerpo. Una aplicación importante del reconocimiento de estas zonas de interés está altamente asociado con el diagnóstico de enfermedades.

Un primer procedimiento estadístico que surgió para el procesamiento de imágenes fue empleado por Worsley en 2004 [22] y Lazar en 2008 [8], con métodos basados en voxeles [1] los cuales toman cada voxel como respuesta y ciertas variables como predictoras. La principal desventaja de este procedimiento es que cada voxel se trata como un solo elemento independiente y espacialmente no correlacionado con los otros, lo cual en la práctica no debería ignorarse. Otro camino hacia la solución del problema es implementado por Caffo en el 2010, con el uso de componentes principales (PCA) (véase en [14] y [15]); una ventaja de este procedimiento es que se realiza en dos pasos, primero una reducción de la dimensionalidad por PCA y luego un ajuste de un modelo de regresión basado en los primeros d componentes principales. Una solución alternativa fue dada por Hua Zhou, Lexin Li & Hongtu Zhu en su artículo *Tensor Regression with Aplications in Neuroimaging Data Analysis* [24], el procedimiento que se implementa está basado en la exploración y la formulación de un modelo matemático-estadístico a partir de arreglos multidimensionales, frecuentemente conocidos como tensores, y se fundamenta en el marco teórico del modelo lineal generalizado de alta dimensionalidad.

El modelo de regresión tensor hace uso de las operaciones del espacio vectorial de los tensores y de sus propiedades de descomposición. Adicionalmente, esta metodología se basa en

[1] El Voxel es un término surgido del inglés "Volumetric píxel" y se refiere a la unidad cúbica que forma parte de un objeto tridimensional, y constituye la más mínima unidad de un conjunto tridimensional que puede ser procesada.

suponer que los datos siguen una distribución perteneciente a la familia exponencial. Así, se desarrolla un algoritmo de estimación en base a la optimización de la función de verosimilitud, que bajo ciertos supuestos, resulta ser en cadena, una estimación por máxima verosimilitud de varios GLM clásicos. Lo anterior se debe a que en cada iteración se fijan todos los parámetros a estimar salvo uno, es decir que, en cada paso el modelo depende de un solo parámetro y el resto se mantienen constantes.

Un nuevo enfoque corresponde a los modelos de regresión para datos funcionales que incorporan imágenes como predictoras. Puntualmente, para el caso en el que el funcional represente una imagen, se han desarrollado modelos en los cuales se supone que la función pertenece a un espacio en particular, tal es el caso de el espacio de funciones cuadrado integrables, denotado por $\mathcal{L}^2$ (véase [16]). El uso de las funciones de variación acotada en este problema, esta fundamentado en el hecho de que permiten modelar funciones con discontinuidades sobre lineas o curvas, pudiendo así representar bien los vértices y contornos de una imagen. Lo anterior produce una ventaja al suponer adicionalmente que, el funcional pertenece a la familia de funciones de variación total finita y acotada, es por este hecho que los investigadores Xiao Wang y Hongtu Zhu en el artículo *Generalized Scalar-on-Image Regression Models via Total Variation* [20] emplean una técnica basada en el problema de minimización de la variación total de una función para resolver el problema de regresión con imágenes como covariables. En la metodología de los modelos de regresión vía variación acotada se desea minimizar la variación total del coeficiente de regresión. Para su estimación, se emplea un procedimiento basado en los algoritmos de dirección alternativa, el Langrangiano aumentado y mínimos cuadrados ponderados estándar.

Debido a que las dos metodologías (modelo de regresión tensor y modelo de variación total) se desarrollaròn de forma independiente, arrojan buenos resultados y nunca se han contrastado, el principal objetivo del trabajo es comparar ambos procesos en el contexto de clasificación binaria de imágenes bidimensionales. De esta manera, la idea es, que a partir del criterio de comparación dado por el área bajo la curva ROC (*Receiver Operating Characteristic*) se pueda decidir cual de los modelos clasifica mejor una base de datos determinada.

El documento se estructura de la siguiente forma: en el capítulo 2 se presenta los conceptos preliminares necesarios para el desarrollo del trabajo. Posteriormente, en los capítulos 3 y 4 se muestran, respectivamente, los modelos de clasificación basados en los objetos tensoriales y los objetos funcionales de variación acotada. La construcción de la curva ROC, y el criterio de contraste que se genera a partir de ella, esta contenido en el capítulo 5. El capítulo 6 tiene como objetivo mostrar los resultados tanto de los estudios de simulación como la aplicación a datos reales y finalmente, en el capítulo 7 se presentan las conclusiones y recomendaciones.

2. Preliminares

En el presente capítulo se presentan los conceptos de familia exponencial, modelo de regresión, modelo lineal generalizado y regresión logística, los cuales son necesarios para el desarrollo de este trabajo. Las nociones y definiciones que se presentan a continuación fueron extraídas de [3].

2.1. Familia exponencial de distribuciones

La familia exponencial es una familia de distribuciones unificadas por conveniencia matemática, basada en algunas propiedades algebraicas útiles, y ampliamente utilizada en espacios euclidianos paramétricos de dimensiones finitas. Esta familia adquiere una considerable importancia en el problema de la estimación puntual de una función paramétrica, debido a que contiene muchas distribuciones frecuentemente usadas. Dada Y una variable aleatoria, se dice que la distribución de Y pertenece a la familia exponencial si su función de densidad está dada por:

$$f(y;\theta,\phi) = exp\left\{\frac{yb(\theta)+c(\theta)}{a(\phi)} + d(y,\phi)\right\},$$

donde ϕ se denomina parámetro de dispersión y usualmente se puede estimar a partir de los datos. Adicionalmente, $b(.)$, $c(.)$, $a(.)$ y $d(.,.)$ son funciones conocidas. Es común encontrar una reparametrización de la familia exponencial planteada como $\eta = b(\theta)$. En particular, se puede ver que,

$$f(y;\eta,\phi) = exp\left\{\frac{y\eta+d_0(\eta)}{a(\phi)} + d(y,\phi)\right\},$$

donde $d_0(\eta) = c(b^{-1}(\eta))$. Algunas propiedades de esta reparametrización son las siguientes:

- $\frac{\partial d_0(\eta)}{\partial \eta} = E(Y) = \mu$
- $a(\phi)\frac{\partial^2 d_0(\eta)}{\partial^2 \eta} = Var(Y)$

Con base en lo anterior, se pueden apreciar los siguientes ejemplos de variables aleatorias que pertenecen a la familia exponencial.

Ejemplo **1** *Se considera el caso en el que Y tenga función de densidad normal de media μ y varianza σ^2:*

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} = exp\left\{\frac{\mu y - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}log(2\pi\sigma^2)\right\}$$

Sea $\eta = \mu$ *y* $\phi = \sigma^2$*, así:*

$$f(y;\eta,\phi) = exp\left\{\frac{\eta y - \eta^2/2}{\phi} - \frac{y^2}{2\phi} - \frac{1}{2}log(2\pi\phi)\right\}.$$

De esta manera si se toma $d_0(\eta) = \frac{-\eta^2}{2}$*,* $a(\phi) = \phi$ *y* $d(y,\phi) = \frac{y^2}{2\phi} - \frac{1}{2}log(2\pi\phi)$*, entonces se observa que la distribución normal de hecho, pertenece a la familia exponencial.*

Ejemplo **2** *Se considera caso en el que:*

$$Y = \begin{cases} 1 & \textit{si la salida es un éxito} \\ 0 & \textit{si la salida es un fracaso} \end{cases}$$

con probabilidades $P(Y=1) = \pi$ *y* $P(Y=0) = 1-\pi$*. Esto corresponde a la distribución Bernoulli de parámetro* π*:*

$$f(y;\theta) = exp\left\{ylog\left(\frac{\pi}{1-\pi}\right) + log(1-\pi)\right\},$$

donde $\theta = \frac{\pi}{1-\pi}$*,* $a(\phi) = 1$*,* $b(\theta) = log\left(\frac{\pi}{1-\pi}\right)$*,* $c(\theta) = log(1-\pi)$ *y* $d(y) = 0$*. Reparametrizando con* $\eta = log\left(\frac{\pi}{1-\pi}\right)$*, se obtiene*

$$f(y;\eta) = exp\left\{y\eta - log(1+e^{\eta})\right\}, \tag{2-1}$$

donde $d_0(\eta) = -log(1+e^{\eta})$*.*

2.2. Modelo Lineal Generalizado-GLM

En las aplicaciones, es bastante común encontrar datos que no siguen una distribución normal. Algunos ejemplos incluyen situaciones en las que se requieren modelos que trabajen con datos dicotómicos, ordinales, categóricos o de elecciones discretas; para modelar estos datos, a menudo son apropiadas las distribuciones Bernoulli, Binomial, Poisson, Exponencial, entre otras. En este sentido, surgen los modelos lineales generalizados (GLM por sus siglas en inglés) los cuales proporcionan herramientas efectivas para manejar dichos datos. Para los datos no gaussianos, la media de la respuesta generalmente no está relacionada linealmente con los predictores. Esto tal vez se ilustra más claramente para los resultados binarios. Los

GLM abordan este problema asumiendo que la media de la respuesta está relacionada linealmente con los predictores solo después de una transformación adecuada (frecuentemente dada por una función de enlace denotada por $g(.)$).

El GLM supone variables aleatorias independientes $Y_1, Y_2, ..., Y_N$ cuyas distribuciones pertenecen a la familia exponencial con función de densidad dada por:

$$f(y_i; \eta_i, \phi_i) = exp\left\{\frac{y_i\eta_i + d_0(\eta_i)}{a(\phi_i)} + d(y_i, \phi_i)\right\}, \tag{2-2}$$

así, la función de densidad de probabilidad conjunta de $Y_1, ..., Y_N$ se escribe como

$$f(y_1, ..., y_N; \eta_1, \cdots, \eta_N, \phi_1, \cdots, \phi_N) = exp\left\{\sum_i^N \frac{y_i\eta_i + d_0(\eta_i)}{a(\phi_i)} + \sum_i^N d(y_i, \phi_i)\right\}.$$

En la especificación del modelo usualmente se está interesado en un pequeño conjunto de parámetros $\beta_1, ..., \beta_p$, con $p < N$. Para esto se supone que se tiene la relación $E[Y_i] = \mu_i$, donde μ_i es una función que depende de η_i y ϕ_i. Ahora, para los modelos lineales generalizados existe una transformación de μ_i tal que:

$$g(\mu_i) = \alpha + \vec{\beta}'\vec{x}_i \tag{2-3}$$

donde $\vec{\beta} = (\beta_1, \cdots, \beta_p)'$ y el vector $\vec{x}_i$ de variables explicativas son de dimensión $p \times 1$, esto para $i = 1, \cdots, N$. La función $g(\cdot)$ se denomina **función de enlace** y es estrictamente creciente y diferenciable. Se denomina a $\vec{x}_i'$ como la i-ésima fila de la matriz de diseño $\mathbf{X}$, y reescribimos la ecuación 2-3 de la manera $g(\vec{\mu}) = \alpha + \mathbf{X}\vec{\beta}$. A continuación, se exhiben algunos ejemplos que ilustran estos conceptos.

Ejemplo **3** *Para el ejemplo 1 de la anterior sección, se tiene que si $Y_i \sim N(\mu_i, \sigma_i^2)$, entonces*

$$f(y_1, ..., y_N; \eta_1, ..., \eta_N) = exp\left\{\sum_i^N \frac{\eta y_i - \eta^2/2}{\phi} - \sum_i^N \frac{y_i^2}{2\phi} - \frac{N}{2}log(2\pi\phi)\right\}, \tag{2-4}$$

donde $d_0(\eta_i) = \frac{-\eta_i^2}{2}$, $\frac{\partial d_0(\eta_i)}{\partial \eta_i} = d_0'(\eta_i) = -\eta_i$. Nótese que en particular, $g(.) := I(.)$, es decir,

$$-d_0(\eta_i) = \eta_i = \mu_i = \alpha + \vec{\beta}'\vec{x}_i,$$

para $i = 1, \cdots, N$. Este modelo se conoce comúnmente como ***Modelo de Regresión Lineal****.*

Ejemplo **4** *Para el ejemplo 2, si $Y_i \sim Bernoulli(\pi_i)$, se tiene que $\mu_i = E(Y_i) = \pi_i$ y además,*

$$f(y_1, ..., y_N; \eta_1, ..., \eta_N) = exp\left\{\sum_i^N y_i\eta_i - \sum_i^N log(1 + e^{\eta_i})\right\}, \tag{2-5}$$

donde $d_0(\eta_i) = -log(1e^{\eta_i})$. Nótese que en particular, $\frac{\partial d_0(\eta_i)}{\partial \eta_i} = d_0'(\eta_i) = -\frac{e^{\eta_i}}{1+e^{\eta_i}} = -\mu_i$, es decir,

$$g(-d_0'(\eta_i)) = g\left(\frac{e^{\eta_i}}{1 + e^{\eta_i}}\right) = g(\mu_i) = \alpha + \vec{\beta}'\vec{x}_i,$$

para $i = 1, \cdots, N$

Por último, se menciona el análisis de **regresión logística (modelo logistic o modelo logit)**, el cual se enmarca en el conjunto de GLM. La regresión logística analiza datos distribuidos binomialmente de la forma:

$$Y_i \sim B(n_i, \pi_i), \text{ para, } i = 1, \cdots, N$$

donde $B(n_i, \pi_i)$ representa la distribución binomial, la cual es una distribución de probabilidad discreta que cuenta el número de éxitos en una secuencia de n_i (conocida) ensayos Bernoulli independientes entre sí, con una probabilidad π_i (desconocida) fija de ocurrencia del éxito entre los ensayos. La función de densidad para cada Y_i estadado por:

$$f(y_i; \theta_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1 - \pi_i)^{n_i - y_i} = exp\left\{y_i log\left(\frac{\pi_i}{1 - \pi_i}\right) + log(1 - \pi_i) + log\binom{n_i}{y_i}\right\},$$

donde $\theta_i = \frac{\pi_i}{1-\pi_i}$, $a(\phi_i) = 1$, $b(\theta_i) = log\left(\frac{\pi_i}{1-\pi_i}\right)$, $c(\theta_i) = log(1 - \pi_i)$ y $d(y_i) = log\binom{n_i}{y_i}$. Reparametrizando con $\eta_i = log\left(\frac{\pi_i}{1-\pi_i}\right)$, se obtiene

$$f(y_i; \eta_i) = exp\left\{y\eta_i - log(1 + e^{\eta_i}) + log\binom{n_i}{y_i}\right\}, \tag{2-6}$$

donde $d_0(\eta_i) = -log(1 + e^{\eta_i})$.
Los coeficientes de regresión generalmente se estiman usando el método de máxima verosimilitud. A diferencia de la regresión lineal con residuos distribuidos normalmente, no es posible encontrar una expresión de forma cerrada para los valores de los coeficientes que maximicen la función de probabilidad en el modelo de regresión logistica, por lo que se debe utilizar un proceso iterativo; por ejemplo el método de Newton. Este procedimiento comienza con una solución tentativa (inicial), la revisa para ver si se puede mejorar y repite esta revisión hasta que no se realice ninguna mejora adicional, momento en el que se dice que el proceso

ha convergido.

Por otro lado, la bondad de ajuste en los modelos de regresión lineal generalmente se mide usando R^2, como no se tiene un análogo directo en la regresión logística, se pueden usar varios métodos como lo es la prueba de desviación y razón de probabilidad frecuentemente conocida como *Deviance*. La medida de *Deviance* es análoga a los cálculos de la suma de cuadrados en regresión lineal y es una medida de la falta de ajuste a los datos en un modelo de regresión logística. Cuando hay disponible un modelo "saturado"(un modelo con un ajuste teóricamente perfecto), la desviación se calcula comparando un modelo dado con el modelo saturado. De esta manera, el valor de *Deviance* corresponde a:

$$D = -2[\ell(\hat{\beta}_{saturado}; y) - \ell(\hat{\beta}_{ajustado}; y)],$$

donde $\hat{\beta}_{saturado}$ y $\hat{\beta}_{ajustado}$ denotan el estimador de máxima verosimilitud de β en el modelo saturado y ajustado respectivamente. Se puede mostrar que esta métrica sigue una distribución chi-cuadrado aproximada. Así, los valores más pequeños indican un mejor ajuste ya que el modelo ajustado se desvía menos del modelo saturado. Por el contrario, un valor de chi-cuadrado grande indica que una cantidad significativa de la varianza no se explica. Cuando el modelo saturado no está disponible. lo que es común, la desviación se calcula simplemente como 2 veces la función de verosimilitud del modelo ajustado.

3. Arreglos multidimensionales tensores

El propósito del siguiente capítulo es mencionar y establecer algunos hechos importantes concernientes a los arreglos multidimensionales y algunas de sus características; la construcción del modelo de regresión basado en los tensores y sus propiedades, y finalmente, el método de estimación que se usará a lo largo del trabajo.

3.1. Definición, principales operaciones y propiedad de descomposición

Un arreglo multidimensional o tensor es un objeto algebraico de varias componentes, que generaliza los conceptos de escalar, vector y matriz de manera que sea independiente de cualquier sistema de coordenadas elegido. Por definición de espacio vectorial sobre $\mathbb{R}^n$ [10], una vez elegida una base vectorial, los componentes de un elemento del espacio en una base pueden ser compilados en un vector de n entradas, este concepto puede extenderse de manera que, los coeficientes de un tensor en una base se podrán guardar en una matriz de varias dimensiones. En ese sentido sea, $\mathbf{M} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ un tensor D-dimensional, denotamos $\mathbf{M} = (m)_{i_1,i_2,...,i_D}$ con $m \in \mathbb{R}$ y $i_1 = 1, 2, ..., p_1$, $i_2 = 1, 2, ..., p_2$,..., $i_D = 1, 2, ..., p_D$ [24]. A continuación se describe la presentación del tensor D-dimensional para algunos valores de D .

- Si $D = 1$, entonces $M = (m)_{i_1} = (m)_{1,2,...,p_1} = (m_1, m_2, \cdots, m_{p_1})$. Lo que representa un vector.

Figura 3-1.: Tensor 1-dimensional con $x = 1, 2, ..., p_1$

- Si $D = 2$, entonces $M = (m)_{i_1,i_2} = (m)_{x,y}$. Lo que representa una matriz.

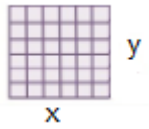

Figura 3-2.: Tensor 2-dimensional con $x = 1, 2, ..., p_1$ y $y = 1, 2, ..., p_2$.

- Si $D = 3$, entonces $M = (m)_{i_1,i_2,i_3} = (m)_{x,y,z}$

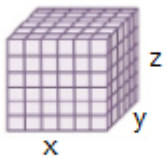

Figura 3-3.: Tensor 3-dimensional con $x = 1, 2, ..., p_1$, $y = 1, 2, ..., p_2$ y $z = 1, 2, ..., p_3$.

- Si $D = 4$, entonces $M = (m)_{i_1,i_2,i_3,i_3} = (m)_{x,y,z,t}$. Lo que representa, por ejemplo, un tensor 3-dimensional medido a través del tiempo.

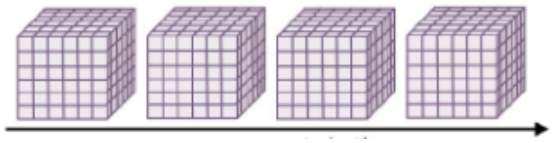

Figura 3-4.: Tensor 4-dimensional con $x = 1, 2, ..., p_1$, $y = 1, 2, ..., p_2$, $z = 1, 2, ..., p_3$ y $t = 1, 2, ..., p_4$.

Una imagen de píxeles es una estructura de datos que representa una malla rectangular de píxeles o puntos de color. Los puntos de color son reducidos a números para su almacenamiento y tratamiento; el conjunto de valores numéricos adopta la forma de una matriz (tensor 2-dimensional) a la que denominamos matriz-imagen. La reproducción en escala de grises utiliza un solo color de representación, pero entre el blanco y el negro pueden darse una serie de tonos intermedios. El tono de cada píxel puede estar representado por un número, por ejemplo de 8 bits. En este caso, caben $2^8 = 256$ posibilidades (de 0 a 255): se asigna un valor 0 al negro y el valor más alto (255), al blanco, los tonos intermedios son numerados del 1 al 254. En infografía 3D (tres dimensiones) el concepto de una rejilla plana de píxeles se extiende a un espacio tridimensional formado por ladrillos cúbicos llamados vóxeles. En este caso, existe una reja tridimensional con elementos (cubos) que contienen la información

del color, por ejemplo RGB. Nótese que lo anterior en el espacio tensorial se representaría por un vector 3-dimensional.

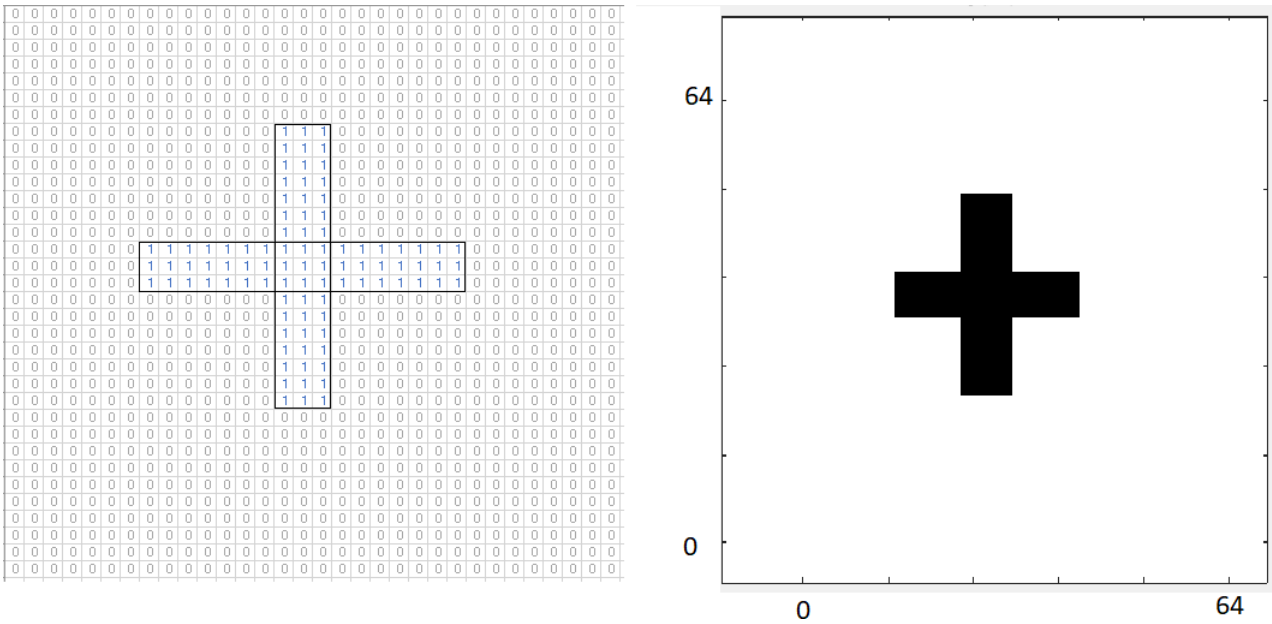

Figura 3-5.: Ejemplo de representación de una imagen 64×64 como un tensor 2-dimensional, en una escala de grises

Para el desarrollo del método de regresión basada en tensores, se deben definir algunas operaciones que se presentan a continuación, las cuales fueron extraídas de [24].

Definición **1** *Dadas dos matrices* $A = [\vec{a}_1, \cdots, \vec{a}_n] \in \mathbb{R}^{m\times n}$ *y* $B = [\vec{b}_1, \cdots, \vec{b}_q] \in \mathbb{R}^{p\times q}$. *El producto de Kronecker se define como la matriz* $mp \times nq$ *resultante de*

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}$$

Definición **2** *Dadas dos matrices* $A = [\vec{a}_1, \cdots, \vec{a}_n] \in \mathbb{R}^{m\times n}$ *y* $B = [\vec{b}_1, \cdots, \vec{b}_n] \in \mathbb{R}^{p\times n}$, *el producto de Khatri-Rao es definido por columnas como el producto de Kronecker, es decir,*

$$A \odot B = [\vec{a}_1 \otimes \vec{b}_1, \cdots, \vec{a}_n \otimes \vec{b}_n].$$

Si además n=q=1, entonces $A \odot B = A \otimes B$.

Definición **3** *Dado* $\mathbf{B} \in \mathbb{R}^{p_1\times\cdots\times p_D}$, *el operador vectorización de B, vec(B), apila las entradas de un tensor D-dimensional a un vector columna.*

Definición **4** *Dados $\vec{b}_1, \vec{b}_2, \cdots, \vec{b}_D$ con $\vec{b}_d \in \mathbb{R}^{p_d}$ para $d = 1, \cdots, D$, el producto outer de $\vec{b}_1, \vec{b}_2, \cdots, \vec{b}_D$, denotado $\vec{b}_1 \circ \vec{b}_2 \circ \cdots \circ \vec{b}_D$, es un arreglo de dimensión $p_1 \times \cdots \times p_D$ con entradas*

$$(\vec{b}_1 \circ \vec{b}_2 \circ \cdots \circ \vec{b}_D)_{i_1,\ldots,i_D} = \prod_{d=1}^{D} b_{di_d}.$$

Para el caso $D = 3$, por ejemplo, el producto outer de $\vec{b}_1, \vec{b}_2, \vec{b}_3$ es un arreglo de dimensión $p_1 \times p_2 \times p_2$ con entradas

$$(\vec{b}_1 \circ \vec{b}_2 \circ \vec{b}_3)_{i_1,i_2,i_3} = \prod_{d=1}^{3} b_{di_d} = b_{1i_1} b_{2i_2} b_{3i_3}.$$

Definición **5** *Dado $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$. El operador $B_{(d)}$ (Mode-d matricization) mapea el tensor B a una matriz $p_d \times \prod_{d' \neq d} p'_d$, de tal manera que al elemento $(i_1, \cdots, i_D)$ del arreglo $\mathbf{B}$ lo envía en el elemento (i_d, j) de la matriz $B_{(d)}$ con $j = 1 + \sum_{d' \neq d}(i_d - 1) \prod_{d'' < d', d'' \neq d} p_{d''}$.*

En el anexo C se presentan algunos ejemplos puntuales para las anteriores definiciones. Ahora bien, los espacios vectoriales poseen la propiedad de que sus elementos se pueden ver como suma de otras componentes del mismo espacio. Es por esto que se establece el siguiente teorema.

Teorema **1** *Un arreglo multidimensional $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ admite una descomposición de rango R si:*

$$\mathbf{B} = \sum_{r=1}^{R} \vec{\beta}_1^{(r)} \circ \cdots \circ \vec{\beta}_D^{(r)},$$

donde $\vec{\beta}_1^{(r)} \in \mathbb{R}^{p_d}$, $d = 1, \cdots, D$, $r = 1, \cdots, R$ y $\mathbf{B}$ no puede ser escrita como una suma de menos elementos que R.

Notación: Se establece que $\mathbf{B} = [\mathbf{B}_1, \cdots, \mathbf{B}_D]$, donde $\mathbf{B}_d = [\vec{\beta}_d^{(1)}, \cdots, \vec{\beta}_d^{(R)}] \in \mathbb{R}^{p_d \times R}$.

Lema **1** *Si un tensor $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ admite una descomposición de rango R entonces, las operaciones mode-d matricization y vectorización se pueden reescribir como:*

$$\mathbf{B}_{(d)} = \mathbf{B}_d(\mathbf{B}_D \odot \cdots \odot \mathbf{B}_{d+1} \odot \mathbf{B}_{d-1} \odot \cdots \odot \mathbf{B}_1) \quad y \quad vec(\mathbf{B}) = (\mathbf{B}_D \odot \cdots \odot \mathbf{B}_1)\mathbf{1}_R$$

Note que, para el caso en el que la descomposición sea de rango 1, se concluye que $\mathbf{B} = \vec{\beta}_1 \circ \vec{\beta}_2 \circ \cdots \circ \vec{\beta}_D$, donde $\vec{\beta}_d \in \mathbb{R}^{p_d}$. Como consecuencia se tendrá que

$$\mathbf{B} = [\vec{\beta}_1, \cdots, \vec{\beta}_D] \quad y \quad vec(\mathbf{B}) = (\mathbf{B}_D \odot \cdots \odot \mathbf{B}_1)\mathbf{1}_1 = (\vec{\beta}_D \odot \cdots \odot \vec{\beta}_1) = \vec{\beta}_D \otimes \cdots \otimes \vec{\beta}_1.$$

3.2. Modelo de Regresión Tensor

Con el fin de introducir el modelo de regresión tensor se enuncia el GLM para el caso en el que la variable independiente sea multidimensional, la cual es una extensión natural del GLM clásico con una covariable vectorial [24]. Este modelo supone que se tienen variables aleatorias independientes $Y_1, Y_2, ..., Y_N$ cuyas distribuciones pertenecen a la familia exponencial, $f(y_i; \eta_i)$ para $i = 1, ..., N$, definidas en el capitulo anterior. En la especificación del modelo se supone que $\mathbf{X}_i$ es un arreglo multidimensional, $\mathbf{X}_i \in \mathbb{R}^{p1\times p2\times\cdots\times p_D}$, $\mathbf{Z_i} \in \mathbb{R}^{p_0}$ para $i = 1, \cdots, N$ y el modelo se sitúa en el margo teórico de los GLM, mencionados en el capitulo 2, cuya función de enlace viene dada por:

$$g(\mu_i) = \vec{\gamma}'\mathbf{Z_i} + \langle \mathbf{B}, \mathbf{X_i} \rangle. \tag{3-1}$$

Los parámetros desconocidos son $\vec{\gamma}' \in \mathbb{R}^{p_0}$ y $\mathbf{B} \in \mathbb{R}^{p1\times p2\times\cdots\times p_D}$, este último denominado coeficiente función-imagen. Se puede observar que $\eta_i(\mathbf{X_i}, \mathbf{Z_i}; \vec{\gamma}, \mathbf{B}) = (d_0')^{-1}[-g^{-1}(\vec{\gamma}'\mathbf{Z_i} + \langle \mathbf{B}, \mathbf{X_i} \rangle)]$, de acuerdo a las nociones dadas en el capitulo 2. Los diferentes enfoques que han surgido difieren principalmente en la forma en como se construye el producto punto $\langle U, V \rangle$ en la ecuación 3-1 para $U \in \mathbb{R}^{p1\times p2\times\cdots\times p_D}$ y $V \in \mathbb{R}^{p1\times p2\times\cdots\times p_D}$. Se define entonces,

$$\langle \mathbf{X}, \mathbf{B} \rangle := \langle vec(\mathbf{X}), vec(\mathbf{B}) \rangle = vec(\mathbf{B})'vec(\mathbf{X}) = \sum_{i_1}\sum_{i_2}\cdots\sum_{i_D} b_{i_1,i_2,\cdots,i_D}\mathbf{X}_{i_1,i_2,\cdots,i_D},$$

la cual es una generalización del producto interno entre matrices (véase en [17]). De esta manera, la ecuación 3-1 se reescribe como:

$$g(\mu_i) = \vec{\gamma}'\mathbf{Z_i} + vec(\mathbf{B})'vec(\mathbf{X}) \tag{3-2}$$

En este modelo $\mathbf{B}$ tiene el mismo número de parámetros de $\mathbf{X}$, es decir, $\prod_{d=1}^{D} p_d$, el cual tiene dimensión en exceso más grande que el tamaño de muestra usual, esto significa un problema pues la idea natural es aproximar $\mathbf{B}$ con menos parámetros (se examina esta situación mas adelante). El modelo de regresión tensor surge para suplir esta necesidad, suponiendo primero que $\mathbf{B}$ admite una descomposición de rango 1 y recordando que para este caso $vec(\mathbf{B}) = \vec{\beta}_D \otimes \cdots \otimes \vec{\beta}_1$. En este sentido, la ecuación 3-2 se establece como:

$$g(\mu_i) = \alpha + \vec{\gamma}'\mathbf{Z}_i + \langle \mathbf{B}, \mathbf{X}_i \rangle = \alpha + \vec{\gamma}'\mathbf{Z}_i + (\vec{\beta}_D \otimes \cdots \otimes \vec{\beta}_1)'vec(\mathbf{X}_i). \tag{3-3}$$

Si ocurre el caso en el que $\mathbf{B}$ admita descomposición de rango R, el modelo de regresión tensor se enmarca bajo las mismas condiciones mencionadas anteriormente, con la particularidad de que su función de enlace esta dada por la reescritura de 3-2 de la siguiente manera:

$$\begin{aligned} g(\mu_i) &= \alpha + \vec{\gamma}'\mathbf{Z}_i + \left\langle \sum_{r=1}^{R} \vec{\beta}_1^{(r)} \circ \vec{\beta}_2^{(r)} \circ \cdots \circ \vec{\beta}_D^{(r)}, \mathbf{X}_i \right\rangle \\ &= \alpha + \vec{\gamma}'\mathbf{Z}_i + \langle (\mathbf{B}_D \odot \cdots \odot \mathbf{B}_1)1_R, vec(\mathbf{X}_i) \rangle, \end{aligned} \tag{3-4}$$

Este modelo da solución al problema de dimensionalidad ya que el número de parámetros del mismo es $p_0 + R\sum_d p_d$, el cual es menor a $p_0 + \prod_{d=1}^{D} p_d$. La diferencia es significativa cuando la calidad de la imagen se mejora, para ilustrar esto se muestra un bosquejo del tamaño en la tabla a continuación.

D	p_0	p_1	p_2	$p_0 + \prod_{d=1}^{D} p_d$	$p_0 + R\sum_d p_d \quad (R \approx 3)$
2	3	28	28	$3 + 784 = 787$	$3 + R * 56 \approx 115$
2	3	64	64	$3 + 4096 = 4099$	$3 + R * 128 \approx 259$
2	3	128	128	$3 + 16384 = 16387$	$3 + R * 256 \approx 515$

Tabla 3-1.: Cambio del número de parámetros a estimar, sin reducción de dimensionalidad (columna 5) y con reducción de dimensionalidad (columna 6), para un modelo de imágenes bidimensionales ($D = 2$) de tamaño 28×28 píxeles, 64×64 píxeles y 128×128 píxeles. Se fija $p_0 = 3$ y un rango $R = 2$.

Finalmente, es importante mencionar que para la identificabilidad del modelo, la regresión lineal clásica requiere que $vec(\mathbf{X}_i) \in \mathbb{R}^{\prod_{d=1}^{D} p_d}$ sea linealmente independiente para de esta forma estimar todos los parámetros, lo que requiere un tamaño de muestra $n \geq \prod_{d=1}^{D} p_d$. El requisito sobre el tamaño de la muestra se reduce en gran medida al imponer la estructura del *lema* 1 en las matrices, ya que solo se requiere independencia lineal de los vectores colapsados $[\mathbf{B}_1, \cdots, \mathbf{B}_D]' vec(\mathbf{X}_i) \in \mathbb{R}^{R(\sum_{d=1}^{D} p_d - D + 1)}$, de esta manera se requiere que $n \geq R(\text{máx}_{d=1,\cdots,D}\, p_d)$.

3.3. Método de estimación

Esta sección se muestra el algoritmo de optimización del modelo de regresión tensor. La estimación de los parámetros se hace por medio de máxima verosimilitud, de modo que si se tienen n datos independientes e idénticamente distribuidos dentro de la familia exponencial $\{(y_i, x_i, \vec{z}_i), i = 1, \cdots, n\}$, entonces la función de log-verosimilitud está dada por:

$$\ell(\alpha, \vec{\gamma}, \mathbf{B}_1, \cdots, \mathbf{B}_D) = \sum_{i=1}^{n} y_i \eta_i - d_0(\eta_i) + \sum_{i=1}^{n} d(y_i),$$

en el cual, la variable η_i está relacionada con los parámetros de la regresión $[\alpha, \vec{\gamma}, \mathbf{B}_1, \cdots, \mathbf{B}_D]$ por la función de enlace $g(.)$ del modelo de Regresión Tensor Generalizado (es decir $\eta_i(\mathbf{X_i}, \mathbf{Z}_i; \vec{\gamma}, \mathbf{B}) = (d_0')^{-1}[-g^{-1}(\vec{\gamma}'\mathbf{Z}_i + \langle \mathbf{B}, \mathbf{X_i} \rangle)]$). Se aprecia que la función $g(\mu_i)$ no es lineal en $(\mathbf{B}_1, \cdots, \mathbf{B}_D)$ conjuntamente, pero sí es lineal en $\mathbf{B}_d$ individualmente. Esto sugiere actualizar $(\alpha, \vec{\gamma})$ y $\mathbf{B}_d$ alternadamente, según el siguiente algoritmo porpuesto en [24]:

Inicializar: $(\alpha^{(0)}, \vec{\gamma}^{(0)}) = argmax_{\alpha, \vec{\gamma}} \ell(\alpha, \vec{\gamma}, \vec{0}, \cdots, \vec{0})$, $\mathbf{B}_d^{(0)} \in \mathbb{R}^{p_d \times R}$ para $d = 1, \cdots, D$.
repíta

para $d = 1, \cdots, D$ **haga**

$$\mathbf{B}_d^{(t+1)} = argmax_{\mathbf{B}_d} \ell(\alpha^{(t)}, \vec{\gamma}^{(t)}, \mathbf{B}_1^{(t+1)}, \cdots, \mathbf{B}_{d-1}^{(t+1)}, \mathbf{B}_d,$$

$$\mathbf{B}_{d+1}^{(t)}, \cdots, \mathbf{B}_D^{(t)})$$

y para

$$(\alpha^{(t+1)}, \vec{\gamma}^{(t+1)}) = argmax_{\alpha,\vec{\gamma}} \ell(\alpha, \vec{\gamma}, \mathbf{B}_1^{(t+1)}, \cdots, \mathbf{B}_D^{(t+1)})$$

hasta $\ell(\vec{\theta}^{(t+1)}) - \ell(\vec{\theta}^{(t)}) < \epsilon$.

De acuerdo al anterior desarrollo se observa que, el problema se vuelve entonces una regresión GLM tradicional, con Rp_d parámetros, y el proceso de estimación se convierte ahora en una secuencia de optimización GLM de baja dimensionalidad. Es importante mencionar que el anterior algoritmo supone el rango R conocido al estimar $\mathbf{B}$, lo que será relevante para el uso de la metodología.

Para finalizar este capítulo, a continuación se describe brevemente el modelo de regresión tensor que se usará para los procesos de estimación y clasificación. Dadas las variables aleatorias independientes $Y_1, Y_2, ..., Y_N$ (variables categóricas 0 y 1 que representan estado del paciente) cuyas distribuciones, $Y_i \sim Ber(\pi_i)$ para $i = 1, ..., N$ y sea $\mathbf{X}_i$ ($\mathbf{X}_i \in \mathbb{R}^{p1 \times p2}$ imágenes diagnosticas) un arreglo 2-dimensional.

Suponemos que $\mathbf{B}$ tiene descomposición de rango R, para $\beta_i^{(r)} \in \mathbb{R}^{p_i}$ con $i = 1, 2$, es decir que el arreglo de coeficientes satisface:

$$\mathbf{B} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \beta_2^{(r)},$$

más aún, si definimos $B_1 = [\beta_1^{(1)}, \cdots, \beta_1^{(R)}] \in \mathbb{R}^{p_1 \times R}$ y $B_2 = [\beta_2^{(1)}, \cdots, \beta_2^{(R)}] \in \mathbb{R}^{p_2 \times R}$ entonces $Vec\mathbf{B}$ se reescribe como:

$$Vec\mathbf{B} = (\mathbf{B}_2 \odot \mathbf{B}_1)1_R.$$

Finalmente, el modelo los GLM tensorial se establece como aquel en el cual la función de enlace viene dada por:

$$g(\pi_i) = logit(\pi_i) = \alpha + \vec{\gamma}' \mathbf{Z}_i + \left\langle \sum_{r=1}^{R} \vec{\beta}_1^{(r)} \circ \vec{\beta}_2^{(r)}, \mathbf{X}_i \right\rangle = \alpha + \vec{\gamma}' \mathbf{Z}_i + \langle (\mathbf{B}_{\odot} \mathbf{B}_1) 1_R, vec(\mathbf{X}_i) \rangle,$$

De esta manera, el objetivo es estimar el arreglo de coeficientes $\mathbf{B}$ estimando B_1 y B_2, con el modelo tensor general ($R > 0$). Optimizando la función de log-verosimilitud dada por:

$$\ell(\alpha, \vec{\gamma}, \mathbf{B}_1, \mathbf{B}_2) = \sum_{i=1}^{n} y_i \eta_i - d_0(\eta_i) + \sum_{i=1}^{n} d(y_i),$$

con $d_0(\eta_i) = -log(1 + e^{\eta_i})$. De esta manera $d_0'(\eta_i) = -\frac{1}{1+e^{-\eta_i}}$, $d_0'(\eta_i)^{-1} = log\left(\frac{-\eta_i}{1+\eta_i}\right)$ y $\eta_i = (d_0')^{-1}[-g^{-1}(\vec{\gamma}'\mathbf{Z}_i + \langle \mathbf{B}, \mathbf{X_i} \rangle)] = log\left(\frac{\pi_i}{1-\pi_i}\right) = logit(\pi_i)$.

4. Funcionales, variación total y espacio de variación acotada

El capítulo que se desarrolla a continuación presenta algunos resultados del análisis de datos funcionales que surgen intentando suplir los métodos estadísticos tradicionales, los cuales pueden presentar dificultades a medida que p (el número de variables) aumenta y resulta ser mucho más grande en comparación a n (el número de individuos). Formalmente, un dato funcional f es una observación de una variable aleatoria funcional $F \in \mathbb{H}$, en el cual $\mathbb{H}$ es un espacio infinito dimensional.

Definición **6** *Sea f una función, se dice que f es p-integrable sobre un dominio S si*

$$\left(\int_S |f(t)|^p dt\right)^{\frac{1}{p}} < \infty.$$

Se denota $\mathcal{L}^p(S)$ al espacio del las funciones p-integrables. De esta manera $\mathcal{L}^1(S)$ corresponde al espacio de las funciones integrables (véase en [18]).

El espacio de las funciones $\mathcal{L}^P(S)$ es una espacio vectorial, pero no resulta ser un espacio de Banach (normado y completo). Para suplir esta carencia, se debe definir, sobre este espacio, cierta relación de equivalencia de tal manera que el espacio obtenido por ella sí constituya un espacio vectorial de Banach. La relación de equivalencia R sobre $\mathcal{L}^P(S)$ que logra este objetivo, es aquella en la que las clases de equivalencias están formadas por funciones que coinciden casi en todas partes y está dada por:

$$fRg \text{ si, y solo si, } \int_S |f(t) - g(t)|^p dt = 0.$$

Se prueba que efectivamente esta es una relación de equivalencia, y se define $L^p = \mathcal{L}^p/R$ el espacio vectorial cuyos elementos son las clases de equivalencia dadas por R (véase en [18]); a menudo, el campo de estudio se establece como el espacio L^2.

Ahora bien, lo anterior se define sobre un enfoque de tipo determinista, es por esta razón que se deben ajustar las nociones dadas para el caso en el que los elementos de L^2 sean aleatorios (véase en [5] y [7]). Al igual que una variable aleatoria, una función aleatoria se define en un espacio de probabilidad, por ejemplo Ω, en donde para cada $\omega \in \Omega$, $f(\omega)$ es una función

determinista. Se asume que todas las realizaciones $f(\omega)$, $\omega \in \Omega$ son elementos de $\mathcal{L}^2$, lo cual significa que para cada $\omega \in \Omega$:

$$||f(\omega)||^2 = \int_S \{f(\omega)(t)\}^2 dt < \infty.$$

De modo que, por definición [7] f es integrable si $E(||f||) = E\left(\int_S \{f(t)\}^2 dt\right)^{\frac{1}{2}} < \infty$, y es cuadrado integrable si

$$E(||f||^2) = E\left(\int_S \{f(t)\}^2 dt\right) < \infty.$$

Se puede notar que, decir que una función aleatoria es cuadrado integrable es equivalente a decir que tiene un segundo momento finito. De manera análoga, dada otra función aleatoria g definida sobre Ω se dice que:

$$fRg \text{ si, y solo si, } E\left(\int_S \{f(t) - g(t)\}^2 dt\right) = 0.$$

Frecuentemente se supone que $\mathbb{H}$ corresponde al espacio de funciones aleatorias de valor real cuadrado integrables sobre un dominio $S \subseteq \mathbb{R}$ ($\mathcal{L}^2$). Paralelamente, f también puede ser situado en la teoría de procesos estocásticos, ya que el dato funcional se puede ver como una trayectoria del proceso $\{f(t), t \in T\}$, donde T es un intervalo de la recta real.

A menudo se cuenta con observaciones discretas de la función y no con el objeto funcional completo, por esta razón la primera tarea consiste en convertir los valores discretos observables en una función. Para reconstruir la forma funcional de las curvas de la muestra a partir de observaciones discretas se pueden usar dos caminos: el primero es asumir que se obtienen los valores discretos sin error y ajustar una función de naturaleza determinista, que en la práctica corresponde a un proceso de interpolación. El segundo camino es, asumir que los puntos poseen un error no observable y ajustar un funcional que implícitamente contenga un ruido. Este proceso se denomina suavizamiento o regularización, y consiste en asumir que la curva de la muestra se puede aproximar por un funcional que pertenece a un espacio generado por una base $\{\phi_1(t), \cdots, \phi_P(t)\}$ de tal manera que

$$x(t) \approx \sum_i^P c_i \phi_i(t),$$

lo cual puede ser escrito en forma matricial como $x(t) = \vec{c}'\vec{\phi}(t)$, donde $\vec{c} = (c_1, \cdots, c_P)'$ y $\vec{\phi}(t) = (\phi_1(t), \cdots, \phi_p(t))'$. Si adicionalmente se considera que las curvas se observan con error, entonces en realidad, la muestra de m observaciones discretizadas [13] estaría compuesta de:

$$y_k = x(t_k) + \epsilon_k = \sum_k^P c_i \phi_i(t_k) + \epsilon_k, \qquad con \qquad k = 1, \cdots, m.$$

Lo primero que se debe establecer es el conjunto de funciones base. Si se considera el caso particular en el que la curva observada con ruido sea suave, una de las bases comúnmente usadas se denomina B-Spline, constituida por funciones polinómicas a trozos, sobre las que se imponen restricciones en los puntos de unión, los cuales se conocen como nodos. Luego de elegir la base, para obtener los coeficientes de la regularización, se puede usar un método de aproximación suave como los que se exhiben a continuación:

1. Regresión spline: aproxima el coeficiente $\vec{c}$ optimizando por el criterio de mínimos cuadrados [5]. Minimizando

$$MSE(\vec{y}|\vec{c}) = \sum_{k=1}^{m}(y_k - \sum_{i=1}^{P} c_i\phi_i(t_k))^2 = (\vec{y} - \Phi\vec{c})'(\vec{y} - \Phi\vec{c}),$$

   con Φ una matriz de tamaño $m \times P$ que contiene $\phi_p(t_k)$, obteniendo como estimador $\hat{\vec{c}} = (\Phi'\Phi)^{-1}\Phi'\vec{y}$.

2. Suavizamiento spline: introduce un término de penalización $\int[D^2x(s)]^2ds = \vec{c}'R_2\vec{c}$ que mide la rugosidad de una función (representa la medida de curvatura de la función). La matriz R_2 contiene las integrales de producto exterior de vectores dada por $\int D^2\phi(s)D^2\phi'(s)ds$. Así, se minimiza

$$CPMSE_2(\vec{y}|\vec{c}) = (\vec{y} - \Phi\vec{c})'(\vec{y} - \Phi\vec{c}) + \lambda\vec{c}'R_2\vec{c},$$

   de lo que resulta $\hat{\vec{c}} = (\Phi'\Phi + \lambda R_2)^{-1}\Phi'\vec{y}$ como estimador. La variable λ se conoce como el parámetro de suavizamiento [5].

3. P-splines: la rugosidad considerada en el método anterior trae consigo un problema computacional debido al cálculo de las integrales del cuadrado de la segundas derivadas para la matriz R_2. P-splines considera un enfoque más sencillo definiendo la rugosidad de una función como la suma de las diferencias de orden d al cuadrado. Este tipo de penalización depende de la base usada y sólo funciona si los puntos de la muestra están igualmente espaciados [5]. Ahora bien, cabe resaltar que el método de P-splines minimiza el error de mínimos cuadrados penalizados:

$$DPMSE_d(\vec{y}|\vec{c}) = (\vec{y} - \Phi\vec{c})'(\vec{y} - \Phi\vec{c}) + \lambda\vec{c}'P_d\vec{c},$$

   donde $P_d = (\triangle^d)'\triangle^d$ con $\triangle^d$ la representación matricial del operador de diferencias de orden d. En este caso, se obtiene como estimador de los coeficientes

$$\hat{\vec{c}} = (\Phi'\Phi + \lambda P_d)^{-1}\Phi'\vec{y}.$$

Las superficies bidimensionales serán los objetos base en la construcción del modelamiento funcional con imágenes. Si se tiene una imagen de tamaño 64×64 píxeles en una escala de

grises, este objeto puede ser visto como una superficie aplicando la siguiente función (campo escalar):

$$\begin{aligned} f\colon [0,63]\times[0,63] &\longrightarrow [0,255] \\ (s,t) &\longmapsto z \end{aligned}$$

donde (s,t) corresponde a la localización del píxel dentro de la imagen y z corresponde al valor en el rango de color. En el rango, 0 corresponde a blanco y 255 a negro.

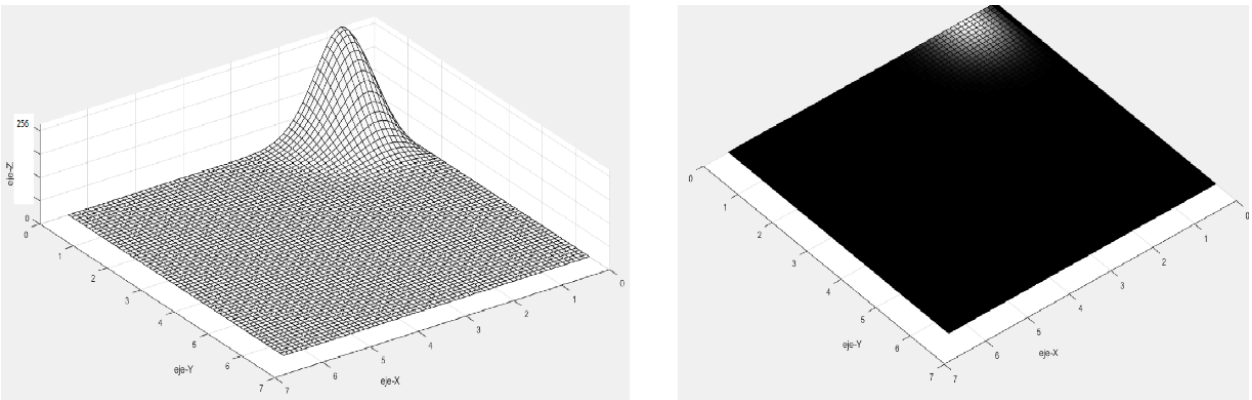

Figura 4-1.: Representación de una imagen 64×64 píxeles (derecha) como una superficie continua (izquierda)

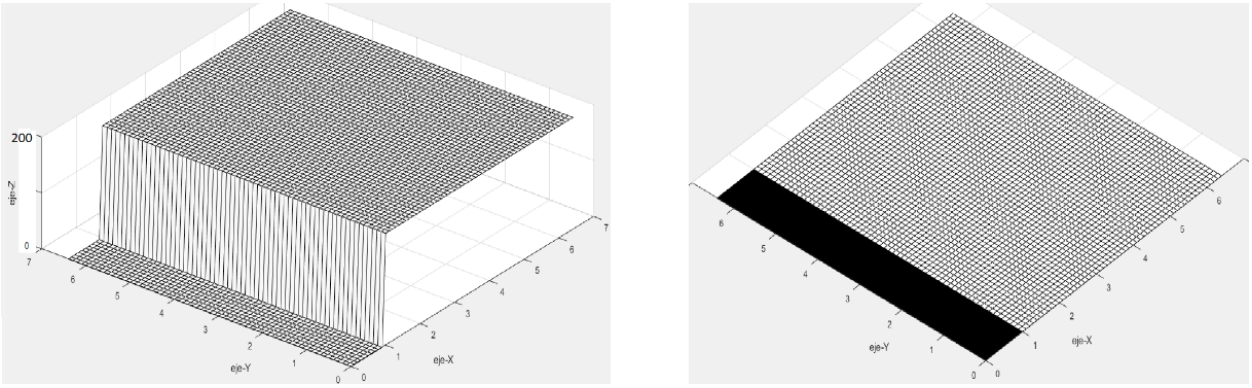

Figura 4-2.: Representación de una imagen de tamaño 64×64 píxeles con un cambio de color abrupto (derecha) como una superficie discontinua (izquierda)

Las técnicas de suavizamiento vistas anteriormente 4 consideran curvas o superficies con la característica de ser suaves (figura **4-1**), por esta razón hacen un fuerte uso de la derivabilidad, lo que produce que el término de penalización para el suavizamiento no sea demasiado sensible y por ende, no logre capturar los picos o saltos que algunas imágenes puedan tener (figura **4-2**) las cuales son frecuentes encontrar en la practica. En 1992, Rudin,

Osher y Fatemi formulan el modelo ROF (véase en [17]), cuyo objetivo principal es reducir la variación total de una señal aproximando la señal original. Este modelo corresponde al modelo de regularización de la variación total, en el cual se preserva los bordes y al mismo tiempo se elimina el ruido en regiones planas. La siguientes definiciones son extraídas de [18] y [6].

Definición **7** *Sea u una función continua de valor real definida sobre $T \subseteq \mathbb{R}$. La variación total de u en T, denotada por $||u||_{TV}$, se define como:*

$$||u||_{TV} = \sup_{p \in \mathcal{P}} \left\{ \sum_{i=0}^{np-1} |u(x_{i+1}) - u(x_i)| \right\},$$

donde $\mathcal{P}$ es una partición de T en la cual $x_i \leq x_{i+1}$ para $0 \leq n \leq np-1$.

Si u es diferenciable y su derivada es Riemann integrable, su variación total es la componente vertical de la longitud de arco de su gráfica, es decir,

$$||u||_{TV} = \int_S \sqrt{(\partial_s u(s))^2} ds = \int_S |u'(s)| ds.$$

La anterior definición se puede extender de manera análoga para funciones cuyo dominio esta contenido en el espacio $\mathbb{R}^2$.

Definición **8** *Sea u una función integrable en $S \subseteq \mathbb{R}^2$, $u \in L^1(S)$. La variación total de u en S, denotada por $||u||_{TV}$, se define como:*

$$||u||_{TV} = \sup \left\{ \int_S u(s,t)\, div f(s,t)\, ds\, dt, \quad con \quad f \in C^\infty(S, \mathbb{R}^2) \;\; y \;\; ||f||_\infty \leq 1 \right\},$$

donde $C^\infty(S, \mathbb{R}^2)$ corresponde al espacio de campos vectoriales con valor en $\mathbb{R}^2$ infinitamente diferenciables con soporte compacto [1] S y $f(s,t) = (f_1(s,t), f_2(s,t))$, $||f||_\infty = \sup_{(s,t)\in S} |f(s,t)|$, $div f(s,t) = \frac{\partial}{\partial u} f_1(s,t) + \frac{\partial}{\partial v} f_2(s,t)$.

Si u es diferenciable en S, entonces

$$||u||_{TV} = \int_S |\Delta u| ds dt = \int_S \sqrt{(\partial_s u)^2 + (\partial_t u)^2} ds\, dt,$$

y en este caso, u pertenece al espacio de Sobolev [2].

[1] El soporte compacto es el menor cerrado que contiene al conjunto de puntos en el dominio donde la función no es cero.

[2] El espacio de Sobolev es denotado por $W^{1,1}$ y resulta ser un sub espacio vectorial de L^2 (el espacio de funciones con derivadas de primer orden integrables).

Definición **9** *Sea u una función con dominio $S \subseteq \mathbb{R}^2$. Decimos que u tiene variación acotada en S si $||u||_{TV} < \infty$ y denotamos $BV(S)$ el espacio vectorial de funciones de variación acotada en S para funciones diferenciables.*

Una imagen f se puede descomponer en una componente u y una parte de textura o ruido v. Ahora, una tarea importante en el procesamiento de imágenes es la restauración o recuperación de la imagen real u. La regularización de las superficies dos dimensionales consiste en resolver el problema lineal inverso $f(s,t) = u(s,t) + v(s,t)$, donde $f \in L^2(S)$, $u \in BV(S)$ (son campos escalares), $S \subseteq \mathbb{R}^2$ y v pertenece a un espacio de Banach adecuado. Debido a que $||u||_{TV} < \infty$, una de las técnicas mas populares y efectivas consiste en resolver el problema de minimización: $\text{mín}_{u \in BV(S)} ||u||_{TV}$ sujeto a uRf, lo que es equivalente a encontrar el mínimo de la función de optimización dada por:

$$\int_S |\Delta u| \, ds \, dt + \frac{\mu}{2}||f-u||_{L^2}^2 \; = \int_S |\Delta u| \, ds \, dt + \frac{\mu}{2}\int_S |f-u|^2 ds \, dt. \tag{4-1}$$

La razón principal del uso de las funciones de variación acotada para la solución del problema esta dado por su ventaja en la retención de vértices o picos y contornos en una imagen. En la práctica, un dato funcional es observado en un número finito de puntos, así, como un abuso de notación, usaremos $u \in \mathbb{R}^{N\times N}$ y $f \in \mathbb{R}^{N\times N}$ para denotar las imágenes discretas con y sin ruido, respectivamente. Ahora bien, aunque u y f son elementos del espacio $\mathbb{R}^{N\times N}$, estos no son tratados bajo la teoría del análisis multivariado. Esto se debe a que el dominio donde los datos funcionales son medidos puede diferir, y para cada observación el orden de sus componentes contiene una cantidad importante de información, por lo cual, no es permitida la permutación de sus elementos. Con respecto a esto, se define el gradiente discreto de un elemento de $\mathbb{R}^{N\times N}$ [1]:

Definición **10** *Dado $u \in \mathbb{R}^{N\times N}$, el gradiente discreto de u, $\nabla : BV(S) \longmapsto \mathbb{R}^{N\times N\times 2}$ está definido como:*

$$(\nabla u)_{jk} = \begin{cases} (u_{j+1,k} - u_{jk}, u_{j,k+1} - u_{jk}) & si \quad 1 \le j,k \le N-1, \\ (0, u_{j,k+1} - u_{jk}) & si \quad j = N, \; 1 \le k \le N-1, \\ (u_{j+1,k} - u_{jk}, 0) & si \quad k = N, \; 1 \le j \le N-1, \\ (0,0) & si \quad j = N = k. \end{cases}$$

Con lo anterior en mente, se menciona el siguiente ejemplo que ilustra el comportamiento de los conceptos introducidos.

Ejemplo **5** *Se consideran $N = 2$ y u la imagen dada en la figura **4-3**, que se representa por una malla de puntos discretos definidos en la matriz $u = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, de esta*

manera:

$$(\nabla u)_{11} = (u_{21} - u_{11}, u_{12} - u_{11}) = (3-1, 2-1) = (2,1)$$
$$(\nabla u)_{21} = (0, u_{22} - u_{21}) = (0, 4-3) = (0,1)$$
$$(\nabla u)_{12} = (u_{22} - u_{12}, 0) = (4-2, 0) = (2,0)$$
$$(\nabla u)_{22} = (0,0)$$

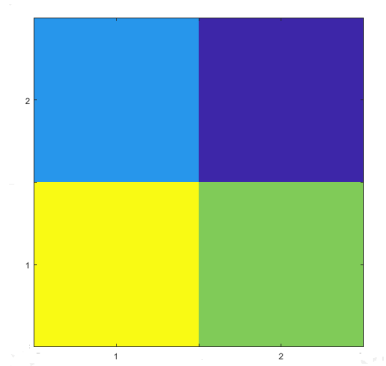

Figura 4-3.: Imagen de tamaño 2×2 en la que cada entrada de posición 11, 12, 21 y 22 toma los valores $1 = Azul\ claro$, $2 = Azul\ oscuro$, $3 = Amarillo$ y $4 = Verde$, respectivamente.

Se puede notar que

$$(\nabla u)_{jk} = ((\nabla u)_{jk,1}, (\nabla u)_{jk,2})^T = \left(\begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix} \right),$$

además, (∇u) estima el cambio de color de un píxel a otro, en ambas direcciones (vertical y horizontal). Es importante notar que, por ejemplo, la transición de azul claro a amarillo tiene una medida de 2, mientras que el cambio de azul claro a azul oscuro tiene una medida de 1. Ahora bien, la variación total discreta dependerá de la norma que se use sobre el espacio $\mathbb{R}^{N \times N}$.

Definición **11** *Dado $u \in \mathbb{R}^{N \times N}$ se define la norma $||u||^*_{TV}$, denominada la versión anisotrópica de la variación total como [1]:*

$$||u||^*_{TV} = ||\nabla u||_1 = \sum_j \sum_k \sqrt{(\nabla u)^2_{jk,1} + (\nabla u)^2_{jk,2}}.$$

Definición **12** *Dado $u \in \mathbb{R}^{N \times N}$ se define la norma $||u||_{TV}$ denominada la versión isotrópica de la variación total como [1]:*

$$||u||_{TV} = ||\nabla u||_2 = \sum_j \sum_k \{|\nabla u)_{jk,1}| + |\nabla u)_{jk,2}|\}.$$

Es posible concluir que estas dos normas son equivalentes ($||u||^*_{TV} \equiv ||u||_{TV}$ [1]) a partir de la siguiente relación:

$$\frac{1}{\sqrt{2}}||u||^*_{TV} \leq ||u||_{TV} \leq \sqrt{2}||u||^*_{TV}.$$

Por facilidad se usará $||u||^*_{TV}$ para el desarrollo del modelo. De esta forma, siendo $f \in \mathbb{R}^{N\times N}$ la imagen con ruido, el modelo de variación total discreto para la reconstrucción de imágenes resulta ser:

$$||u||^*_{TV} + \frac{\mu}{2}||u-f||^2 = \sum_j \sum_k \{|(\nabla u)_{jk,1}| + |(\nabla u)_{jk,2}|\} + \frac{\mu}{2}||u-f||_2^2. \tag{4-2}$$

Si se supone que todas las imágenes 2-D están vectorizadas por columnas, concisamente $u \in \mathbb{R}^n$ con $n = N^2$, entonces $||u||^*_{TV}$ puede ser expresado como $\sum_i ||D_i u||_2$ donde $D_i\beta \in \mathbb{R}^2$ es el gradiente discreto de u en el i-ésimo píxel. Para nuestro ejemplo anterior, se considera $N = 2$, $vec(u) := \vec{u} = (u_{11}, u_{12}, u_{21}, u_{22})^T = (1, 2, 3, 4)^T$ y

$$D_1 = \begin{pmatrix} -1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 \end{pmatrix}, D_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}, D_3 = \begin{pmatrix} 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ y } D_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Se puede observar que $D_1\vec{u} = (\nabla u)_{11}$, $D_2\vec{u} = (\nabla u)_{12}$, $D_3\vec{u}) = (\nabla u)_{21}$ y $D_4\vec{u} = (\nabla u)_{22}$ (véase en [1]). Si se supone también que existe $A \in \mathbb{R}^{m\times n}$ (con $m < n$) matriz de medida, que representa un operador de desenfoque o convolución, entonces para $\vec{f} \in \mathbb{R}^m$, la ecuación 4-2 resulta ser:

$$\sum_i ||D_i\vec{u}||_2 + \frac{\mu}{2}||A\vec{u} - \vec{f}||_2^2. \tag{4-3}$$

Adicionalmente, si se introduce una variable auxiliar $\vec{w}_i \in \mathbb{R}^2$ para cada pixel, para que $D_i\vec{u}$ se desacople de su término no diferenciable $\sum_i ||.||$, entonces se obtiene una aproximación del problema 4-3 y, por consiguiente, el problema se reduce a encontrar el mínimo ($\text{mín}_{u,w}$) de la función de optimización:

$$\sum_i ||\vec{w}_i||_2 + \frac{\mu}{2}||A\vec{u} - \vec{f}||_2^2, \quad sujeto\ a \quad \vec{D}_i\vec{u} = \vec{w}_i. \tag{4-4}$$

Con lo anterior en mente, aplicando el método TVAL3 *(Total variation Minimizationby Augmented LagrangianAlternating DirectionAlgorithm)* enunciado en el anexo A, para resolver 4-4, se reescribe la función de optimización como:

$$\sum_i (||\vec{w}_i||_1 + \frac{\delta^{(k)}}{2}||D_i\vec{\beta} - w_i||^2 - \lambda_i'(D_i\vec{u} - w_i)) + \frac{\mu}{2}||A\vec{u} - \vec{f}||^2. \tag{4-5}$$

El algoritmo alternativo, presentado por [1], en el paso k consiste en ($\delta^{(k)} = 2^5$ y $\mu = 2^8$):

1. Inicialice $\vec{u}^{(0)}$ y $\lambda_i^{(0)}$
2. Dado $\vec{u}^{(k)}$ y $\lambda_i^{(k)}$ calcule $w^{(k+1)}$ como,

$$\vec{w}_i^{(k+1)} = \text{máx}\left\{||D_i\vec{u}^{(k)} - \frac{\lambda_i}{\delta}|| - \frac{1}{\delta}, 0\right\} \frac{D_i\vec{u}^{(k)} - \frac{\lambda_i}{\delta}}{||D_i\vec{u}^{(k)} - \frac{\lambda_i}{\delta}||}$$

3. Fije $\vec{w}_i = \vec{w}_i^{(k+1)}$ y $\lambda_i^{(k)}$ aplique descenso del gradiente para:

$$-\lambda_i'(D_i - \vec{w}_i) + \frac{\delta}{2}||D_i\vec{u} - \vec{w}_i||^2 + \frac{\mu}{2}||A\vec{u} - \vec{f}||^2$$

   * Solución: $\vec{u}^{(k+1)} = \alpha D'\lambda_i + \delta D'(D\vec{u}^{(k)} - w_i^{(k)}) + \mu A'(A\vec{u}^{(k)} - f)$

4. Dado $= \vec{u}^{(k+1)}$ y $\vec{w}_i = w_i^{(k+1)}$, actualizamos el multiplicador de Lagrange por:

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} - \delta(D_i\vec{u} - w_i)$$

5. Acepte el actual $\vec{u}$ como el parámetro inicial para ejecutar nuevamente el ciclo, e itere hasta convergencia.

En este punto, es clave mencionar que el proceso de restauración de imágenes que se menciona anteriormente es importante no solo por su enorme uso en el procesamiento de imágenes. En el desarrollo del trabajo, el método TVAL3 adquiere una enorme importancia, pues este procedimiento, bajo ciertos ajustes, se emplea para estimar los coeficientes óptimos en los modelos de regresión lineal funcional con respuesta escalar, y más aún, en los modelos de regresión lineal generalizados de tipo escalar-función que se describen en la siguiente sección.

4.1. Modelo Lineal Escalar-Función de valor real

La regresión lineal es una técnica común de análisis estadístico de datos. Se utiliza para determinar la medida en que existe una relación lineal entre una variable dependiente y una o más variables independientes. Los modelos de regresión funcional se subdividen en tres categorías amplias [7], dependiendo de si las respuestas o los regresores, o ambos, son funcionales:

1. Regresión escalar-función: los regresores son funcionales, pero la respuesta es escalares.

$$Y_i = \int \beta(s)X_i(s)ds + \epsilon_i.$$

   El parámetro es la función de regresión β.

2. Regresión función-escalar: la respuesta es funcional, pero los regresores son escalares.

$$Y_i(t) = \sum_{k=1}^{p} x_{ik}\beta_k(t) + \epsilon_i(t).$$

   Los parámetros son las funciones coeficientes β_k.

3. Regresión función-función: la respuesta es funcional, y también lo son los regresores.

$$Y_i(t) = \int \beta(t,s)X_i(s)ds + \epsilon_i(t).$$

   El parámetro es el kernel (superficie) β.

Es importante resaltar que en todos los casos los parámetros son objetos infinitos dimensionales que deben estimarse a partir de una muestra finita. Por lo tanto, la estimación a menudo implica restricciones adicionales, ya sea imponiendo condiciones de suavidad, o restringiendo la acción de los operadores correspondientes a los subespacios apropiados. En el presente trabajo los modelos de regresión que se usaran serán de tipo escalar-función.

El modelo lineal escalar función supone que la variable respuesta y es escalar, mientras que las variables regresoras son de tipo funcional, como se había mencionado. De esta forma, para $j = 1, 2, ..., m$ el modelo matemático sigue la siguiente relación:

$$y_j = \alpha + < \beta, \chi_j > + \epsilon_j = \alpha + \int_T \beta(s)\chi_j(s)ds + \epsilon_j.$$

La primera vía de solución para el problema se hace mediante el uso de la aproximación en términos de funciones base $\{\phi_1, \cdots, \phi_K\}$ y sus respectivos coeficientes, es decir, haciendo uso de la descomposición de β:

$$\beta(t) \approx \sum_{k=1}^{K} b_k\phi_k(t),$$

así,

$$< \beta, \chi_j >= \int_T \beta(s)\chi_j(s)ds \approx \int_T \left(\sum_{k=1}^{K} b_k\phi_k(s)\right)\chi_j(s)ds = \sum_{k=1}^{K} b_k \int_T \phi_k(s)\chi_i(s)ds = \sum_{k=1}^{K} b_k\chi_{jk},$$

en el cual $\chi_{jk} = \int_S \phi_k(s)\chi_j(s)ds$. En este orden de ideas, el modelo será:

$$y_j = \alpha + \sum_{k=1}^{K} b_k\chi_{jk} + \epsilon_j \quad para \quad j = 1, 2, ..., m.$$

Matricialmente el modelo se puede expresar como:

$$\vec{Y} = X\vec{b} + \vec{\epsilon},$$

por lo tanto:

$$\vec{\hat{b}} = (X'X)^{-1}(X'Y).$$

Ahora bien, es lógico preguntarse qué sucede si se supone que tanto χ como β son funcionales de tipo bidimensional y si se reescribe el modelo de regresión de la siguiente manera:

$$y_j = \alpha + < \beta, \chi_j > + \epsilon_j = \alpha + \int_S \int_T \beta(s,t)\chi_j(s,t)dsdt + \epsilon_j, \quad \text{con } j = 1, 2, \cdots, m.$$

Se supone que tanto β como χ pertenecen a el espacio de funciones bivariadas cuadrado integrables $\mathcal{L}^2$ sobre su dominio. Para la estimación de este modelo se considera que: $X_j \in \mathbb{R}^{N\times N}$, Y es una respuesta escalar y $\beta \in \mathbb{R}^{N\times N}$. Adicionalmente, para $U \in \mathbb{R}^{p_1\times p_2}$ y $V \in \mathbb{R}^{p_1\times p_2}$ se asume que

$$< U, V >:= \sum_{i_1}\sum_{i_2} U_{i_1,i_2}V_{i_1,i_2},$$

corresponde a una aproximación de la doble integral, es decir,

$$< \chi_j, \beta >:= \int_0^1 \int_0^1 \beta(s,t)\chi_i(s,t)dsdt \approx \sum_{i_1}\sum_{i_2} \beta_{i_1,i_2}x_{i_1,i_2}.$$

De este modo, si ambas X y Y están centradas, el modelo aproximado quedará expresado como:

$$y_j = \alpha + \sum_{i_1}\sum_{i_2} \beta_{i_1,i_2}x_{i_1,i_2} + \epsilon_j.$$

Ahora bien, para incluir la variación total como herramienta en la solución del problema, se considera que el funcional $\beta \in BV$. En este contexto, $||\beta||_{TV} < \infty$, por lo que el problema de regresión lineal se puede abordar optimizando la variación total de β, que es finita siempre y cuando ϵ_j sea lo mas pequeño posible (véase en [20]), lo que es equivalente a encontrar el mínimo de la función de optimización dada por:

$$||\beta||_{TV} + \dot{\lambda}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{4-6}$$

donde $\dot{\lambda}$ es el parámetro de optimización (multiplicador de Lagrange). Sea A_x una matriz de diseño de tamaño $m \times n$, con $n = N^2$, de tal manera que la $j - ésima$ fila es la vectorización de X_j. Usando entonces la noción de variación total discreta de β, el problema se reduce a encontrar el valor óptimo de:

$$\sum_i ||D_i\beta||_1 + \dot{\lambda}||\mathbf{y} - A_x\beta||_2. \tag{4-7}$$

Nótese que, si se toma $\dot{\lambda} = \frac{\mu}{2}$, $A_x = A$ y $\vec{y} = \vec{f}$, 4-7 es equivalente a resolver el problema de minimización dado en 4-4, por esta razón se puede usar el método **TVAL3** para dar solución a la optimización.

4.2. Modelos de regresión lineal generalizado Escalar-Función

EL modelo de regresión lineal, mencionado anteriormente, puede ser considerado como una extensión sobre el campo funcional de los modelos de regresión desarrollados en la estadística clásica. Del mismo modo, las nociones de modelos de regresión lineal generalizado puede ser extendida para los modelos funcionales de tipo escalar-función. Formalmente, sean Y_i, $i = 1, \cdots, N$, variables aleatorias independientes cuya distribución pertenece a la familia exponencial con función de densidad de probabilidad dada por 2-2. La especificación del modelo está dada por la función de enlace 2-3 reemplazando X_i por $X_i(t)$, y adicionando una variable aleatoria escalar Z_i, es decir,

$$g(\mu_i) = \gamma^T Z_i + \langle X_i(t), \beta \rangle. \tag{4-8}$$

4.2.1. Modelo de variación acotada

El modelo de variación acotada hace referencia al modelo de regresión generalizado para $\mathbf{Y_i}|(\mathbf{X_i}, \mathbf{Z_i})$ $(i = 1, \cdots, n)$ definido por 4-8 con el supuesto adicional de que $\beta \in BV(S)$. Denotando $\xi = (\gamma, \beta)$ y dadas las realizaciones $X_i \in \mathbb{R}^{N\times N}$ y $Z_i \in \mathbb{R}^{p_0}$ y $y_i \in \mathbb{R}$ de las variables $\mathbf{X_i}$, $\mathbf{Z_i}$ y $\mathbf{Y_i}$ respectivamente, se halla el estimador de ξ , denotado $\hat{\xi}$, minimizando una reescritura de 4-7, la cual es la función de logverosimilitud penalizada (véase en [19] y [2]) dada por :

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i\eta_i - d_0(\eta_i)}{a(\phi_i)} + d(y_i, \phi_i)\right\} + \lambda||\beta||_{TV} = \frac{1}{n}\sum_{i=1}^{n}\{y_i\eta_i - d_0(\eta_i)\} + \lambda||\beta||_{TV}. \tag{4-9}$$

Note que, como se había mencionado previamente, $\eta_i = d_0'^{-1}(g(\mu_i)) = d_0'^{-1}(\gamma' Z_i + \langle \beta, X_i \rangle)$ para $i = 1, \cdots, n$. Se usará el algoritmo basado en mínimos cuadrados iterativos re-ponderados estándar para GLM (véase en [21]), que consiste en:

Pesos iterativos $\quad w_i^{(k)} = \frac{\partial^2 d_0(\eta_i^{(k)})}{\partial^2 \eta_i^{(k)}}$ Luego $\hat{w}_i^{(k)} = \left.\frac{\partial^2 d_0(\eta_i^{(k)})}{\partial^2 \eta_i^{(k)}}\right|_{\hat{\eta}_i^{(k)}}$

Variable dependiente $\quad \hat{Y}_i^{(k)} = g(\hat{\mu}_i^{(k)}) + (Y_i - \hat{\mu}_i^{(k)})\dot{g}(\hat{\mu}_i^{(k)})$

con $\quad \hat{\mu}_i^{(k)} = \mu(X_i, Z_i; \hat{\xi}_I) \qquad \dot{g}(\mu) = \frac{\partial g(\mu)}{\partial \mu} \qquad \dot{\mu}_i = \frac{\partial \mu_i}{\partial \xi} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \xi}$

$$\begin{aligned}\hat{\xi}^{(k+1)} =& arg\min_{\xi}\left\{\sum_{i=1}^{n} w_i^{(k)}[\hat{Y}_i^{(k)} - \dot{\mu}_i(\hat{\xi}^{(k)})\xi]^2 + \lambda||\beta||_{TV}\right\} \\ =& arg\min_{\xi}\left\{||W^{\frac{1}{2}(k)}(\hat{\mathbf{Y}}^{(k)} - \dot{\mu}\xi')|| + \lambda||\beta||_{TV}\right\},\end{aligned} \tag{4-10}$$

donde:

$$W^{\frac{1}{2}(k)} = \begin{pmatrix} \sqrt{w_1^{(k)}} & 0 & \cdots & 0 \\ 0 & \sqrt{w_2^{(k)}} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sqrt{w_n^{(k)}} \end{pmatrix}, \quad \hat{\mathbf{Y}}^{(k)} = \begin{pmatrix} \hat{Y}_1^{(k)} \\ \hat{Y}_2^{(k)} \\ \vdots \\ \hat{Y}_n^{(k)} \end{pmatrix} \quad \text{y} \quad \dot{\mu} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \gamma} & \frac{\partial \mu_1}{\partial \beta} \\ \frac{\partial \mu_2}{\partial \gamma} & \frac{\partial \mu_2}{\partial \beta} \\ \vdots & \vdots \\ \frac{\partial \mu_n}{\partial \gamma} & \frac{\partial \mu_n}{\partial \beta} \end{pmatrix}\Bigg|_{\hat{\xi}^{(k)}}.$$

Algoritmo

1. Inicializar: $\xi^{(0)} = (\gamma^{(0)}, \beta^{(0)})$.
2. Para cada k defina los pesos iterativos $\hat{w}_i^{(k)}$, la variable dependiente $\hat{y}_i^{(k)}$ y la función objetivo 4-10. Use TVAL3 para encontrar $\xi^{(k+1)}$
3. Itere el paso 2 hasta convergencia.

Para finalizar este capítulo, a continuación se describe brevemente el modelo de variación total que será empleado para los procesos de estimación y clasificación. Dadas las variables aleatorias independientes $Y_1, Y_2, ..., Y_n$ (variables que toman solo los valores 0 y 1, y que representan estado del paciente) cuyas distribuciones siguen $Y_i \sim Bernoulli(\pi_i)$ para $i = 1, ..., n$ y $\mathbf{X}_i$, con $\mathbf{X}_i \in \mathbb{R}^{N \times N}$. Para el modelo logístico $d_0(\eta_i) = -log(1 + e^{\eta_i})$ y

$$\eta_i = log\left(\frac{\mu_i}{1-\mu_i}\right) = \langle X_i, \beta \rangle + \gamma^T Z_i, \quad \text{para } i = 1, \cdots, n$$

De esta manera, el objetivo es estimar la imagen coeficiente $\beta \in \mathbb{R}^{N \times N}$ con la función de optimización:

$$n^{-1} \sum_{i=1}^{n} \left\{ \frac{y_i \eta_i - d_0(\eta_i)}{a(\phi_i)} + d(y_i, \phi_i) \right\} + \lambda ||\beta||_{TV} = n^{-1} \sum_{i=1}^{n} \{y_i \eta_i - d_0(\eta_i)\} + \lambda ||\beta||_{TV}.$$

Pesos iterativos $\hat{w}_i^{(k)} = \frac{e^{\hat{\eta}_i^{(k)}}}{(1+e^{\hat{\eta}_i^{(k)}})^2}$ **Variable dependiente** $\hat{Y}_i^{(k)} = \hat{\eta}_i^{(k)} + \frac{Y_i - \hat{\mu}_i^{(k)}}{\hat{\mu}_i^{(k)}(1-\hat{\mu}_i^{(k)})}$

Por lo tanto la estimación de $\hat{\xi}_{I+1}$ puede ser obtenido resolviendo mínimos cuadrados ponderados penalizados (4-10).

5. Prevalencia en los modelos de clasificación: curva ROC

En esta sección se describe la herramienta que se usará para elegir el mejor modelo clasificación en las diferentes simulaciones y sobre la aplicación a datos reales. Dentro del conjunto de herramientas de predominio, existen dos tipos de indicadores que pueden utilizarse para estimar la calidad de los modelos de clasificación:

1) Indicadores de calidad cuantitativa: estadísticas que expresan la calidad de la clasificación mediante valores numéricos.

2) Indicadores gráficos: la calidad de la clasificación se representa en un gráfico que combina indicadores cuantitativos seleccionados. Los métodos gráficos simplifican la evaluación de la calidad del modelo y visualizan los resultados de la clasificación. Dichos indicadores incluyen:

   a Matriz de confusión.

   b Curva ROC (Receiver Operating Characteristic o Característica Operativa del Receptor).

A continuación se definen algunas nociones básicas que serán de utilidad en la construcción de los indicadores de prevalencia de modelos para la clasificación binaria 0-1 (véase [10]).

- **TP-Verdadero positivo:** el número de observaciones correctamente asignadas a la clase 1.

- **TN-Verdadero negativo:** el número de observaciones correctamente asignadas a la clase 0.

- **FP-Falso positivo:** el número de observaciones asignadas por el modelo a la clase positiva, que en realidad pertenecen a la clase 0.

- **FN-Falso negativo:** el número de observaciones asignadas por el modelo a la clase negativa, que en realidad pertenecen a la clase 1.

El número de observaciones que en realidad pertenecen a la clase 1 denotado Pos corresponde a $Pos = TP + FN$, y el número de observaciones que en realidad pertenecen a la clase 0 denotado Neg se establece como $Neg = FP + TN$.

↓ Real\Predicho →	Clase 1	Clase 0
Clase 1	TP	FN
Clase 0	FP	TN

Si todas las observaciones han sido clasificadas correctamente, entonces

↓ Real\Predicho→	Clase 1	Clase 0
Clase 1	n_p	0
Clase 0	0	n_N

con n_p y n_N siendo la cantidad total de observaciones que pertenecen a la clase 1 y 0 respectivamente. A partir de los indicadores definidos anteriormente se derivan:

- TPR (*True Positive Rate*): refleja la capacidad del clasificador para detectar miembros de la clase positiva.
$$TPR = \frac{TP}{TP + FN}.$$
También se conoce como SE (sensibilidad).

- TNR (*True Negative Rate*): refleja la capacidad del clasificador para detectar miembros de la clase negativa.
$$TNR = \frac{TN}{TN + FP}.$$
También se conoce como SP (especificidad).

- FPR (*False Positive Rate*): refleja la frecuencia con la que el clasificador comete un error al clasificar el estado negativo como positivo.
$$FPR = \frac{FP}{FP + TN}.$$
Nota: $1 - SP = FPR$.

- FNR (*False Negative Rate*): refleja la frecuencia con la que el clasificador comete un error al clasificar el estado positivo como negativo.
$$FNR = \frac{FN}{FN + TP}$$
Nota: $1 - SE = FNR$.

- ACC: refleja la exactitud de predicción general del clasificador, es decir, la probabilidad de hacer la predicción correcta, la cual es igual a la relación entre el número de decisiones correctas y el número total de decisiones.
$$ACC = \frac{TP + TN}{TP + TN + FP + FN}.$$

El punto de corte es un cierto valor umbral α que se puede usar para determinar si una observación pertenece a una clase en particular, es decir, si $P(clase(x) = 1) \geq \alpha$, entonces x se asigna a la clase 1. Como se mencionó anteriormente, la curva ROC es uno de los métodos usados para visualizar la calidad de clasificación. Dicha curva muestra la dependencia entre SE (proporción de positivos en la población clase 1) y $1 - SP$ (proporción de positivos en la población clase 0) (véase [10]). La información contenida en la curva permite notar que:

- Si la prueba fuera perfecta, es decir, cualquier punto de corte tiene sensibilidad y especificidad iguales a 1, entonces la curva solo tiene el punto (0,1).
- Si la prueba fuera inútil, esto es, ambas coinciden y la sensibilidad es igual a la proporción de falsos positivos, entonces la curva sería la diagonal de (0,0) a (1,1).
- Las pruebas habituales tienen curvas intermedias. Cuanto más convexa sea la curva, mejor será el clasificador.

Los pasos para la construcción de la curva ROC se presentan a continuación:

i) Calculamos los valores de la función de decisión.

ii) Probamos el clasificador para diferentes umbrales α.

iii) Para cada clasificación con un valor del umbral α obtenemos un par (SE, 1-SP), que corresponde a un punto en la curva ROC como se muestra en la figura **5-1** (para cada clasificación con un valor del umbral α también tenemos la matriz de confusión correspondiente).

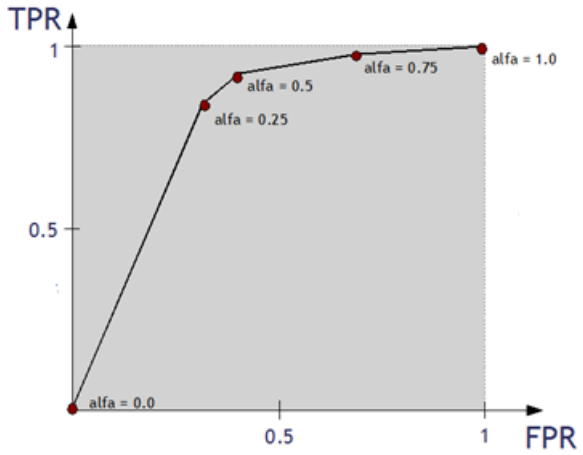

Figura 5-1.: Curva ROC, señalando la trayectoria en función de los diferentes valores de α. Tomado de $https : //algolytics.com/wp - content/uploads/2018/05/roc3_en.png$

Finalmente, para evaluar el clasificador sobre la base de la curva ROC se puede calcular el coeficiente ABC, que corresponde al área bajo la curva ROC. Cuanto mayor sea el valor del coeficiente ABC, mejor. $ABC = 1$ significa un clasificador perfecto, $ABC = 0{,}5$ se obtiene para clasificadores puramente aleatorios. $ABC < 0{,}5$ significa que el clasificador funciona peor que uno aleatorio. Lo anterior se muestra visualmente en la figura **5-2**. Ésta área puede interpretarse como la probabilidad de que ante un par de individuos, uno de la clase 1 y el otro de la clase 0, la prueba los clasifique correctamente.

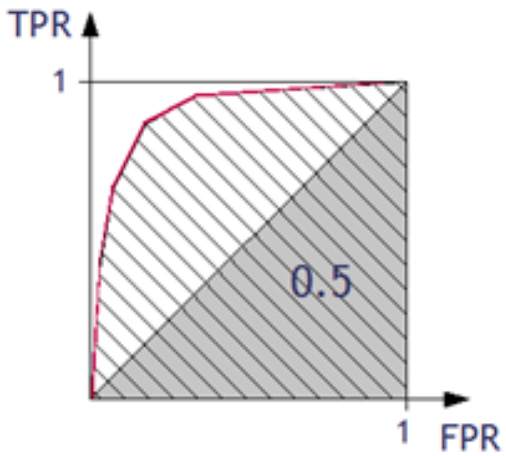

Figura 5-2.: Curva ROC, denotando el área bajo la curva (ABC). Detallando que la diagonal de divide el cuadrado de tamaño 1×1 tendra como ABC 0.5. Tomado de $https://algolytics.com/wp-content/uploads/2018/05/roc5.png$

El contraste de las dos metodologías: el modelo de regresión tensor y el modelo vía variación acotada se hará por medio de la clasificación de un mismo conjunto de imágenes bidimensionales cuyas categorías estarán previamente establecidas por medio de la curva ROC, puntualmente a través de la media de área bajo esta curva.

6. Resultados

Esta sección tiene como objetivo presentar los resultados de los procesos de simulación y la aplicación de datos reales. Todo el procedimiento de estimación y compilación de salidas se realizó por medio de la plataforma y del lenguaje propio de programación de MATLAB. El capítulo está dividido en tres secciones, la primera parte muestra un estudio de simulación en donde se muestra el desempeño de las dos metodologías a lo largo de varias simulaciones. La segunda parte contiene un ejercicio preliminar, en el cual las variables explicativas son imágenes reales y las respuestas son simuladas, con el fin de contrastar los resultados de estimación del coeficiente imagen. Finalmente, con el fin de examinar y comparar la aplicabilidad de los modelos, se mostrará una aplicación a datos reales sobre un conjunto de imágenes diagnósticas.

Es importante mencionar que para llevar a cabo las estimaciones de los modelos que hacen uso de imágenes reales se hizo un ajuste en el tamaño de estas covariables. Esto debido a que las imágenes podían no ser cuadradas y, adicionalmente, el modelo de regresión tensorial precisa más observaciones a medida que el tamaño del tensor aumenta, y más aún, tiene un límite inferior para el tamaño de observaciones en función de la dimensión del arreglo dos-dimensional y el rango R (Tabla 3.2).

El reajuste de tamaño se realizó por medio del comando *array_resize* de MATLAB (véase en [23]). Esta función cambia el tamaño de una matriz multidimensional usando 4 métodos diferentes: interpolación (predeterminado), transformada discreta del coseno, PCA mediante descomposición de valor singular de orden superior (HOSVD) y PCA a través de descomposición de valores singulares (SVD) marginal (cuando el método es HOSVD o SVD, solo es permitida la reducción del tamaño). Finalmente se redujo cada una de las imágenes a un tamaño 128×128 ó 64×64 por el método 1: interpolación.

6.1. Proceso de simulación

El proceso de simulación de las covariables se realiza usando la representación de Haar discreta (véase en el anexo B). Dada una dirección fija d_0, donde X se puede escribir de la forma

$$X(u,v) = \sum_j \sum_{\mathbf{k}} w_{j,k}\xi_{j,\mathbf{k}}\phi_{j,\mathbf{k}}(u,v),$$

en la cual $\{\phi_{j,\mathbf{k}}\}$ es la colección de bases de Haar bivariadas [4] con $j \geq 0$, $\mathbf{k} = (k_1, k_2) \in \{0,1,2,\ldots,2^6-1\} \times \{0,1,2,\ldots,2^6-1\}$, tomando $w_{j,k} = p_j$ y $\xi_{j,k} = \xi_j$ variables aleatorias independientes e idénticamente distribuidas gaussianas con media 0 y varianza unitaria. Reescribiremos, por facilidad, $\mathbf{X_i}(u,v) = \sum_j p_j\xi_{j,i}\phi_j(u,v)$, la cual representa la $i-ésima$ imagen de tamaño 64×64. Para las simulaciones de $\mathbf{B}$ (el coeficiente-imagen) usaremos dos alternativas que se describen a continuación:

1. B tomará la forma de:

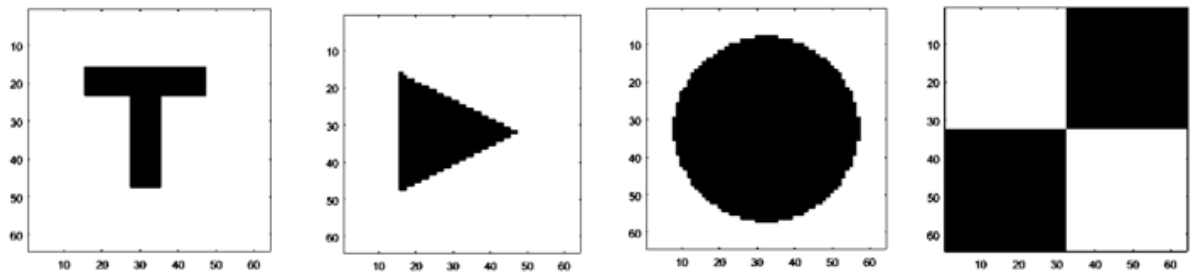

Figura 6-1.: Imágenes coeficientes $\mathbf{B}$, con una forma específica, con las cuales se realizará el proceso de estimación ($Forma-T$, $Triángulo$, $Círculo$ y $Combinado$).

2. B tomará la forma de:

$$\mathbf{B} = B(u,v) = \sum_j p_j\xi_j\phi_j(u,v),$$

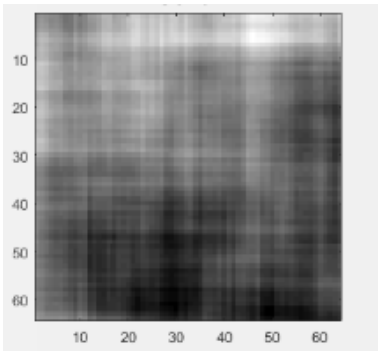

Figura 6-2.: Imagen coeficiente $\mathbf{B}$, simulada de la misma manera que las covariables X, con la cual también se realizará el proceso de estimación. ($Simulado$)

Se obtienen las realizaciones de respuestas univariadas $y_i \in \mathbb{R}$, con $Y_i|\mathbf{X_i}, \mathbf{Z_i} \sim Bernoulli(\pi_i)$, siguiendo

$$\pi_i = \frac{1}{1+exp(-\eta_i)} \quad y \quad \eta_i = \vec{\gamma}^T Z_i + \langle \mathbf{B}, X_i \rangle, \quad \text{tomando } \vec{\gamma} = \vec{1}_5 \text{ y } \langle \mathbf{B}, \mathbf{X} \rangle = \langle Vec\mathbf{B}, Vec\mathbf{X} \rangle.$$

El anterior procedimiento nos da como resultado los datos conformados por las triplas (y_i, Z_i, X_i). Posteriormente, se conforma un conjunto de datos de n observaciones para cada una de las covariables junto con su respectivas respuestas que será dividida en una muestra de entrenamiento y prueba con n_e y $n_p = n - n_e$ número de observaciones respectivamente. Finalmente, se procede a estimar el coeficiente imagen B para ambas metodologías, con la muestra de tamaño n_e, y realizar el proceso de clasificación para la muestra de prueba. El proceso de estimación para ambas regresiones se describe brevemente a continuación:

- **Regresión Tensor:** en este estudio $D = 2$, $p_1 = 128$, $p_2 = 128$, $p_0 = 5$. Se supone que $\mathbf{B}$ tiene descomposición de rango R, para $\beta_i^{(r)} \in \mathbb{R}^{64}$ con $i = 1, 2$, es decir que el arreglo de coeficientes satisface:

  $$\mathbf{B} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \beta_2^{(r)},$$

  más aún, si definimos $B_1 = [\beta_1^{(1)}, \cdots, \beta_1^{(R)}] \in \mathbb{R}^{64\times R}$ y $B_2 = [\beta_2^{(1)}, \cdots, \beta_2^{(R)}] \in \mathbb{R}^{64\times R}$ entonces $Vec\mathbf{B}$ se reescribe como:

  $$Vec\mathbf{B} = (\mathbf{B}_2 \odot \mathbf{B}_1)1_R.$$

  De esta manera, el objetivo es estimar el arreglo de coeficientes $\mathbf{B}$ (matriz 128 x 128) estimando B_1 y B_2, con el modelo general tensor ($R > 0$).Optimizando la función de log-verosimilitud dada por:

  $$\ell(\alpha, \vec{\gamma}, \mathbf{B}_1, \mathbf{B}_2) = \sum_{i=1}^{n} y_i\eta_i - d_0(\eta_i) + \sum_{i=1}^{n} d(y_i),$$

  descrita en la sección 3.3 con $d_0(\eta_i) = -log(1 + e^{\eta_i})$.

- **Regresión Variación Total:** en el modelo de regresión vía variación acotada se tiene que $\mathbf{B} = \beta$, el cual corresponde a una imagen 64×64 en escala de grises y se supone que su representación funcional posee una medida de variación total finita ($||\mathbf{B}||_{TV} < \infty$). De esta manera, el objetivo es estimar la imagen coeficiente $\mathbf{B}$ con la función de optimización:

  $$n^{-1} \sum_{i=1}^{n} \left\{ \frac{y_i\eta_i - d_0(\eta_i)}{a(\phi_i)} + d(y_i, \phi_i) \right\} + \lambda ||\mathbf{B}||_{TV} = n^{-1} \sum_{i=1}^{n} \{y_i\eta_i - d_0(\eta_i)\} + \lambda ||\mathbf{B}||_{TV},$$

  en la cual $d_0(\eta_i) = -log(1 + e^{\eta_i})$.

6.1.1. Resultados

Los resultados arrojados provienen de la simulación de n observaciones, (y_i, Z_i, X_i) para $i = 1, \cdots, n$, de los cuales n_e corresponde a la muestra de entrenamiento y con la cual se estimó el vector de coeficientes para las dos metodologías: el modelo tensor y la variación total. Los $n_p = n - n_e$ individuos restantes conforman la muestra de prueba y para estos se realizó el proceso de clasificación y se calculó el TP, TN, FP y FN con cada uno de los coeficientes obtenidos anteriormente. Con todo esto, posteriormente se computó el ABC.

El proceso se replicó 100 veces y se calculó el promedio de ABC y su correspondiente varianza, para los tamaños $n = 200, 625$ y 1255. Las tablas que se presentan a continuación muestran el valor del área bajo la curva (ABC) obtenido del proceso de clasificación, tanto por el modelo de regresión tensor como el modelo de variación total y sus respectivas variaciones fijando **B** como cada una de las figuras dadas en **6-1** y **6-2**. Adicionalmente, se muestra el tiempo promedio computacional (en segundos) para el modelo de regresión tensor y para el modelo de variación total.

- $n = 200$

	$Forma - T$	$Triángulo$	$Círculo$	$Combinado$	$Simulado$
ABC_{Tensor}	0.4975	0.5115	0.5134	0.5279	0.5010
σ^2_{Tensor}	0.0222	0.0259	0.0433	0.0247	0.0175
$ABC_{Var.T}$	0.9356	0.9337	0.9307	0.8831	0.8220
$\sigma^2_{Var.T}$	0.0025	0.0025	0.0018	0.0036	0.0051

Tabla 6-1.: Valor del ABC de clasificación y su respectivo variación, con un tamaño de muestra de $n = 200$, dividido en $n_e = 160$ observaciones de entrenamiento y $n_p = 40$ observaciones de prueba, para cada uno de las diferentes formas de **B** (el coeficiente imagen)

	Tensor	VT
$Forma - T$	11.6432	3.8594
$Triangulo$	11.3472	4.2813
$Circulo$	19.6600	3.7656
$Combinado$	11.8087	1.8750
$Simulado$	11.3815	3.900

Tabla 6-2.: Tiempo promedio computacional (en segundos), con un tamaño de muestra de $n = 200$, dividido en $n_e = 160$ observaciones de entrenamiento y $n_p = 40$ observaciones de prueba, para cada uno de las diferentes formas de **B** (el coeficiente imagen)

- $n = 625$

	$Forma - T$	$Triángulo$	$Círculo$	$Combinado$	$Simulado$
ABC_{Tensor}	0.5737	0.5223	0.5160	0.7289	0.5906
σ^2_{Tensor}	0.0159	0.0251	0.0286	0.0048	0.0252
$ABC_{Var.T}$	0.9943	0.9616	0.9809	0.9436	0.8500
σ^2_{Tensor}	0.00002	0.0006	0.0002	0.0007	0.0017

Tabla 6-3.: Valor del ABC de clasificación y su respectivo variación, con un tamaño de muestra de $n = 625$, dividido en $n_e = 500$ observaciones de entrenamiento y $n_p = 125$ observaciones de prueba, para cada uno de las diferentes formas de **B** (el coeficiente imagen)

	Tensor	VT
$Forma - T$	61.7588	3.3125
$Triangulo$	62.2800	3.9219
$Circulo$	62.2800	3.4844
$Combinado$	54.2650	2.7118
$Simulado$	64.1761	2.5156

Tabla 6-4.: Tiempo promedio computacional (en segundos), con un tamaño de muestra de $n = 625$, dividido en $n_e = 500$ observaciones de entrenamiento y $n_p = 125$ observaciones de prueba, para cada uno de las diferentes formas de **B** (el coeficiente imagen)

- $n = 1250$

	$Forma - T$	$Triángulo$	$Círculo$	$Combinado$	$Simulado$
ABC_{Tensor}	0.8560	0.5301	0.5197	0.8378	0.7597
σ^2_{Tensor}	0.0002	0.0226	0.0333	0.0066	0.0076
$ABC_{Var.T}$	0.9996	0.9918	0.9904	0.9963	0.8531
σ^2_{Tensor}	0.00005	0.00006	0.00005	0.00002	0.0006

Tabla 6-5.: Valor del ABC de clasificación y su respectivo variación, con un tamaño de muestra de $n = 1250$, dividido en $n_e = 1000$ observaciones de entrenamiento y $n_p = 250$ observaciones de prueba, para cada uno de las diferentes formas de **B** (el coeficiente imagen)

	Tensor	VT
$Forma - T$	74.9034	2.2813
$Triangulo$	84.2650	2.6563
$Circulo$	104.7641	5.093
$Combinado$	54.2650	2.6525
$Simulado$	93.4304	3.3226

Tabla 6-6.: Tiempo promedio computacional (en segundos), con un tamaño de muestra de $n = 1250$, dividido en $n_e = 1000$ observaciones de entrenamiento y $n_p = 250$ observaciones de prueba, para cada uno de las diferentes formas de $\mathbf{B}$ (el coeficiente imagen)

De los anteriores resultados se puede observar que, en los tres casos, el modelo de variación total precisa de menos tiempo de computo en comparación al modelo de regresión tensor, lo que resulta importante puesto que este valor se puede adicionar como criterio en la elección del mejor modelo. En general, el tiempo promedio usado por el modelo tensorial es por lo menos el doble que el empleado por el modelo TV. Adicionalmente, se puede notar que para ambos modelos, el aumento del número de observaciones incrementa el valor medio de ABC y reduce la medida de varianza.

Es importante mencionar que con una muestra de prueba de tamaño 40 (muestra de entrenamiento de 160 observaciones), el ABC promedio para el modelo de variación total para la $Forma - T$, $Triángulo$ y $Círculo$ resulta ser superior a 0,9, lo que es bastante bueno, a diferencia del modelo tensor que permanece cercano a 0,5 e inclusive para la primera forma se encuentra por debajo de este valor (es peor que una clasificación aleatoria). Aunque en el segundo modelo, para las formas $Combinado$ y $Simulado$ no se alcanza a superar el nivel de 0.9, su correspondiente valor no está tan distante. Se puede observar también que los resultados del modelo TV mejoran para la muestra de prueba de tamaño 1225 (muestra de entrenamiento de 500 observaciones) puesto que 4 de las 5 formas de $\mathbf{B}$, salvo $Combinado$, alcanzan a sobrepasar el umbral de 0,9 y su varianza se reduce en gran cantidad.

Finalmente, en este último tamaño de muestra, el modelo de variación total es notoriamente superior al modelo de regresión tensor. Sin embargo, vale la pena resaltar que este último también mejora en gran medida para la $Forma - T$, $Combinado$ y $Simulado$. Esto se evidencia en que el modelo tensorial aumenta el valor medio de ABC (y reduce la varianza promedio) con una diferencia superior a 0,07, en comparación a la anterior muestra. En este proceso se resalta la forma $Combinado$ para la cual el ABC es 0,7289.

La representación de las estimaciones (en una de las 100 iteraciones) de las imágenes coeficientes $\mathbf{B}$ para los tamaños $n = 200$ ($n_e = 160$) y $n = 1250$ ($n_e = 1000$) las figuras **6-3** y **6-3** respectivamente. La información obtenida también evidencia que el modelo de regresión tensor no se comporta totalmente bien para para algunos $\mathbf{B}$ fijados. Tal es el caso para

Triángulo y *Círculo*.

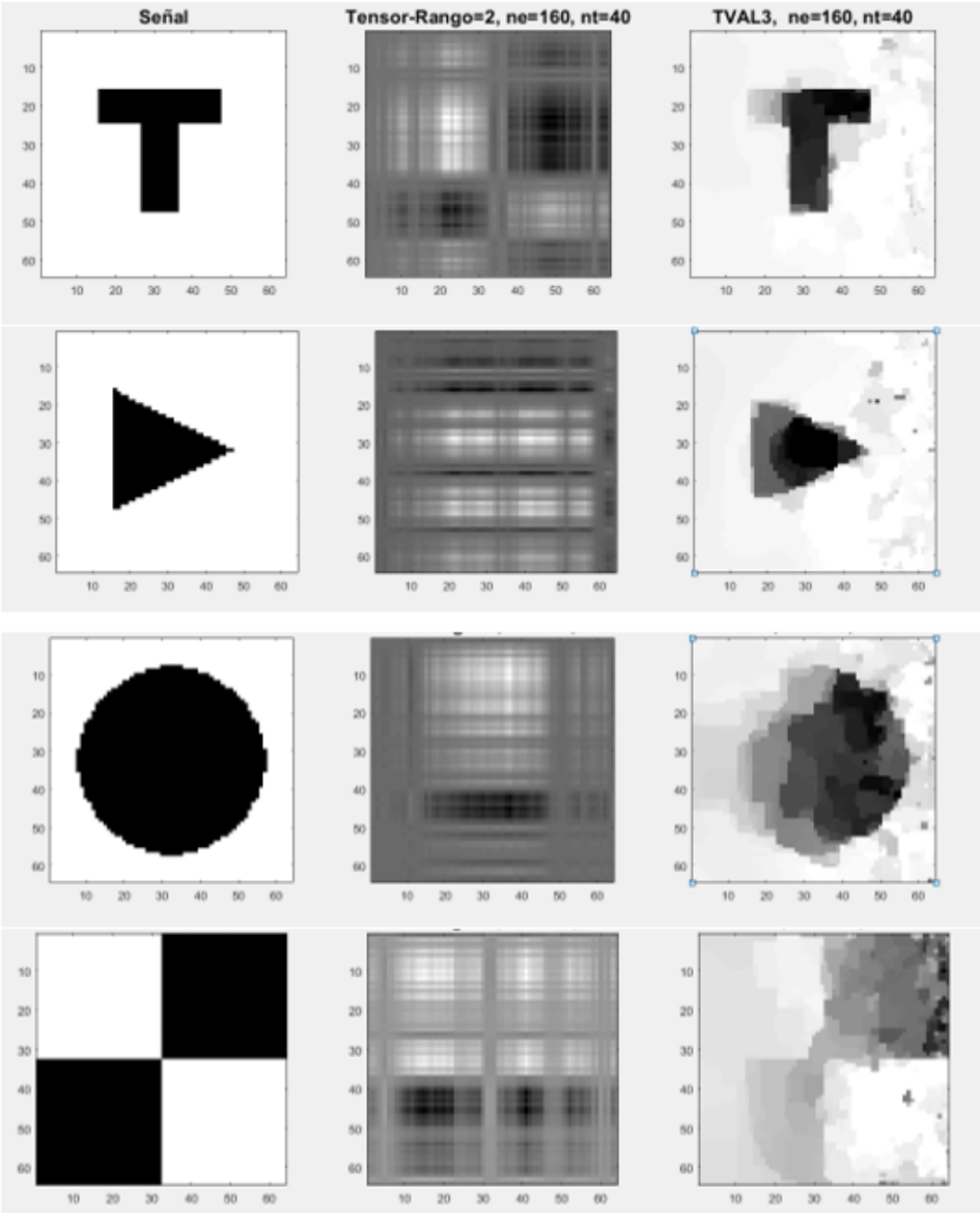

Figura 6-3.: $B - coeficientes$, para una de las 100, iteraciones con $n = 200$ ($n_e = 160$). Fijando el rango de la regresión tensor en 2.

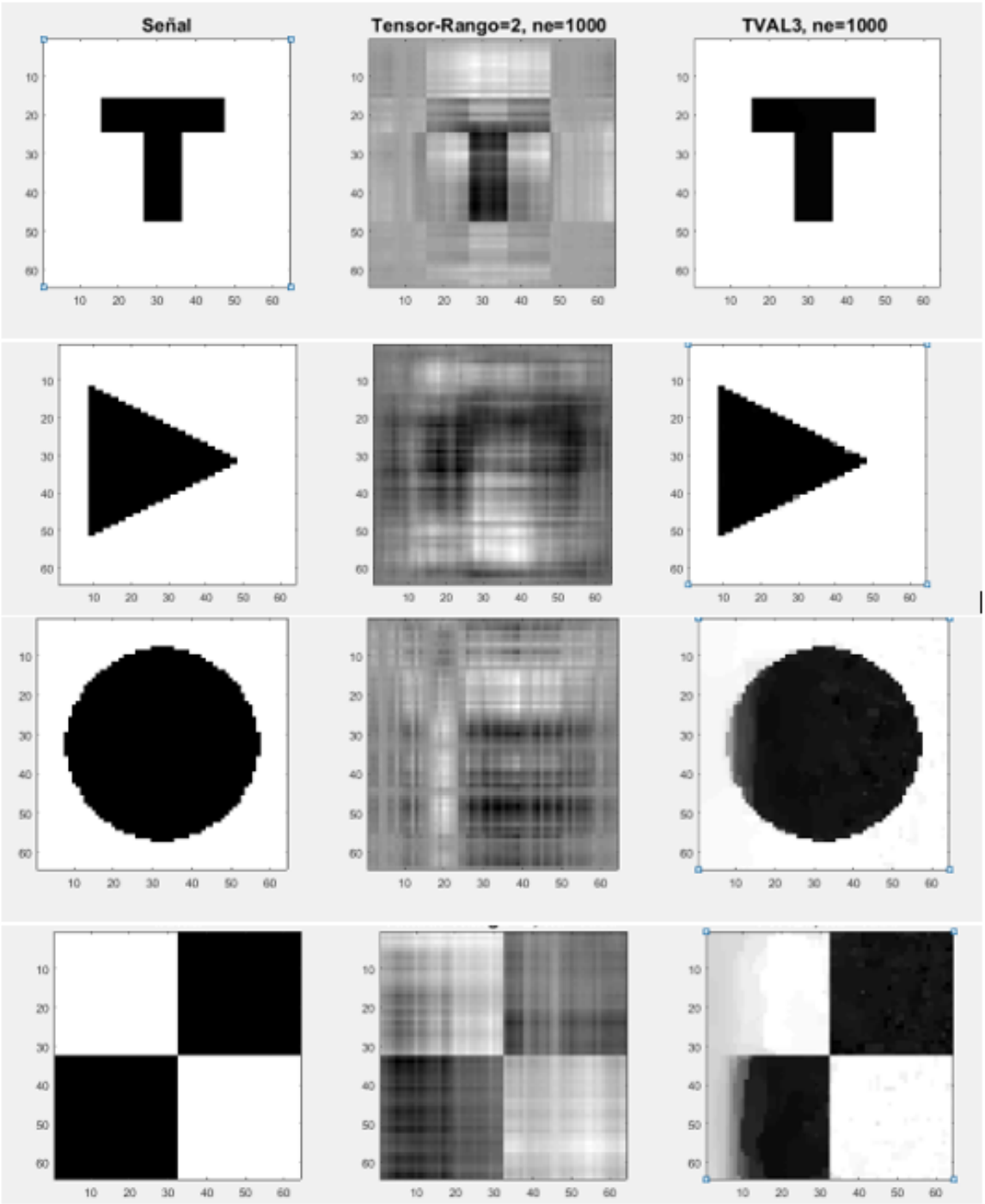

Figura 6-4.: $B - coeficientes$ estimados, para una de las 100 iteraciones, con $n = 1250$ $(n_e = 1000)$. Fijando el rango de la regresión tensor en 2.

A continuación, en la figura **6-5**, se presentan las curvas ROC para una de las 100 iteraciones, con un tamaño de muestra de $n = 1250$ $(n_e = 1000)$ para las formas dadas en la

figura **6-2**. De acuerdo a las gráficas se pude afirmar que, la curva correspondiente al modelo de variación total (roja) esta por encima a la obtenida por el modelo tensorial (azul) en tres de las cuatro formas de **B**, esta diferencia es más pronunciada en las formas de *triángulo* y *círculo* lo que corrobora los resultados obtenidos en la figura **6-4** de $B - coeficientes$ estimados. Para finalizar, las curvas obtenidas para la forma *combinado* presenta varios puntos de intersección y no se observa gran distanciamiento, esto complementa los resultados obtenidos para la figura **6-4**, en la cual esta forma, visualmente es similar a la señal real para ambos modelos.

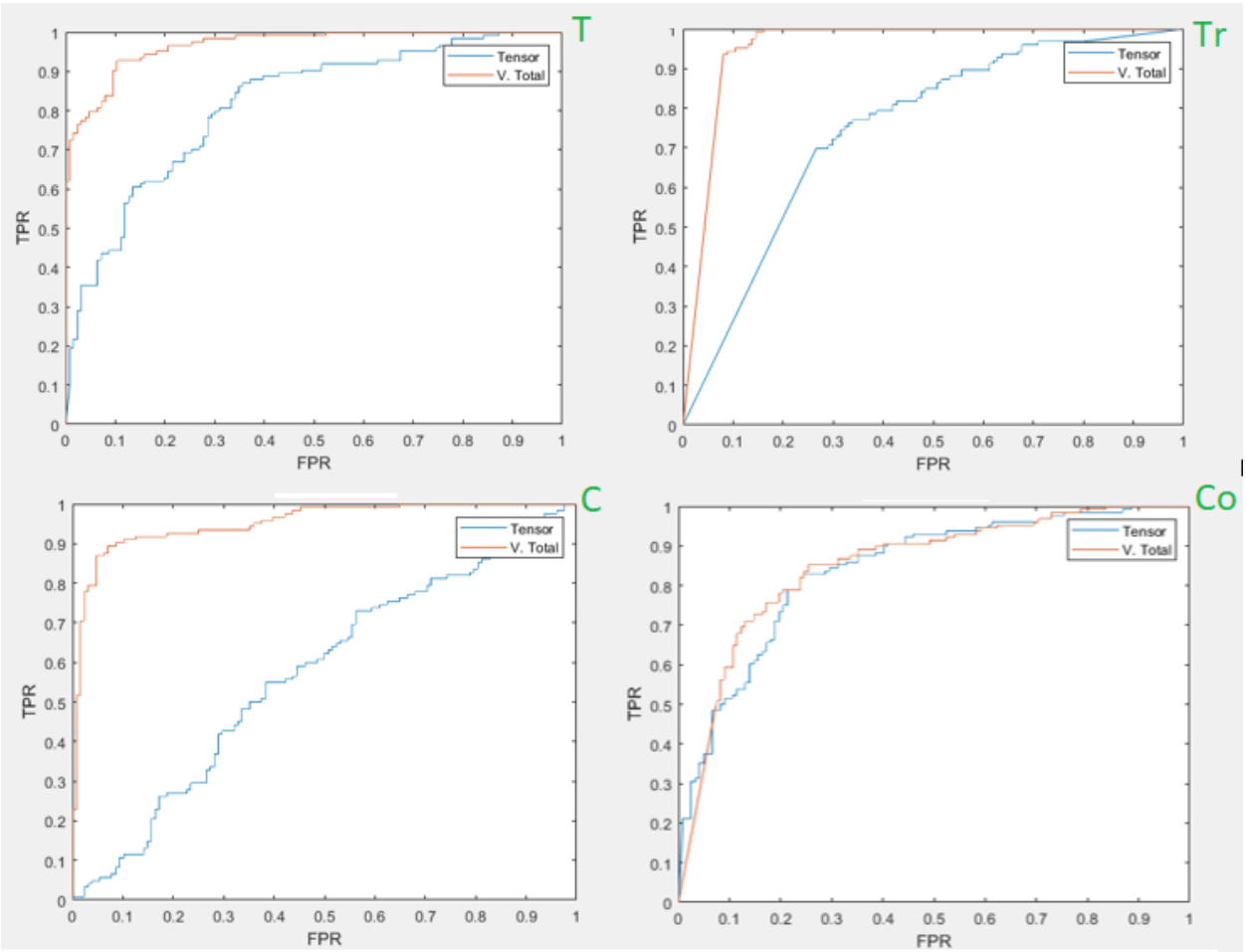

Figura 6-5.: Curvas ROC, para una de las 100 iteraciones, con $n = 1250$ ($n_e = 1000$). Fijando el rango de la regresión tensor en 2. Las etiquetas corresponden a: (T) $Forma - T$, (Tr) *triángulo*, (C) *Círculo*, (Co) *Combinado*

6.2. Ejercicio preliminar

El objetivo de este primer proceso será comparar las estimaciones del coeficiente imagen de ambas metodologías para diferentes tamaños de muestra. El fin principal de este estudio es examinar visualmente las estimaciones de $\mathbf{B}$ para los dos modelos a lo largo de los diferentes tamaños de muestra, cuando las covariables son imágenes reales. Debido a esto, en este ejercicio preliminar no se tendrá en cuenta como tal la calidad de la clasificación, es por esta razón que la base de datos solo presenta individuos sanos y la variable respuesta es simulada.

El ejercicio que se presenta en esta sección consiste en el uso la base de datos de $MRI\ Healthy\ Dataset$ tomada de la página $www.kaggle.com$, que consta de 2560 imágenes cerebrales dos-dimensionales de individuos sanos. De la muestra de tamaño original $n_1 = 2560$ se tomaron sub muestras de tamaños $n_2 = 1280$, $n_3 = 640$ y $n_4 = 320$. Se fija el tamaño de muestra n sobre el conjunto $\{n_1, n_2, n_3, n_4\}$ y en este sentido, las covariables X_i corresponderán a las imágenes cerebrales de individuos sanos, ajustadas al tamaño de 128×128 píxeles, provenientes del conjunto de datos con n observaciones. El coeficiente-imagen $\mathbf{B}$ está dado por la figura **6-6**. Posteriormente, se calcula la variable respuesta univariada tomando en consideración la siguiente relación:

$$y_i \sim Bernoulli\left(\frac{1}{1 + exp(-\eta_i)}\right) \quad con \quad \eta_i = \langle \mathbf{B}, X_i \rangle.$$

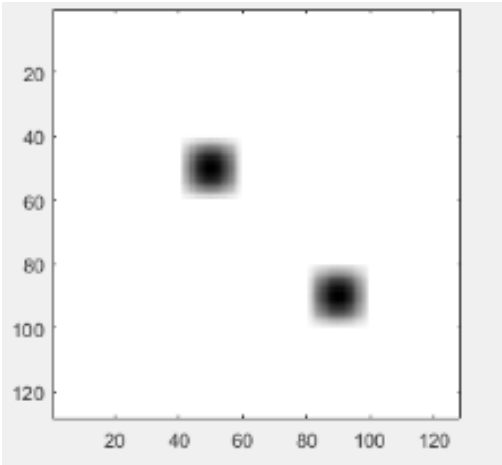

Figura 6-6.: Imagen coeficiente $\mathbf{B}$ que será fijada. Es simulada como un producto de dos rejillas con el rango de una función senosoidal, en el intervalo [0,128].

Finalmente con las duplas $(\mathbf{X_i}, y_i)$ para $i = 1, \cdots, n$ obtenidas, se procede a estimar el coeficiente $\mathbf{B}$ por medio de los dos métodos. La superposición de una de las covariables y el coeficiente imagen de tamaño 128×128 se representa, para un ejemplo, en la siguiente figura.

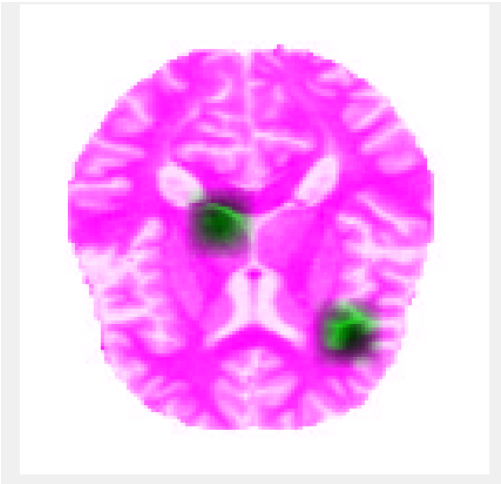

Figura 6-7.: Superposición de un ejemplo de las imágenes cerebrales de la base de datos de individuos sanos ($\mathbf{X_i}$ en fucsia) con la imagen señal fija ($\mathbf{B}$ en verde). Ambas imágenes tienen un tamaño de 128×128 píxeles.

En la figura 6-8, se muestran las imágenes que representan la estimaciones del coeficiente por el modelo tensor y el modelo de variación total para cada una de las muestras de tamaño n_0, n_1, n_2, n_3 y n_4. El tiempo promedio computacional para el modelo de regresión tensor y para el modelo de variación total, en segundos, se muestra en la tabla a continuación.

	Tensor	VT
$n_1 = 2560$	690.9713	11.0016
$n_2 = 1280$	326.8191	9.9063
$n_3 = 640$	177.0209	7.2813
$n_4 = 320$	6.9745	6.0506

Tabla 6-7.: La cantidad de observaciones n se establece como $n_1 = 2560$, $n_2 = 1280$, $n_3 = 640$ y $n_4 = 320$. Para cada uno de estos tamaños de muestra se estima el coeficiente imagen. La tabla muestra el tiempo promedio computacional que requiere cada uno de los dos modelos, en segundos, para la estimación de $\mathbf{B}$, de tamaño 128×128 píxeles, empleando un tamaño n observaciones.

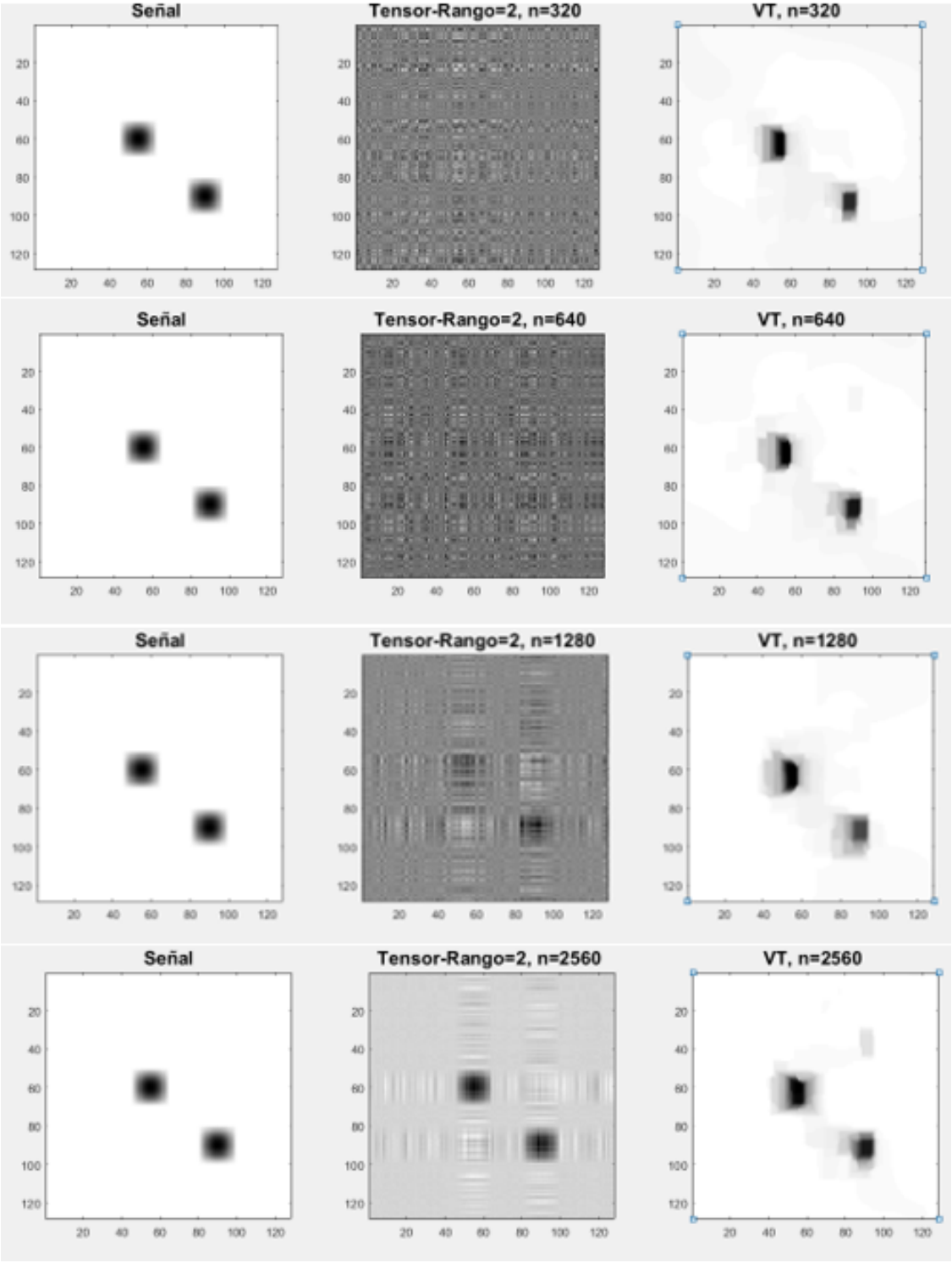

Figura 6-8.: Representación de las estimaciones de $\mathbf{B}$, de tamaño 128×128 píxeles. $\mathbf{B}$ original (izquierda), estimación vía regresión tensor de rango 2 (centro) y estimación vía variación total (derecha). Cada fila de imágenes emplea un tamaño de muestra de $n_0 = 2560$, $n_1 = 2560$, $n_2 = 1280$, $n_3 = 640$ y $n_4 = 320$ para las estimaciones respectivamente

Se evidencia que el tamaño de muestra necesario para la obtención de una estimación visualmente favorable es mayor para el modelo de regresión tensor. Es importante examinar la relación que existe entre el tamaño de muestra n y la estimación del coeficiente imagen en el modelo de regresión tensorial. Con el objetivo de evaluar la consistencia entre el aumento de observaciones y la mejora de la estimación en este modelo, se simulan 8000 covariables X_i siguiendo los parámetros de las bases de Haar usadas en la sección 6.1. Se consideran cuatro casos cada uno con un tamaño de muestra de $n = 500$, $n = 1000$, $n = 5000$ y $n = 8000$. Posteriormente se fija $\mathbf{B}$ (de la $Forma - T$ y $Combinado$) y se computa la variable respuesta univariada para cada conjunto de observaciones. Finalmente para cada sub muestra se estima el coeficiente imagen $\mathbf{B}$ por medio del modelo de regresión tensor. Las figuras a continuación muestran la representación de las estimaciones fijando $\mathbf{B}$ como la $Forma - T$ y $Combinado$ respectivamente.

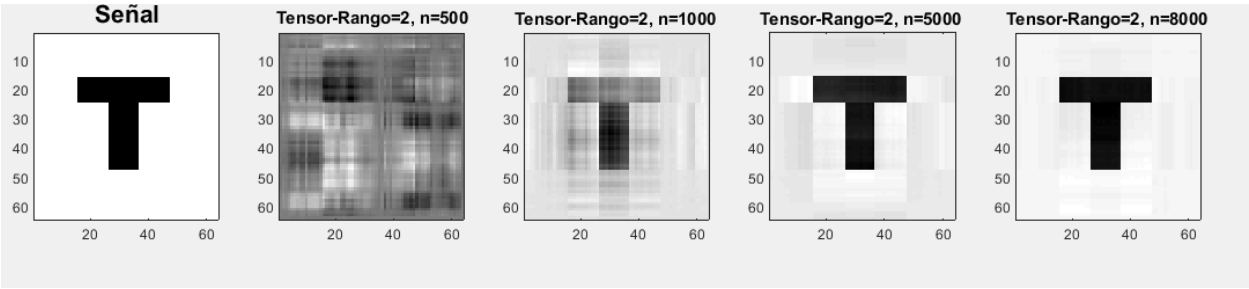

Figura 6-9.: Se simulan 8000 variables X_i siguiendo los parámetros de las bases de Haar, luego se estima el modelo para sub muestras de tamaño $n = 500$, $n = 1000$, $n = 5000$ y $n = 8000$ para $\mathbf{B}$ de la $Forma - T$. En la figura se muestran las estimaciones del coeficiente imagen.

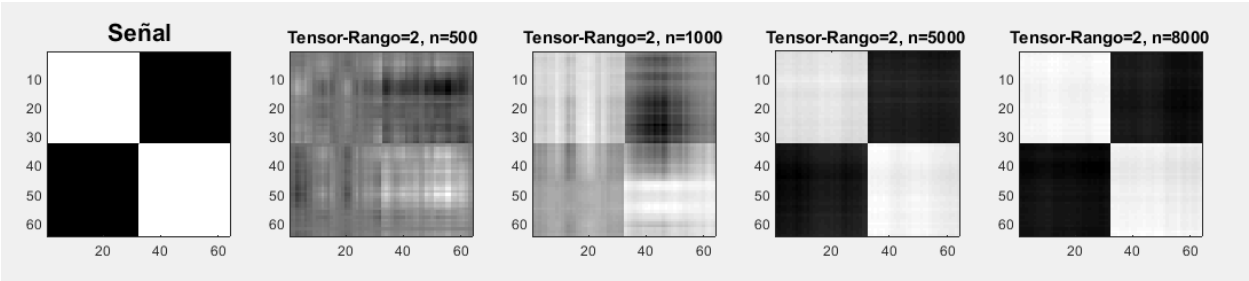

Figura 6-10.: Se simulan 8000 variables X_i siguiendo los parámetros de las bases de Haar, luego se estima el modelo para sub muestras de tamaño $n = 500$, $n = 1000$, $n = 5000$ y $n = 8000$ para $\mathbf{B}$ de la $Combinado$. En la figura se muestran las estimaciones del coeficiente imagen.

Se corrobora en ambos casos que, la relación entre el tamaño de muestra n y la calidad del estimador del coeficiente $\mathbf{B}$ es directamente proporcional y se evidencia que el modelo

de regresión tensor precisa de un tamaño de muestra "grande" para obtener una estimación visualmente favorable. Es importante mencionar que las imágenes son de tamaño 64×64, por ende, al aumentar la resolución de la imagen también se debe aumentar el tamaño de muestra lo que impactaría aun mas en el número de observaciones necesarias para el mejoramiento de los resultados.

6.3. Aplicación a datos reales

Esta sección contiene la aplicación del presente trabajo a los datos reales con uso la base de datos de *brain tumor dataset* tomada de *www.kaggle.com*, que consta de 253 imágenes $T1 - weighted$ inhaladas por contraste. El conjunto de datos se divide en 98 imágenes de individuos sanos y 155 imágenes de pacientes con algún tipo de tumor cerebral.

El objetivo es primero estimar, para cada uno de los dos modelos, el coeficiente de regresión con una sub muestra de entrenamiento constituidas por las imágenes cerebrales como covariables y el estado del paciente como respuesta. Posteriormente, con las imágenes diagnosticas restantes, la idea es aplicar cada una de las dos metodologías de clasificación y estimar la condición del paciente (sano y enfermo). Finalmente, para contrastar los modelos sobre la muestra de prueba, se calcula el área bajo la curva ROC ABC y el tiempo computacional asociados al anterior proceso de clasificación. Un ejemplo de imágenes para cada una de las dos clases se muestra a continuación:

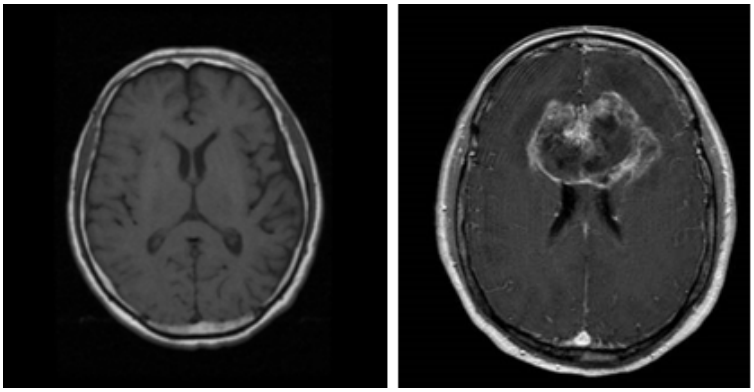

Figura 6-11.: Dos ejemplos de imágenes cerebrales (MRI dos-dimensional) pertenecientes a la base de datos: brain tumor dataset. A la izquierda se presenta una imagen de un paciente sano y a la derecha una imagen de un paciente enfermo.

Las imágenes diagnósticas cerebrales de la base de datos corresponden a las variables explicativas $\mathbf{X_i}$ y las respuestas univariadas y_i representan si el paciente esta enfermo o sano (1 o 0 respectivamente). El tamaño de muestra $n = 253$ fue dividido en $n_e = 202$ observaciones de entrenamiento (78 sanos y 124 enfermos) y $n_p = 51$ observaciones de prueba (20 sanos y 31 enfermos). Es importante mencionar que debido al tamaño de muestra, las imágenes se

redimensionaron a 128×128 píxeles, haciendo uso de la función $array_resize$ mencionada anteriormente. Además, se calcula el modelo de regresión tensor para un rango $R = 1$ debido a que para rangos superiores, se necesita reducir más las imágenes o alternativamente aumentar la muestra. El modelo de variación total no es restrictivo en la relación del número de observaciones y el tamaño de las imágenes, ya que los elementos de este modelo se sitúan en el espacio de funcionales infinito dimensionales.

En la siguiente tabla se presentan los resultados de ABC obtenidos para la clasificación por las dos metodologías y su respectivo tiempo computacional:

	ABC	$Tiempo\ com.(sg)$
$Tensor$	0.5527	17.87
$V.\ Total$	0.8164	6.22

Tabla 6-8.: Valor del ABC de clasificación y su respectivo tiempo computacional con tamaño de muestra de $n = 253$, dividido en $n_e = 202$ observaciones de entrenamiento y $n_p = 51$ observaciones de prueba, para cada uno de las dos metodologías regresión tensor (Rango $R = 1$) y variación total.

A continuación se presenta la curva ROC para la clasificación realizada:

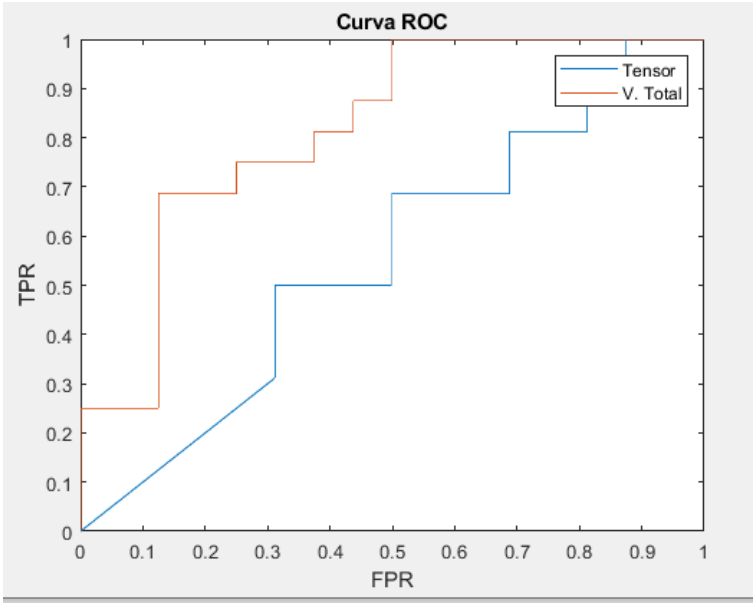

Figura 6-12.: Curva ROC para la clasificación para tamaño de muestra de $n = 253$, dividido en $n_e = 202$ observaciones de entrenamiento y $n_p = 51$ observaciones de prueba, para cada uno de las dos metodologías regresión tensor (Azul) y variación total (naranja).

De acuerdo a los resultados obtenidos, se evidencia el predominio del modelo de variación

total sobre el modelo de regresión tensor bajo el criterio de ABC (figura **6-12** la trayectoria del modelo V. total siempre esta por encima de la curva Tensor, $ABC_{VT} = 0{,}8164 > 0{,}5527 = ABC_{Tensor}$). Asi mismo, bajo la medida del tiempo computacional requerido para el proceso de clasificación, nuevamente el modelo de variación supera al modelo tensorial. Ahora bien, si se reducen todas las imágenes, de entrenamiento y prueba, a un tamaño de 64×64 para poder obtener los resultados del modelo de regresión tensor en un rango $R = 2$ y se realiza el proceso de clasificación, entonces los resultados son los siguientes:

	ABC	$Tiempo\ com.(sg)$
$Tensor$	0.5210	2.871
$V.\ Total$	0.7900	2.265

Tabla 6-9.: Valor del ABC de clasificación y su respectivo tiempo computacional si se reducen las imágenes a un tamaño de 64×64, con tamaño de muestra de $n = 253$, dividido en $n_e = 202$ observaciones de entrenamiento y $n_p = 51$ observaciones de prueba, para cada uno de las dos metodologías regresión tensor (Rango $R = 2$) y variación total.

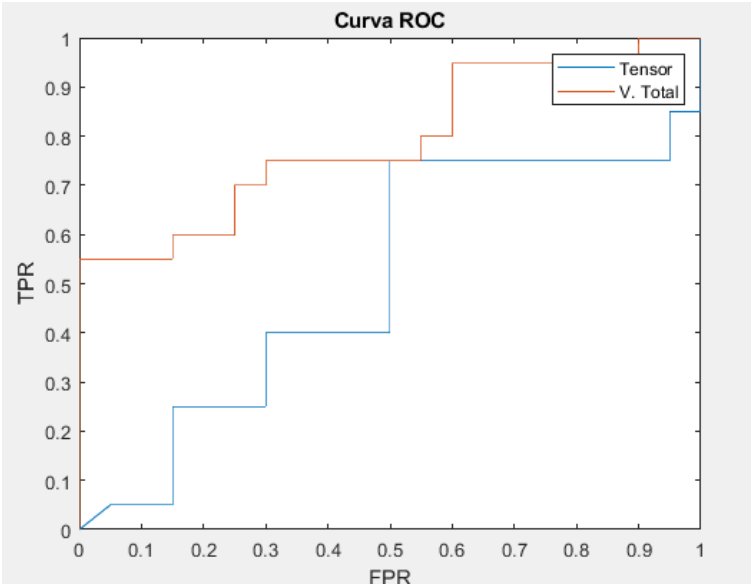

Figura 6-13.: Curva ROC para la clasificación para tamaño de muestra de $n = 253$ si se reducen las imágenes a un tamaño de 64×64, dividido en $n_e = 202$ observaciones de entrenamiento y $n_p = 51$ observaciones de prueba, para cada uno de las dos metodologías regresión tensor (Azul) y variación total (naranja).

Una vez más, se puede observar en la figura **6-13** que la curva ROC del modelo de V. Total resulta tener una concavidad mas pronunciada a lo largo de la trayectoria del umbral

α (salvo en un intervalo pequeño donde es igual). Lo que permite afirmar que el área bajo curva del modelo de variación total es mayor al área bajo la curva del modelo tensorial, como se evidencia en la tabla 6-9 ($ABC_{VT} = 0{,}7900 > 0{,}5210 = ABC_{Tensor}$)). De igual manera el tiempo computacional requerido es mayor en el primer modelo, sin embargo se debe hacer la observación de que la disminución de la resolución permite que esta diferencia de tiempos sea cercana a 0. Así mismo, se puede notar que la reducción en el tamaña de las imágenes covariables resta valor en el criterio ABC para ambos modelos, lo que de hecho puede ser considerado un suceso natural.

7. Conclusiones y Recomendaciones

7.1. Conclusiones

En este trabajo se ha comparado el rendimiento de clasificación del modelo de regresión tensor y el modelo de variación total, bajo el indicadores gráfico de la curva ROC y explícitamente, el criterio dado por la cuantía del área bajo esta curva. Lo que permite decidir que modelo posee una mejor calidad de clasificación visual y de forma cuantitativa. Este proceso se llevo a cabo mediante diferentes estudios de simulación y a través de una aplicación con datos reales.

Los resultados de la simulación arrojaron que el modelo de variación total genera buenos resultados. El ABC obtenido por este siempre está por encima del arrojado por la regresión tensor y su varianza es muy pequeña, inclusive con tamaños de muestra relativamente pequeños. Esto difiere de la estimación arrojada por la regresión tensor, la cual presenta un impacto considerable al aumentar el tamaño de muestra.

Aunque el modelo de regresión tensor mejora sus resultados notablemente a medida que el tamaño de muestra aumenta (el ABC aumenta y la varianza se estabiliza), también se incrementa cuantiosamente el tiempo computacional que se requiere para la estimación de los coeficientes del modelo. La relación que existe entre el tamaño de muestra y la estimación del coeficiente imagen en el modelo de regresión tensorial se analiza también en el ejercicio preliminar, corroborando la anterior afirmación. Esto lleva a concluir que, si la muestra es suficientemente grande, entonces el modelo tensorial es relativamente bueno. Sin embargo, en comparación con el modelo funcional, este último obtiene mejores resultados en menor tiempo.

Adicionalmente, a través de los ejercicios con datos reales se observó que el proceso de clasificación (para ambos modelos) se ve influenciado cuando la muestra no se encuentra balanceada, es decir, el número de clases no es proporcional. El modelo de variación total presenta un buen resultado inclusive disminuyendo la resolución de las observaciones, esto se ve reflejado en que su curva ROC presenta una convexidad más pronunciada a lo largo de la trayectoria con respecto al modelo tensor. De esta forma, para un valor α fijo (probabilidad de encontrarse enfermo) se puede entonces afirmar que, si la frecuencia con la que ambos clasificadores cometen un error al clasificar el enfermo como sano es la misma, entonces la curva ROC refleja que la capacidad del clasificador para detectar individuos enfermos es mayor

con el modelo propuesto por la variación total en comparación con el modelo de regresión tensor. También vale la pena mencionar que existen valores de α en los cuales las curvas se solapan (**6-13**), más específicamente, la capacidad de detección de individuos de clase 1 es igual para ambos modelos.

Pese a todos los resultados arrojados, que muestran una gran superioridad del modelo de variación total sobre el modelo de regresión tensor, la metodología tensorial presenta una ventaja que es importante mencionar. El modelo de variación total supone funcionales bidimensionales sobre un dominio de igual cardinalidad, mientras que el modelo de regresión tensor tiene a consideración arreglos $D - dimensionales$ no necesariamente cuadrados. Lo anterior permite el uso del modelo tensorial en imágenes en escalas de tipo RGB, MRI en 3D o inclusive, sobre bases de datos compuestas por FMRI (Imagen por resonancia magnética funcional).

7.2. Recomendaciones

Antes de finalizar, se sugieren algunos recomendaciones con base en los resultados y conclusiones. Se recomienda la continuación del trabajo incluyendo técnicas como el uso de penalizaciones que sean útiles para la identificación y la obtención de resultados más robustos. Adicionalmente, es importante explorar el campo de limpieza y estandarización de imágenes, esto con el fin de reducir el ruido y retener un mayor porcentaje de información que permita mejores estimaciones y una optimización de los procesos de computo.

Por último, es fundamental la exploración y el desarrollo del modelo de variación total especialmente en dos campos: en primer lugar, sobre funcionales $D - dimensionales$ o por lo menos de dimensión $D = 3$ o $D = 4$, y en segundo lugar, sobre un modelo de clasificación basado en el espacio de funcionales de variación acotada con múltiples clases. En ambos casos, esto permitiría ampliar el espectro de uso de la metodología.

A. Anexo: Total variation Minimization by Augmented Lagrangian Alternating Direction Algorithms-TVAL3

En el presente anexo exhibimos la minimización de la variación total por el método de lagrangiano aumentado y los algoritmos de dirección propuestos en la construcción de TVAL3. Dado el problema de optimización:

$$\min_{x} F(x) \quad sujeto\, a \quad Dx = b, \tag{A-1}$$

Para resolver A-1 existen variedad de caminos en la literatura, el primer método de solución se da al optimizar la función Lagrangiana:

$$\mathcal{L}(x, \lambda) = F(x) - \lambda'(Dx - b),$$

Donde λ corresponde a el multiplicador de Lagrange, de esta forma para dar solución calculamos:

$$\nabla_x \mathcal{L}(x, \lambda) = 0 \implies \nabla F(x) - D'\lambda = 0 \quad y \quad \nabla_\lambda \mathcal{L}(x, \lambda) = 0 \implies Dx - b = 0. \tag{A-2}$$

Sin embargo, otra estrategia para resolver A-1, es cambiar la solución del problema de optimización con restricciones a la solución de una secuencia de problemas sin restricciones (iteraciones). Lo más sencillo es cambiar la restricción por una penalización, la cual se hace muy grande si se viola la convergencia del punto actual. D esta manera A-1 se transforma en:

$$\mathcal{L}(x, \delta^{(k)}) = F(x) + \frac{\delta^{(k)}}{2}||Dx - b||^2, \quad \text{para} \quad k = 1, 2, \ldots \tag{A-3}$$

El parámetro $\delta^{(k)}$ se incrementa conforme pasan las iteraciones. Con lo anterior en mente para la iteración $k - esima$, mediante el método de penalización cuadrática se puede hallar los puntos óptimos x para los cuales $\mathcal{L}(x, \delta^{(k)}) = 0$ con $\delta^{(k)}$ fijo. Para la siguiente iteración como condición inicial se usa $x = x^{(k)}$ y posteriormente se computa:

1. $x^{(k+1)} = \text{mín}_x \mathcal{L}(x, \delta^{(k)})$.

2. $\delta_{(k+1)} = \eta\delta_{(k)}$, con $\eta > 1$

Por otro lado, el gradiente de la función de penalización esta dado por

$$\nabla_x \mathcal{L}(x, \delta^{(k)}) = \nabla F(x) + D'[\delta(Dx - b)] = 0 \tag{A-4}$$

Es importante notar que el óptimo de la función de penalización debe ser el mismo que el óptimo del problema con restricciones, mediante la igualación de términos de A-4 con A-2 notamos que $\lambda \approx \delta(Ax - b)$, desafortunadamente esta aproximación es en general muy mala, esto se debe a que estamos multiplicando un termino muy grande con uno muy pequeño.El método del lagrangiano Aumentado se propuso con la finalidad de solventar la deficiencia de la penalización cuadrática para estimar los multiplicadores de Lagrange. La motivación es estimar simultáneamente x y λ en el proceso iterativo. Esto se logra combinando los términos que involucran a las restricción en el lagrangiano y en la penalización Cuadrática.

El algoritmo denominado *Total variation Minimization by Augmented Lagrangianand Alternating Direction Algorithms*, resulta ser una modificación de los algoritmos de búsqueda lineal no monotopicos, denominados *An Alternating Direction Method ADAL*, aplicado en un problema de minimización de variación total (véase en [9] y [12]). La principal razón de su uso se debe a que los algoritmos $ADAL$ no imponen diferenciabilidad en la función objetivo o sus derivadas.

El método de dirección alternativa, pertenecen a la familia del método clásico lagrangiano aumentado, que iterativamente resuelven el problema lineal restringido en la $k - esima$ iteración:

$$\mathcal{L}(x, \lambda, \delta^{(k)}) = F(x) + \frac{\delta^{(k)}}{2}||Dx - b||^2 - \lambda'(Dx - b),$$

con $\delta^{(k)}$ el parámetro de penalización del $k - esimo$ paso y λ el multiplicador de Lagrange.

B. Anexo: bases de Haar

El análisis de wavelets esta especialmente indicado para el estudio de señales con pulso o intermitencias, sucesos que ocurren de manera no periódica. Adicionalmente, los problemas que presentan discontinuidades, singularidades o gradientes altos requieren funciones de interpolación que tengan una buena localización. Es por esto que, descomposiciones de tipo Fourier devuelven muy pocos datos, debido a que estas pierden casi toda la información temporal. En 1909 Alfred Haar propone una secuencia de funciones como ejemplo de un sistema ortonormal contable para el espacio de las funciones de cuadrado integrable en la recta real. Algunas diferencias con la secuencia de funciones de las bases de Fourier son:

Fourier	**Wavelets**
Se eliminan las frecuencias más altas	Se eliminan los menores coeficientes
No se anulan en $[0,1)$ y son C^∞	Son funciones escalonadas y soportadas en $[0,1)$

Las bases Haar permiten representar en formas eficiente funciones con segmentos suaves, con discontinuidades o picos y se conforma de una función wavelet madre (véase en [11]):

$$\psi(t) = \begin{cases} 1 & si \quad 0 \leq t < 1/2 \\ -1 & si \quad 1/2 \leq t \leq 1 \\ 0 & eoc \end{cases} \qquad \text{y de una función indicadora} \qquad \phi(t) = I_{[0,1)}.$$

Cualquier función real continua se puede aproximar por combinaciones lineales de $\phi(t)$, $\phi(2t)$, $\phi(4t)$,$\cdots$, $\phi(2^k t)$, $\cdots$. Análogamente, el anterior resultado resultado también se tiene para $\psi(\cdot)$.

Ahora bien, este resultado se puede extender para funciones sobre el plano $\mathbb{R}^2$, una base para las funciones bivariadas (véase [4]) estará dado por suma de funciones de la forma:

$$\phi^d(u,v) = \phi^{d_1}(u)\phi^{d_2}(v), \quad con \quad d \in \{(0,1),(1,0),(1,1)\}.$$

Se puede notar que esta base de funciones incluye la función indicadora $I_{[0,1)\times[0,1)} = I_{[0,1)^2}$ y otras funciones definidas como

$$\phi^d_{j,\mathbf{k}}(\mathbf{u}) = \phi^d_{j,\mathbf{k}}(u,v) = 2^{j/2}\phi^d(2^j\mathbf{u} - \mathbf{k}),$$

con $j \geq 0, \mathbf{k} = (k_1, k_2) \in \{0, 1, 2, \ldots, 2^m - 1\} \times \{0, 1, 2, \ldots, 2^n - 1\}$ y $d \in \{(0,1), (1,0)\ y\ (1,1)\}$, este ultimo representa la dirección de la descomposición *vertical*, *horizontal* y *diagonal* respectivamente. De esta manera, dada una función en dos variables $X(.,.)$ entonces

$$X(u,v) = \sum_d \sum_j \sum_k w_{j,k}^d \phi_{j,k}^d(u,v),$$

donde $w_{j,k}^d$ contiene información acerca de X a una escala proporcional a 2^j, dirección d y ubicación $2^{-j}\mathbf{k}$.

C. Anexo: ejemplo puntual del calculo de las operaciones tensoriales definidas

En esta sección, dadas matrices y vectores particulares, se calculan las operaciones definidas en el capitulo 3 con el fin de presentar de una manera más clara como actúan cada de estas funciones. Los ejemplos se presentan a continuación:

Dado $m = 2$, $n = 2$, $p = 3$, $q = 2$ y las matrices $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ y $B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \\ 9 & 10 \end{pmatrix}$, el **producto de Kronecker** esta dado por:

$$A \otimes B = \begin{pmatrix} 1 \begin{pmatrix} 5 & 6 \\ 7 & 8 \\ 9 & 10 \end{pmatrix} & 2 \begin{pmatrix} 5 & 6 \\ 7 & 8 \\ 9 & 10 \end{pmatrix} \\ 3 \begin{pmatrix} 5 & 6 \\ 7 & 8 \\ 9 & 10 \end{pmatrix} & 4 \begin{pmatrix} 5 & 6 \\ 7 & 8 \\ 9 & 10 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 5 & 6 & 10 & 12 \\ 7 & 8 & 14 & 16 \\ 9 & 10 & 18 & 20 \\ 15 & 18 & 20 & 24 \\ 21 & 24 & 28 & 32 \\ 27 & 30 & 36 & 40 \end{pmatrix}.$$

La matriz resulta ser de tamaño $6 \times 4 = mp \times nq$.
Debido a que $n = 2$, $q = 2$, es posible calcular el **producto de Khatri-Rao** y este da como resultado:

$$A \odot B = \begin{pmatrix} (1,3) \otimes (5,7,9) & (2,4) \otimes (6,8,10) \end{pmatrix} = \begin{pmatrix} 5 & 12 \\ 7 & 16 \\ 9 & 20 \\ 15 & 24 \\ 21 & 32 \\ 27 & 40 \end{pmatrix}.$$

La vectorización de la matriz B da como resultado un vector cuyas componentes son las columbas de B apiladas una tras otra, es decir, $vec(B) = \begin{pmatrix} 5 & 7 & 9 & 6 & 8 & 10 \end{pmatrix}$. Para una mayor claridad de esta operación sobre tensores de dimensión $D > 2$, exponemos el ejemplo en el que se tiene un arreglo $3 - dimensional$ denotado B^* con $p_1 = 2$, $p_2 = 3$ y $p_3 = 2$, dado por $B^* = \begin{pmatrix} B_1^* & B_2^* \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 11 & 12 \end{pmatrix} & \begin{pmatrix} 5 & 6 \\ 7 & 8 \\ 9 & 10 \end{pmatrix} \end{pmatrix}$, visualmente B^* representa tomar la

primera matriz B_1^* fijándolo en el espacio y justo detrás de ella posicionar la matriz B_2^*. La vectorización del arreglo corresponde a una extensión del caso matricial, en el que se apila la primera matriz y luego la matriz siguiente a esta hasta finalizar, es decir,

$$vec(B^*) = (1, 3, 11, 2, 4, 12, 5, 7, 9, 6, 8, 10)\,.$$

Finalmente, para definir el *producto outer* se considera $p_1 = 3$, $p_2 = 2 = p_3$ con $\vec{b_1} = (5, 6, 7)$, $\vec{b_2} = (3, 4)$ y $\vec{b_3} = (1, 2)$, de este modo las componentes de $(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})$ están dadas por:

$$\begin{aligned}
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{111} &= 5 \times 3 \times 1 = 15 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{112} &= 5 \times 3 \times 2 = 30 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{121} &= 5 \times 4 \times 1 = 20 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{122} &= 5 \times 4 \times 2 = 40 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{211} &= 6 \times 3 \times 1 = 18 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{221} &= 6 \times 4 \times 1 = 24 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{212} &= 6 \times 3 \times 2 = 36 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{222} &= 6 \times 4 \times 2 = 48 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{321} &= 7 \times 4 \times 1 = 28 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{322} &= 7 \times 4 \times 2 = 56 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{311} &= 7 \times 3 \times 1 = 21 \\
(\vec{b_1} \circ \vec{b_2} \circ \vec{b_3})_{312} &= 7 \times 3 \times 2 = 42
\end{aligned}$$

Bibliografía

[1] Chambolle, A.: An algorithm for total variation minimization and application. En: *Comput. Imaging Vision.* 20 (2004), Nr. 1, p. 89–97

[2] Chen, C. ; He, L. ; Li, H. ; Huang, J.: Fast Iteratively Reweighted Least Squares Algorithms for Analysis-Based Sparsity Reconstruction. En: *Med. Image Anal.* 49 (2018), Nr. 1, p. 141–151

[3] Dobson, A.: *An Introduction to Generalized Linear Models 2nd ed.* Boca Raton : Chapman Hall-CRC, 2002

[4] Eckley, I. ; G., Nason: LS2W: Implementing the Locally Stationary 2D Wavelet Process Approach in R. En: *J. Stat. Softw.* 43 (2011), Nr. 3, p. 1–23

[5] Horváth, L. ; Kokoszka, P.: *Inference for Functional Data with Applications.* New York : Springer, 2012

[6] Hsing, T. ; Eubank, R.: *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators.* Chennai : Wiley Series in Probability and Statistics, 2015

[7] Kokoszka, P. ; Reimherr, M.: *Introduction to Functional Data Analysis.* New York : Chapman and Hall-CRC, 2017

[8] Lazar, N.: *The Statistical Analysis of Functional MRI Data.* New York : Springer, 2008

[9] Li, C. ; Yin, W. ; Jiang, H. ; Zhang, Y.: An Efficient Augmented Lagrangian Methodwith Applications to Total Variation Minimization. En: *Comput. Optim. Appl.* 56 (2013), Nr. 3, p. 507–530

[10] Metz, C.: Basic principles of ROC analysis. En: *Semin. Nucl. Med.* 8 (1978), Nr. 4, p. 283–298

[11] Nason, G.: *Wavelet Methods in Statistics with R.* New York : Springer, 2011

[12] Qin, Z. ; Ma, S. ; Goldfarb, D.: An Alternating Direction Method for Total Variation Denoising. En: *Optim. Methods Softw.* 30 (2015), Nr. 3, p. 594–615

[13] Ramsay, J. ; Silverman, B.: *Functional Data Analysis.* New York : Springer, 2005

[14] Reiss, PT. ; Ogden, RT.: Fuctional Principal Component Regression and Functional Partial Least Squares. En: *J. Amer. Statist. Assoc.* 102 (2007), Nr. 479, p. 984–996

[15] Reiss, PT. ; Ogden, RT.: Fuctional generalized Linear Models with Images as Predictors. En: *Biometrics* 66 (2010), Nr. 1, p. 61–69

[16] Reiss, T. ; Todd, R.: Functional Principal Component Regression and Functional Partial Least Squares. En: *J. Amer. Statist. Assoc.* 102 (2007), p. 984–996

[17] Rudin, L. ; Osher, S. ; Fatemi, E.: Nonlinear total variation based noise removal algorithms. En: *Physica D* 60 (1992), Nr. 1, p. 259–268

[18] Rudin, W.: *Principles of Mathematical Analysis.* New York : McGraw-Hill, 1976

[19] Toshio, S.: *Applied Matrix and Tensor Variate Data Analysis.* Tokyo : Springer, 2016

[20] Wang, X. ; Zhu, H.: Generalized Scalar-on-Image Regression Models via Total Variation. En: *J. Amer. Statist. Assoc.* 112 (2017), Nr. 519, p. 1156–1168

[21] Wang, Y. ; Yang, J. ; Yin, W. ; Zhang, Y.: A new alternating minimization algorithm for totalvariation image reconstruction. En: *SIAM J. Imaging Sci.* 1 (2008), Nr. 3, p. 248–272

[22] Worsley, K.: Developments in Random Field Theory. En: *Human Brain Function: Second Edition* 2 (2004)

[23] Zhou, H.: Matlab TensorReg Toolbox Version 1.0, Available online. (2017)

[24] Zhou, H. ; Li, L. ; Zhu, H.: Tensor Regression with Applications in Neuroimaging Data Analysis. En: *J. Amer. Statist. Assoc.* 108 (2013), Nr. 502, p. 540–552