Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Matemáticas

Detección comprimida

Autor: Beatriz Gómez Martín

Tutor: Eustasio del Barrio Tellado

Índice general

Capítulo 1

Introducción.

El Premio Princesa de Asturias de Investigación Científica y Técnica volvió a ser otorgado, tras muchos años, en el ámbito matemático. Este premio, representante de contribuciones importantes a la humanidad en los campos científico y tecnológico, fue concedido en 2020 a los matemáticos Ingrid Daubechies, Yves Meyer, Emmanuel Candès y Terence Tao. Los dos primeros fueron galardonados por sus importantes aportaciones a la teoría de wavelets, fundamentales en el procesamiento de datos. A su vez, Emmanuel Candès y Terence Tao recibieron el premio por su aportación al ámbito de compresión de señales con una nueva técnica bautizada como *Compressed sensing* o *detección comprimida*, objeto de estudio de este trabajo.

En los últimos años, los avances en compresión de datos han sido de vital importancia. Esto es debido a que la cantidad de información y de datos se ha multiplicado de forma significativa por el uso de internet y el desarrollo y demanda de contenido multimedia, entre otras razones. Por ello, surge la necesidad de investigar nuevas técnicas de compresión que nos permitan almacenar y transportar señales de forma eficiente y sin pérdida de información en el proceso.

El método clásico de reducción de la dimensión en tratamiento de datos es el análisis de componentes principales (PCA). Se trata de una técnica que permite transformar matrices de gran dimensión en otras que contienen información más condensada, facilitando la búsqueda de patrones o la visualización de los datos. Sin embargo, este método presenta una serie de desventajas, entre ellas el hecho de que las direcciones finales de las proyecciones dependen de los datos originales. Esta dependencia puede ocasionar problemas si queremos añadir datos al estudio, aunque sea en una cantidad

pequeña, puesto que sería necesario calcular de nuevo las componentes principales. Esto hace que se trate de un método computacionalmente costoso en grandes dimensiones. Por último, el PCA requiere un acceso total a los datos iniciales, acceso que en algunos casos puede estar restringido de tal forma que solamente se muestren unos pocos datos en cada periodo de tiempo.

Para salvar las desventajas mencionadas anteriormente, surge la técnica del compressed sensing, que fue introducida entre 2005 y 2006 por Candes y Tao [1] y por Donoho [2].

Supongamos que queremos almacenar una señal de gran dimensión como por ejemplo una imagen o una señal de audio. En lugar de tomar grandes cantidades de datos para seguidamente desechar la mayor parte de ellos en su compresión, el compressed sensing requiere solamente una pequeña cantidad de medidas repartida de forma aleatoria y permite reconstruir la señal original sin apenas pérdida de información en el proceso. A partir de ahora se hablará de señales dispersas, lo que se refiere a señales con un número pequeño de entradas no nulas. El compressed sensing asume la dispersión de la señal en una base determinada. La señal original $x \in \mathbb{R}^d$ es K-dispersa en alguna base si su representación en dicha base presenta $K \ll d$ entradas no nulas. Matricialmente, escribimos

$$x = U\alpha$$

siendo α la representación dispersa de x en la base formada por las columnas de la matriz $U \in \mathcal{M}^{d,d}(\mathbb{R})$. Como se ilustra en la figura (1.1), el vector de mediciones $y \in \mathbb{R}^n$, con $K < n \ll d$ viene dado por la expresión

$$y = Wx \tag{1.1}$$

donde $W \in \mathcal{M}^{n,d}(\mathbb{R})$ es la matriz de compressed sensing. La elección de esta matriz es un aspecto clave del proceso, que abordaremos en la el capítulo 4. La opción más habitual es tomar una matriz de proyecciones aleatorias, en cuyo caso las entradas de W son variables aleatorias siguiendo una distribución normal o de Bernoulli.

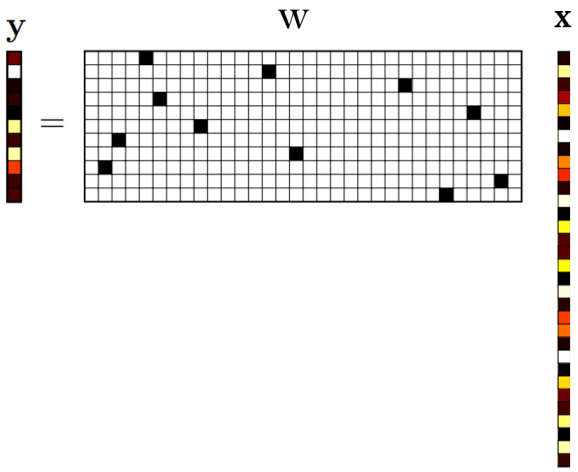

Figura 1.1: Esquema inicial del compressed sensing

Teniendo la información adicional de que la señal x es dispersa en cierta base formada por las columnas de una matriz U, podemos reconstruirla a partir de (1.1). Así, el objetivo es encontrar el vector α más disperso posible que cumpla

$$y = Wx = WU\alpha,$$

como se muestra en la figura (1.2). Además, denotando por Z a la matriz $Z = WU$ obtenemos el sistema representado en (1.3).

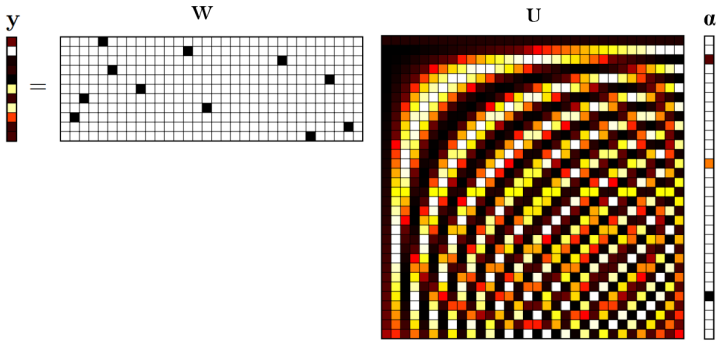

Figura 1.2: Esquema de compressed sensing tomando una expresión dispersa α de x en la base formada por las columnas de U.

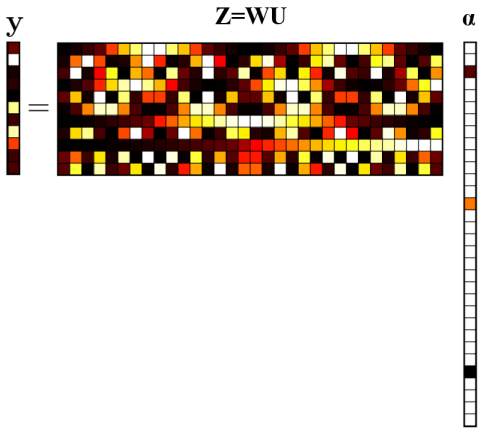

Figura 1.3: Esquema final del compressed sensing.

El sistema anterior es indeterminado, puesto que hay infinitos vectores α que satisfacen la condición pedida. La solución más dispersa entre todas las posibles la encontramos resolviendo el siguiente problema de optimización

$$\underset{\alpha}{argmin}\|\alpha\|_0 \text{ tal que } y = WU\alpha \tag{1.2}$$

donde definiremos $\|.\|_0$ como el número de entradas no nulas de un vector. Veremos en el capítulo 5 que el problema (1.2) es no convexo, y en gene-

ral solamente podemos encontrar la solución recorriendo todas las posibles opciones. Este problema es computacionalmente intratable. Es por esto que, probaremos que bajo ciertas condiciones sobre la matriz W o la matriz WU, es posible relajar el problema de optimización (1.2) a un problema convexo de minimización en l^1:

$$\underset{\alpha}{argmin}\|\alpha\|_1 \text{ tal que } y = WU\alpha.$$

Un problema convexo se puede resolver mediante métodos numéricos clásicos. Estudiaremos el descenso por coordenadas, con el cual podremos recuperar α completando así el proceso de compressed sensigng.

1.1. Objetivos y estructura.

El objetivo de este trabajo es describir y estudiar en detalle las distintas etapas de desarrollo del compressed sensing. Además, se programará el proceso mediante R y se propondrán algunos ejemplos de recuperación de la señal en imágenes.

El trabajo se estructura de la siguiente forma. En primer lugar se introducirá la técnica del análisis de componentes principales, que es uno de los antecedentes del compressed sensing. Seguidamente se propondrán en el capítulo 3 dos opciones de construcción de matrices ortogonales que permitan expresiones dispersas de la señal. En el capítulo 4 se estudiarán las propiedades de las matrices de proyecciones aleatorias, para proponer en el capítulo 5 un problema de minimización cuya resolución permita recuperar la señal dispersa. Además, se estudiarán una serie de condiciones que debe cumplir la matriz de compressed sensing para poder relajar el problema de optimización a un problema convexo. En último lugar, en el capítulo 6 se estudiará el método del descenso por coordenadas para la resolución del problema de minimización convexo y finalmente se mostrarán los resultados obtenidos en la programación del proceso.

Capítulo 2

Antecedentes del Compressed Sensing: PCA.

2.1. Introducción.

El análisis de componentes principales (PCA) es una técnica de reducción de la dimensión muy utilizada en campos como el procesamiento de imágenes o el tratamiento de datos. Se trata de una técnica que utiliza una transformación ortogonal para convertir un conjunto de observaciones posiblemente correlacionadas en un conjunto de variables sin correlación llamadas componentes principales. El nuevo conjunto es de dimensión mucho menor que el inicial, lo que es de gran utilidad cuando tratamos con grandes cantidades de datos. Supongamos que queremos comprimir m vectores x_1,...$x_m \in \mathbb{R}^d$ de forma que los vectores comprimidos sean de dimensión $n < d$. Podemos realizar la compresión mediante una matriz ortogonal $W \in \mathcal{M}^{n,d}(\mathbb{R})$ de tal forma que se envíe cada vector $x \in \{x_1, ..., x_m\}$ en el vector comprimido $v \in \mathbb{R}^n$

$$x \mapsto v = Wx.$$

Seguidamente, la descompresión se lleva a cabo mediante otra transformación de matriz $U \in \mathcal{M}^{d,n}(\mathbb{R})$, teniendo

$$v = Wx \mapsto \tilde{x} = Uv = UWx.$$

Veremos más adelante que esta transformación está definida de manera que la primera componente principal, que será la primera columna de la matriz U, tiene la mayor varianza posible. Cada una de las columnas, o

componentes principales sucesivas tendrá a su vez la mayor varianza posible bajo la condición de ortogonalidad con las componentes anteriores.

Un aspecto muy interesante es que el conjunto de vectores obtenido tras el PCA constituye una base ortogonal, que se utiliza en diversos textos, como en [11] para lograr la dispersión necesaria en en desarrollo de la técnica de compressed sensing.

2.2. Interpretación geométrica.

Ilustremos en qué consiste el PCA con un ejemplo en dos dimensiones. Supongamos que tenemos un conjunto de observaciones en $\mathbb{R}^2$. El vector que define la primera componente principal (PC1) sigue la dirección en la que las observaciones varían más. La proyección de cada punto sobre esa dirección equivale al valor de la primera componente para dicho punto.

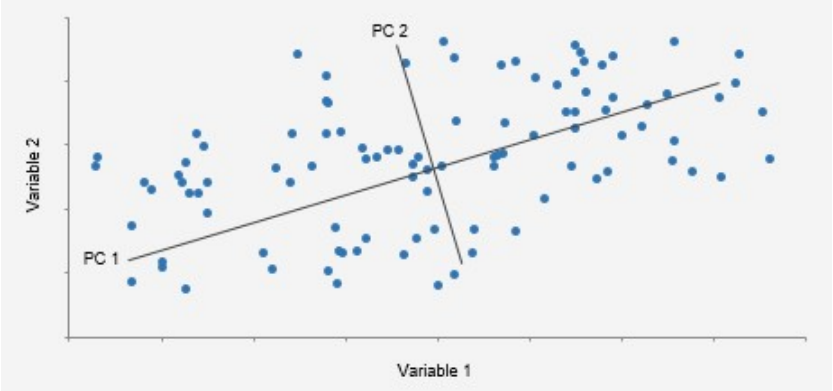

Figura 2.1: PCA en dos dimensiones

La segunda componente (PC2) sigue la segunda dirección en la que los datos muestran mayor varianza y que no está correlacionada con la primera componente. La condición de no correlación entre componentes principales equivale a decir que son ortogonales, por lo que como estamos trabajando en dimensión 2 tenemos que los vectores de PC1 y PC2 forman una base ortogonal de $\mathbb{R}^2$. En el caso de aplicar el análisis de componentes principales en un espacio de dimensión mayor, las componentes se ordenan de mayor a

menor cantidad de varianza reflejada.

El proceso de PCA identifica aquellas direcciones en las que la varianza es mayor. Por lo tanto, como la varianza de una variable se mide en su misma escala elevada al cuadrado, si antes de calcular las componentes no se estandarizan todas las variables para que tengan media 0 y desviación estándar 1, aquellas variables cuya escala sea mayor dominarán al resto. De ahí que sea recomendable estandarizar siempre los datos.

2.3. Base teórica.

Veamos a continuación dos interpretaciones equivalentes del PCA que podemos encontrar en [3]. La primera construcción se realiza mediante la minimización del error entre los datos originales y su reconstrucción imponiendo la condición de ortogonalidad en la transformación, mientras que la segunda procede mediante la maximización de la varianza explicada por cada componente.

El motivo por el que desarrollamos las dos contrucciones es que en la primera aseguramos que el error cometido en la transformación pequeño, y por otra parte la segunda proporciona la base para entender el proceso geométricamente.

2.3.1. PCA como minimización del error.

Consideramos m observaciones, $x_1, x_2, ..., x_m$ vectores de $\mathbb{R}^d$ con $m \gg d$. Queremos reducir la dimensión de los vectores anteriores mediante una matriz $W \in \mathcal{M}^{n,d}(\mathbb{R})$, $n < d$ correspondiente a una transformación lineal. Buscamos también la matriz de descompresión $U \in \mathcal{M}^{d,n}(\mathbb{R})$ de tal forma que se minimice el error entre el vector original y el recuperado tras el proceso, es decir, para cada $x \in \{x_1, ..., x_m\}$ queremos resolver el problema

$$\underset{U \in \mathcal{M}^{d,n}(\mathbb{R}), W \in \mathcal{M}^{n,d}(\mathbb{R})}{argmin} \|x - UWx\|_2^2. \tag{2.1}$$

El siguiente lema muestra que la solución al problema anterior se da cuando U es ortonormal y W su matriz traspuesta.

Lema 2.3.1. *Sea (U, W) una solución de (2.1). Entonces las columnas de U son ortonormales, es decir U^TU es la identidad de dimensión n, y $W = U^T$.*

Demostración. Fijamos U y W cualquiera y consideramos la aplicación $x \longmapsto UWx$. El subespacio $Im = \{UWx : x \in \mathbb{R}^d\}$ es un subespacio de $\mathbb{R}^d$ lineal por ser imagen de una aplicación lineal, y tiene dimensión a lo sumo n. Elegimos una base ortonormal $\{v_1, v_2, ..., v_n\}$ del subespacio imagen, y consideramos la matriz $V \in \mathcal{M}^{d,n}(\mathbb{R})$ cuyas columnas son los vectores $v_i, i = 1, ..., n$. Tenemos que $V^TV = I$ y cada vector $\tilde{x}$ en Im se puede expresar como $\tilde{x} = Vy$ con $y \in \mathbb{R}^n$. Entonces, para todo $x \in \mathbb{R}^d$ e $y \in \mathbb{R}^n$ tenemos

$$\|x - Vy\|_2^2 = (x - Vy)^T(x - Vy) = (x^T - y^TV^T)(x - Vy) = \\ \|x\|^2 + y^TV^TVy - 2y^TV^Tx = \|x\|^2 + \|y\|^2 - 2y^T(V^Tx). \quad (2.2)$$

Minimizamos la expresión anterior igualando el gradiente respecto a y a 0 obtenemos que $y = V^Tx$. Por lo tanto para todo x tenemos que

$$VV^Tx = \underset{\tilde{x} \in Im}{argmin} \|x - \tilde{x}\|_2^2. \quad (2.3)$$

y podemos sustituir U, W por V, V^T puesto que para todo (U, W) se verifica

$$\|x - VV^Tx\|_2^2 \leq \|x - UWx\|_2^2.$$

□

Utilizando el lema anterior, reescribimos el problema de optimización (2.1) de la siguiente forma

$$\underset{U \in \mathcal{M}^{d,n}(\mathbb{R}): UU^T = I}{argmin} \sum_{i=1}^{m} \|x_i - UU^Tx_i\|_2^2. \quad (2.4)$$

Simplificando la expresión, se verifica que para todo $x \in \mathbb{R}^d$

$$\|x - UU^Tx\|^2 = (x - UU^Tx)^T(x - UU^Tx)$$

$$= (x^T - x^TUU^T)(x - UU^Tx) = \|x\|^2 - 2x^TUU^Tx + x^TUU^TUU^Tx$$

$$= \|x\|^2 - x^TUU^Tx = \|x\|^2 - tr(U^Txx^TU),$$

donde el operador tr denota la traza de la matriz, es decir la suma de los elementos de su diagonal principal.

El problema de optimización (2.4) es por lo tanto equivalente a

$$\underset{U\in\mathcal{M}^{d,n}(\mathbb{R}):UU^T=I}{argmin} \sum_{i=1}^{m}\|x_i\|^2 - tr(U^T x_i x_i^T U) \tag{2.5}$$

y puesto que $\sum_{i=1}^{m}\|x_i\|^2$ no depende de la matriz U y la traza de una matriz es un operador lineal escribimos

$$\underset{U\in\mathcal{M}^{d,n}(\mathbb{R}):UU^T=I}{argmax} tr(U^T \sum_{i=1}^{m} x_i x_i^T U)$$

Sea $A = \sum_{i=1}^{m} x_i x_i^T$. A es simétrica, por lo que por la versión real del teorema espectral para operadores simétricos podemos descomponerla como $A = VDV^T$, donde D es diagonal y $V^TV = I$. Los elementos de la diagonal de D son los autovalores de A y las columnas de V los autovectores correspondientes. Asumimos sin pérdida de generalidad que $D_{1,1} \geq D_{2,2} \geq ... \geq D_{d,d}$ y como las entradas de A son mayores o iguales a 0, A es semidefinida positiva con lo que $D_{d,d} \geq 0$.

Veamos en el siguiente resultado que la solución al problema (2.4) es la matriz U cuyas columnas son los n autovectores de A correspondientes a los n mayores autovalores.

Teorema 2.3.2. *Sean $x_1, x_2, ..., x_m$ vectores arbitrarios de $\mathbb{R}^d$, $A = \sum_{i=1}^{m} x_i x_i^T$ y sean $u_1, ..., u_n$ n autovectores de la matriz A correspondientes a los n mayores autovalores. Entonces la solución al problema de optimización* (2.4) *es la matriz U cuyas columnas son $u_1, ..., u_n$.*

Demostración. Sea VDV^T la descomposición espectral de A. Fijamos una matriz $U \in \mathcal{M}^{d,n}(\mathbb{R})$ con columnas ortonormales y la matriz $B = V^TU$. Tenemos $VB = VV^TU = U$, y

$$U^TAU = B^TV^TVDV^TVB = B^TDB,$$

con lo que

$$tr(U^TAU) = \sum_{j=1}^{d} D_{j,j} \sum_{i=1}^{n} B_{j,i}^2.$$

Tenemos que $\sum_{j=1}^{d}\sum_{i=1}^{n} B_{j,i}^2 = n$, puesto que $B^TB = U^TVV^TU = U^TU = I$. Además, si tomamos otra matriz $\tilde{B} \in \mathcal{M}^{d,d}(\mathbb{R})$ tal que las primeras n

columnas son las columnas de B y tal que $\tilde{B}^T\tilde{B} = I$ tenemos $\sum_{i=1}^{d} \tilde{B}_{j,i}^2 = 1$, lo que implica que $\sum_{i=1}^{n} B_{j,i}^2 \leq 1$. Por consiguiente, tenemos

$$tr(U^T AU) \leq \max_{\beta\in[0,1]^d:\|\beta\|_1\leq n} \sum_{j=1}^{d} D_{j,j}\beta_j.$$

Se comprueba fácilmente que el lado derecho de la desigualdad es igual a $\sum_{j=1}^{n} D_{j,j}$. Hemos probado que para toda matriz U con columnas ortonormales se cumple que $tr(U^T AU) \leq \sum_{j=1}^{n} D_{j,j}$. Si tomamos U la matriz cuyas columnas son los n mayores autovectores de A obtenemos la igualdad, con lo que se concluye la prueba.

□

A los vectores $u_1, ..., u_n$ se les conoce con el nombre de componentes principales. Toda señal de entrada Z se puede expresar como combinación lineal de estas n componentes de forma que:

$$Z = \phi_1 u_1 + ... + \phi_n u_n.$$

Escribiendo esta expresión componente a componente tenemos:

$$Z_i = \phi_{1,i} u_{1,i} + ... + \phi_{n,i} u_{n,i}$$

Los coeficientes $\phi_{i,j}$ reciben el nombre de loadings o pesos, e indican la importancia que tiene la entrada j de la componente principal i sobre nuestra señal Z.

Observación:

Sabemos que la matriz de covarianzas de un vector x puede calcularse como

$$S = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)(x_i - \mu)^T,$$

donde $\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$.
Puesto que estamos trabajando con datos centrados, esta matriz S será

$$S = \frac{1}{m}\sum_{i=1}^{m} x_i^2,$$

que es la matriz A considerada en el teorema (2.3.2) multiplicada por una constante $1/m$. Esto muestra que para datos centrados las componentes principales se obtienen calculando los autovectores de la matriz de covarianzas.

2.3.2. PCA como maximización de la varianza.

En la observación anterior se introduce otra interpretación del análisis de componentes principales como maximización de la varianza, que se detalla a continuación. Esta segunda interpretación consiste en que dados $x_1, ..., x_m \in \mathbb{R}^d$ y otro vector aleatorio x con la misma distribución que $x_1, ..., x_m$ y con $\mathbb{E}[x] = 0$, queremos encontrar un vector unitario $w \in \mathbb{R}^d$ tal que la variable aleatoria $\langle w, x\rangle$ tenga varianza máxima.
Buscamos resolver el problema

$$\underset{w:\|w\|=1}{argmax} Var[\langle w, x\rangle] = \underset{w:\|w\|=1}{argmax} \frac{1}{m}\sum_{i=1}^{m}(\langle w, x_i\rangle)^2.$$

Utilizando que para todo vector unitario $w \in \mathbb{R}^d$ se verifica que

$$(\langle w, x_i\rangle)^2 = tr(w^T x_i x_i^T w)$$

llegamos al mismo problema de optimizacion estudiado en (2.5) tomando $n = 1$, $U \in \mathcal{M}^{d,1}(\mathbb{R})$. Por lo tanto la solución al problema de maximización de la varianza es el primer autovector de la matriz $A = \sum_{i=1}^{m} x_i x_i^T$. Ahora teniendo el autovector anterior w_1 como primera componente principal, buscamos $w_2 \in \mathbb{R}^d$ unitario que maximice la varianza de $\langle w_2, x\rangle$ y que no esté correlacionado con $\langle w_1, x\rangle$. La solución al problema de optimización

$$\underset{w:\|w\|=1,\mathbb{E}[(\langle w_1,x\rangle)(\langle w,x\rangle)]=0}{argmax} Var[\langle w, x\rangle]$$

será la segunda componente principal, o segundo autovector de la matriz A. Veámoslo: Tenemos en primer lugar que

$$\mathbb{E}[(\langle w_1, x\rangle)(\langle w, x\rangle)] = w_1^T \mathbb{E}[xx^T]w = m w_1^T A w.$$

Como w es autovector de A, tenemos que $Aw = \lambda_w w$ siendo λ_w el autovalor de A asociado a w. Por tanto,

$$m w_1^T A w = 0 \iff \langle w_1, w\rangle = 0$$

Nuestro problema de optimización es entonces

$$w* = \underset{w:\|w\|=1,\langle w_1,w\rangle=0}{argmax} \frac{1}{m}\sum_{i=1}^{m}(\langle w, x_i\rangle)^2 = \underset{w:\|w\|=1,\langle w_1,w\rangle=0}{argmax} tr(w^T \sum_{i=1}^{m} x_i x_i^T w).$$

Recordemos que el problema de PCA en el caso $n = 2$ es equivalente a encontrar $W \in \mathcal{M}^{d,2}(\mathbb{R})$ tal que se maximice la expresión

$$W^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T W.$$

Denotamos por w_1, w_2 las dos primeras columnas de W, que sabemos que son las dos primeras componentes principales. Veamos que $w* = w_2$: Tenemos

$$W^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T W = w_1^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w_1 + w_2^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w_2$$

y como $w*$ y w_1 son ortonormales

$$w_1^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w_1 + w_2^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w_2 \geq w_1^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w_1 + {w*}^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w*,$$

con lo que concluimos que $w* = w_2$.

Podemos proceder de la misma forma para $w_3, ..., w_n$ hasta obtener todas las componentes principales requeridas.

Esta interpretación del PCA justifica la interpretación geométrica explicada anteriormente.

Capítulo 3

Expresión de la señal en forma dispersa.

Las señales normalmente presentan una cantidad enorme de datos, entre los cuales la información relevante es extremadamente difícil de encontrar. Por ello, el procesamiento de una señal se vuelve mucho más simple y rápido cuando trabajamos con una representación dispersa de la misma, en la cual unos pocos coeficientes revelan la información que estamos buscando. Debido a lo anterior, la dispersión es la estructura en la cual se apoyan muchos de los modelos de compresión, en concreto el compressed sensing.

Para introducir la noción de dispersión nos basamos en la representación de nuestra señal en una base dada $\{u_j\}_{=1}^{d}$ de $\mathbb{R}^d$. Toda señal x $\in \mathbb{R}^d$ se puede representar en términos de la nueva base de tal forma que

$$x = \sum_{j=1}^{d} u_j \alpha_j$$

donde u_j son las columnas de la matriz U y $\alpha = [\alpha_1, \alpha_2, ..., \alpha_d]^T$ un vector de coeficientes de $\mathbb{R}^d$, el vector disperso que utilizaremos en el compressed sensing. Puesto que los vectores de la base son linealmente independientes, tenemos que la matriz U es inversible, y

$$x = U\alpha \longleftrightarrow \alpha = U^{-1}x.$$

Para encontrar un método de compresión con error mínimo es conveniente que la matriz anterior U sea ortogonal, es decir, que $U^TU = I$, donde I es la

matriz identidad. Tenemos entonces que $U^{-1} = U^T$ y cada coeficiente α_j se obtiene proyectando la señal en el vector j-ésimo de la base U, es decir,

$$\alpha_j = \langle x, u_j \rangle = \sum_{i=1}^{n} \langle x_i u_{i,j} \rangle$$

y en forma matricial $\alpha = U^T x$.

Definición 3.0.1. *Decimos que x es K-disperso en la base U si existe un vector α con $K \ll n$ entradas distintas de 0 que satisfaga la ecuación $x = U\alpha$.*

Una señal K-dispersa se puede comprimir de forma eficiente preservando solamente las entradas no nulas y su posición en el vector, para lo que se utilizan $O(K \log_2 n)$ bits: $\log_2 n$ para expresar la posición de cada entrada en forma binaria (puesto que con un número de bits l se pueden representar los números del 1 al 2^l) y una cantidad de bits independiente de n para almacenar cada coeficiente. Este proceso se llama codificación por transformación, y está basado en la existencia de una base adecuada en la que nuestra señal sea dispersa. Para señales que no son exactamente dispersas la compresión depende de la cantidad de coeficientes de α que queramos preservar, tomando aquellos con mayor valor absoluto.

En el desarrollo de la técnica de compressed sensing, nuestro primer objetivo es encontrar una representación dispersa del vector original x. Por lo tanto, en la siguiente sección desarrollaremos algunas de las posibles bases ortogonales en las cuales se ha comprobado que la mayoría de las señales de la naturaleza son dispersas.

3.1. Métodos para encontrar bases ortogonales adecuadas.

3.1.1. Transformada de coseno discreta (DCT).

La DCT [4] (transformada de coseno discreta) es una transformación comunmente utilizada en el procesamiento de datos que nos permite expresar

nuestra señal en términos de una suma de sinusoides con diferentes frecuencias y amplitudes. Existen varias expresiones de la DCT, la que emplearemos a continuación por ser la más común es la DCTII.

DCT en señales de una dimensión.

Partimos de una señal en forma de vector, que podría ser una señal de audio, un electrograma o un vector que contenga los valores tomados por cierta variable. Tenemos que la señal $x \in \mathbb{R}^n$ está formada por n puntos $x_1, ...x_n$ correspondientes a las entradas del vector x. Los puntos $X_1, ..., X_n$ tras aplicar la DCT se obtienen mediante la siguiente transformación:

$$X_k = 2\sum_{i=0}^{n-1} x_i \cos(\frac{\pi}{n}(i+\frac{1}{2})k).$$

La versión ortonormal de la expresión anterior se consigue introduciendo un reescalado de factor $\sqrt{\frac{1}{4n}}$ cuando $k = 0$ y $\sqrt{\frac{1}{2n}}$ para $k > 0$, con lo que

$$X_k = \sqrt{\frac{1}{n}}x_0 + \sum_{i=0}^{n-1} x_i\sqrt{\frac{2}{n}} \cos(\frac{\pi}{n}(i+\frac{1}{2})k). \tag{3.1}$$

En forma matricial, podemos escribir la DCT como

$$X = Cx$$

donde $x = (x_1, x_2, ..., x_n)$, $X = (X_1, X_2, ..., X_n)$ y C es una matriz real $n \times n$ cuyo elemento (k, i) se define como:

$$C_{k,i} = \begin{cases} \sqrt{\frac{1}{n}} & \text{si } k = 0 \\ \sqrt{\frac{2}{n}} \cos(\frac{\pi}{n}(i+\frac{1}{2})k) & \text{si } k > 0 \end{cases}$$

C es la matriz de la DCT. Se trata de una matriz invertible, con lo que podemos afirmar que sus columnas forman una base de $\mathbb{R}^n$. Además, se cumple $C * C^T = I_n$, como se demuestra en [4].

DCT en señales de dos dimensiones.

Si nuestro interés se centra en el análisis de imágenes, necesitamos una transformada para señales de dos dimensiones [5]. Para ello, basta con aplicar la DCT de la sección anterior sobre las filas y luego sobre las columnas (o viceversa) de nuestra matriz de datos. Sea M una matriz $n \times n$, su transformada X en forma matricial es

$$X = C^T M C$$

donde $C \in \mathcal{M}^{n,n}(\mathbb{R})$.

Dicho de otra forma, la transformada viene dada por la expresión

$$X_{k_1,k_2} = \sum_{i_1=0}^{n-1}(\sum_{i_2=0}^{n-1} G(i_1)G(i_2)\ M_{i_1,i_2}\ \cos[\frac{\pi}{n}(i_2+\frac{1}{2})\ k_2])\ \cos[\frac{\pi}{n}(i_1+\frac{1}{2})\ k_1]),$$

donde hemos utilizado la función

$$G(u) = \begin{cases} \frac{1}{\sqrt{4n}} & \text{si } u = 0 \\ \frac{1}{\sqrt{2n}} & \text{si } u > 0 \end{cases}$$

para expresar el reescalado que garantiza la ortogonalidad visto en (3.1).

En la práctica, el coste computacional de la DCT bidimensional es demasiado grande, por lo que se ha desarrollado una técnica de cálculo más eficiente, denominada transformada rápida de coseno discreta que se detalla en [6].

3.1.2. Bases de Wavelets.

La transformada Wavelet se propuso como una alternativa al análisis mediante la transformada de Fourier. La señal de entrada no se escribe como combinación lineal de sinusoides, sino como una combinación de escalamientos y traslaciones de una función prototipo llamada generalmente wavelet madre. La primera wavelet madre surgió en 1910, cuando Haar construyó la

siguiente función constante a trozos llamada wavelet de Haar:

$$\Psi(t) = \begin{cases} 1 \text{ si } 0 \leq t < 1/2 \\ \\ -1 \text{ si } 1/2 \leq t < 1 \\ \\ 0 \text{ en otro caso} \end{cases}$$

Se trata del wavelet más simple posible, en el que la no continuidad de Ψ constituye una ventaja para el análisis de señales con transiciones repentinas, tales como el monitoreo de fallo de una herramienta en una fábrica. A partir de la propuesta de Haar, se diseñaron otras wavelets madre útiles para distintos tipos de señales.

Además, en los sistemas wavelet, las wavelet madre vienen acompañadas de una función auxiliar llamada función de escala, que se denota ϕ. La función de escala del wavelet de Haar se define como

$$\phi(t) = \begin{cases} 1 \text{ si } 0 \leq t < 1 \\ \\ 0 \text{ en otro caso} \end{cases}$$

En la figura (3.1) se representan las funciones wavelet y de escala del wavelet de Haar.

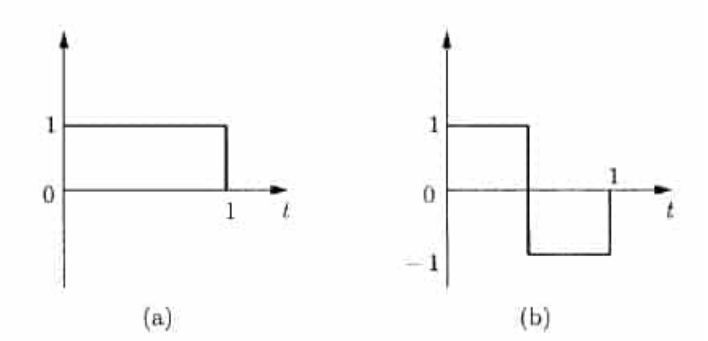

Figura 3.1: (a)Función de escala ϕ. (b)Función wavelet Ψ.

El aspecto más importante de las wavelet madre es que mediante traslaciones y dilataciones de las mismas se origina una base ortonormal de $L^2(\mathbb{R})$.

Si Ψ es la wavelet madre considerada, la familia de funciones que forman la base está definido de la siguiente manera:

$$\Psi_{m,n}(t) = \frac{1}{\sqrt{m}}\Psi\left(\frac{t-n}{m}\right)$$

donde $m \in \mathbb{Z}$ se denomina parámetro de escala y $n \in \mathbb{Z}$ parámetro de traslación. En muchos casos la wavelet madre esá dilatada o re-escalada en escalas que son potencias de 2, es decir,

$$\Psi_{m,n}(t) = 2^{-m/2}\Psi(2^{-m}t - n),$$

donde $m, n \in \mathbb{Z}$. Veamos un ejemplo de dilataciones y traslaciones en el wavelet de Haar en la figura (3.2).

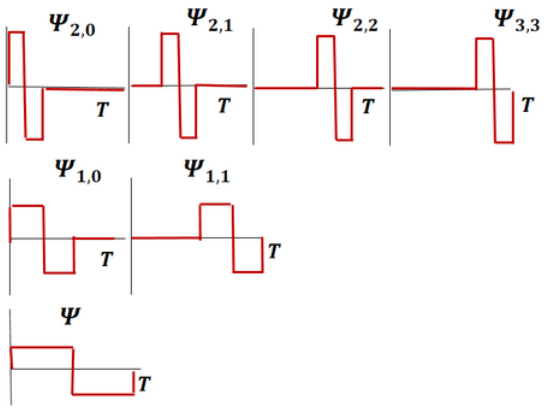

Figura 3.2: Dilataciones y traslaciones del wavelet de Haar Ψ.

Transformada wavelet discreta (DWT).

A partir de la wavelet madre es posible definir, para una función $f \in L^2(\mathbb{R})$ la transformada wavelet discreta (DWT) de la siguiente manera:

$$\tilde{f}(m,n) = \int_{-\infty}^{\infty} \bar{\Psi}_{m,n}(x)f(x)dx = \langle \Psi_{m,n}, f\rangle,$$

siendo $\langle . \rangle$ el producto escalar en L^2.

En el caso de que los parámetros m, n se reemplacen por números reales cualquiera obtenemos la transformada wavelet continua. Para más detalles sobre la transformada wavelet continua se puede consultar [8].

Teoría de bancos de filtros para el desarrollo de la DWT.

La forma de estudio más común de la transformada wavelet es el estudio mediante bancos de filtros. Por ello, a continuación se van a introducir nociones superficiales de filtros. Para más detalle se puede consultar [9].

- Un filtro es una secuencia de valores que se emplea para destacar o suavizar ciertos aspectos en una señal. El filtro es desplazado sobre la señal calculando un producto interno entre los coeficientes del filtro y aquellos de la señal sobre los que se encuentra.

- Los filtros pueden ser finitos o infinitos. Aquellos que tienen un número finito de coeficientes son llamados filtros de respuesta de impulso finita o FIR (Finite Impulse Response).

- Dadas la función de escala ϕ y la wavelet madre Ψ existen dos tipos de filtros h y g que cumplen

$$\phi(t) = 2\sum_{k=0}^{L-1} g(k)\phi(2t-k)$$

$$\Psi(t) = 2\sum_{k=0}^{L-1} h(k)\Psi(2t-k)$$

  donde k toma valores discretos en $0, 1, ...L-1$, siendo L la longitud del filtro elegido. Las ecuaciones anteriores son llamadas la ecuación de dilatación y la ecuación wavelet, respectivamente. Para unos coeficientes $h(k)$ y $g(k)$ que cumplan unas condiciones determinadas podemos crear las funciones de escala y wavelet correspondientes. En la práctica no es necesario construir las funciones y basta con trabajar con los coeficientes de los filtros.

- Un conjunto de filtros constituye un banco de filtros. Un ejemplo simple de banco de filtros es el formado por los filtros definidos en el punto anterior: un filtro de paso-bajo g, que permite el paso de las frecuencias más bajas y atenúa las frecuencias más altas, y un filtro de paso-alto h, que permite el paso de las frecuencias más altas. Si a una señal se le aplica este banco de filtros, se tiene como resultado dos nuevas señales, una con las frecuencias bajas de la señal y la otra con las frecuencias altas.

- En 1976 A.Croisier, D.Esteban y C.Galand [10] introdujeron el concepto de filtro espejo en cuadratura (QMF del inglés *quadrature mirror filter*). Dos filtros que cumplen que el filtro de paso-alto es calculado como el filtro espejo en cuadratura del filtro de paso-bajo garantizan una compresión sin pérdidas, como se muestra en la referencia citada. Esto quiere decir que si se satisface la siguiente fórmula:

$$g[L-1-i] = (-1)^i h[i],$$

siendo L la longitud de los filtros, podemos aplicar a la transformada su inversa y recuperar completamente la señal de inicio.

Forma matricial de la DWT.

En esta sección buscamos, al igual que para la DCT, la forma matricial de la transformada wavelet discreta. En [7] (pág.136) se propone el siguiente método de cálculo de una matriz ortogonal para la expresión de la DWT. La recuperación de la señal original se hallará utilizando la matriz inversa, la cual será igual a la traspuesta por la ortogonalidad de la matriz.

Para entender la formación de la matriz de la DWT, comenzaremos con un ejemplo de transformada en el que intervienen los elementos siguientes:

- La señal $x \in \mathbb{R}^8$.
- Dos filtros de longitud 6, uno paso-alto (h) y otro paso-bajo (g) que cumplen la propiedad de QMF.

Consideramos la matriz T siguiente:

$$Tx = \begin{bmatrix} h[1] & h[0] & 0 & 0 & 0 & 0 & 0 & 0 \\ g[1] & g[0] & 0 & 0 & 0 & 0 & 0 & 0 \\ h[3] & h[2] & h[1] & h[0] & 0 & 0 & 0 & 0 \\ g[3] & g[2] & g[1] & g[0] & 0 & 0 & 0 & 0 \\ h[5] & h[4] & h[3] & h[2] & h[1] & h[0] & 0 & 0 \\ g[5] & g[4] & g[3] & g[2] & g[1] & g[0] & 0 & 0 \\ 0 & 0 & h[5] & h[4] & h[3] & h[2] & h[1] & h[0] \\ 0 & 0 & g[5] & g[4] & g[3] & g[2] & g[1] & g[0] \\ 0 & 0 & 0 & 0 & h[5] & h[4] & h[3] & h[2] \\ 0 & 0 & 0 & 0 & g[5] & g[4] & g[3] & g[2] \\ 0 & 0 & 0 & 0 & 0 & 0 & h[5] & h[4] \\ 0 & 0 & 0 & 0 & 0 & 0 & g[5] & g[4] \end{bmatrix} \begin{bmatrix} x[0] \\ x[1] \\ x[2] \\ x[3] \\ x[4] \\ x[5] \\ x[6] \\ x[7] \end{bmatrix} = \begin{bmatrix} y[0] \\ y[1] \\ y[2] \\ y[3] \\ y[4] \\ y[5] \\ y[6] \\ y[7] \\ y[8] \\ y[9] \\ y[10] \\ y[11] \end{bmatrix}$$

Observamos que de esta forma T tiene dimension 12×8, con lo que el vector transformado y tiene dimensión 12. Con el objetivo de obtener como salida un vector de igual dimensión que el de entrada, vamos a considerar la matriz truncada eliminando las dos primeras y las dos últimas filas. Este truncado no es único, elegimos esta forma para conservar la "forma simétrica"de la matriz.

$$T' = \begin{bmatrix} h[3] & h[2] & h[1] & h[0] & 0 & 0 & 0 & 0 \\ g[3] & g[2] & g[1] & g[0] & 0 & 0 & 0 & 0 \\ h[5] & h[4] & h[3] & h[2] & h[1] & h[0] & 0 & 0 \\ g[5] & g[4] & g[3] & g[2] & g[1] & g[0] & 0 & 0 \\ 0 & 0 & h[5] & h[4] & h[3] & h[2] & h[1] & h[0] \\ 0 & 0 & g[5] & g[4] & g[3] & g[2] & g[1] & g[0] \\ 0 & 0 & 0 & 0 & h[5] & h[4] & h[3] & h[2] \\ 0 & 0 & 0 & 0 & g[5] & g[4] & g[3] & g[2] \end{bmatrix} = M = \begin{bmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \end{bmatrix},$$

donde $m_k,\ k = 0, 1, ..., 7$ es la k-ésima fila de la matriz.
Las filas de T' son linealmente independientes, lo cual es obvio por la propiedad QMF de los filtros, y podemos entonces aplicar el proceso de ortogonalización de Gram-Schmidt. Las filas m_2 a m_5 ya son ortogonales, puesto que no han sido truncadas. En primer lugar ortogonalizamos m_0 respecto a $m_2,\ m_3,\ m_4,\ m_5$.

$$m_0' = m_0 - \sum_{i=2}^{5} \frac{m_i m_0^T}{\|m_i\|^2} m_i,$$

y ortogonalizamos m_1 respecto a m_0' y a $m_2,\ m_3,\ m_4,\ m_5$

$$m_1' = m_1 - \frac{m_0' m_1^T}{\|m_0'\|^2} m_0' - \sum_{i=2}^{5} \frac{m_i m_1^T}{\|m_i\|^2} m_i.$$

Hacemos lo mismo con m_6 y m_7 teniendo en cuenta que ya son ortogonales a m_0 y m_1 puesto que los ceros se sitúan en distinto lugar. Así

$$m_7' = m_7 - \sum_{i=2}^{5} \frac{m_i m_7^T}{\|m_i\|^2} m_i,$$

$$m_6' = m_6 - \frac{m_7' m_6^T}{\|m_7'\|^2} m_7' - \sum_{i=2}^{5} \frac{m_i m_6^T}{\|m_i\|^2} m_i.$$

El resultado es entonces la matriz ortogonal

$$M' = \begin{bmatrix} m'_0 \\ m'_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m'_6 \\ m'_7 \end{bmatrix},$$

con la que podremos representar la señal x de forma dispersa como queríamos.

Caso general:
En el caso general, si tenemos una señal $x \in \mathbb{R}^N$ con N par, para cualquier filtro siempre es posible truncar la matriz T de forma que el resultado sea una matriz $N \times N$. Este truncado siempre origina una matriz M de rango máximo. Si tenemos filtros de longitud L, Los m_i serán filtros truncados por la izquierda para $i = 0, 1, ..., L/2-1$, filtros enteros para $i = L/2, N-L/2$, y filtros truncados por la derecha en las $L/2-1$ últimas filas. Ortogonalizamos mediante Gram-Schmidt las filas desde m_0 hasta $m_{L/2-2}$ respecto a ellas mismas:

$$m'_k = m_k - \sum_{i=0}^{k-1} \frac{m_i m_k^T}{\|m_i\|^2} m_i, \quad k = 0, 1, ..., L/2-2,$$

y de la misma forma los vectores desde $m_{N-L/2+2}$ hasta m_{N-1}.

DWT en dos dimensiones.

Consideramos ahora la transformada wavelet discreta en una señal bidimensional, en nuestro caso una imagen. Supongamos que tenemos la matriz M almacenando los valores numéricos de los píxeles de la imagen. $M \in \mathcal{M}^{n,n}(\mathbb{R})$, donde $n = 2^R$ para algún R. Hallamos como en la sección anterior la matriz ortogonal $T \in \mathcal{M}^{n,n}(\mathbb{R})$ de la DWT unidimensional, y aplicamos la transformación primero por filas y luego por columnas (o viceversa). En forma matricial

$$X = TMT^T.$$

En conclusión, hemos presentado dos métodos distintos para la obtención de una base ortogonal en la cual la señal de entrada x tenga una representación dispersa. La elección de un método u otro dependerá de la naturaleza de la señal y del coste computacional de cada uno de los métodos.

Capítulo 4

Proyecciones aleatorias.

En la sección anterior hemos visto la forma de transformar una señal x calculando su expresión en una base ortogonal de matriz $U \in \mathcal{M}^{d,d}(\mathbb{R})$ de tal forma que la señal resultante α sea dispersa. Una forma simple de comprimir el vector x sería multiplicarlo por U^T y almacenar los pares (coeficiente, posición) indicando las entradas no nulas del vector α resultante. Sin embargo, esto requiere en primer lugar medir x, almacenarlo y después multiplicarlo por U^T.

Entonces, la pregunta que se nos plantea ahora es, ¿por qué medir tantos datos si una enorme cantidad de ellos va a ser desechada en el proceso de comprimido?

Esta pregunta es la base del método de compressed sensing. Concretamente, en lugar de adquirir directamente el vector x, el método se basa en tomar $n \ll d$ medidas de la forma $y = Wx$ (repartidas de forma aleatoria) utilizando como matriz de compresión $W \in \mathcal{M}^{n,d}(\mathbb{R})$. W será la matriz de compressed sensing correspondiente a una proyección aleatoria.

El objetivo de esta sección será estudiar las propiedades de las proyecciones aleatorias, para concluir que una matriz aleatoria nos permite comprimir y descomprimir la señal con poca pérdida de información en el proceso.

Además, veremos en el capítulo 5 que bajo ciertas condiciones sobre la matriz de compressed sensing W podemos pasar de un problema de minimización que no se resuelve con los métodos numéricos clásicos, a otro equivalente y que se puede resolver mediante métodos de descenso del gradiente. En concreto, una de las condiciones que veremos es la propiedad de isometría

restringida (RIP), propiedad que satisfacen las matrices aleatorias con probabilidad alta cuando la dimensión es lo suficientemente grande.

El método de las proyecciones aleatorias se basa en proyectar los datos en direcciones aleatorias que son independientes de los datos en sí, lo que es computacionalmente sencillo y eficiente sobre todo a medida que la dimensión de las señales aumenta. Un aspecto interesante de las proyecciones aleatorias es que con una probabilidad alta preservan la distancia entre los datos originales y los transformados. En el plano teórico, el método de las proyecciones aleatorias se basa en el lema de Johnson-Lindestrauss [12], propuesto en 1984, que afirma que *un conjunto de puntos en un espacio de dimensión alta puede ser proyectado en otro espacio de dimensión inferior de forma que las distancias relativas entre los puntos se preservan.* El espacio de dimensión menor es seleccionado de forma aleatoria basándose en una distribución dada.

A continuación, analizaremos los resultados teóricos que respaldan esta técnica de transformación de los datos, demostrando entre otros resultados el lema de Johnson-Lindestrauss.

Definición 4.0.1. *Una proyección aleatoria de una señal $x \in \mathbb{R}^d$ es una transformación lineal $x \mapsto Wx$ donde W es una matriz aleatoria.*

Nuestro objetivo es encontrar una transformación de matriz W que distorsione los datos en la menor medida posible, es decir, para x_1 y x_2 vectores de $\mathbb{R}$ la distancia entre ambos vectores antes y después de la transformación debe ser casi la misma. Buscamos una matriz tal que el cociente

$$\frac{\|Wx_1 - Wx_2\|}{\|x_1 - x_2\|}$$

sea próximo a 1. Denotando por x a la diferencia $x = x_1 - x_2$, nos centraremos en el estudio de $\frac{\|Wx\|}{\|x\|}$.

4.1. Lema de Johnsons-Lindestrauss.

El lema de Johnsons-Lindestrauss [12] es un resultado que establece que un conjunto de puntos en un espacio de alta dimensión se puede proyectar

en un espacio de dimensión mucho más baja de tal manera que las distancias entre los puntos se conservan casi por completo, es decir, la inmersión utilizada es casi una isometría (Ver la figura (4.1)). Esta inmersión es al menos Lipschitz, y también puede tomarse como una proyección ortogonal.

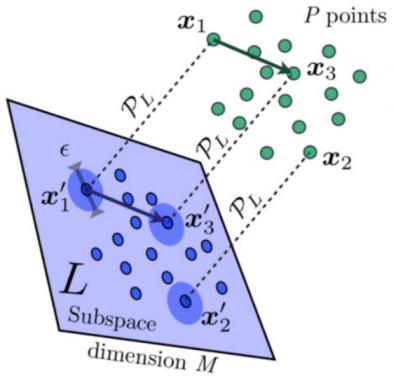

Figura 4.1: Representación gráfica del lema de Johnsons-Lindestrauss.

Utilizaremos una variante de este resultado para demostrar que la distorsión provocada por la aplicación de una proyección aleatoria distribuida normalmente sobre nuestra señal es pequeña con probabilidad $1-\delta$, donde la tolerancia δ se puede elegir en $(0,1)$.

Lema 4.1.1 (Johnsons-Lindestrauss)**.** *Sea Q un conjunto finito de vectores de $\mathbb{R}^d$, $\delta \in (0,1)$ y n un entero tal que*

$$\epsilon = \sqrt{\frac{6\log(2|Q|/\delta)}{n}} \leq 3$$

Entonces, si $W \in \mathcal{M}^{n,d}(\mathbb{R})$ es una matriz aleatoria con entradas independientes siguiendo una distribución normal de media 0 y varianza 1/n, con probabilidad al menos 1-δ sobre la elección de W tenemos que

$$\sup_{x\in Q}\left|\frac{\|Wx\|^2}{\|x\|^2}-1\right| < \epsilon.$$

Introducimos el siguiente lema intermedio que facilitará la prueba del lema de Johnsons-Lindestrauss. En la demostración de este lema se utilizará la propiedad de concentración de las variables que siguen una distribución χ^2. Esta propiedad se puede encontrar detallada en el anexo final.

Lema 4.1.2. *Sea $x \in \mathbb{R}^n$ y sea $W \in \mathbb{R}^{n,d}$ una matriz aleatoria con entradas independientes siguiendo una normal estándar. Entonces, para cada $\epsilon \in (0,3)$ tenemos*

$$\mathbb{P}\left[\left|\frac{\|(1/\sqrt{n})Wx\|^2}{\|x\|^2} - 1\right| > \epsilon\right] \leq 2e^{-\epsilon^2 n/6}$$

Demostración. Podemos asumir sin pérdida de generalidad que $\|x\|^2 = 1$. Reformulamos entonces el problema como

$$\mathbb{P}[|\|(1/\sqrt{n})Wx\|^2 - 1| > \epsilon] \leq 2e^{-\epsilon^2 n/6},$$

o lo que es lo mismo

$$\mathbb{P}[(1-\epsilon)n \leq \|Wx\|^2 \leq (1+\epsilon)n] \geq 1 - 2e^{-\epsilon^2 n/6}$$

Sea w_i la i-ésima fila de la matriz W. El producto $< w_i, x >$ es suma de variables aleatorias distribuidas normalmente, por lo que sigue una distribución de media 0 y varianza $\sum_j x_j^2 = \|x\| = 1$. La variable aleatoria $\|Wx\|^2 = \sum_{i=1}^n (< w_i, x >)^2$, es suma de distribuciones $N(0,1)$ elevadas al cuadrado, por lo que Wx sigue una distribución χ^2. Aplicando las desigualdades de concentración para variables χ^2 vistas en el anexo (desigualdad (8.4)) obtenemos el resultado buscado.

□

Demostración (Lema de Johnsons-Lindestrauss). Por el lema anterior, tenemos que para cada $\epsilon \in (0,3)$:

$$\mathbb{P}\left[\sup_{x \in Q}\left|\frac{\|Wx\|^2}{\|x\|^2} - 1\right| > \epsilon\right] \leq 2|Q|e^{-\epsilon^2 n/6}$$

Denotamos por δ al lado derecho de la desigualdad. Despejando ε obtenemos que

$$\epsilon = \sqrt{\frac{6\log(2|Q|/\delta)}{n}}$$

□

Cabe remarcar que ϵ no depende de la dimensión d de x, con lo que el lema se aplica también a espacios de Hilbert de dimensión infinita.

En conclusión, el lema de Johnsons-Lindestrauss garantiza que la distorsión provocada por una proyección aleatoria distribuida normalmente sobre nuestra señal es pequeña con probabilidad alta. Esto quiere decir que utilizar una matriz aleatoria es buena opción si buscamos proyectar la señal original en un subespacio de dimensión menor, y además, esta reducción de la dimensión de los datos es computacionalmente sencilla.

Capítulo 5

Formulación del problema de minimización.

La finalidad del compressed sensing es reconstruir una señal completa a partir de unas pocas mediciones de la misma distribuidas aleatoriamente, por lo que en nuestro caso el vector y de mediciones es conocido. El objetivo de esta sección es encontrar el problema de minimización que nos permita recuperar la señal $x \in \mathbb{R}^d$ a partir de su proyección $y \in \mathbb{R}^n$.

Además, el método de detección comprimida se basa en la existencia de una representación dispersa de la señal, por lo que el estudio de la dispersión de los datos motiva la siguiente definición.

Definición 5.0.1. *Definimos la cantidad $\|x\|_0$ como el número de coeficientes no nulos de un vector x. Esto es:*

$$\|x\|_0 := \{\#k : x_k \neq 0\}$$

$\|x\|_0$ se denomina norma 0, aunque en realidad no es una norma. Esta denominación está justificada por el hecho de que $\|x\|_q \to \|x\|_0$ cuando $q \to 0$, como se muestra en la figura siguiente:

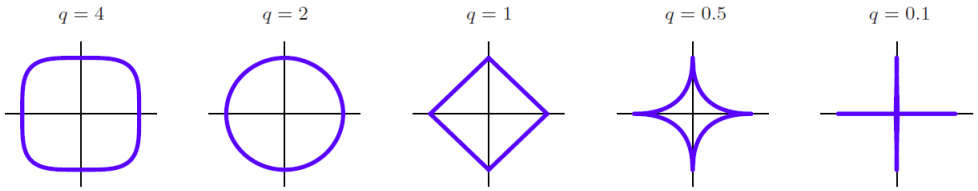

Figura 5.1: Valores de la norma $\|x\|_q$ para distintos valores de q en el caso de un espacio de dos dimensiones.

Nuestro objetivo es entonces recuperar la señal x a partir de su compresión y, sin embargo, el hecho de que $n < d$ implica que la matriz W no puede tener rango máximo. Por el teorema de rango-nulidad el núcleo de la transformación ha de ser no trivial, lo que implica que para una señal $x_0 \in \mathbb{R}$ existen infinitos vectores x tales que $y = Wx = Wx_0$ para la matriz dada W.

La clave del compressed sensing reside en que si tenemos la información adicional de que $x = U\alpha$ donde α es disperso, podemos recuperar x de forma exacta aunque el sistema sea altamente indeterminado. Para ello, buscamos el vector más disperso posible que cumpla $\alpha = U^T x$. Queremos resolver el problema l^0 siguiente:

$$min\|\alpha\|_0 \quad \text{tal que} \quad y = Wx = WU\alpha. \tag{5.1}$$

A partir de ahora consideraremos la matriz Z obtenida mediante el producto de la matriz de la proyección aleatoria W y la matriz U de la base ortogonal que nos permite trabajar con el vector α disperso, $Z = WU$.

5.1. Convexificación del problema.

Queremos resolver ahora el problema de optimización (5.1), pero tenemos un problema con norma l^0 que no es convexo, lo que ocasiona dificultades en su resolución. Un problema no convexo es cualquier problema de optimización donde la función objetivo o la región delimitada por las restricciones no es convexa. Los problemas no convexos pueden presentar múltiples mínimos locales. Puede llevar un tiempo exponencial verificar que un problema no

convexo no tiene soluciones factibles, que la función objetivo es no acotada o que la solución óptima encontrada es única (pág.84 de [3]). Para verificar la unicidad, se tendrían que barrer todos los subconjuntos que puedan generar subsistemas $y_S = Sx_S$, y luego verificar todas las posibles soluciones. El costo de realizar esta búsqueda exhaustiva crece exponencialmente a medida que la dimensión aumenta.

La solución a este problema se ha estudiado en los últimos 20 años, con el avance de las técnicas de estadística y matemática aplicada a tratamiento de señales. Para encontrar una solución de forma sencilla, podemos aproximar el problema en l^0 por otro en l^q para un $q > 0$. Cuanto más pequeño sea q mejor será la aproximación, sin embargo, tal y como se ve en la figura (5.1), si tomamos $q < 1$ el problema de optimización no es convexo.

Por ello, vamos a considerar el problema relajado l^1, es decir, buscamos

$$min\|\alpha\|_1 \quad \text{tal que} \quad y = Z\alpha. \tag{5.2}$$

A continuación veremos que bajo ciertas condiciones sobre la matriz de detección comprimida Z se puede garantizar la equivalencia entre los problemas (5.1) y (5.2).

5.2. Equivalencia entre los problemas l_0 y l_1.

En esta sección daremos algunos resultados relativos a dos propiedades sobre la matriz Z que nos permiten sustituir el problema no convexo con norma 0 por el problema con norma 1. Estas propiedades son la propiedad de isometría restringida (RIP) y la propiedad de núcleo restringido (RN), y ambas implican que los problemas (5.1) y (5.2) son equivalentes. Además, si se satisface la equivalencia entre los dos problemas se cumple la propiedad de núcleo restringido.

5.2.1. RIP, propiedad de isometría restringida.

La propiedad de isometría restringida (RIP, en inglés *restricted isometry property*) caracteriza las matrices que son casi ortonormales cuando operan sobre vectores dispersos. El concepto fue introducido por Emmanuel Candès

y Terence Tao [1] y como veremos a continuación, las matrices con esta condición garantizan la equivalencia entre los problemas (5.1) y (5.2).

Definición 5.2.1. *Una matriz $W \in \mathcal{M}^{n,d}(\mathbb{R})$ es (ϵ,s)-RIP (satisface la propiedad de isometría restringida) si para todo $x \neq 0$ con $\|x\|_0 \leq s$ tenemos*

$$|\frac{\|Wx\|_2^2}{\|x\|_2^2} - 1| \leq \epsilon.$$

El siguiente teorema muestra que para un vector disperso α, las matrices RIP proporcionan un esquema de compresión de la señal α y la norma en l^0 sin pérdida de información en el proceso, es decir, que

Teorema 5.2.2. *Sea $\epsilon < 1$ y W una matriz (ϵ,2s)-RIP. Sea α un vector con $\|\alpha\|_0 \leq s$, $y = W\alpha$ la compresión de α, y sea*

$$\tilde{\alpha} \in \underset{v:Wv=y}{argmin}\|v\|_0$$

un vector reconstruido. Entonces $\tilde{\alpha} = \alpha$.

Demostración. Supongamos por reducción al absurdo que $\tilde{\alpha} \neq \alpha$. Puesto que α satisface $\alpha \in \underset{v:Wv=y}{argmin}\|v\|_0$, tenemos $\|\tilde{\alpha}\|_0 \leq \|\alpha\|_0 \leq s$. Además, $\|\alpha - \tilde{\alpha}\|_0 \leq 2s$. Por ser W RIP y puesto que $W\alpha = W\tilde{\alpha} = y$, $W(\alpha - \tilde{\alpha}) = 0$ tenemos que $|0 - 1| \leq \epsilon$, lo que lleva a contradicción.

□

Ahora que hemos garantizado que la solución al problema l^0 conduce a una reconstrucción perfecta de la señal dispersa α, veamos que esta solución es la misma que la del problema en l^1. Por lo tanto, podremos sustituir $\|v\|_0$ por $\|v\|_1$ lo cual nos permite convertir nuestro problema inicial en un problema convexo que puede ser resuelto de forma eficiente.

Teorema 5.2.3. *Supongamos que se verifican las condiciones del teorema (5.2.2) y que $\epsilon < \dfrac{1}{1+\sqrt{2}}$. Entonces,*

$$\alpha = \underset{v:Wv=y}{argmin}\|v\|_0 = \underset{v:Wv=y}{argmin}\|v\|_1.$$

Demostraremos un resultado aún más fuerte que el anterior, que se verifica para un vector cualquiera x no necesariamente disperso.

Teorema 5.2.4. *Sea $\epsilon < \frac{1}{1+\sqrt{2}}$, W una matriz $(\epsilon, 2s)$-RIP y x un vector arbitrario. Denotamos por x_s al vector igual a x en los s mayores elementos de x y nulo en el resto de entradas, esto es*

$$x_s \in argmin_{v:\|v\|_0 \leq s} \|x - v\|_1. \tag{5.3}$$

Sea $y = Wx$ la compresión de x y sea

$$x^* \in argmin_{v:Wv=y} \|v\|_1.$$

el vector reconstruido. Entonces,

$$\|x^* - x\|_2 \leq 2\frac{1+\rho}{1-\rho} s^{1/2} \|x - x_s\|_1,$$

donde $\rho = \sqrt{2}\epsilon/(1-\epsilon)$.

Demostración. Seguimos una demostración propuesta por Candès en 2008 [13]. Sea $h = x* - x$. Dados un vector v y un conjunto de índices I, denotamos por v_i el vector cuya i-ésima componente es el elemento v_i si $i \in I$ y 0 sino. Lo primero que haremos será partir el conjunto de índices $[d] = \{1, ..., d\}$ en subconjuntos de tamaño s disjuntos, $[d] = T_0 \sqcup T_1 \sqcup ... \sqcup T_{d/s-1}$ donde $|T_i| = s$ para todo i y asumimos por simplicidad que d/s es entero. En T_0 ponemos los s índices correspondientes a los s mayores elementos de x en valor absoluto. Sea $T_0^C = [d] \setminus T_0$, T_1 será el conjunto de los s índices correspondientes a los mayores elementos de $h_{T_0^C}$ en valor absoluto. Sea $T_{0,1} = T_0 \cup T_1$ y $T_{0,1}^C = [d] \setminus T_{0,1}$. T_2 contendrá los s índices de los elementos de $h_{T_{0,1}^C}$ con mayor valor absoluto y de igual forma construimos T_3, T_4... Para probar el resultado necesitamos el siguiente lema.

Lema 5.2.5. *Sea W una matriz $(\epsilon, 2s)$-RIP. Entonces, para dos conjuntos disjuntos I, J, los dos de tamaño a lo sumo s, y para todo vector u tenemos que $\langle Wu_I, Wu_J \rangle \leq \epsilon \|u_I\|_2 \|u_J\|_2$.*

Demostración. Sin pérdida de generalidad podemos asumir $\|u_I\|_2 = \|u_J\|_2 = 1$.

$$\langle Wu_I, Wu_J \rangle = \frac{\|Wu_I + Wu_J\|_2^2 - \|Wu_I - Wu_J\|_2^2}{4}.$$

Como $|J \cup I| \leq 2s$ tenemos que por la condición RIP $\|Wu_I + Wu_J\|_2^2 \leq (1+\epsilon)(\|u_I\|_2^2 + \|u_J\|_2^2) = 2(1+\epsilon)$ y $-\|Wu_I - Wu_J\|_2^2 \leq -(1-\epsilon)(\|u_I\|_2^2 + \|u_J\|_2^2) = -2(1-\epsilon)$. Con lo que

$$\langle Wi_I, Wu_J \rangle \leq \frac{2(1+\epsilon) - 2(1-\epsilon)}{4} = \epsilon$$

y concluimos el resultado. □

Ahora ya podemos continuar la demostración. Tenemos

$$\|h\|_2 = \|h_{T_{0,1}} + h_{T_{0,1}^C}\|_2 \leq \|h_{T_{0,1}}\|_2 + \|h_{T_{0,1}^C}\|_2. \tag{5.4}$$

Para probar el teorema vamos a mostrar las dos afirmaciones siguientes:

$$\text{Afirmación 1: } \|h_{T_{0,1}^C}\|_2 \leq \|h_{T_0}\|_2 + 2s^{-1/2}\|x - x_s\|_1.$$
$$\text{Afirmación 2: } \|h_{T_{0,1}}\|_2 \leq \frac{2\rho}{1-\rho} s^{-1/2}\|x - x_s\|_1.$$

Combinando las dos afirmaciones anteriores con (5.4) tenemos

$$\begin{aligned}\|h\|_2 &\leq \|h_{T_{0,1}}\|_2 + \|h_{T_{0,1}^C}\|_2 \leq 2\|h_{T_{0,1}}\|_2 + 2s^{-1/2}\|x - x_s\|_1 \\ &\leq 2(\frac{2\rho}{1-\rho} + 1)s^{-1/2}\|x - x_s\|_1 \\ &= 2\frac{1+\rho}{1-\rho} s^{-1/2}\|x - x_s\|_1\end{aligned}$$

como queríamos demostrar.

Afirmación 1: Para probar esta afirmación utilizamos el hecho de que $x*$ minimiza la norma l^1. Tomamos $j > 1$. Para cada $i \in T_j$ e $i' \in T_{j-1}$ tenemos que $|h_i| \leq |h'_i|$, con lo que $\|h_{T_j}\|_\infty \leq \|h_{T_{j-1}}\|_1 / s$. Así, tenemos

$$\|h_{T_j}\|_2 \leq s^{1/2}\|h_{T_j}\|_\infty \leq s^{-1/2}\|h_{T_{j-1}}\|_1.$$

Escribiendo lo anterior para $j = 2, 3, ...$ y utilizando la desigualdad triangular se obtiene

$$\|h_{T_{0,1}^C}\|_2 \leq \sum_{j \geq 2} \|h_{T_j}\|_2 \leq s^{-1/2}\|h_{T_0^C}\|_1 \tag{5.5}$$

Vamos a ver ahora que $\|h_{T_0^C}\|_1$ no puede ser grande. Por la definición de $x*$ tenemos que $norm x_1 \geq \|x*\|_1 = \|x+h\|_1$, y utilizando la desigualdad triangular

$$\|x\|_1 \geq \|x+h\|_1 = \sum_{i\in T_0^C} |x_i+h_i| + \sum_{i\in T_0} |x_i+h_i| \geq \|x_{T_0}\|_1 - \|h_{T_0}\|_1 + \|x_{T_0^C}\|_1 - \|x_{T_0^C}\|_1 \tag{5.6}$$

y como $\|x_{T_0^C}\|_1 = \|x - x_s\|_1 = \|x\|_1 - \|x_{T_0}\|_1$ tenemos que

$$\|h_{T_{0,1}^C}\|_2 \leq \|h_{T_0}\|_1 + 2\|x_{T_0^C}\|_1. \tag{5.7}$$

Combinando esto con (5.5) tenemos que

$$\|h_{T_{0,1}^C}\|_2 \leq s^{-1/2}(\|h_{T_0}\|_1 + 2\|x_{T_0^C}\|_1) \leq \|h_{T_0}\|_2 + 2s^{-1/2}\|x_{T_0^C}\|_1,$$

lo que concluye la prueba de la afirmación 1.

Afirmación 2: Para esta afirmación vamos a utilizar la condición RIP

$$(1-\epsilon)\|h_{T_{0,1}}\|_2^2 \leq \|Wh_{T_{0,1}}\|_2^2. \tag{5.8}$$

Puesto que $Wh_{T_{0,1}} = Wh - \sum_{j\geq 2} Wh_{T_j} = -\sum_{j\geq 2} Wh_{T_j}$ tenemos que

$$\|Wh_{T_{0,1}}\|_2^2 = -\sum_{j\geq 2} \langle Wh_{T_{0,1}}, Wh_{T_j}\rangle = -\sum_{j\geq 2} \langle Wh_{T_0} + Wh_{T_1}, Wh_{T_j}\rangle.$$

Por la condición RIP en los productos internos tenemos que para todo $i = 1, 2$ y $j \leq 2$

$$|\langle Wh_{T_i}, Wh_{T_j}\rangle| \leq \epsilon \|h_{T_i}\|_2 \|h_{T_j}\|_2.$$

Como $\|h_{T_0}\|_2 + \|h_{T_1}\|_2 \leq \sqrt{2}\|h_{T_{0,1}}\|_2$, obtenemos que

$$\|Wh_{T_{0,1}}\|_2^2 \leq \sqrt{2}\epsilon \|h_{T_{0,1}}\|_2 \sum_{j\geq 2} \|h_{T_j}\|_2.$$

Combinando esto con (5.5) y (5.8)

$$(1-\epsilon)\|h_{T_{0,1}}\|_2^2 \leq \sqrt{2}\epsilon \|h_{T_{0,1}}\|_2 s^{-1/2} \|h_{T_0^C}\|_1$$

con lo que

$$\|h_{T_{0,1}}\|_2 \leq \frac{\sqrt{2}\epsilon}{1-\epsilon} s^{-1/2} \|h_{T_0^C}\|_1.$$

Finalmente, utilizando (5.7) obtenemos

$$\|h_{T_{0,1}}\|_2 \leq \rho s^{-1/2}(\|h_{T_0}\|_1 + 2\|x_{T_0^C}\|_1) \leq \rho\|h_{T_0}\|_2 + 2\rho s^{-1/2}\|x_{T_0^C}\|_1,$$

y como $\|h_{T_0}\|_2 \leq \|h_{T_{0,1}}\|_2$ esto implica

$$\|h_{T_{0,1}}\|_2 \leq \frac{2\rho}{1-\rho} s^{-1/2}\|x_{T_0^C}\|_1,$$

con lo que concluimos el resultado. □

Nuestro objetivo es comprimir la señal α con una proyección aleatoria, ya que hemos probado anteriormente que este tipo de transformación nos garantiza un esquema con poca distorsión. Por ello vamos a probar que las matrices aleatorias $W \in \mathcal{M}^{n,d}(\mathbb{R})$ con dimensión n de orden mayor que $s\log(d)$ son $(\epsilon, s)-$RIP con probabilidad al menos $1-\delta$, donde podemos elegir la tolerancia δ como queramos en (0,1). Con este resultado, podremos sustituir el problema de minimización no convexo por el convexo en l^1 siempre que estemos trabajando con una matriz aleatoria lo suficientemente grande.

Observación. .
El resultado también muestra que la multiplicación de una matriz ortogonal por una aleatoria lo suficientemente grande da como resultado una matriz RIP.
Esta afirmación es de mucha utilidad, puesto que en nuestro problema inicial (5.1) la matriz $Z = WU$ se obtiene mediante el producto de W, que es una matriz de compresión aleatoria, y U que es la matriz ortogonal utilizada para representar el vector x de la forma dispersa α.

□

Teorema 5.2.6. *Sea $U \in \mathcal{M}^{d,d}(\mathbb{R})$ una matriz ortonormal arbitraria y $\epsilon, \delta \in (0,1)$ escalares. Sea s un entero en $[d]$ y sea n un entero que satisface*

$$n \geq \Omega(s\log(d)).$$

Sea $W \in \mathcal{M}^{n,d}(\mathbb{R})$ una matriz con entradas siguiendo una distribución normal de media 0 y varianza $1/n$. Entonces, con probabilidad $1-\delta$ sobre la elección de W, la matriz WU es $(\epsilon, s)-$RIP.

Demostración. La demostración sigue la prueba propuesta por Baraniuk, Davenport, De-Vore y Wakin en 2008 [15].

La idea es combinar el lema de Johnsons-Lindestrauss con la propiedad de recubrimiento de la bola unidad enunciada en el siguiente lema (la demostración de este lema se puede ver en [14]). Se define el número de recubrimiento $N(T, d, \epsilon)$ de un conjunto T en R^d como el número mínimo de bolas de radio ϵ necesarias para cubrir T. El lema nos dice que

$$N(\bar{B}_d(0,1), d, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^d,$$

cantidad que utilizaremos para acotar el cardinal del conjunto Q al aplicar el lema de Johnsons-Lindestrauss (lema 4.1.1).

Lema 5.2.7. *Sea $\epsilon \in (0,1)$. Existe un conjunto finito $Q \in \mathbb{R}^d$ de cardinal $|Q| \leq \left(\frac{3}{\epsilon}\right)^d$ tal que*

$$\sup_{x:\|x\|\leq 1} \min_{v\in Q} \|x - v\| \leq \epsilon.$$

Sea x un vector que puede ser expresado como $x = U\alpha$ donde U es una matriz ortonormal y $\|\alpha\|_0 \leq s$. Combinando el lema anterior con el lema de Johnsons-Lindestrauss podemos probar que nuestra matriz aleatoria W no distorsiona x. Introducimos el siguiente lema:

Lema 5.2.8. *Sea $U \in \mathcal{M}^{d,d}(\mathbb{R})$ una matriz ortonormal y sea $I \subset [d]$ un conjunto de índices de cardinal $|I| = s$. Sea S el subespacio generado por $\{u_i : i \in I\}$, donde u_i es la columna i-ésima de U. Sea $\delta \in (0,1)$, $\epsilon \in (0,1)$ y $n \in \mathbb{N}$ tal que*

$$n \geq 24 \frac{\log(2/\delta) + s\log(12/\epsilon)}{\epsilon^2}.$$

Entonces, si W es una matriz con entradas independientes siguiendo una distribución normal $N(0, 1/n)$, con probabilidad al menos $1 - \delta$ tenemos que

$$\sup_{x\in S} \left| \frac{\|Wx\|}{\|x\|} - 1 \right| < \epsilon$$

Demostración. Es suficiente probar el lema para todo $x \in S$ con $\|x\| = 1$. Podemos escribir $x = U_I\alpha$ donde $\alpha \in \mathbb{R}^s$, $\|\alpha\|_2 = 1$, con U_I la matriz cuyas

columnas son $\{U_i : i \in I\}$. Tomando $\epsilon/4 > 0$ y utilizando el lema (5.2.7) sabemos que existe un conjunto Q de cardinal $|Q| \leq (12/\epsilon)^s$ tal que

$$\sup_{\alpha:\|\alpha\|=1} \min_{v\in Q} \|\alpha - v\| \leq (\epsilon/4).$$

Como U es ortogonal preserva la norma y $\|\alpha - v\| = \|U_I(\alpha - v)\| = \|U_I\alpha - U_I v\|$, por lo que también tenemos que

$$\sup_{\alpha:\|\alpha\|=1} \min_{v\in Q} \|U_I\alpha - U_I v\| \leq (\epsilon/4).$$

Aplicando el lema de Johnson-Lindestrauss (lema 4.1.1) en el conjunto $\tilde{Q} = \{U_I v : v \in Q\}$ obtenemos que se cumple con probabilidad al menos $1-\delta$ que

$$\sup_{v\in Q} \left| \frac{\|WU_I v\|^2}{\|U_I v\|^2} - 1 \right| \leq \epsilon/2, \tag{5.9}$$

lo que implica que

$$\sup_{v\in Q} \left| \frac{\|WU_I v\|}{\|U_I v\|} - 1 \right| \leq \epsilon/2. \tag{5.10}$$

Para que se pueda aplicar el lema de Johnsons-Lindestrauss y se verifique la desigualdad (5.9) necesitamos que $\epsilon/2$ satisfaga la condición pedida en dicho lema, es decir,

$$\epsilon/2 = \sqrt{\frac{6\log(2|\tilde{Q}|/\delta)}{n}} \leq 3.$$

Operando

$$n = \frac{6\log(2|\tilde{Q}|/\delta)}{\epsilon^2/4} = 24\ \frac{\log(2/\delta) + \log(|\tilde{Q}|)}{\epsilon^2}$$

Como $|\tilde{Q}| \geq sup(|Q|) = (12/\epsilon)^s$ tenemos que n tiene que cumplir

$$n \geq 24\ \frac{\log(2/\delta) + s\log(12/\epsilon)}{\epsilon^2}.$$

Ahora, teniendo (5.10) queremos probar que $(1-\epsilon) \leq \|Wx\| \leq (1+\epsilon)$. Sea a el número más pequeño que cumple que

$$\forall x \in S,\ \frac{\|Wx\|}{\|x\|} \leq 1 + a.$$

Nuestro objetivo es probar que $a \leq \epsilon$. Esto viene del hecho de que para cualquier $x \in S$ de norma 1 existe $v \in Q$ tal que $\|x - U_I v\| \leq \epsilon/4$ y por lo tanto

$$\|Wx\| \leq \|WU_I v\| + \|W(x - U_I v)\| \leq 1 + \epsilon/2 + (1+a)\epsilon/4.$$

Por lo tanto

$$\forall x \in S,\ \frac{\|Wx\|}{\|x\|} \leq 1 + (\epsilon/2 + (1+a)\epsilon/4).$$

Pero la definición de a implica que

$$a \leq \epsilon/2 + (1+a)\epsilon/4 \Rightarrow a \leq \frac{\epsilon/2 + \epsilon/4}{1 - \epsilon/4} \leq \epsilon.$$

Esto prueba que para todo $x \in S$ tenemos $\dfrac{\|Wx\|}{\|x\|} - 1 \leq \epsilon$ y que

$$\|Wx\| \geq \|WU_I v\| - \|W(x - U_I v)\| \geq 1 - \epsilon/2 - (1+\epsilon)\epsilon/4 \geq 1 - \epsilon.$$

□

El lema anterior nos dice que para $x \in S$ de norma 1 tenemos

$$(1 - \epsilon) \leq \|Wx\| \leq (1 + \epsilon).$$

La demostración del teorema viene de la unión de todas las posibles elecciones de I. Sean Q_I todos los posibles conjuntos que se pueden formar a partir de los conjuntos de índices I de cardinal s. Como

$$Q = \cup_{I:|I|=s} Q_I.$$

Hay $\binom{d}{s}$ posibles subconjuntos de cardinal s, y podemos acotar esta cifra por

$$\binom{d}{s} = \frac{d(d-1)(d-2)...(d-s+1)}{s!} \leq \frac{d^s}{s!} \leq \left(\frac{ed}{s}\right)^s,$$

donde la última desigualdad procede de la desigualdad de Stirling $s! \geq (s/e)^s$. Así, tenemos $|Q| \leq \left(\dfrac{3}{\epsilon}\right)^s \left(\dfrac{ed}{s}\right)^s = \left(\dfrac{3ed}{\epsilon s}\right)^s$ y razonando como en la demostración del lema intermedio, de la unión de todos los posibles I se obtiene que $n \geq \Omega(s \log(d/s)) \geq \Omega(s \log(d))$, como queríamos.

□

En conclusión, hemos probado que las matrices aleatorias cuya dimensión es de orden superior a $s\log(d)$ son RIP con probabilidad alta. Sin embargo, cuando no se satisface esta condición sobre la dimensión de la matriz aleatoria, la propiedad RIP es difícil de verificar. Por ello se introduce también la propiedad de núcleo restringido.

5.2.2. RN, propiedad de núcleo restringido.

La propiedad de núcleo restringido es una condición menos fuerte que la de isometría restringida y que se satisface si y sólo si los problemas en l^0 (5.1) y l^1 (5.2) son equivalentes.
Veremos en esta sección que una condición suficiente para que se cumpla la propiedad de núcleo restringido es la de incoherencia mutua, que tiene la ventaja de ser fácilmente computable.

Supongamos que el problema en l^0 tiene una única solución $\tilde{\alpha}$ tal que $y = Z\tilde{\alpha}$. Además, suponemos que $\tilde{\alpha}$ tiene soporte en un subconjunto $S \subseteq \{1, 2, ..., d\}$, es decir $\tilde{\alpha}_j = 0$ para $j \in S^C$.

Definimos el cono convexo

$$\mathbb{C}(S) := \{\beta \in \mathbb{R}^d / \|\beta_{S^C}\|_1 \leq \|\beta_S\|_1\}.$$

que contiene todos los vectores cuya norma l_1 en el soporte es mayor que la norma fuera de él.

Se define el núcleo de Z como $Ker(Z) = \{\beta \in \mathbb{R}^d / Z\beta = \mathbf{0}\}$. La equivalencia entre el problema en l_1 y l_0 dependerá de la siguiente propiedad:

Definición 5.2.9. *La matriz Z satisface la propiedad de núcleo restringido (RN(S) del iglés restricted nullspace property) si*

$$\mathbb{C}(S) \cap Ker(Z) = \{0\}. \tag{5.11}$$

Teorema 5.2.10. *Supongamos que $\tilde{\alpha}$ es la única solución del problema en l_0 (5.1) y tiene soporte en S. Entonces si la matriz Z satisface la propiedad RN respecto de S, el problema relajado en l_1 tiene una única solución igual a $\tilde{\alpha}$.*

Demostración. En primer lugar, supongamos que Z satisface la propiedad de núcleo restringido. Sea $\alpha^* \in \mathbb{R}^d$ una solución óptima de y definimos el error

Δ como el vector $\Delta := \alpha^* - \tilde{\alpha}$. El objetivo es probar que $\Delta = 0$ y para ello basta mostrar que $\Delta \in Ker(Z) \cap \mathbb{C}(S)$. Tenemos que α^* y $\tilde{\alpha}$ son dos soluciones óptimas de los problemas en l^1 y l^0 respectivamente, por lo que se verifica que $Z\tilde{\alpha} = y = Z\alpha^*$ y entonces $Z\Delta = 0$. Por otra parte, como $\tilde{\alpha}$ es también una solución posible del problema l^1 y α^* es la solución óptima tenemos que $\|\alpha^*\|_1 \leq \|\tilde{\alpha}\|_1 = \|\tilde{\alpha}_S\|_1$. Escribiendo $\alpha^* = \tilde{\alpha} + \Delta$ tenemos

$$\begin{aligned}\|\tilde{\alpha}_S\|_1 \geq \|\alpha^*\|_1 &= \|\tilde{\alpha}_S + \Delta_S\|_1 + \|\Delta_{S^C}\|_1 \\ &\geq \|\tilde{\alpha}_S\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^C}\|_1,\end{aligned}$$

donde hemos utilizado la desigualdad triangular. Reorganizando los términos de la desigualdad anterior llegamos a que

$$\|\Delta_{S^C}\|_1 \leq \|\Delta_S\|_1$$

con lo que $\Delta \in \mathbb{C}(S)$ y puesto que por hipótesis Z satisface la propiedad de núcleo restringido, concluimos que $\Delta = 0$ como queríamos.

□

La condición más simple que garantiza la propiedad de núcleo restringido es la de incoherencia. Veremos en la proposición (5.2.12) que una coherencia baja es suficiente para garantizar la equivalencia entre los problemas en l_0 y l_1.

Definición 5.2.11. *En álgebra lineal , la coherencia mutua de una matriz Z se define como*

$$\nu(Z) = \max_{j,j'=1,2,\ldots,d} \frac{|\langle z_j, z_{j'}\rangle|}{\|z_j\|_2 \|z_{j'}\|_2}.$$

Si las columnas z_j están centradas, se trata del valor absoluto máximo de los correlaciones cruzadas entre las columnas de Z. Cuando las columnas tienen norma 1 tenemos que $\nu = max_{j\neq j'}|\langle z_j, z_{j'}\rangle|$, lo cual nos permite definir la incoherencia mutua también como la distancia de la matriz de Gram $Z^T Z$ a la matriz identidad elemento a elemento. Una definición alternativa para la incoherencia mutua es el menor δ tal que

$$\|\frac{Z^T Z}{n} - I\|_\infty \leq \delta.$$

Proposición 5.2.12 (Incoherencia mutua implica RN)**.** .

Supongamos que para algún entero $k \in \{1, 2, ..., d\}$ *tenemos* $\nu(Z) < \frac{1}{3k}$.

Entonces Z satisface la propiedad RN de orden k, y por tanto los problemas l_0 y l_1 son equivalentes para todos los vectores con soporte a lo sumo k.

Demostración. Podemos asumir sin pérdida de generalidad que $\|z_j\|_2 = 1$ para todo $j = 1, 2, ..., p$. Para simplificar la notación, supongamos que $\nu(Z) < \frac{\delta}{k}$ para algún $\delta > 0$, y probemos que es suficiente con tomar $\delta = 1/3$. Para un subconjunto arbitrario S de cardinal k, supongamos que $\alpha \in \mathbb{C}(S) \setminus \{0\}$. Basta con probar que $\|Z\beta\|_2^2 > 0$:

$$\|Z\beta\|_2^2 > \|Z_S\beta_S\|_2^2 + 2\beta_S^T Z_S^T Z_{S^C}\beta_{S^C}.$$

Por un lado tenemos

$$2|\beta_S^T Z_S^T Z_{S^C}\beta_{S^C}| \leq |\sum_{i\in S}\sum_{j\in S^C} ||\beta_i||\beta_j||\langle z_i, z_j\rangle|$$

$$(i) \leq 2\|\beta_S\|_1\|\beta_{S^C}\|_1\nu(Z)$$

$$(ii) \leq \frac{2\delta\|\beta_S\|_1^2}{k}$$

$$(iii) \leq 2\delta\|\beta_S\|_2^2$$

donde en (i) hemos utilizado la definición de incoherencia mutua, en (ii) utilizamos que $\beta \in \mathbb{C}(S)$ y $\nu(Z) < \frac{\delta}{k}$ para algún $\delta > 0$ y en (iii) utilizamos que $\|\beta_S\|_1^2 \leq \|\beta_S\|_2^2$ por Cauchy-Swartz, puesto que el cardinal de S es a lo sumo k.

Por lo tanto, hemos establecido que

$$\|Z\beta\|_2^2 \geq \|Z_S\beta_S\|_2^2 - 2\delta\|\beta_S\|_2^2.$$

Sólo queda acotar inferiormente $\|Z_S\beta_S\|_2^2$. Denotamos por $\|.\|_{op}$ al máximo de los valores singulares de una matriz, y tenemos

$$\|Z_S^T Z_S - I_k\|_{op} \leq \max_{i\in S}\sum_{j\in S\setminus\{i\}} \|\langle z_i, z_j\rangle\| \leq k\frac{\delta}{k} = \delta.$$

En consecuencia $\|Z_S\beta_S\|_2^2 \geq (1-\delta)\|\beta_S\|_2^2$ y podemos concluir que $\|Z\beta\|_2^2 > (1-3\delta)\|\beta_S\|_2^2$, por lo que es suficiente con tomar $\delta = 1/3$ como queríamos demostrar.

□

Capítulo 6

Recuperación de la señal.

Hagamos una breve recapitulación de lo que hemos visto hasta ahora. En primer lugar partimos de una matriz de proyecciones aleatorias $W \in \mathcal{M}^{n,d}(\mathbb{R})$ con $n < d$, a partir de la cual se ha captado el vector $y \in \mathbb{R}^n$. Este vector y es la compresión de la señal original $x \in \mathbb{R}^d$, de forma que tenemos que

$$y = Wx.$$

Como vimos en el capítulo 5, para poder resolver el sistema anterior y recuperar la señal x necesitamos una forma dispersa de la misma. Para ello se supone que la señal x admite una expresión dispersa en una base ortogonal formada por las columnas de una matriz $U \in \mathcal{M}^{d,d}(\mathbb{R})$ conocida. Hemos visto en el capítulo 3 varias técnicas para obtener U, con lo que obtenemos en forma matricial

$$y = WU\alpha = Z\alpha,$$

donde α es un vector disperso. Para recuperar α debemos buscar el vector más disperso posible que cumpla el sistema, con lo que el problema de minimización a resolver es

$$min\|\alpha\|_0 \text{ tal que } y = Z\alpha.$$

Se trata de un problema no convexo que no se puede resolver mediante los métodos numéricos clásicos. Sin embargo, hemos probado en el capítulo 5 que si la matriz Z satisface la propiedad de isometría restringida o la de núcleo restringido, podemos resolver de forma equivalente el problema con norma 1

$$min\|\alpha\|_1 \text{ tal que } y = Z\alpha,$$

que sí es convexo.

Por lo tanto, para completar el proceso de compressed sensing solamente falta resolver el problema relajado a norma 1. Para ello se distingue entre señales con y sin presencia de ruido. Un problema sin presencia de ruido se puede escribir como un problema lineal y por tanto resolverse con algoritmos como el símplex o los métodos de punto interior. Para las señales con presencia de ruido, que son las más comunes en la práctica, se utilizará el método del descenso por coordenadas que detallaremos en esta sección.

Cabe mencionar que únicamente se realiza una descripción del procedimiento. Para más detalle se puede consultar la sección 5.5 de [3]. Además, el algoritmo del descenso por coordenadas converge al óptimo y es competitivo frente a otras implementaciones del método del descenso del gradiente como por ejemplo el descenso con aceleración de Nesterov [3].

6.1. Señales sin presencia de ruido.

En 2001 Chen, Donoho y Saunders introdujeron el nombre de *Basis pursuit* [16] para denotar un problema de la forma

$$min\|\alpha\|_1 \text{ tal que } y = Z\alpha.$$

El problema de *Basis pursuit* se puede expresar bajo la forma de un problema lineal. Veámoslo:

Definición 6.1.1. *Un problema lineal en la llamada forma estándar es un problema de optimización con restricciones definido en términos de una variable $x \in \mathbb{R}^m$ por*

$$min\ c^T x \quad sujeto\ a \quad Ax = b,\ x \geq 0,$$

donde $c^T x$ es la función objetivo.

Nuestro problema puede ser escrito como un problema lineal estándar. Escribimos el vector α como resta de dos vectores $\alpha = u - v$, ambos con entradas mayores o iguales a 0. Así, el problema de minimización será $min\ (c^T u + c^T v)$, donde $c^T = [1, 1, ..., 1]$. El mínimo se dará cuando u contenga las entradas positivas de α con el resto de entradas nulas, y v las entradas negativas con el resto de entradas nulas. Entonces tendremos el problema

$$min\ (c^T u + c^T v) \text{ tal que } Z(u - v) = y$$

o expresado de otra forma

$$min\ (c^T u + c^T v) \quad \text{tal que} \quad [Z\ -Z] \begin{bmatrix} u \\ v \end{bmatrix} = y$$

donde $u, v \geq 0$. Por lo tanto podemos encontrar la solución resolviendo un problema lineal equivalente (la equivalencia entre los problemas de minimización de la norma l^1 y los problemas lineales es sabida desde los años 50 [19]).

Algoritmos de resolución: Para resolverlo en [20] se proponen dos de los algoritmos más utilizados en programación lineal. El método símplex propuesto por Dantzig [18] es el método clásico, y procede examinando vértices adyacentes del poliedro de soluciones. Además, en los últimos años ha habido grandes avances en programación lineal debido a la utilización de los llamados métodos de punto interior propuestos p or Karmarkar, N. (1984) [21], que basan su estrategia en la búsqueda del óptimo a través de caminos que recorren la zona interior de la región factible.

6.2. Señales con presencia de ruido.

Como en la práctica la recuperación no es perfecta debido a la presencia de ruido en las señales, Chen, Donoho y Saunders [**?**] introdujeron el problema de *Basis pursuit denoising*(BPDN), una adaptación del BP que toma una forma similar al LASSO [17].

Asumimos que aparece un factor de error ρ en la igualdad

$$y = Z\alpha + \mu\rho,$$

donde $\mu > 0$ es una constante y ρ un vector que contiene ruido siguiendo una distribución normal estándar.
En nuestro problema y es un vector conocido y α un vector disperso desconocido, y lo que queremos conseguir es que el error entre el vector y y su reconstrucción sea lo más pequeño posible. En lugar de minimizar $\|\alpha\|_1$ minimizamos el error $\|y - Z\alpha\|_2$ imponiendo como condición que $\|\alpha\|_1$ sea menor que un $\epsilon > 0$ fijado, $\|\alpha\|_1 < \epsilon$, o lo que es lo mismo

$$min \|y - Z\alpha\|_2^2 \ \text{ tal que } \ \|\alpha\|_1 < \epsilon.$$

El problema anterior es equivalente al *basis pursuit denoising*, nombre que recibe el problema

$$\min_{\alpha} \frac{1}{2}\|y - Z\alpha\|_2^2 + \lambda\|\alpha\|_1.$$

El parámetro de regularización λ ajusta el grado de dispersión, un parámetro λ grande proporciona una solución con pocas componentes distintas de cero, mientras que si $\lambda = 0$ obtenemos solución no dispersa.
Se trata de un problema de optimización con funciones convexas, con lo que es posible encontrar la solución. Sin embargo, como la norma l^1 no es diferenciable en ningún punto donde alguna coordenada del vector sea nula, tenemos que estudiar la forma de resolver problemas convexos con funciones no diferenciables. Utilizaremos el método del descenso por coordenadas.

6.2.1. Descenso por coordenadas.

El método del descenso coordinado (desarrollado en [3] consiste en un algoritmo de optimización para problemas convexos no diferenciables como el que queremos resolver en este caso. En cada iteración, el algoritmo realiza una búsqueda unidimensional a lo largo de una dirección de coordenadas para encontrar el mínimo local de la función.

Supongamos en primer lugar que tenemos solamente una variable predictora, es decir, que $\alpha \in \mathbb{R}$ con $d = 1$ y la matriz $Z \in \mathcal{M}^{n,d}(\mathbb{R})$ es un vector $\tilde{z}$. En lugar de utilizar $\tilde{z}$ utilizaremos su vector traspuesto que denotaremos por z. Queremos resolver

$$\min_{\alpha} \|y - \alpha z\|^2 + \lambda|\alpha|.$$

El problema es sobre una variable, con lo que si denotamos $f(\alpha) = \|y - \alpha z\|^2 + \lambda|\alpha|$ podemos obtener una solución igualando la derivada de f a 0, teniendo en cuenta que en $\alpha = 0$ la función valor absoluto no es diferenciable. Así, tenemos

$$f'(\alpha) = \begin{cases} -2(y - \alpha z)^T z - \lambda = -2\langle y, z\rangle + 2\alpha - \lambda & \text{si } \alpha < 0 \\ -2(y - \alpha z)^T z + \lambda = -2\langle y, z\rangle + 2\alpha + \lambda & \text{si } \alpha > 0 \end{cases}$$

Igualando $f'(\alpha)$ a 0 obtenemos

$$\hat{\alpha} = \begin{cases} \langle y,z\rangle + \dfrac{\lambda}{2} & \text{si } \langle y,z\rangle < -\dfrac{\lambda}{2} \\ 0 & \text{si } -\dfrac{\lambda}{2} \leq \langle y,z\rangle \leq \dfrac{\lambda}{2} \\ \langle y,z\rangle - \dfrac{\lambda}{2} & \text{si } \langle y,z\rangle > \dfrac{\lambda}{2} \end{cases}$$

Esto es debido a que si $\langle y,z\rangle < \dfrac{\lambda}{2}$ entonces $f(\langle y,z\rangle + \dfrac{\lambda}{2}) > f(0)$ y que si $\langle y,z\rangle > \dfrac{\lambda}{2}$ entonces $f(\langle y,z\rangle - \dfrac{\lambda}{2}) > f(0)$. Probemos la primera afirmación, la segunda se prueba de forma análoga.
Suponemos que el problema está estandarizado, con lo que $\|z\|^2 = 1$

$$\begin{aligned} f(\langle y,z\rangle + \frac{\lambda}{2}) &= \|y - (\langle y,z\rangle + \frac{\lambda}{2})z\|^2 + \lambda(\langle y,z\rangle + \frac{\lambda}{2}) \\ &= \|y\|^2 + \|\langle y,z\rangle + \frac{\lambda}{2}\|^2 - 2(\langle y,z\rangle + \frac{\lambda}{2})\langle y,z\rangle + \lambda(\langle y,z\rangle + \frac{\lambda}{2}) \\ &= f(0) + (\langle y,z\rangle + \frac{\lambda}{2})((\langle y,z\rangle + \frac{\lambda}{2}) - 2(\langle y,z\rangle - \frac{\lambda}{2})) \\ &= f(0) + (\langle y,z\rangle + \frac{\lambda}{2})^2 > f(0). \end{aligned}$$

Se define el operador

$$S_\lambda(x) = sign(x)(|x| - \lambda)_+$$

S_λ, denominado operador *Soft thresholding*, que actúa reduciendo o aumentando el valor de x en λ en valor absoluto, dependiendo del signo de x.

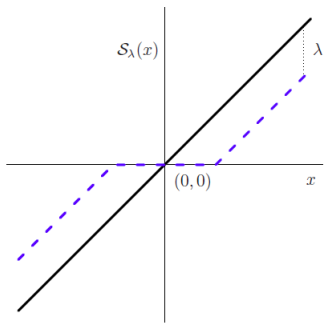

Figura 6.1: Representación del operador Soft thresholding.

Utilizando este operador podemos expresar $\hat{\alpha}$ de la forma $\hat{\alpha} = S_{\lambda/2}(\langle y, z\rangle)$.

Si trasladamos el procedimiento anterior al caso en el que tenemos d variables predictoras, se tratará de minimizar la función objetivo en α_k para cada $k = 1, ..., d$ manteniendo fijas el resto de variables. Así, consideramos la siguiente función objetivo dependiente sólo de α_k.

$$f(\alpha) = f(\alpha_k) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1, j\neq k}^{d} z_{i,j}\alpha_j - z_{i,k}\alpha_k \right)^2 + \lambda \sum_{j=1, j\neq k}^{d} |\alpha_j| + \lambda|\alpha_k|$$

$$= \sum_{i=1}^{n} (r_i^k - z_{i,k}\alpha_k) + \lambda \sum_{j=1, j\neq k}^{d} |\alpha_j| + \lambda|\alpha_k|$$

donde se denomina residuo a la cantidad $r_i^k = y_i - \sum_{j=1, j\neq k}^{d} Z_{i,j}\alpha_j$.

Minimizar el problema anterior equivale a minimizar la misma función que en el caso en el que tan solo existía una variable predictora con $y_i = r_i^k$ y $z_i = z_{i,k}$. Con esto obtenemos que el mínimo se alcanza en

$$\alpha_k = S_{\lambda/2}(\langle r^k, z_{:,k}\rangle). \tag{6.1}$$

El algoritmo se itera un número determinado de veces, y en cada una de ellas k recorre cada una de las variables predictoras para actualizar la componente α_k, obteniendo finalmente un vector próximo al mínimo $\hat{\alpha}$.

En el caso de una matriz Z ortogonal, el operador soft-thresholding juega un papel central en la resolución del problema. Para comprobarlo, vemos que el esquema de minimización toma una forma más simple si los predictores z_j son ortogonales, es decir, si $\langle z_{:,j}, z_{:,k}\rangle = 0$ para cada $j \neq k$. En este caso, la expresión (6.1) se simplifica, puesto que $\langle r^k, z_{:,k}\rangle = \langle y, z_{:,j}\rangle$ y entonces $\tilde{\alpha}$ es simplemente el operador soft-thresholding aplicado a $\langle y, z_{:,j}\rangle$. En consecuencia, en el caso de una matriz Z ortogonal la solución buscada tiene una forma explícita y no es necesario realizar más iteraciones.

En la práctica, normalmente estamos interesados en encontrar la solución no solamente para un λ, sino que buscamos la solución para distintos valores de λ hasta encontrar el óptimo. Además, si elegimos $\lambda = 0$ en (6.2.1) el descenso por coordenadas nos lleva a la solución de un problema de mínimos cuadrados. Cuando la matriz Z tiene rango máximo, el punto de convergencia es la solución del problema de mínimos cuadrados, es decir $\lambda = 0$ (ver pág 16 de [3]).

Capítulo 7

Compressed sensing en imágenes con R.

En esta sección se expondrán los resultados obtenidos tras la programación del proceso de compressed sensing en R. Nos centraremos en el caso de imágenes en escala de grises. Para la implementación del compressed sensing en color, únicamente se requiere realizar el proceso de forma separada para cada uno de los parámetros RGB.

7.1. Código.

Para la representación de la señal en forma dispersa, se utilizará la forma matricial de la DCT, descrita en el capítulo 3. En primer lugar, debemos cargar la imagen, que se supondrá cuadrada por simplicidad. El cambio a imágenes rectangulares requiere una modificación mínima del código.

```
im <- load.image("ubicacion_de_la_imagen.jpg")
im<-grayscale(im)   #Convierte a escala de grises
M <- as.matrix(im)   #Convierte la imagen a una matriz
d<-dim(M)[1]
```

El siguiente código muestra la implementación de la DCT en forma matricial.

```
C<-matrix(nrow=d,ncol=d)
for (k in 0:d-1){
```

```
  for (i in 0:d-1){
    if( k == 0){
      C[k+1,i+1]<-sqrt(1/d)
    }
    else
      C[k+1,i+1]<- sqrt(2/d)*cos((pi*k*(1/2+i))/d)
  }
}
```

El siguiente paso es la creación de una matriz $W \in \mathcal{M}^{n,d}(\mathbb{R})$ con entradas independientes siguiendo una distribución normal de media 0 y varianza $1/d$. Además, computamos la matriz de compressed sensing $Z = WC$, donde C es la matriz de la transformada.

La dimensión n refleja el número de mediciones de la imagen original que queremos tomar. Se construye entonces la matriz $Y \in \mathcal{M}^{n,d}(\mathbb{R})$, $Y = WM$ de mediciones. Intentaremos recuperar la imagen de matriz M a partir de Y.

```
n<-100
W<-matrix(rnorm(n*d,mean=0,sd=1/d), n, d)
Z<-W%*%C
Y<-W%*%M
```

Para la recuperación de la señal implementamos una función que efectúa el descenso por coordenadas, que trabaja sobre un vector y previamente definido (los vectores y son las columnas de la matriz Y). Se repite el algoritmo un número fijo de iteraciones *maxit*.

```
coord.descent <- function(Z, maxit, lambda){
  p <- dim(Z)[2]
  theta <- matrix(0, ncol=p, nrow=maxit)
  for(k in 1:maxit){
    for(i in 1:p) {
      iter = ifelse(k>1, yes=k-1, 1)
      theta[k,i] <- t(Z[,i])%*% (y - Z[,-i]%*%theta[iter,-i])/
       (t(Z[,i])%*%Z[,i])
      theta[k,i] <- softthresh(theta[k,i], lambda)
    }
  }
```

```
  return(theta)
}
```

La salida *theta* de esta función es una matriz que contiene los valores para cada una de las coordenadas obtenidos en cada iteración.

Puesto que las columnas de Z son ortogonales, se realiza solamente una iteración del algoritmo, y además vimos que el óptimo se alcanza para $\lambda = 0$. Reconstruimos la señal operando por columnas en una matriz que llamamos *IMAGEN*.

```
IMAGEN<-matrix(0,d,d)
maxiter <- 1
for(j in 1:d){
  y<-Y[,j]
  out_cd = coord.descent(Z, maxiter, lambda=0)
  IMAGEN[,j]<-C%*%out_cd[1,]    #Deshacemos la transformada
}
```

Ya se ha terminado el proceso de recuperación de la imagen original. El proceso se ha realizado por columnas, aunque también se podría trabajar por filas. Veremos los resultados obtenidos en la siguiente sección.

7.2. Resultados.

7.2.1. DCT.

En primer lugar veamos la eficacia de la transformada mediante la DCT. Para ello, se calcula la transformada de la imagen original y se guardan un porcentaje pequeño de las entradas, tomando aquellas con mayor valor absoluto. Después, se realiza la transformada inversa y se observa el resultado en la figura 7.1. Para cada caso, se ha calculado el RMSE entre la imagen original y la reconstrucción.

(a) 2 %

(b) 5 %

(c) 10 %

(d) Original

Figura 7.1: Reconstrucción de la imagen original tomando distintos porcentajes de coeficientes de la DCT.

7.2.2. Compressed sensing.

Se han realizado varias pruebas tomando en cada vez proyecciones aleatorias de distintas dimensiones sobre la señal original. En el primer ejemplo se toma una proyección aleatoria que convierte el conjunto de datos en otro cuyo tamaño es un 35 % del original. Se computa el proceso primero por columnas y luego por filas, con lo que se obtienen las dos imágenes representadas en la

figura (7.2).

Figura 7.2: Compressed sensing por columnas y por filas tomando la dimensión de las proyecciones un 35 % de la original.

Figura 7.3: Media de las imágenes de la figura (7.2) comparada con la imagen original.

Como se puede observar, aparecen unas marcas verticales y horizontales.

Para solventar este problema se obtendrá una reconstrucción final a partir de la media de las dos salidas. Al realizar la media píxel a píxel se consigue reducir el error cometido por cada imagen. En la figura (7.3) se muestra el resultado obtenido.

La figura (7.4) muestra el RMSE cometido entre la imagen original y la reconstruida frente a la dimensión de las proyecciones en tanto por ciento sobre dimensión inicial. Se observa que a partir del 40 % un aumento de los datos tomados no produce mejora significativa de la recuperación.

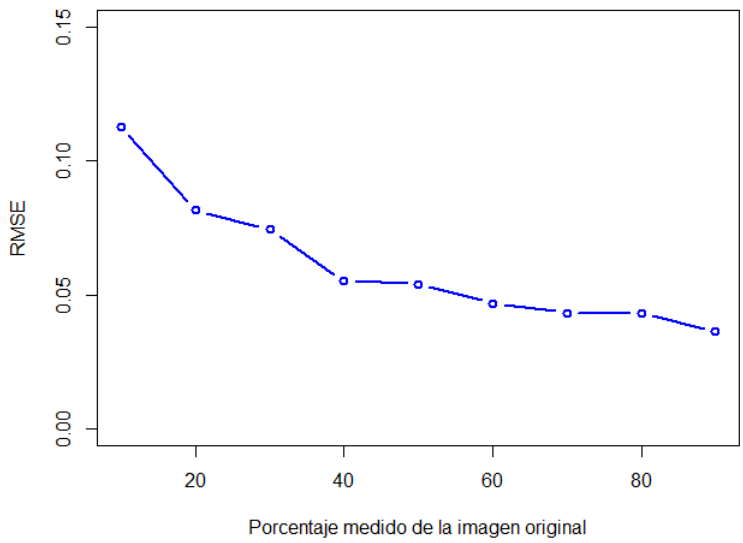

Figura 7.4: RMSE cometido en la reconstrucción.

En la figura (7.5) se presentan las reconstrucciones para proyecciones de tamaño un 10 %, 20 % 40 %, 50 % y 60 % de la dimensión original. Se comparan con la imagen inicial. Se puede apreciar que con un 20 % ya se puede distinguir la imagen original. Además, a partir de un 40 % el resultado obtenido no mejora significativamente al aumentar la cantidad de mediciones, como se refleja en (7.4).

(a) 10 %

(b) 20 %

(c) 40 %

(d) 50 %

(e) 60 %

(f) Original

Figura 7.5: Reconstrucción de la imagen original con proyecciones de distinto tamaño.

Capítulo 8

Conclusiones.

La técnica del compressed sensing ha supuesto uno de los mayores avances de las últimas décadas en todos los aspectos relacionados con el ámbito del procesamiento de señales.

Desde el punto de vista matemático, lo que se ha logrado en este trabajo es una reconstrucción fiel de los datos iniciales tomando una cantidad pequeña de mediciones. Además, utilizando cotas probabilísticas se garantiza la reconstrucción a partir de proyecciones aleatorias. El compressed sensing está relacionado con diversos ámbitos matemáticos, desde las transformadas discretas, la búsqueda de solución de un problema no convexo, las proyecciones aleatorias o los métodos de descenso del gradiente, entre otros. Su amplio alcance hace que continuamente aparezcan nuevos avances que hacen el método más eficiente o mejoran la reconstrucción de la señal.

La principal dificutad del compressed sensing es la variedad de métodos disponibles en cada etapa de su desarrollo. Además, la elección de un método u otro dependerá de la naturaleza de la señal a tratar. Es por esto que uno de los retos de mejora del compressed sensing es plantear un esquema general que se pueda aplicar a señales de distintos tipos. En sí, determinar una buena base que facilite la expresión en forma dispersa es la tarea más compleja. Esta tarea requiere un conocimiento profundo de la naturaleza de los datos y por tanto no se puede generalizar a todos los tipos de señales. El compressed sensing es especialmente eficiente cuando las señales son expresables en forma dispersa en bases de wavelets, sinusoidales o una combinación de las mismas, aunque como hemos comprobado también permite obtener una reconstrucción buena en señales (en concreto en imágenes) genéricas.

En conclusión, el compressed sensing se muestra como una herramienta

prometedora dentro del ámbito del tratamiento de señales, no sólo por su capacidad de compresión sino por su capacidad de extraer información relevante del conjunto de datos inicial. Por lo tanto, esta técnica ofrece grandes retos tanto a la comunidad científica como a la industria y un amplio margen de mejora, con que lo sin duda continuará evolucionando en el transcurso de los próximos años.

Anexo.

Cotas de Chernoff.

En esta sección vamos a desarrollar las cotas de Chernoff, que son de utilidad para la demostración del lema (8.0.4). Se trata de cotas exponencialmente decrecientes para las distribuciones de sumas de variables aleatorias independientes. Son una cota más fina que las conocidas cotas basadas en el primer y segundo momento tales como la inecuación de Markov o la inecuación de Chebyshev, las cuales solo obtienen cotas de nivel exponencial cuando la distribución decrece. Las cotas de Chernoff requieren que las variables sean independientes, una condición que ni las inecuaciones de Markov ni de Chebyshev requieren y que en nuestro caso se verifica.

Sean Z_1,...,Z_m variables aleatorias independientes siguiendo una distribución de Bernouilli, donde para cada i, $\mathbb{P}[Z_i = 1] = p_i$ y $\mathbb{P}[Z_i = 0] = 1 - p_i$. Sean $p = \sum_{i=1}^{m} p_i$ y $Z = \sum_{i=1}^{m} Z_i$. Utilizando la monotonía de la exponencial y la desigualdad de Markov, tenemos que para todo $t > 0$

$$\mathbb{P}[Z > (1+\delta)p] = \mathbb{P}[e^{tZ} > e^{t(1+\delta)p}] \leq \frac{\mathbb{E}[e^{tZ}]}{e^{(1+\delta)tp}}. \tag{8.1}$$

Además, utilizando la independencia de las variables aleatorias en (1) y la

desigualdad $1 + x \leq e^x$ en (2) se obtiene

$$\begin{aligned}
\mathbb{E}[e^{tZ}] = \mathbb{E}[e^{t\sum_i Z_i}] &= \mathbb{E}[\prod_i e^{tZ_i}] \\
(1) \quad &= \prod_i \mathbb{E}[e^{tZ_i}] \\
&= \prod_i (p_i e^t + (1-p_i)e^0) \\
&= \prod_i (1 + p_i(e^t - 1)) \\
(2) \quad &\leq \prod_i e^{p_i(e^t-1)} \\
&= e^{\sum_i p_i(e^t-1)} \\
&= e^{(e^t-1)p}.
\end{aligned}$$

Combinando lo anterior con la ecuación (8.1) y tomando $t = log(1+\delta)$ se obtiene el siguiente lema.

Lema 8.0.1. *Sean Z_1,...,Z_m variables aleatorias independientes siguiendo una distribución de Bernouilli, donde para cada i, $\mathbb{P}[Z_i = 1] = p_i$ y $\mathbb{P}[Z_i = 0] = 1 - p_i$. Sean $p = \sum_{i=1}^m p_i$ y $Z = \sum_{i=1}^m Z_i$. Entonces, para cualquier $\delta > 0$,*

$$\mathbb{P}[Z > (1+\delta)p] \leq e^{-h(\delta)p},$$

donde

$$h(\delta) = (1+\delta)log(1+\delta) - \delta.$$

Utilizando la desigualdad $h(a) \geq a^2/(2 + 2a/3)$ obtenemos el siguiente resultado.

Lema 8.0.2. *Utilizando la notación del lema anterior, tenemos también que*

$$\mathbb{P}[Z > (1+\delta)p] \leq e^{-p\frac{\delta^2}{2+2\delta/3}}.$$

Hemos conseguido una cota para el caso en el que $Z > (1+\delta)p$. En la otra dirección se aplican cálculos similares detallados a continuación:

$$\mathbb{P}[Z < (1-\delta)p] = \mathbb{P}[-Z > -(1-\delta)p] = \mathbb{P}[e^{-tZ} > e^{-t(1-\delta)p}] \leq \frac{\mathbb{E}[e^{-tZ}]}{e^{-(1-\delta)tp}},$$

y además utilizando la independencia de las variables aleatorias en (1) y la desigualdad $1 + x \leq e^x$ en (2)

$$\begin{aligned}
\mathbb{E}[e^{-tZ}] = \mathbb{E}[e^{-t\sum_i Z_i}] &= \mathbb{E}[\prod_i e^{-tZ_i}] \\
(1) \quad &= \prod_i \mathbb{E}[e^{-tZ_i}] \\
&= \prod_i (p_i e^{-t} + (1-p_i)e^0) \\
&= \prod_i (1 + p_i(e^{-t} - 1)) \\
(2) \quad &\leq \prod_i e^{p_i(e^{-t}-1)} \\
&= e^{(e^{-t}-1)p}.
\end{aligned}$$

Puesto que $t = -log(1-\delta)$ podemos simplificar la expresión de la cota anterior

$$\mathbb{P}[Z < (1-\delta)p] \leq \frac{e^{-\delta p}}{e^{(1-\delta)log(1-\delta)p}} = e^{-ph(-\delta)}.$$

Se verifica que $h(-\delta) \geq h(\delta)$ y por lo tanto

Lema 8.0.3. *Utilizando la notación de los lemas anteriores,*

$$\mathbb{P}[Z < (1-\delta)p] \leq e^{-ph(-\delta)} \leq e^{-ph(\delta)} \leq e^{-p\frac{\delta^2}{2+2\delta/3}}.$$

En conclusión, hemos obtenido las dos cotas siguientes

$$\mathbb{P}[Z > (1+\delta)p] \leq e^{-p\frac{\delta^2}{2+2\delta/3}},$$

$$\mathbb{P}[Z < (1-\delta)p] \leq e^{-p\frac{\delta^2}{2+2\delta/3}}.$$

Estas son las llamadas cotas de Chernoff. Recordamos que teníamos $Z = \sum_i^m Z_i$ donde Z_1,...,Z_m variables aleatorias independientes siguiendo una distribución de Bernouilli de parámetros p_i y donde $p = \sum_i^m p_i$.
Puesto que se verifica $p = \mathbb{E}[\mathbb{Z}]$, las cortas de Chernoff indican que la variable aleatoria Z se encuentra concentrada en torno a su media:

$$\mathbb{P}[(1+\delta)p < Z < (1-\delta)p] \leq 2e^{-p\frac{\delta^2}{2+2\delta/3}}$$

o lo que es lo mismo

$$\mathbb{P}[(1-\delta)p < Z < (1+\delta)p] \leq 1 - 2e^{-p\frac{\delta^2}{2+2\delta/3}}.$$

Gráficamente se ilustra en la figura (8.1).

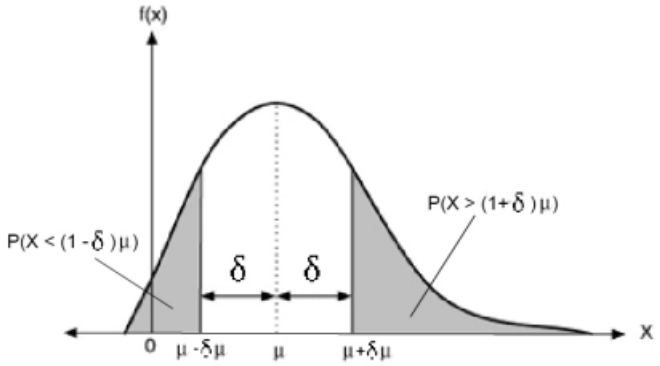

Figura 8.1: Descripción gráfica de las cotas de Chernoff.

Concentración de variables χ^2.

Sean $X_1, ..., X_k$ variables aleatorias independientes normalmente distribuidas, $X_i \sim N(0,1)$ para $i = 1, ..., k$. La variable X_i^2 sigue una distribución χ^2 (ji cuadrado), y $Z = X_1^2 + ... + X_k^2$ sigue una distribución χ^2 con k grados de libertad. Tenemos que $\mathbb{E}[X_i^2] = \mathbb{E}[X_i]^2 = 1$, y $\mathbb{E}[Z] = \mathbb{E}[\sum_{i=1}^k X_i] =$

$\sum_{i=1}^{k} \mathbb{E}[X_i] = k$. En el siguiente lema se muestra que la variable aleatoria χ_k^2 está concentrada alrededor de su media.

Lema 8.0.4. *Sea $Z \sim \chi_k^2$. Entonces para todo $\epsilon > 0$ tenemos*

$$\mathbb{P}[Z \leq (1-\epsilon)k] \leq e^{-\epsilon^2 k/6}, \tag{8.2}$$

y para $\epsilon \in (0,3)$ tenemos

$$\mathbb{P}[Z \geq (1+\epsilon)k] \leq e^{-\epsilon^2 k/6} \tag{8.3}$$

Demostración. Queremos acotar la cantidad $\mathbb{P}[Z \leq (1-\epsilon)k]$ para todo $\epsilon > 0$, para lo que utilizaremos el método de las cotas de Chernoff que hemos visto anteriormente.

$$\begin{aligned}
\mathbb{P}[Z \leq (1-\epsilon)k] &= \mathbb{P}[-Z \geq -(1-\epsilon)k] \\
&= \mathbb{P}[e^{-\lambda Z} \geq e^{-(1-\epsilon)k\lambda}] \\
&\leq \frac{\mathbb{E}[e^{\lambda Z}]}{e^{-(1-\epsilon)k\lambda}} \\
&= e^{(1-\epsilon)k\lambda}\mathbb{E}[e^{-\lambda\sum_{i=1}^{k} X_i^2}] \\
&= e^{(1-\epsilon)k\lambda}\mathbb{E}[\prod_{i=1}^{k} e^{-\lambda X_i^2}] \\
&= e^{(1-\epsilon)k\lambda}\prod_{i=1}^{k}\mathbb{E}[e^{-\lambda X_i^2}] \\
&= e^{(1-\epsilon)k\lambda}(\mathbb{E}[e^{-\lambda X_1^2}])^k.
\end{aligned}$$

En primer lugar hemos aprovechado la monotonía de la exponencial y la multiplicación por una constante $\lambda > 0$ que determinaremos a continuación. Después se ha utilizado la desigualdad de Markov y el hecho que los X_i son variables aleatorias independientes e igualmente distribuidas.

Queremos ahora acotar $\mathbb{E}[e^{\lambda X_1^2}]$. Utilizando que $e^{-a} \leq 1 - a + \dfrac{a^2}{2}$ para todo $a > 0$:

$$\mathbb{E}[e^{\lambda X_1^2}] \leq 1 - \lambda\mathbb{E}[X_1^2] + \frac{\lambda^2}{2}\mathbb{E}[X_1^4].$$

Ahora, calculamos

$$\mathbb{E}[X_1^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\, dx = 1$$

$$\mathbb{E}[X_1^4] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^4 e^{-\frac{x^2}{2}}\, dx = 3$$

y utilizamos que $1 - a \leq e^{-a}$ para obtener

$$\mathbb{E}[e^{-\lambda X_1^2}] \leq 1 - \lambda + \frac{3}{2}\lambda^2 \leq e^{-\lambda + \frac{3}{2}\lambda^2}.$$

Tenemos entonces

$$\begin{aligned}
\mathbb{P}[Z \leq (1-\epsilon)k] &\leq e^{(1-\epsilon)k\lambda}(\mathbb{E}[e^{-\lambda X_1^2}])^k \\
&\leq e^{(1-\epsilon)k\lambda} e^{-\lambda k + \frac{3}{2}\lambda^2 k} \\
&= e^{-\epsilon\lambda k + \frac{3}{2}\lambda^2 k},
\end{aligned}$$

y tomando $\lambda = \epsilon/3$ obtenemos la desigualdad (8.2)

Para la desigualdad (8.3) utilizaremos la siguiente expresión de la función generadora de momentos de la distribución χ^2

$$\forall \lambda < \frac{1}{2}, \mathbb{E}[e^{\lambda Z^2}] = (1-2\lambda)^{-k/2}.$$

Utilizando el método de Chernoff tenemos

$$\begin{aligned}
\mathbb{P}[Z \geq (1+\epsilon)k] &= \mathbb{P}[e^{\lambda Z} \geq e^{(1+\epsilon)k\lambda}] \\
&\leq e^{-(1+\epsilon)k\lambda}\mathbb{E}[e^{\lambda Z}] \\
&= e^{-(1+\epsilon)k\lambda}(1-2\lambda)^{-k/2} \\
&\leq e^{-(1+\epsilon)k\lambda} e^{k\lambda} = e^{-\epsilon k\lambda}.
\end{aligned}$$

En la última desigualdad hemos utilizado que $(1-a) \leq e^{-a}$. Tomando $\lambda = \epsilon/6$ obtenemos la desigualdad buscada. Observamos que ϵ tiene que estar comprendido entre 0 y 3 para que se verifique la condición impuesta sobre λ. □

Combinando las dos desigualdades anteriores es inmediato afirmar que para todo $\epsilon \in (0, 3)$ tenemos

$$\mathbb{P}[(1-\epsilon)k \leq Z \leq (1+\epsilon)k] \geq 1 - 2e^{-\epsilon^2 k/6}, \tag{8.4}$$

Recordamos que $\mathbb{E}[Z] = k$, lo que indica que la distribución χ^2 se encuentra distribuida en torno a su media.

Bibliografía

[1] Candes, E. y Tao, T. (2005), 'Decoding by linear programming', IEEE Trans. on Information Theory 51, 4203-4215.

[2] D. L. Donoho, 'Compressed sensing,', IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289-1306, abril 2006, doi: 10.1109/TIT.2006.871582.

[3] Hastie, T., Tibshirani, R. y Wainwright, M. (2015). "Statistical Learning with Sparsity: The Lasso and Generalizations". CRC Press.

[4] K. R. Rao and P. Yip (1990). 'Discrete Cosine Transforms'. Academic Press, New York.

[5] V. Britanak and K. R. Rao. 'Two-dimensional DCT/DST universal computational structure for 2m × 2n block sizes', IEEE Transactions on Signal Processing, Vol. 48, No. 11, November 2000, pp. 3250—3255.

[6] Makhoul, J. (1980). 'A fast cosine transform in one and two dimensions'. IEEE Transactions on Acoustics, Speech, and Signal Processing. 28 (1): 27-34

[7] A.Jensen and A.la Cour-Harbo (2001). 'Ripples in mathematics: The discrete wavelet transform'. Berlin: Springer.

[8] Daubechies, Ingrid (1999). 'Ten Lectures on Wavelets'. 1a ed. 6a reimp. Philadelphia [etc: Society for Industrial and Applied Mathematics.

[9] Joab R. Winkler. 'Orthogonal Wavelets via Filter Banks Theory and Applications'. The University of Sheffield, 2000.

[10] B. A.Croisier, D.Esteban, C.Galand (1976). 'Perfect channel splitting by use of interpolation/decimation tree decomposition techniques'. First International Conference on Sciences and Systems, Patras, pp.443-446.

[11] R. Masiero, G. Quer, D. Munaretto, M. Rossi, J. Widmer and M. Zorzi, "Data Acquisition through Joint Compressive Sensing and Principal Component Analysis,"GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference, 2009, pp. 1-6, doi: 10.1109/GLOCOM.2009.5425458.

[12] Johnson, W. B. and Lindenstrauss, J. (1984). 'Extensions of Lipschitz mappings into a Hilbert space'. Contemporary Mathematics 26, 189–206.

[13] Candès, E. (2008). 'The restricted isometry property and its implications for compressed sensing'. Comptes Rendus Mathematique.

[14] Shalev-Shwartz, Shai and Ben-David, Shai (2014). 'Understanding Machine Learning - From Theory to Algorithms'. Cambridge University Press.

[15] Baraniuk, R., Davenport, M., DeVore, R (2008). 'A Simple Proof of the Restricted Isometry Property for Random Matrices'.

[16] S.S. Chen, D.L. Donoho, M.A. Saunders. 'Atomic Decomposition by Basis Pursuit'. SIAM Journal on Scientific Computing 20(1), p.33-61, 1998. https://doi.org/10.1137/S1064827596304010

[17] R. Tibshirani 'Regression Shrinkage and Selection via the lasso'. Journal of the Royal Statistical Society: Series B 58(1), p.267–88, 1996. JSTOR 2346178

[18] G. B. Dantzig. 'Linear Programming and Extensions'. Princeton University Press, Princeton, NJ, 1963.

[19] Peter Bloomfiels and William Steiger. 'Least Absolute Deviations: Theory, Applications, and Algorithms'. Birkhauser, Boston, 1983.

[20] Chen, Scott Shaobing and Donoho, David L. and Saunders, Michael A. (1998). 'Atomic Decomposition by Basis Pursuit'. SIAM Journal on Scientific Computing.

[21] Karmarkar, N. (1984). 'A new polynomial-time algorithm for linear programming'. Combinatorica. 4 (4): 373–395.