El Condicionamiento en Matrices Singulares y el Jacobiano normal de la Parametrización desde la Descomposición en Valores Singulares

(Condition number on Singular Matrices and the normal Jacobian of Parameterization from Singular Value Decomposition)

Luis Miguel Soto Sánchez

Trabajo de Fin de Grado

para acceder al

Grado en Matemáticas

Facultad de Ciencias

Universidad de Cantabria

Director: Luis Miguel Pardo Vasallo

Junio - 2022

Agradecimientos

En primer lugar me gustaría dar las gracias a mi director Luis Miguel, por todo su esfuerzo, dedicación y sobretodo paciencia. Durante este TFG con él tengo la sensación de que he aprendido muchas matemáticas nuevas y que además me han parecido muy interesantes y elegantes. Por último agradecerle su buen trabajo haciéndome sentir cómodo en todo momento.
Me gustaría también agradecer al resto de profesores de la Universidad de Cantabria por el trato recibido y por todo lo que he aprendido.
Agradecer a mi familia por el apoyo durante estos años, especialmente a mi hermano, por el cual seguramente hice esta carrera y que me ha ayudado cuando lo he necesitado.

Abstract. This paper begins by presenting some basic results on the Turing-von Neumann ($\kappa(A)$) and Edelmann-Smale ($\mu(A)$) condition numbers for matrices with or without rank deficit. The classical justification of the condition numbers for regular matrices is usually the control of the relative error caused by the perturbation of a system of linear equations, in the case of non-square or non-maximum rank matrices, this work is tackled, with self-contained proof, by Wedin's Theorem, que muestra how conditioning of singular or rank-deficit matrices is related to accuracy in solving least squares problems. The next fundamental result that is treated is the Schmidt-Mirsky-Eckart-Young Theorem, which proves the equivalence between the conditioning and the inverse of a distance on a projective space (or a sphere), this Theorem is also proved in a self-contained way.
After the completion of said demonstration, the second part of the work begins, which deals with the problem of the behavior in average of these condition numbers. For this, we make use of the Co-area Formula, and following an excellent work by C. Beltrán of 2009, we carry out the complete procedures to obtain the Normal Jacobian of the parameterization coming from the singular value decomposition, among other results.

Key words: Condition number, singular values, normal Jacobian.

Resumen. Este trabajo comienza presentando algunos resultados básicos sobre los números de condicionamiento de Turing-von Neumann ($\kappa(A)$) y de Edelmann-Smale ($\mu(A)$) para matrices con o sin déficit de rango. La justificación clásica de los números de condicionamiento para matrices regulares suele ser el control del error relativo causado por la perturbación de un sistema de ecuaciones lineales, en el caso de matrices no cuadradas o que no son de rango máximo, este trabajo abordará, con demostración autocontenida, el Teorema de Wedin que muestra cómo el condicionamiento de matrices singulares o con déficit de rango se relaciona con la precisión en la resolución de los problemas de mínimos cuadrados. El siguiente resultado fundamental que se aborda es el Teorema de Schmidt-Mirsky-Eckart-Young que prueba la equivalencia entre el condicionamiento y el inverso de una distancia sobre un espacio proyectivo (o una esfera) y que también es demostrado de manera autocontenida.
Tras la finalización de dicha demostración comienza la segunda parte del trabajo, la cual trata el problema del comportamiento en promedio de estos números de condicionamiento. Para ello, hacemos uso de la Fórmula de la Co-área, y siguiendo un excelente trabajo de C. Beltrán de 2009 realizamos los procedimientos completos para obtener el Jacobiano Normal de la parametrización proveniente de la descomposición en valores singulares entre otros resultados.

Palabras clave: Condicionamiento, valores singulares, jacobiano normal.

Índice

Capítulo 0. Introducción y Resumen de Contenidos de la Memoria i
0.1. Introducción i
0.2. Resumen de los contenidos de la Memoria ii
0.2.1. Resumen del Capítulo 1 ii
0.2.2. Resumen del Capítulo 2 v
0.2.3. Apéndices Finales viii
0.2.4. Sobre el estilo y la ortografía usados en este TFG ix

Capítulo 1. Condicionamiento Lineal: Propiedades Básicas 1
1.1. Introducción y Resumen del Contenido del Capítulo 1
1.1.1. Resumen de los contenidos del Capítulo 2
1.1.2. Precisión, perturbaciones y condicionamiento: una cota superior relativa 5
1.2. Números de Condicionamiento del Álgebra Lineal: nociones básicas 6
1.3. Condicionamiento, Mínimos Cuadrados y Perturbaciones: Teorema de Wedin 10
1.4. El teorema de Schmidt-Mirsky-Eckart-Young 18
1.4.1. Métricas en el espacio proyectivo y en esferas 18
1.4.2. La métrica en $\mathbb{P}_{N-1}(\mathbb{R})$ y S^{N-1} 19
1.4.3. Las Métricas en $\mathbb{P}_{N-1}(\mathbb{C})$ 21
1.4.4. El Teorema de Schmidt-Mirsky-Eckart-Young 22

Capítulo 2. Elementos de la geometría Integral en algunas variedades de matrices. 29
2.1. Introducción y Resumen del Contenido del Capítulo 29
2.2. La esfera como variedad de Riemann 32
2.3. Estructura Riemanniana del grupo ortogonal 33
2.4. El jacobiano normal de la parametrización desde la SVD 42
2.5. El volumen de la fibra de la parametrización SVD de una matriz 45
2.6. El uso de la fórmula de la Co-área para la parametrización SVD 46

Apéndice A. Algunos Conocimientos previos al condicionamiento de matrices 49
A.1. Pseudo-inversa de Moore-Penrose 49
A.2. Estratificación por co-rango del espacio del matrices 52
A.3. Existencia de Descomposición en Valores Singulares 53
A.4. El Teorema Espectral 54
A.5. Teoremas de Transversalidad 55
A.6. Algunas propiedades destacables de la Fórmula de la Co-área 56
A.6.1. Integración en espacios euclídeos. 56
A.6.2. Integración en variedades diferenciables. 57
A.6.3. La Fórmula de la Co-área. 58
A.6.4. Probabilidad e integrales. 59

Apéndice. Bibliografía 61

CAPíTULO 0

Introducción y Resumen de Contenidos de la Memoria

Índice

0.1. Introducción **i**
0.2. Resumen de los contenidos de la Memoria **ii**
0.2.1. Resumen del Capítulo 1 ii
0.2.2. Resumen del Capítulo 2 v
0.2.3. Apéndices Finales viii
0.2.4. Sobre el estilo y la ortografía usados en este TFG ix

0.1. Introducción

Uno de los grandes expertos mundiales en estabilidad del Análisis Numérico, N. Highman [**High, 02**] responde en su blog (*https://nhigham.com*) a la siguiente pregunta: *"Who Invented the Matrix Condition Number?"*. El uso del término *"ill-conditioned"* se remonta al año 1933, en un trabajo de R.R. Mallock [**Mal, 1933**]. Pero el término y los conceptos subyacentes no son desarrollados por Mallock, cuyo trabajo tiene un perfil poco matemático. J.H. Wilkinson en su libro [**Wilk, 1963**] atribuye las primeras nociones de condicionamiento en Álgebra Lineal a A. Turing: A la sazón, A. Turing investigaba, durante 1945, en un proyecto al British National Physical Laboratory (NPL), con el que pretendía construir un "automatic electronic digital computer with internal program storage". El proyecto denominado Automatic Computing Engine (ACE), nunca llegó a ser completado. Turing acabó dejando el NPL y se trasladó a la Universidad de Manchester en 1948.
J.H. Wilkinson trabajó con A. Turing en el NPL entre 1946 y 1948. El propio Wilkinson afirma, en la charla de su medalla Turing de 1970, que Turing le convirtió desde un analista clásico a un analista numérico, haciéndole abandonar la idea de retornar a Cambridge. No olvidemos la enorme influencia de J.H. Wilkinson en el desarrollo ulterior en el Álgebra Lineal Numérica (ver también [**Wilk, 1965**]). Wilkinson hace referencia al trabajo de A. Turing [**Turing, 1948**], que aunque publicado en 1948, el artículo se habría escrito con anterioridad, según Wilkinson. En ese artículo A. Turing se plantea los siguientes aspectos:
This paper contains descriptions of a number of methods for solving sets of linear simultaneous equations... but its main concern is with the theoretical limits of accuracy...

En el trabajo, Turing pretente explicar no sólo la noción de mal-condicionado, sino también experimentos de autores como H. Hotelling, que mostraban que la eliminación gaussiana y otros métodos directos conducen a errores de redondeo exponenciales. Turing pretende clasificar casos "normales", distinguiéndolos de casos "extremadamente pesimistas". Para ello, A. Turing introduce los números de condicionamiento M y N, que evolucionaron con el tiempo a los más actuales que trataremos en este Trabajo Fin de Grado.

En una visita a Princeton, en 1947, A. Turing tiene una conversación con J. von Neumann que, en [**Turing, 1948**], se resume en la siguiente frase:

In the meantime another theoretical investigation was being carried out by J.v. Neumann, who reached conclusions similar to those of this paper for the [particular] case of positive definite matrices, and communicated them to the writer [A. Turing] at Princeton in January 1947 before the proofs given here were complete.

Como el trabajo de von Neumann vió la luz más rápidamente que el de Turing, las versiones más "conformistas" de la historia suele asignar "la prioridad" a J. von Neumann y H. Goldstine (ver su monumental trabajo en [**vNG, 1947**]). Alguna que otra leyenda urbana cuenta que, dada la debilidad profesional de Turing frente al gigante matemático von Neumann, el flujo de información entre ambos no fue bidireccional, por más que Turing le de a von Neumann el desarrollo independiente de la teoría.
En cualquier caso, sin entrar en polémicas históricas, en este Trabajo Fin de Grado atribuiremos el origen de los números de condicionamiento a los desarrollos independientes y simultáneos de [**Turing, 1948**] y [**vNG, 1947**]. Más discusiones sobre la historia del origen del condicionamiento (siempre en discusión) puede seguirse en la versión extendida, escrita por L. Blum, de sus trabajos [**Blum, 04**] y [**Blum, 12**].
El objetivo último del Trabajo Fin de Grado es doble: de una parte recordaremos algunas nociones de condicionamiento lineal (especialmente en el caso de déficit de rango) y trataremos de explicar esa idea de "normalidad" frente a "extremadamente pesimistas" (que ya avanzaba A. Turing) a través de resultados sobre la distribución de probabilidad del condicionamiento. En este segundo análisis nos centraremos en el trabajo de [**Be, 09**] y sus brillantes ideas. A continuación describiremos estos objetivos a través de la descripción de los dos Capítulos.

0.2. Resumen de los contenidos de la Memoria

En esta Sección vamos a exponer cuáles son los contenidos de los diferentes Capítulos de esta memoria. También se incluye una mención del Apéndice que contiene esta memoria y otra en la que se trata el estilo que sigue la memoria.

0.2.1. Resumen del Capítulo 1. El Capítulo 1 comenzará recordando dos de las versiones básicas del condicionamiento lineal. Supongamos $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$:

- *El condicionamiento de Turing-von Neumann-Wilkinson (entre otros):* Dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ definimos el número de condicionamiento de Turing-von Neumann-Wilkinson et al. mediante:
$$\kappa(A) := \| A \|_2 \| A^+ \|_2,$$
donde A^+ es la pseudo-inversa de Moore-Penrose (véase la Sección A.1 del Apéndice A) y $\| \cdot \|_2$ hace referencia a la norma como operador lineal.
Nótese que en el caso de matrices cuadradas regulares $A \in GL(n, \mathbb{K})$ este es el mismo condicionamiento que usualmente se discute en el grado.
- *El condicionamiento de Demmel-Edelman-Smale y otros:* Para una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ definimos el número de condicionamiento mediante:
$$\mu(A) := \| A \|_F \| A^+ \|_2,$$
donde $\| \cdot \|_F$ es la norma de Frobenius de la matriz (véase el Lema 1.2.1), mientras que $\| \cdot \|_2$ y A^+ son las introducidas en el item anterior.

Aunque hay otras posibles variantes, esencialmente equivalentes, en este Trabajo de Fin de grado nos limitaremos a trabajar con estas dos que aparecen recogidas en la Definición 3 del Capítulo 1.
Ambas nociones están relacionadas con los valores singulares de la matriz (véase Proposición 1.2.8 y la Sección A.3 del Apéndice A para un sucinto recordatorio de la SVD). De hecho, la Descomposición en Valores Singulares subyace a buena parte de esta memoria y es el "leitmotiv" de su desenlace. La justificación clásica del uso de los números de condicionamiento para matrices regulares $A \in GL(n, \mathbb{K})$ suele ser el control del error relativo causado por la perturbación de un sistema de ecuaciones lineales que se pretende resolver (véase Proposición 1.1.1 por ejemplo). En el caso de matrices que no son cuadradas o no tienen un rango máximo, el número de condicionamiento, tal y como lo hemos definido, tiene su justificación en la resolución del problema de mínimos cuadrados (o proyeccion ortogonal). El resultado fundamental, mucho menos conocido en este contexto, es el Teorema de Wedin (cf. [**Wed, 1973**]) que recogemos como Teorema 1.3.8:

Teorema 0.2.1. *Sea $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ una matriz de rango máximo con $n \geq m$. Sea $b \in \mathbb{K}^n$ un vector. Sea $x \in \mathbb{K}^m$ un punto que resuelve el problema de mínimos cuadrados con datos (A, b),*

es decir,

$$\| Ax - b \|_2 = \min\{\| Az - b \|_2 : z \in \mathbb{K}^m\}.$$

Denotamos por $r := b - Ax \in \mathbb{K}^n$*, el residuo asociado a la solución* x *dada.*
Consideramos una serie de perturbaciones de los datos $\delta A \in \mathcal{M}_{n\times m}(\mathbb{K})$*,* $\delta b \in \mathbb{K}^n$ *y sea* $\epsilon > 0$ *un número real positivo. Supongamos que las perturbaciones satisfacen las siguientes propiedades:*

- *El número real* ϵ *acota los errores relativos, esto es, se tiene:*

$$\frac{\| \delta A \|_2}{\| A \|_2} \leq \epsilon \ , \ \frac{\| \delta b \|_2}{\| b \|_2} \leq \epsilon$$

- *La matriz* $A + \delta A \in \mathcal{M}_{n\times m}(\mathbb{K})$ *es también una matriz de rango máximo.*

Sea $y \in \mathbb{K}^m$ *una solución del problema de mínimos cuadrados con datos (perturbados)* $(A + \delta A, b + \delta b)$*, es decir, supongamos*

$$\|(A + \delta A)y - (b + \delta b)\|_2 = \min\{\|(A + \delta A)z - (b + \delta b)\|_2 : z \in \mathbb{K}^m\},$$

y denotemos por $\delta x = y - x$ *y por* $s = (b + \delta b) - (A + \delta A)y \in \mathbb{K}^m$ *el residuo asociado a estos datos perturbados y a la solución* $y \in \mathbb{K}^m$*.*
Entonces, si $\kappa(A) \cdot \epsilon < 1$*, se tienen las siguientes desigualdades:*

$$\frac{\| \delta x \|_2}{\| x \|_2} = \frac{\| y - x \|_2}{\| x \|_2} \leq \frac{\epsilon \cdot \kappa(A)}{1 - \epsilon \cdot \kappa(A)} \left(2 + (\kappa(A) + 1) \frac{\| r \|_2}{\| A \|_2 \| x \|_2} \right),$$

$$\frac{\| \delta r \|_2}{\| b \|_2} = \frac{\| r - s \|_2}{\| b \|_2} \leq (1 + 2\kappa(A)) \cdot \epsilon,$$

donde hemos usado $\delta r = s - r$*.*

Lo que aporta esta memoria es una demostración con los detalles esenciales requeridos para su prueba. Algunos, por no salirnos del espacio máximo, han ido al apéndice.

La siguiente sección del Capítulo 1 se dedica a dar una prueba completa y autocontenida del Teorema de Schmidt-Mirsky-Eckart-Young (ver [**Schm, 1967**], [**Mrs, 1960**], [**E&Y, 1936**]). Una observación natural de los números de condicionamiento introducidos anteriormente es que ambos son funciones homogéneas de grado 0, esto es,

$$\forall A \in \mathcal{M}_{n\times m}(\mathbb{K}), \forall \lambda \in \mathbb{K} \setminus \{0\}$$

$$\kappa(\lambda A) = \kappa(A)$$

$$\mu(\lambda A) = \mu(A).$$

Esto significa que son, por naturaleza, funciones definidas en las correspondientes esferas unidad $S^{nm-1}(\mathbb{K})$ donde

- $S^{nm-1}(\mathbb{R}) = S^{nm-1}$ es la esfera real usual.
- $S^{nm-1}(\mathbb{C}) = S^{2(nm)-1}$ es la esfera unidad compleja (en $\mathbb{C}^{nm}$).

Tiene sentido, por tanto, interpretar ambos condicionamientos como funciones de las esferas del espacio de matrices

$$\kappa, \mu : S^{nm-1}(\mathbb{K}) \longrightarrow \mathbb{R}_+.$$

De hecho, son funciones naturalmente definidas en los espacios proyectivos respectivos $\mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{R})) = \mathbb{P}_{nm-1}(\mathbb{R})$ o $\mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{C})) = \mathbb{P}_{nm-1}(\mathbb{C})$.
La sorpresa surge de un clásico resultado histórico, atribuido a varios autores, y que aquí presentamos como Teorema de Schmidt-Mirsky-Eckart-Young (cf. [**Schm, 1967**], [**Mrs, 1960**], [**E&Y, 1936**]). Para entender el resultado, consideremos la esfera $S^{n-1} \subseteq \mathbb{R}^n$. La esfera S^{n-1} es una subvariedad de Riemann de dimensión $n-1$ con una natural función de distancia geodésica basada en el ángulo y que denominamos *distancia riemanniana en la esfera.* Dicha función distancia viene dada por:

$$\begin{array}{rccl} d_R\colon & S^{n-1} \times S^{n-1} & \longrightarrow & \mathbb{R}_+ \\ & (u, v) & \longmapsto & \arccos\left(|\langle u, v\rangle|\right), \end{array}$$

donde $\langle \cdot, \cdot \rangle$ es el producto euclídeo usual en $\mathbb{R}_+$.
Una distancia suplementaria en S^{n-1} es denominada la *distancia del seno* por algunos autores y la *distancia proyectiva* por otros y viene dada por la siguiente identidad:

$$\begin{array}{rccc} d_{\mathbb{P}}: & S^{n-1} \times S^{n-1} & \longrightarrow & \mathbb{R}_+ \\ & (u, v) & \longmapsto & \sin(d_R(u, v)) = \sqrt{1 - \langle u, v \rangle^2}. \end{array}$$

Ambas métricas, junto con la métrica euclídea usual, son comparables como se destribe en la figura siguiente, donde se describen las métricas que acabamos de mencionar, de izquiera a derecha:

i) Métrica Euclídea: $\| u - v \|_2$
ii) Métrica Riemanniana: $d_R(u, v)$
iii) Métrica Riemanniana: $d_{\mathbb{P}}(u, v)$

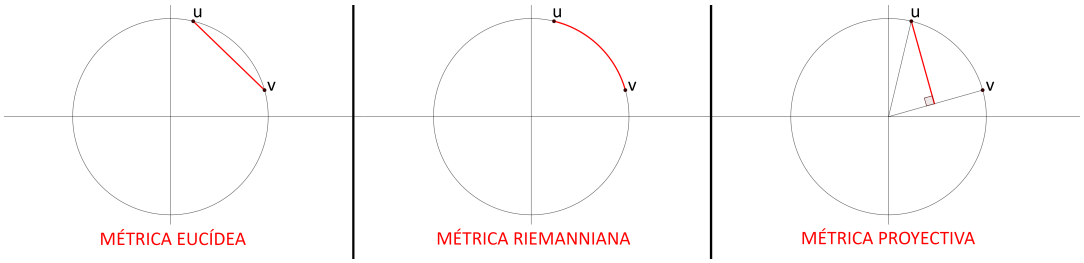

FIGURE 1. Métricas sobre el espacio proyectivo.

Las métricas de la esfera se trasladan de manera natural a los espacios proyectivos como se muestra en la Subsección 1.4.2. Para el caso complejo estas métricas se redefinen de manera natural a partir de la métrica de Fubini-Study tanto en $S^{2n-1} = \{x \in \mathbb{C}^n : \| z \|^2 = 1\}$ como en el espacio proyectivo complejo $\mathbb{P}_{n-1}(\mathbb{C})$.
Si consideramos $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$ y el espacio afín $\mathbb{A}^{nm} = \mathcal{M}_{n \times m}(\mathbb{K})$ de las matrices coordenadas en $\mathbb{K}$, las anteriores métricas se construyen de manera natural en las esferas S^{nm-1} (caso real), S^{2nm-1} (caso complejo) y los espacios proyectivos $\mathbb{P}_{nm-1}(\mathbb{R})$ (caso real) y $\mathbb{P}_{nm-1}(\mathbb{C})$ (caso complejo). Si acaso, resaltemos que la métrica euclídea (caso real) o hermítica (caso complejo) parte del producto euclídeo o hermítico de Frobenius:

$$\langle A, B \rangle_F := Tr(AB^*),$$

donde B^* es la traspuesta conjugada de B. Estas construcciones son las que se exponen en las subsecciones 1.4.1, 1.4.2, 1.4.3. Así, tenemos la distancia del seno para matrices que, a continuación, expresamos reescribiendo para espacios proyectivos:

$$\begin{array}{rccc} d_{\mathbb{P}}: & \mathbb{P}_{nm-1}(\mathbb{K}) \times \mathbb{P}_{nm-1}(\mathbb{K}) & \longrightarrow & \mathbb{R}_+ \\ & (A, B) & \longmapsto & \sqrt{1 - \frac{\langle A, B \rangle_F^2}{\| A \|_F^2 \, \| B \|_F^2}}. \end{array}$$

El Teorema de Schmidt-Mirsky-Eckart-Young clásico para matrices cuadradas de rango máximo $A \in \mathbb{P}_{n^2-1}(\mathbb{K})$ mediante:

$$\mu(A) = \frac{1}{d_{\mathbb{P}}(A, \sum)},$$

donde $\sum \subseteq \mathbb{P}_{n^2-1}(\mathbb{K})$ es la subvariedad algebraica proyectiva formada por aquellos puntos de $\mathbb{P}_{n^2-1}(\mathbb{K})$ determinados por matrices singulares. Esto es,

$$\sum := \{A \in \mathbb{P}_{n^2-1}(\mathbb{K}) : \det(A) = 0\},$$

donde $\det(\cdot)$ es el determinante de la matriz A. La relevancia más destacable de este resultado clásico es que transforma una estimación de la estabilidad, precisión (y hasta la complejidad) de algoritmos matriciales en un objeto de naturaleza puramente geométrica: el inverso de la distancia a la variedad de matrices singulares. Nótese que, en este sentido, el condicionamiento de Demmel-Edelman-Smale tiene una formulación más elegante como invariante geométrico.
El resultado clásico es fácilmente generalizable a situaciones donde las matrices no son cuadradas y ni siquiera son de rango máximo, lo que muestra lo intrínsecamente geométrico que resulta el condicionamiento.

En el Capítulo 1 nos dedicamos a ofrecer una prueba auto-contenida completa de este resultado. Para establecerlo, consideremos el espacio afín de las matrices $\mathcal{M}_{n\times m}(\mathbb{K})$, con coordenadas en $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$. Consideremos el respectivo espacio proyectivo $\mathbb{P}_{nm-1}(\mathbb{K})$ y consideremos la estratificación por rango (véase el texto [**AGV, 1985**] o la Sección A.2 del Apéndice A, para más información).
Denotemos por $\sum_t' \subseteq \mathcal{M}_{n\times m}(\mathbb{K})$ la variedad algebraica de las matrices cuyo rango es a lo sumo t para $0 \leq t \leq \min\{n, m\}$. Es decir,

$$\sum_t' := \{A \in \mathcal{M}_{n\times m}(\mathbb{K}) : \operatorname{rank}(A) \leq t\}.$$

Consideramos la proyección canónica $\pi : \mathcal{M}_{n\times m}(\mathbb{K}) \setminus \{0\} \longrightarrow \mathbb{P}_{nm-1}(\mathbb{K})$. Obsérvese que $\sum_t'$ es una variedad algebraica afín definida mediante ecuaciones polinomiales homogéneas de grados entre $t+1$ y $\min\{n, m\}$: son los determinantes $k \times k$ aplicados a los menores de orden k de A, con $t+1 \leq k \leq \min\{n, m\}$. Por tanto, tiene sentido considerar la variedad algebraica proyectiva

$$\sum_t := \pi(\sum_t' \setminus \{0\}) \subseteq \mathbb{P}_{nm-1}(\mathbb{K}).$$

Para una matriz (o punto proyectivo) $\pi(A) \in \sum_t$ consideremos su condicionamiento (antes definido) $\mu(A) = \| A \|_F \| A^+ \|_2$. De hecho, podemos elegir $A \in \sum_t'$ con $\| A \|_F = 1$ como representante y considerar $\mu(A) = \| A^+ \|_2$.
El Teorema de Schmidt-Mirsky-Eckart-Young que probamos es el siguiente:

Teorema 0.2.2. *Con las notaciones precedentes, sea $A \in \sum_t', 1 \leq t \leq \min\{n, m\}$ y $\alpha = \pi(A) \in \mathbb{P}_{nm-1}(\mathbb{K})$. Se tiene:*

$$\mu(A) = \frac{1}{d_{\mathbb{P}}(\alpha, \sum_{t-1})} = \frac{1}{d_F(A, \sum_{t-1}')},$$

donde $d_F(A, \sum_{t-1}')$ es la distancia de cualquier representante A de α con $\| A \|_F = 1$ al cono proyectante $\sum_{t-1}'$ con respecto a la métrica de Frobenius.

La subsección 1.4.4 es la culminación de la Sección 1.4 dedicada a la prueba de este Teorema.

0.2.2. Resumen del Capítulo 2. Como ya se señala en el Teorema de Wedin o en la Subsección 1.1.1, el condicionamiento lineal es un parámetro muy necesario para controlar la resolución de problemas (ecuaciones lineales, mínimos cuadrados, etc.) tras perturbar los datos del mismo. Es, por tanto, esencial, en la aplicación de algoritmos numéricos, el disponer "a priori" de buenas razones (tranquilizadoras) que nos garanticen que el problema particular que se trata está bien condicionado. Sin embargo, no resulta muy eficiente exigir que cada aplicación de nuestro algoritmo calcule el condicionamiento del input, lo que puede resultar en un alto aumento del tiempo de ejecución.
Surge así la pregunta de conocer el comportamiento en promedio de los números de condicionamiento lineales. Algunas referencias que citan el problema (y son citadas en la tesis de A. Edelman) son [**vN, 1963**], [**Birk&Gul, 1963**], [**Smale, 1985**], [**Demmel, 1988**]. De hecho, será A. Edelman quien presente los resultados más "ajustados" (en especial en el caso complejo) del comportamiento en promedio del condicionamiento lineal en [**Edelman, 1988**] y en su Ph. D. Thesis [**Edelman, 1989**].
Estos resultados vienen a garantizar un resultado del tipo siguiente:

> "La probabilidad de que una matriz $A \in \mathcal{M}_n(\mathbb{K})$ tenga un condicionamiento $\mu(A) > \epsilon^{-1}$ para $\epsilon > 0, \epsilon \in \mathbb{R}$, está acotada para la distribución gaussiana por una cantidad de la forma
>
> $$\varphi(n) \cdot \epsilon^k \text{ "}$$

donde $k \in \{1, 2\}$ según sea el caso real o el complejo y $\varphi(n)$ es una cantidad "razonable".
Este tipo de resultados (y muchos otros que se generaron a posteriori), tienen la siguiente pseudo-interpretación algorítmica: dado un algoritmo que recibe como input una matriz A aleatoriamente elegida, si su comportamiento depende del condicionamiento de A, este comportamiento será "bueno" con alta probabilidad. No es pues necesario obligar a nuestros algoritmos a una pre-computación del condicionamiento del input que empeora nuestro rendimiento.
De otro lado, el resultado meramente como resultado matemático es interesantísimo por su propia naturaleza y los "elementos" que lo envuelven. Y sólo ese argumento estético bastaría para ser interesante.
Sin embargo hay un error de principio (o, si se prefiere, una sutileza) que debe ponernos en estado de atención y cuidado. En el mundo real, los problemas para los que se diseñan los

algoritmos no suelen aplicarse a inputs genéricos sino, por ejemplo, a subclases de matrices y problemas. Pongamos tres ejemplos:

- Nuestro algoritmo trata solamente con matrices A estocásticas en $\mathcal{M}_n(\mathbb{R})$.
- Nuestro algoritmo trata solamente con matrices A tri-diagonales en $\mathcal{M}_n(\mathbb{R})$.
- Nuestro algoritmo trata solamente con matrices singulares A en $\mathcal{M}_n(\mathbb{R}) \setminus GL(n,\mathbb{R})$.

Estos tres ejemplos, y muchos otros fáciles de entender, transportan una característica común:

> Todos ellos tienen inputs en un conjunto $\mathscr{L} \in \mathcal{M}_n(\mathbb{R})$ que tiene medida de Lebesgue nula en el espacio afín $\mathbb{R}^{n^2} = \mathcal{M}_n(\mathbb{R})$.

Por tanto, los resultados de probabilidad genérica (i.e. en dimensión n^2) de Edelman, Demmel o Smale no se les aplican: no se puede heredar un resultado de distribución de probabilidad sobre un espacio probabilístico (X,ν) a un subespacio $(Y,\widetilde{\nu})$ cuando $\nu(Y)=0$.
Esto supone, de facto, la necesidad de estudios "ad hoc" del comportamiento en promedio del condicicionamiento lineal restringido a "clases especiales" de matrices. La dificultad es más que evidente: Tomemos como ejemplo el comprender el condicionamiento de matrices $A \in \mathcal{M}_n(\mathbb{C})$, con rango a lo sumo $n-1$. Es decir, sea $\sum'_{n-1} \subseteq \mathcal{M}_n(\mathbb{C})$ la clase de matrices con determinante nulo. Se trata de una matriz A que satisface una ecuación polinomial homogénea de grado n:

$$\textstyle\sum'_{n-1} := \{A \in \mathcal{M}_n(\mathbb{C}) : \det(A) = 0\}.$$

De hecho ni siquiera es una subvariedad diferenciable de $\mathbb{C}^{n^2}$ dado que es un conjunto con abundantes puntos singulares ($\sum'_{n-1}$ son los puntos singulares de $\sum'_n$ (ver [**AGV, 1985**] o Sección A.2 del Apédice A)). Si ahora supongo una distribución gaussiana $(\mathcal{M}_n(\mathbb{C}),\gamma)$ y un resultado sobre la probabilidad, es inaplicable al comportamiento de matrices aleatoriamente elegidas $A \in \sum'_n$.
Ha habido varias aproximaciones a entender este problema. Lo que se desarrolla en el Capítulo 2 es la elegante estrategia de [**Be, 09**]. Consideremos la estratificación por rango ya citada

$$\textstyle\sum'_1 \subsetneq \sum'_2 \subsetneq \ldots \subsetneq \sum'_t \subseteq \mathcal{M}_{n\times m}(\mathbb{K}),$$

con $t = \min\{n,m\}$. Nuestra noción de condicionamiento (tanto κ como μ) se adaptan bien a esta estratificación por usar la pseudo-inversa de Moore-Penrose. Consideremos las proyectivizaciones de esas variedades donde los condicionamientos tienen, además, el significado geométrico que les confiere el Teorema de Schmidt-Mirsky-Eckart-Young.

$$\textstyle\sum_1 \subsetneq \sum_2 \subsetneq \ldots \subsetneq \sum_t \subseteq \mathbb{P}_{nm-1}(\mathbb{K}).$$

Tanto en el caso $\mathbb{K}=\mathbb{R}$ como en el caso $\mathbb{K}=\mathbb{C}$, $\sum_r \setminus \sum_{r-1}$ es una subvariedad de Riemann compacta de $P_{nm-1}(\mathbb{K})$ de volumen finito. Luego hay una distribución de probabilidad natural que da sentido a la cuestión:

PROBLEMA 1. *"Estimar la Probabilidad de elegir aleatoriamente una matriz (un punto proyectivo) $A \in \sum_r \setminus \sum_{r-1}$ tal que $\mu(A) > \epsilon^{-1}$ ".*

De eso trata el Capítulo 2. Además, la idea de [**Be, 09**] es muy interesante por calcular (objetivo esencial de este Trabajo de Fin de Grado) la siguiente estrategia: La describiremos para el caso $\mathbb{K}=\mathbb{R}$. Consideremos la esfera unidad $S^{r-1} \subseteq \mathbb{R}^r$ y definamos el conjunto de puntos de un "octante" estrictamente positivo:

$$S_r^+ := \{(\sigma_1,\ldots,\sigma_r) \in S^{r-1} : \sigma_1 > \sigma_2 > \ldots > \sigma_r > 0\}.$$

Denotemos también por S_r^+ al conjunto de matrices en $\mathcal{M}_n(\mathbb{R})$ cuya diagonal está en S_r^+. Es decir,

$$S_r^+ := \left\{ \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array} \right) \in \mathcal{M}_n(\mathbb{R}) : \sigma_1^2 + \ldots + \sigma_r^2 = 1, \sigma_1 > \ldots > \sigma_r > 0 \right\}.$$

Ahora consideremos la parametrización "natural" de $\left(\sum'_r \setminus \sum'_{r-1}\right) \cap S^{nm-1}$ dada mediante la Descomposición en Valores Singulares representada por S_r^+:

$$\begin{array}{rccc} \varphi\colon & \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m) & \longrightarrow & \left(\sum'_r \setminus \sum'_{r-1}\right) \cap S^{nm-1} \\ & (U,D,V) & \longmapsto & U \cdot D \cdot V^*. \end{array}$$

La clásica fórmula de la Co-áera (ver Subsección A.6.3 del Apéndice A), nos da una regla natural para trasladar esperanzas y probabilidades de la compleja variedad de Riemann $\sum_r \setminus \sum_{r-1}$ a la (más simple) variedad producto $\mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m)$.
Dicha fórmula se resume del modo siguiente:

TEOREMA 0.2.3. [**Fórmula de la Co-área**] *Sean M y N dos variedades Riemannianas. Sean m, n sus dimensiones respectivas, con $m \geq n$. Sea $F : M \to N$ una aplicación diferenciable sobreyectiva tal que la diferencial $d_x F$ es sobreyectiva para casi todo punto $x \in M$. Sea $\phi : M \to [0, \infty]$ una función integrable. Entonces, se tienen las dos siguientes igualdades:*

$$\int_{x \in M} \phi(x) dM = \int_{y \in N} \int_{x \in F^{-1}(y)} \phi(x) \frac{1}{NJ_x F} dF^{-1}(y) dN,$$

$$\int_{x \in M} \phi(x) NJ_x F dM = \int_{y \in N} \int_{x \in F^{-1}(y)} \phi(x) dF^{-1}(y) dN.$$

Es una forma general del Teorema de Cambio de Variable en integración donde el usual jacobiano es reemplazado por el jacobiano normal $NJ_x F$. (ver la Definición 19 de la Subsección A.6.3 del Apéndice A).
En nuestro caso, supongamos $\phi : \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m) \longrightarrow \mathbb{R}$ es una función y $\varphi : \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m) \longrightarrow \sum_r \setminus \sum_{r-1}$ es la parametrización basada en la descomposición en valores singulares. Entonces, el comportamiento de ϕ se puede estudiar mediante la identidad siguiente:

$$I = \int_{\mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m)} \phi(U, D, V) d(\mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m)) = \tag{0.2.1}$$

$$= \int_{A \in \sum_r \setminus \sum_{r-1}} \int_{B \in \varphi^{-1}(A)} \frac{\phi(B)}{NJ_B \varphi} d\varphi^{-1}(A) d(\textstyle\sum_r \setminus \sum_{r-1})$$

A primera vista si ϕ sólo depende de D (como es el caso del condicionamiento) la expresión se simplifica más:

$$I = \int_{\mathscr{O}(n) \times \mathscr{O}(m)} \left(\int_{S_r^+} \phi(D) dS_r^+ \right) d(\mathscr{O}(n) \times \mathscr{O}(m)).$$

Pero además, los valores singulares asociados a una matriz A son únicos por lo que $\phi(U, D, V) = \phi(D)$ es constante en $\varphi^{-1}(A)$ para cualquier A. Así, vamos reduciendo la expresión a

$$I = \int_{A \in \sum_r \setminus \sum_{r-1}} \phi(A) \left(\int_{B \in \varphi^{-1}(A)} \frac{1}{NJ_B \varphi} d\varphi^{-1}(A) \right) d(\textstyle\sum_r \setminus \sum_{r-1}).$$

Resulta, por tanto, crucial comprender el **jacobiano normal de la parametrización por la SVD**, esto es, $NJ_{(U,D,V)}\varphi$ con la notación precedente.
No es extraño observar que el jacobiano normal $NJ_{(U,D,V)}\varphi$ es invariante por la acción de ambos grupos ortogonales o, si se prefiere, es constante en la fibra $\varphi^{-1}(A)$ para cada $A \in \sum_r \setminus \sum_{r-1}$. Así, la identidad toma la forma (para un ϕ dependiente solamente de los valores singulares):

$$I = \int_{\sum_r \setminus \sum_{r-1}} \frac{\phi(D)}{NJ_{(U,D,V)}\varphi} \left(\int_{B \in \varphi^{-1}(A)} d\varphi^{-1}(A) \right) d(\textstyle\sum_r \setminus \sum_{r-1}).$$

En otras palabras, la identidad 0.2.1 para funciones ϕ que sólo dependen de los valores singulares resulta

$$I = \mathrm{Vol}[\mathscr{O}(n)] \cdot \mathrm{Vol}[\mathscr{O}(m)] \int_{S_r^+} \phi(D) d\nu_S \tag{0.2.2}$$

$$= \int_{\sum_r \setminus \sum_{r-1}} \frac{\phi(D)}{NJ_A \varphi} \mathrm{Vol}[\varphi^{-1}(A)] d(\textstyle\sum_r \setminus \sum_{r-1}).$$

donde $NJ_A\varphi$ sería el jacobiano normal en cualquier punto de la fibra $\varphi^{-1}(A)$.
Dado que los volúmenes de los grupos ortogonales $\mathscr{O}(n)$ y $\mathscr{O}(m)$ son conocidos, si elegimos $\phi(D) = \frac{1}{\sigma_r}$ (que es el condicionamiento que nos ocupa) hay solamente dos ingredientes técnicos a controlar y calcular:

- El jacobiano normal $NJ_{(U,D,V)}\varphi$ en cualquier punto $(U, D, V) \in \varphi^{-1}(A)$.
- El volumen de la fibra $\varphi^{-1}(A)$.

Ambos son los resultados fundamentales que se describen en [**Be, 09**] y que demostramos en este Trabajo Fin de Grado. El primero es el Teorema 2.4.5 de la Sección 2.4 y que reproducimos aquí:

TEOREMA 0.2.4. *Sean $U \in \mathscr{O}(n)$, $V \in \mathscr{O}(m)$, $D \in S_r^+$ y sea $A = UDV^* \in \sum_r \setminus \sum_{r-1}$. Sea $NJ_{(U,D,V)}\varphi$ el jacobiano normal de la parametrización φ de la Descomposición en Valores Singulares. Entonces,*

$$NJ_{(U,D,V)}\varphi = \frac{\det(D)^{n+m-2r}\Delta(D)}{\left(\sqrt{2}\right)^{r(n+m-r-1)}}.$$

donde $\det(D) = \prod_{i=1}^r \sigma_i$ es el determinante y $\Delta(D) = \prod_{i<j}(\sigma_i^2 - \sigma_j^2)$ es el discriminante.

El segundo resultado fundamental de [**Be, 09**] es el cálculo del volumen de la fibra $\varphi^{-1}(A)$ para $A \in \sum_r \setminus \sum_{r-1}$. Es el Teorema 2.5.1 de la Sección 2.5 y que reproducimos a continuación:

TEOREMA 0.2.5. *La cantidad $\mathrm{Vol}[\varphi^{-1}(A)]$ es independiente de $A \in \sum_r \setminus \sum_{r-1}$, excepto para un conjunto de medida cero. En efecto, sea $A \in \sum_r \setminus \sum_{r-1}$ tal que todos sus r valores singulares distintos de cero sean diferentes. Entonces tenemos que*

$$\mathrm{Vol}[\varphi^{-1}(A)] = 2^r \mathrm{Vol}[\mathscr{O}(n-r)]\mathrm{Vol}[\mathscr{O}(m-r)].$$

Con estos resultados ya podemos establecer la identidad final fundamental para cualquier función $\phi(A)$ que sólo dependa de los valores singulares de A (escribiremos $\phi(\sigma(A))$ y que, despejando en (0.2.2) se transforma en el siguiente resultado correspondiente al Teorema 2.6.1:

TEOREMA 0.2.6. *Sea $A \in \sum_r \setminus \sum_{r-1}$ el punto proyectivo asociado a una matriz de rango r. Sea $\sigma(A) = (\sigma_1(A), ..., \sigma_r(A)) \in S_r^+$ la lista de sus valores singulares. Sea $\phi : \mathbb{R}^r \to \mathbb{R}$ una función medible. Entonces, se tiene:*

$$\int_{A \in \sum_r \setminus \sum_{r-1}} \phi(\sigma(A))\mathrm{d}(\sum_r \setminus \sum_{r-1}) = H(n,m,r) \cdot \int_{D \in S_r^+} \phi(D)(\det(D)^{n+m-2r}\Delta(D))\mathrm{d}S_r^+,$$

donde

$$H(n,m,r) = \frac{\mathrm{Vol}[\mathscr{O}(n)]\mathrm{Vol}[\mathscr{O}(m)]}{\mathrm{Vol}[\mathscr{O}(n-r)]\mathrm{Vol}[\mathscr{O}(m-r)]\left(\sqrt{2}\right)^{r(n+m-r)+r}}.$$

Por último, aplicaremos este resultado que junto a los resultados de Chen y Dongarra [**Chen&Dongarra, 2005**] y estableceremos unas cotas para la función de distribución de κ para matrices pertenecientes a $\sum_r \setminus \sum_{r-1}$ con la distribución uniforme, esto se realiza en el Teorema 2.6.2 que se enuncia como sigue:

TEOREMA 0.2.7. *Con la distribución uniforme en $\sum_r \setminus \sum_{r-1}$ se tiene la siguiente igualdad para cada t:*

$$P_U\left[\kappa(A) > t : A \in \sum_r \setminus \sum_{r-1}\right] = P_\gamma[\kappa(A) > t : A \in GL(r \times (n+m-r), \mathbb{R})].$$

donde $GL(r \times (n+m-r), \mathbb{R})$ es el abierto Zariski en $\mathcal{M}_{r\times(n+m-r)}(\mathbb{R})$ formado por las matrices de rango máximo y γ es la distribución Gaussiana.
Más aún, se verifican las siguientes cotas

$$\frac{1}{(2\pi)^{1/2}}\left(\frac{0.245}{t}\right)^{(n+m-2r+1)} \leq P_U\left[\frac{\kappa(A)}{\frac{n+m-r}{n+m-2r+1}} > t : A \in \sum_r \setminus \sum_{r-1}\right],$$

$$P_U\left[\frac{\kappa(A)}{\frac{n+m-r}{n+m-2r+1}} > t : A \in \sum_r \setminus \sum_{r-1}\right] \leq \frac{1}{(2\pi)^{1/2}}\left(\frac{6.414}{t}\right)^{(n+m-2r+1)}.$$

0.2.3. Apéndices Finales. El trabajo presentado se finaliza con el Apéndice A que contiene parte de la terminología básica entre la usada en esta Memoria así como algunos resultados teóricos que son o bien básicos o son usados de manera transversal en las demostraciones exhibidas en la Memoria.

0.2.4. Sobre el estilo y la ortografía usados en este TFG. En algún caso precedente se ha discutido el estilo y la ortografía de las memorias presentadas como Trabajo de Fin de Grado en Matemáticas. En evitación de intervenciones innecesarias, queremos clarificar algunos aspectos relativos al estilo elegido en este texto. Se ha elegido el formato de libro (book) de la American Mathematical Society (AMS). Aunque el idioma utilizado es el español, hemos tratado de seguir lo más fielmente posible las recomendaciones del Libro de Estilo de esta asociación [1] , juntamente con las reglas de estilo recomendadas por D. E. Knuth y co-autores para la Mathematical Association of America (MAA) [2].
Específicamente, hemos tratado de seguir atentamente las siguientes dos reglas:

- *"Numbered theorems, lemmas, etc. are proper nouns and, thus, are capitalized: Theorem 2.3, Lemma 3.1, Figure 4.5"* (p. 79 del AMS Style Guide).
- *"Rule 19. Capitalize names like Theorem 1, Lemma 2, Algorithm 3, Method 4"* (en D. E. Knuth et al.).

[1] M. Letourneau, J. Wright Sharp, AMS Style Guide, Journals, October 2017, AMS, Providence, 2017
[2] D. E. Knuth, T. Larrabee, P. M. Roberts, Mathematical Writing, MAA, 1989

CAPíTULO 1

Condicionamiento Lineal: Propiedades Básicas

Índice

1.1. Introducción y Resumen del Contenido del Capítulo 1
1.1.1. Resumen de los contenidos del Capítulo 2
1.1.2. Precisión, perturbaciones y condicionamiento: una cota superior relativa 5
1.2. Números de Condicionamiento del Álgebra Lineal: nociones básicas 6
1.3. Condicionamiento, Mínimos Cuadrados y Perturbaciones: Teorema de Wedin 10
1.4. El teorema de Schmidt-Mirsky-Eckart-Young 18
1.4.1. Métricas en el espacio proyectivo y en esferas 18
1.4.2. La métrica en $\mathbb{P}_{N-1}(\mathbb{R})$ y S^{N-1} 19
1.4.3. Las Métricas en $\mathbb{P}_{N-1}(\mathbb{C})$ 21
1.4.4. El Teorema de Schmidt-Mirsky-Eckart-Young 22

1.1. Introducción y Resumen del Contenido del Capítulo

Uno de los grandes expertos mundiales en estabilidad del Análisis Numérico, N. Highman [**High, 02**] responde en su blog (*https://nhigham.com*) a la siguiente pregunta: *"Who Invented the Matrix Condition Number?"*. El uso del término *"ill-conditioned"* se remonta al año 1933, en un trabajo de R.R. Mallock [**Mal, 1933**]. Pero el término y los conceptos subyacentes no son desarrollados por Mallock, cuyo trabajo tiene un perfil poco matemático. J.H. Wilkinson en su libro [**Wilk, 1963**] atribuye las primeras nociones de condicionamiento en Álgebra Lineal a A. Turing: A la sazón, A. Turing investigaba, durante 1945, en un proyecto al British National Physical Laboratory (NPL), con el que pretendía construir un "automatic electronic digital computer with internal program storage". El proyecto denominado Automatic Computing Engine (ACE), nunca llegó a ser completado. Turing acabó dejando el NPL y se trasladó a la Universidad de Manchester en 1948.
J.H. Wilkinson trabajó con A. Turing en el NPL entre 1946 y 1948. El propio Wilkinson afirma, en la charla de su medalla Turing de 1970, que Turing le convirtió desde un analista clásico a un analista numérico, haciéndole abandonar la idea de retornar a Cambridge. No olvidemos la enorme influencia de J.H. Wilkinson en el desarrollo ulterior en el Álgebra Lineal Numérica (ver también [**Wilk, 1965**]). Wilkinson hace referencia al trabajo de A. Turing [**Turing, 1948**], que aunque publicado en 1948, el artículo se habría escrito con anterioridad, según Wilkinson. En ese artículo A. Turing se plantea los siguientes aspectos:
This paper contains descriptions of a number of methods for solving sets of linear simultaneous equations... but its main concern is with the theoretical limits of accuracy...

En el trabajo, Turing pretente explicar no sólo la noción de mal-condicionado, sino también experimentos de autores como H. Hotelling, que mostraban que la eliminación gaussiana y otros métodos directos conducen a errores de redondeo exponenciales. Turing pretende clasificar casos "normales", distinguiéndolos de casos "extremadamente pesimistas". Para ello, A. Turing introduce los números de condicionamiento M y N, que evolucionaron con el tiempo a los más actuales que trataremos en este Trabajo Fin de Grado.

En una visita a Princeton, en 1947, A. Turing tiene una conversación con J. von Neumann que, en [**Turing, 1948**], se resume en la siguiente frase:

In the meantime another theoretical investigation was being carried out by J.v. Neumann, who reached conclusions similar to those of this paper for the [particular] case of positive definite matrices, and communicated them to the writer [A. Turing] at Princeton in January 1947 before the proofs given here were complete.

Como el trabajo de von Neumann vió la luz más rápidamente que el de Turing, las versiones más "conformistas" de la historia suele asignar "la prioridad" a J. von Neumann y H. Goldstine (ver su monumental trabajo en [**vNG, 1947**]). Alguna que otra leyenda urbana cuenta que, dada la debilidad profesional de Turing frente al gigante matemático von Neumann, el flujo de información entre ambos no fue bidireccional, por más que Turing le de a von Neumann el desarrollo independiente de la teoría.
En cualquier caso, sin entrar en polémicas históricas, en este Trabajo Fin de Grado atribuiremos el origen de los números de condicionamiento a los desarrollos independientes y simultáneos de [**Turing, 1948**] y [**vNG, 1947**]. Más discusiones sobre la historia del origen del condicionamiento (siempre en discusión) puede seguirse en la versión extendida, escrita por L. Blum, de sus trabajos [**Blum, 04**] y [**Blum, 12**].

1.1.1. Resumen de los contenidos del Capítulo. Comenzaremos el Capítulo recordando dos de las versiones básicas del condicionamiento lineal. Supongamos $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$:

- *El condicionamiento de Turing-von Neumann-Wilkinson (entre otros):* Dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ definimos el número de condicionamiento de Turing-von Neumann-Wilkinson et al. mediante:

$$\kappa(A) := \| A \|_2 \| A^+ \|_2,$$

  donde A^+ es la pseudo-inversa de Moore-Penrose (véase la Sección A.1 del Apéndice A) y $\| \cdot \|_2$ hace referencia a la norma como operador lineal.
  Nótese que en el caso de matrices cuadradas regulares $A \in GL(n, \mathbb{K})$ este es el mismo condicionamiento que usualmente se discute en el grado.
- *El condicionamiento de Demmel-Edelman-Smale y otros:* Para una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ definimos el número de condicionamiento mediante:

$$\mu(A) := \| A \|_F \| A^+ \|_2,$$

  donde $\| \cdot \|_F$ es la norma de Frobenius de la matriz (véase el Lema 1.2.1), mientras que $\| \cdot \|_2$ y A^+ son las introducidas en el item anterior.

Aunque hay otras posibles variantes, esencialmente equivalentes, en este Trabajo de Fin de grado nos limitaremos a trabajar con estas dos que aparecen recogidas en la Definición 3 del Capítulo.
Ambas nociones están relacionadas con los valores singulares de la matriz (véase Proposición 1.2.8 y la Sección A.3 del Apéndice A para un sucinto recordatorio de la SVD). De hecho, la Descomposición en Valores Singulares subyace a buena parte de esta memoria y es el "leitmotiv" de su desenlace. La justificación clásica del uso de los números de condicionamiento para matrices regulares $A \in GL(n, \mathbb{K})$ suele ser el control del error relativo causado por la perturbación de un sistema de ecuaciones lineales que se pretende resolver (véase Proposición 1.1.1 por ejemplo). En el caso de matrices que no son cuadradas o no tienen un rango máximo, el número de condicionamiento, tal y como lo hemos definido, tiene su justificación en la resolución del problema de mínimos cuadrados (o proyeccion ortogonal). El resultado fundamental, mucho menos conocido en este contexto, es el Teorema de Wedin (cf. [**Wed, 1973**]) que recogemos como Teorema 1.3.8:

TEOREMA 1.1.1. *Sea $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ una matriz de rango máximo con $n \geq m$. Sea $b \in \mathbb{K}^n$ un vector. Sea $x \in \mathbb{K}^m$ un punto que resuelve el problema de mínimos cuadrados con datos (A, b), es decir,*

$$\| Ax - b \|_2 = \min\{\| Az - b \|_2 : z \in \mathbb{K}^m\}.$$

Denotamos por $r := b - Ax \in \mathbb{K}^n$*, el residuo asociado a la solución* x *dada.*
Consideramos una serie de perturbaciones de los datos $\delta A \in \mathcal{M}_{n\times m}(\mathbb{K})$*,* $\delta b \in \mathbb{K}^n$ *y sea* $\epsilon > 0$ *un número real positivo. Supongamos que las perturbaciones satisfacen las siguientes propiedades:*

- *El número real* ϵ *acota los errores relativos, esto es, se tiene:*

$$\frac{\| \delta A \|_2}{\| A \|_2} \leq \epsilon \ , \ \frac{\| \delta b \|_2}{\| b \|_2} \leq \epsilon$$

- *La matriz* $A + \delta A \in \mathcal{M}_{n\times m}(\mathbb{K})$ *es también una matriz de rango máximo.*

Sea $y \in \mathbb{K}^m$ *una solución del problema de mínimos cuadrados con datos (perturbados)* $(A + \delta A, b + \delta b)$*, es decir, supongamos*

$$\|(A + \delta A)y - (b + \delta b)\|_2 = \min\{\|(A + \delta A)z - (b + \delta b)\|_2 : z \in \mathbb{K}^m\},$$

y denotemos por $\delta x = y - x$ *y por* $s = (b + \delta b) - (A + \delta A)y \in \mathbb{K}^m$ *el residuo asociado a estos datos perturbados y a la solución* $y \in \mathbb{K}^m$*.*
Entonces, si $\kappa(A) \cdot \epsilon < 1$*, se tienen las siguientes desigualdades:*

$$\frac{\| \delta x \|_2}{\| x \|_2} = \frac{\| y - x \|_2}{\| x \|_2} \leq \frac{\epsilon \cdot \kappa(A)}{1 - \epsilon \cdot \kappa(A)} \left(2 + (\kappa(A) + 1) \frac{\| r \|_2}{\| A \|_2 \| x \|_2} \right),$$

$$\frac{\| \delta r \|_2}{\| b \|_2} = \frac{\| r - s \|_2}{\| b \|_2} \leq (1 + 2\kappa(A)) \cdot \epsilon,$$

donde hemos usado $\delta r = s - r$*.*

Lo que aporta esta memoria es una demostración con los detalles esenciales requeridos para su prueba. Algunos, por no salirnos del espacio máximo, han ido al apéndice.

La siguiente sección del Capítulo se dedica a dar una prueba completa y autocontenida del Teorema de Schmidt-Mirsky-Eckart-Young (ver [**Schm, 1967**], [**Mrs, 1960**], [**E&Y, 1936**]). Una observación natural de los números de condicionamiento introducidos anteriormente es que ambos son funciones homogéneas de grado 0, esto es,

$$\forall A \in \mathcal{M}_{n\times m}(\mathbb{K}), \forall \lambda \in \mathbb{K} \setminus \{0\}$$
$$\kappa(\lambda A) = \kappa(A)$$
$$\mu(\lambda A) = \mu(A).$$

Esto significa que son, por naturaleza, funciones definidas en las correspondientes esferas unidad $S^{nm-1}(\mathbb{K})$ donde

- $S^{nm-1}(\mathbb{R}) = S^{nm-1}$ es la esfera real usual.
- $S^{nm-1}(\mathbb{C}) = S^{2(nm)-1}$ es la esfera unidad compleja (en $\mathbb{C}^{nm}$).

Tiene sentido, por tanto, interpretar ambos condicionamientos como funciones de las esferas del espacio de matrices

$$\kappa, \mu : S^{nm-1}(\mathbb{K}) \longrightarrow \mathbb{R}_+.$$

De hecho, son funciones naturalmente definidas en los espacios proyectivos respectivos $\mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{R})) = \mathbb{P}_{nm-1}(\mathbb{R})$ o $\mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{C})) = \mathbb{P}_{nm-1}(\mathbb{C})$.
La sorpresa surge de un clásico resultado histórico, atribuido a varios autores, y que aquí presentamos como Teorema de Schmidt-Mirsky-Eckart-Young (cf. [**Schm, 1967**], [**Mrs, 1960**], [**E&Y, 1936**]). Para entender el resultado, consideremos la esfera $S^{n-1} \subseteq \mathbb{R}^n$. La esfera S^{n-1} es una subvariedad de Riemann de dimensión $n-1$ con una natural función de distancia geodésica basada en el ángulo y que denominamos *distancia riemanniana en la esfera*. Dicha función distancia viene dada por:

$$\begin{array}{rccl} d_R\colon & S^{n-1} \times S^{n-1} & \longrightarrow & \mathbb{R}_+ \\ & (u, v) & \longmapsto & \arccos\left(|\langle u, v\rangle|\right), \end{array}$$

donde $\langle \cdot, \cdot \rangle$ es el producto euclídeo usual en $\mathbb{R}_+$.
Una distancia suplementaria en S^{n-1} es denominada la *distancia del seno* por algunos autores y la *distancia proyectiva* por otros y viene dada por la siguiente identidad:

$$\begin{array}{rccl} d_{\mathbb{P}}\colon & S^{n-1} \times S^{n-1} & \longrightarrow & \mathbb{R}_+ \\ & (u, v) & \longmapsto & \sin(d_R(u, v)) = \sqrt{1 - \langle u, v\rangle^2}. \end{array}$$

Ambas métricas, junto con la métrica euclídea usual, son comparables como se destribe en la figura siguiente, donde se describen las métricas que acabamos de mencionar, de izquiera a derecha:

i) Métrica Euclídea: $\| u - v \|_2$
ii) Métrica Riemanniana: $d_R(u, v)$
iii) Métrica Riemanniana: $d_{\mathbb{P}}(u, v)$

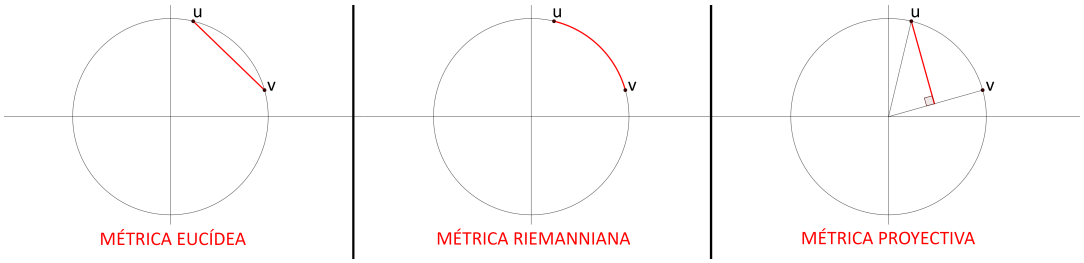

FIGURE 1. Métricas sobre el espacio proyectivo.

Las métricas de la esfera se trasladan de manera natural a los espacios proyectivos como se muestra en la Subsección 1.4.2. Para el caso complejo estas métricas se redefinen de manera natural a partir de la métrica de Fubini-Study tanto en $S^{2n-1} = \{x \in \mathbb{C}^n : \| z \|^2 = 1\}$ como en el espacio proyectivo complejo $\mathbb{P}_{n-1}(\mathbb{C})$.
Si consideramos $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$ y el espacio afín $\mathbb{A}^{nm} = \mathcal{M}_{n\times m}(\mathbb{K})$ de las matrices coordenadas en $\mathbb{K}$, las anteriores métricas se construyen de manera natural en las esferas S^{nm-1} (caso real), S^{2nm-1} (caso complejo) y los espacios proyectivos $\mathbb{P}_{nm-1}(\mathbb{R})$ (caso real) y $\mathbb{P}_{nm-1}(\mathbb{C})$ (caso complejo). Si acaso, resaltemos que la métrica euclídea (caso real) o hermítica (caso complejo) parte del producto euclídeo o hermítico de Frobenius:

$$\langle A, B \rangle_F := Tr(AB^*),$$

donde B^* es la traspuesta conjugada de B. Estas construcciones son las que se exponen en las subsecciones 1.4.1, 1.4.2, 1.4.3. Así, tenemos la distancia del seno para matrices que, a continuación, expresamos reescribiendo para espacios proyectivos:

$$\begin{array}{rccc} d_{\mathbb{P}}\colon & \mathbb{P}_{nm-1}(\mathbb{K}) \times \mathbb{P}_{nm-1}(\mathbb{K}) & \longrightarrow & \mathbb{R}_+ \\ & (A, B) & \longmapsto & \sqrt{1 - \frac{\langle A,B\rangle_F^2}{\| A \|_F^2 \| B \|_F^2}}. \end{array}$$

El Teorema de Schmidt-Mirsky-Eckart-Young clásico para matrices cuadradas de rango máximo $A \in \mathbb{P}_{n^2-1}(\mathbb{K})$ mediante:

$$\mu(A) = \frac{1}{d_{\mathbb{P}}(A, \sum)},$$

donde $\sum \subseteq \mathbb{P}_{n^2-1}(\mathbb{K})$ es la subvariedad algebraica proyectiva formada por aquellos puntos de $\mathbb{P}_{n^2-1}(\mathbb{K})$ determinados por matrices singulares. Esto es,

$$\sum := \{A \in \mathbb{P}_{n^2-1}(\mathbb{K}) : \det(A) = 0\},$$

donde $\det(\cdot)$ es el determinante de la matriz A. La relevancia más destacable de este resultado clásico es que transforma una estimación de la estabilidad, precisión (y hasta la complejidad) de algoritmos matriciales en un objeto de naturaleza puramente geométrica: el inverso de la distancia a la variedad de matrices singulares. Nótese que, en este sentido, el condicionamiento de Demmel-Edelman-Smale tiene una formulación más elegante como invariante geométrico.
El resultado clásico es fácilmente generalizable a situaciones donde las matrices no son cuadradas y ni siquiera son de rango máximo, lo que muestra lo intrínsecamente geométrico que resulta el condicionamiento.
En este Capítulo nos dedicamos a ofrecer una prueba auto-contenida completa de este resultado. Para establecerlo, consideremos el espacio afín de las matrices $\mathcal{M}_{n\times m}(\mathbb{K})$, con coordenadas en $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$. Consideremos el respectivo espacio proyectivo $\mathbb{P}_{nm-1}(\mathbb{K})$ y consideremos la estratificación por rango (véase el texto [**AGV, 1985**] o la Sección A.2 del Apéndice A, para más información).

Denotemos por $\sum_t' \subseteq \mathcal{M}_{n\times m}(\mathbb{K})$ la variedad algebraica de las matrices cuyo rango es a lo sumo t para $0 \leq t \leq \min\{n,m\}$. Es decir,

$$\sum_t' := \{A \in \mathcal{M}_{n\times m}(\mathbb{K}) : \operatorname{rank}(A) \leq t\}.$$

Consideramos la proyección canónica $\pi : \mathcal{M}_{n\times m}(\mathbb{K}) \setminus \{0\} \longrightarrow \mathbb{P}_{nm-1}(\mathbb{K})$. Obsérvese que $\sum_t'$ es una variedad algebraica afín definida mediante ecuaciones polinomiales homogéneas de grados entre $t+1$ y $\min\{n,m\}$: son los determinantes $k \times k$ aplicados a los menores de orden k de A, con $t+1 \leq k \leq \min\{n,m\}$. Por tanto, tiene sentido considerar la variedad algebraica proyectiva

$$\sum_t := \pi(\sum_t' \setminus \{0\}) \subseteq \mathbb{P}_{nm-1}(\mathbb{K}).$$

Para una matriz (o punto proyectivo) $\pi(A) \in \sum_t$ consideremos su condicionamiento (antes definido) $\mu(A) = \| A \|_F \| A^+ \|_2$. De hecho, podemos elegir $A \in \sum_t'$ con $\| A \|_F = 1$ como representante y considerar $\mu(A) = \| A^+ \|_2$.
El Teorema de Schmidt-Mirsky-Eckart-Young que probamos es el siguiente:

Teorema 1.1.2. *Con las notaciones precedentes, sea $A \in \sum_t', 1 \leq t \leq \min\{n,m\}$ y $\alpha = \pi(A) \in \mathbb{P}_{nm-1}(\mathbb{K})$. Se tiene:*

$$\mu(A) = \frac{1}{d_{\mathbb{P}}(\alpha, \sum_{t-1})} = \frac{1}{d_F(A, \sum_{t-1}')},$$

donde $d_F(A, \sum_{t-1}')$ es la distancia de cualquier representante A de α con $\| A \|_F = 1$ al cono proyectante $\sum_{t-1}'$ con respecto a la métrica de Frobenius.

La subsección 1.4.4 es la culminación de la Sección 1.4 dedicada a la prueba de este Teorema.

1.1.2. Precisión, perturbaciones y condicionamiento: una cota superior relativa. Como en el trabajo original de A. Turing, consideraremos el siguiente problema de referencia.

Problema 2. *Sea $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$, el cuerpo de los números reales o el de los complejos. Sean $A \in \mathcal{M}_n(\mathbb{K})$ una matriz y sea $b \in \mathbb{K}^n$ un vector. Resolver el sistema de ecuaciones*

$$A \cdot x = b, \tag{1.1.1}$$

donde hemos usado las notaciones por columnas y

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

El fenómeno de la pérdida de precisión puede interpretarse como la existencia de una perturbación o error tanto en la posible solución como en los datos del problema. Así, supongamos $x \in \mathbb{K}^n$ un punto y $\delta x \in \mathbb{K}^n$ una perturbación (un error). Supongamos también $\delta b \in \mathbb{K}^n$ una perturbación de b y supongamos que tenemos la siguiente relación:

$$A \cdot (x + \delta x) = b + \delta b \tag{1.1.2}$$

Definición 1 (**Norma como operador**). *Dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$; consideremos el operador lineal que define*

$$\begin{array}{rccc} A\colon & \mathbb{K}^m & \longrightarrow & \mathbb{K}^n \\ & x & \longmapsto & A \cdot x, \end{array}$$

donde $x \in \mathbb{K}^m$ es dado como vector columna. Llamaremos norma de A como operador lineal y lo denotaremos mediante $\| A \|_2$, a la cantidad

$$\| A \|_2 := \sup_{x \in \mathbb{K}^m \setminus \{0\}} \frac{\| Ax \|_2}{\| x \|_2} = \sup_{x \in S^{m-1}(\mathbb{K})} \| Ax \|_2,$$

donde las normas $\| \cdot \|_2$ son las introducidas para $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$ y $S^{m-1}(\mathbb{K})$ denota la esfera unidad (siendo $S^{m-1}(\mathbb{R}) = S^{m-1}$ la esfera unidad usual y $S^{m-1}(\mathbb{C}) = S^{2m-1}$.

Proposición 1.1.1. *Con las notaciones precedentes, asumiendo que A es una matriz no singular, x es una solución de la Ecuación (1.1.1) y supongamos $b \neq (0, ..., 0)$. Entonces, se tiene:*

$$\frac{\| \delta x \|_2}{\| x \|_2} \leq \kappa(A) \cdot \frac{\| \delta b \|_2}{\| b \|_2}$$

donde $\|\cdot\|_2$ *es la norma canónica (euclídea si* $\mathbb{K}=\mathbb{R}$*, hermítica si* $\mathbb{K}=\mathbb{C}$*),* $\kappa(A)=\|A\|_2\cdot\|A^{-1}\|_2$ *y* $\|M\|_2$ *es la norma como operador lineal para cualquier* $M\in\mathcal{M}_n(\mathbb{K})$

DEMOSTRACIÓN. Probemos la desigualdad del modo siguiente: Eliminaremos el subíndice 2 para no hacer recargada la prueba. Como $b\neq(0,...,0)$, necesariamente $x\neq(0,...,0)$ y tiene sentido considerar:

$$\frac{\|\delta x\|}{\|x\|}=\frac{\|\delta x\|\,\|b\|}{\|b\|\,\|x\|},$$

como $A\cdot x=b$, entonces $A(x+\delta x)=b+\delta b$ implica $A\delta x=\delta b$. Por ello, suponiendo que A es no singular, tenemos:

$$\frac{\|\delta x\|}{\|x\|}=\frac{\|A^{-1}\delta b\|\,\|A\cdot x\|}{\|b\|\,\|x\|}.$$

Como la norma de una matriz como operador lineal satisface $\|Mx\|\leq\|M\|\,\|x\|$ (por definición) concluiremos:

$$\frac{\|\delta x\|}{\|x\|}\leq\|A^{-1}\|\cdot\|A\|\,\frac{\|\delta b\|\,\|x\|}{\|b\|\,\|x\|}=\kappa(A)\cdot\frac{\|\delta b\|}{\|b\|}.$$

□

En conclusión, bajo las hipótesis de esta Proposición, el error relativo al calcular una solución de la Ecuación (1.1.1) está acotado por el valor relativo de la perturbación de b que se genera y por una cantidad que sólo depende de la matriz de coeficientes. Este será el origen del concepto *número de condicionamiento lineal*.

Debemos indicar que la idea de número de condicionamiento lineal como parámetro de la complejidad recibirá un impulso inesperado a mediados de los años 90 del pasado siglo con los trabajos de M. Shub y S. Smale. En estas investigaciones, que no trataremos en este Trabajo Fin de Grado, M. Shub y S. Smale generalizan el condicionamiento para problemas no lineales y mostrarán que no sólo es un parámetro importante para el control de la precisión numérica o la estabilidad. Los trabajos de estos autores demuestran que el condicionamiento permite ofrecer cotas superiores para la complejidad (números de operaciones aritméticas) en tiempo de algoritmos de resolución numérica (caso lineal y no lineal) basados en técnicas (gaussianas, en origen) de deformación homotópica (también llamadas "path following methods"). Como el objetivo de este Trabajo Fin de Grado se restringe al caso lineal, no profundizaremos en esos temas. Muchos son los trabajos y textos de referencia sobre el condicionamiento que se podrían citar. Algunos se irán introduciendo conforme avancemos este Trabajo Fin de Grado.

1.2. Números de Condicionamiento del Álgebra Lineal: nociones básicas

DEFINICIÓN 2. *Con las notaciones precedentes, designaremos por* $(\mathbb{K},\langle\cdot,\cdot\rangle_{\mathbb{K}})$ *los espacios de Hilbert de dimensión finita que se precisan a continuación*

i) Si $\mathbb{K}=\mathbb{R}$*,* $(\mathbb{K}^n,\langle\cdot,\cdot\rangle_{\mathbb{K}})=(\mathbb{R}^n,\langle\cdot,\cdot\rangle_{\mathbb{R}})$ *será el espacio de Hilbert real cuyo producto escalar es el canónico que viene dado mediante:*

$$\langle a,b\rangle_{\mathbb{R}}:=\sum_{i=1}^{n}a_i\cdot b_i\ ,\ \forall a,b\in\mathbb{R}^n,$$

donde $a=(a_1,...,a_n)$ *y* $b=(b_1,...,b_n)$*. La norma asociada, como es habitual, vendrá dada por:*

$$\|a\|_2:=(\langle a,a\rangle_{\mathbb{R}})^{1/2},\forall a\in\mathbb{R}^n.$$

ii) Si $\mathbb{K}=\mathbb{C}$*,* $(\mathbb{K}^n,\langle\cdot,\cdot\rangle_{\mathbb{K}})=(\mathbb{C}^n,\langle\cdot,\cdot\rangle_{\mathbb{C}})$ *será el espacio de Hilbert complejo de dimensión* n *cuya forma hermítica viene dada por:*

$$\langle a,b\rangle_{\mathbb{C}}:=\sum_{i=1}^{n}a_i\cdot\overline{b_i}\ ,\ \forall a,b\in\mathbb{C}^n,$$

donde $\overline{\cdot}$ *denota conjugación compleja,* $a=(a_1,...,a_n)$ *y* $b=(b_1,...,b_n)$*. De nuevo, la norma asociada viene dada por la forma usual:*

$$\|a\|_2:=(\langle a,a\rangle_{\mathbb{C}})^{1/2},\forall a\in\mathbb{C}^n.$$

Diremos simplemente que $(\mathbb{K}^n, \langle\cdot,\cdot\rangle_{\mathbb{K}})$ es la estructura métrica canónica para referirnos a ambos, especificando en los casos en los que haya que discutir aspectos específicos para $\mathbb{K} = \mathbb{R}$ o para $\mathbb{K} = \mathbb{C}$.

Consideremos ahora el espacio de matrices $\mathcal{M}_{n\times m}(\mathbb{K})$ con n filas y m columnas. Dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ denotaremos por A^* la matriz siguiente:

- Si $\mathbb{K} = \mathbb{R}$, $A^* = A^t \in \mathcal{M}_{n\times m}(\mathbb{R})$ es la matriz traspuesta de la matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$.
- Si $\mathbb{K} = \mathbb{C}$, $A^* \in \mathcal{M}_{n\times m}(\mathbb{C})$ es la matriz traspuesta conjugada de la matriz $A \in \mathcal{M}_{n\times m}(\mathbb{C})$, es decir,
  si $A = (a_{i,j})$ es la descripción de las coordenadas de la matriz A, entonces,
  $$A^* := (\overline{a_{i,j}})^t,$$
  donde $\overline{\cdot}$ denota conjugación en $\mathbb{C}$.

LEMA 1.2.1. *Con las notaciones precedentes, $(\mathcal{M}_{n\times m}(\mathbb{K}), \langle\cdot,\cdot\rangle_F)$ es un espacio de Hilbert (real o complejo según $\mathbb{K} = \mathbb{R}\vee\mathbb{C}$), donde $\langle\cdot,\cdot\rangle_F$ es el producto (escalar o hermítico, respectivamente) dado mediante:*

$$\begin{array}{rccc} \langle\cdot,\cdot\rangle_F\colon & \mathcal{M}_{n\times m}(\mathbb{K})\times\mathcal{M}_{n\times m}(\mathbb{K}) & \longrightarrow & \mathbb{K} \\ & (A,B) & \longmapsto & Tr(A\cdot B^*), \end{array}$$

donde $Tr(M)$ es la traza de la matriz $M \in \mathcal{M}_n(\mathbb{K})$. A este producto (escalar o hermítico, según el caso) se le denomina producto de Frobenius. La norma asociada $\|\cdot\|_F$ viene dada mediante

$$\| A \|_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{i,j}|^2},\ \forall A \in \mathcal{M}_{n\times m}(\mathbb{K}),$$

donde la matriz A satisface

$$A = (a_{i,j})_{\substack{1\le i\le n\\ 1\le j\le m}}.$$

DEMOSTRACIÓN. Es un sencillo ejercicio de comprobación verificar que $\langle\cdot,\cdot\rangle_{\mathbb{R}}$ es una forma bilineal simétrica no degenerada y que $\langle\cdot,\cdot\rangle_{\mathbb{C}}$ es una forma sesquilineal no degenerada. □

PROPOSICIÓN 1.2.2. *Tanto la norma como operador $(\|\cdot\|_2)$ como la norma de Frobenius $(\|\cdot\|_F)$ sobre el espacio de matrices $\mathcal{M}_{n\times m}(\mathbb{K})$ son cantidades invariantes por la acción del grupo ortogonal (en el caso $\mathbb{K} = \mathbb{R}$) y el grupo unitario (en el caso $\mathbb{K} = \mathbb{C}$).*
Es decir, se tiene:

i) Si $\mathbb{K} = \mathbb{R}$, dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y dadas dos matrices ortogonales $U \in \mathcal{O}(n)$, $V \in \mathcal{O}(m)$, se tiene

$$\| UAV^* \|_2 = \| A \|_2,$$

$$\| UAV^* \|_F = \| A \|_F.$$

ii) Si $\mathbb{K} = \mathbb{C}$, dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{C})$ y dadas dos matrices unitarias $U \in \mathcal{U}(n)$, $V \in \mathcal{U}(m)$, se tiene

$$\| UAV^* \|_2 = \| A \|_2,$$

$$\| UAV^* \|_F = \| A \|_F.$$

DEMOSTRACIÓN. Hagamos el caso real. El caso complejo es análogo.
Para la norma como operador $\|\cdot\|_2$, recordemos que si $U \in \mathcal{O}(n)$ y $V \in \mathcal{O}(m)$, estas matrices definen, respectivamente, isometrías sobre $\mathbb{R}^n$ y $\mathbb{R}^m$ para el producto canónico. Es decir, las citadas matrices definen aplicaciones lineales:

$$U\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

$$V^*\colon \mathbb{R}^m \longrightarrow \mathbb{R}^m,$$

de tal modo que $\forall x, y \in \mathbb{R}^n$ y $\forall u, v \in \mathbb{R}^m$ se tiene

$$\langle Ux, Uy\rangle = \langle x, y\rangle,$$

$$\langle V^*u, V^*v\rangle = \langle u, v\rangle.$$

Por tanto, V^* deja invariante la esfera unidad $S^{m-1} \subseteq \mathbb{R}^m$, y, para cualquier matriz $M \in \mathcal{M}_{n\times m}(\mathbb{R})$ se tendrá:

$$\sup_{x\in S^{m-1}} \| M \cdot x \|_2 = \sup_{x\in S^{m-1}} \| M \cdot (V^* x) \|_2 .$$

De otro lado, si $M = U \cdot A$, tendremos que

$$\| U \cdot A \cdot y \|_2^2 = \langle U(Ay), U(Ay)\rangle = \langle Ay, Ay\rangle.$$

Por tanto,

$$\| U \cdot A \cdot V^* \|_2 = \sup_{x\in S^{m-1}} \|(UAV^*)x \|_2 = \sup_{x\in S^{m-1}} \|(AV^*)x \|_2 = \sup_{x\in S^{m-1}} \| A(V^*x) \|_2 = \sup_{x\in S^{m-1}} \| Ax \|_2 = \| A \|_2 .$$

Para la norma de Frobenius, la prueba es aún más simple. Supongamos $A \in \mathcal{M}_{n\times m}(\mathbb{R})$, $U \in \mathscr{O}(n)$ y $V \in \mathscr{O}(m)$. Tenemos que:

$$(UAV^*) \cdot (UAV^*)^* = (UAV^*) \cdot (VA^*U^*) = UAA^*U^*.$$

Ahora, como $U \in \mathscr{O}(n)$, entonces $U^* = U^{-1}$. Y como la traza es un invariante de las clases de semejanza (es el segundo coeficiente del polinomio característico de una matriz), tendremos

$$\| UAV^* \|_F = (Tr((UAV^*)(UAV^*)^*))^{1/2} = \left(Tr(U(AA^*)U^{-1})\right)^{1/2} = Tr(AA^*)^{1/2} = \| A \|_F .$$

□

Proposición 1.2.3. *Con las notaciones precedentes, dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$, se tiene*

$$\| A \|_2 \leq \| A \|_F \leq \sqrt{r}\, \| A \|_2,$$

donde $r = \operatorname{rank}(A)$ es el rango de la matriz A.

Demostración. Usaremos la existencia de descomposición en valores singulares de matrices con coordenadas reales o complejas (ver Sección A.3 del Apéndice A).
Supongamos $\mathbb{K} = \mathbb{C}$ y la prueba será idéntica en el caso $\mathbb{K} = \mathbb{R}$.
Por la existencia de descomposición en valores singulares, dada $A \in \mathcal{M}_{n\times m}(\mathbb{C})$, existen matrices unitarias $U \in \mathscr{U}(n)$ y $V \in \mathscr{U}(m)$ y números reales $\sigma_1 \geq ... \geq \sigma_r > 0$ tales que se da la igualdad n

$$UAV^* = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right),$$

donde $r = \operatorname{rank}(A)$, $\sigma_1 = \| A \|_2$. Por la Proposición precedente, la norma de Frobenius es invariante por la acción del grupo unitario en lo que

$$\| A \|_F = \| UAV^* \|_F = \| \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) \|_F = (\sigma_1^2 + ... + \sigma_r^2)^{1/2}.$$

Por tanto, es claro que $\| A \|_2 = \sigma_1 \leq \| A \|_F$. Y, dado que $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$, es también inmediato que

$$\| A \|_F = (\sigma_1^2 + ... + \sigma_r^2)^{1/2} \leq (r \cdot \sigma_1^2)^{1/2} = \sqrt{r} \cdot \| A \|_2 .$$

□

Nótese que, en realidad, hemos probado el siguiente corolario.

Corolario 1.2.4. *Con las notaciones precedentes, si $\sigma_1 \geq ... \geq \sigma_r > 0$ son los valores singulares de una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ se tiene*

$$\| A \|_2 = \sigma_1 \ , \ \| A \|_F = (\sigma_1^2 + ... + \sigma_r^2)^{1/2},$$

Adicionalmente, tenemos el siguiente corolario,

COROLARIO 1.2.5. *La norma de Frobenius es consistente con la norma canónica tanto en el caso real como en el complejo. Es decir, si* $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$ *y* $A \in \mathcal{M}_{n\times m}(\mathbb{K})$*, se tiene*

$$\| Ax \|_2 \leq \| A \|_2 \| x \|_2 \ , \ \| Ax \|_2 \leq \| A \|_F \| x \|_2 \ , \forall x \in \mathbb{K}^n.$$

DEMOSTRACIÓN. Si bien la primera desigualdad ($\| Ax \|_2 \leq \| A \|_2 \| x \|_2$) es obvia por la propia definición de la norma como operador lineal, la segunda se sigue de la Proposición precedente:

$$\| Ax \| \leq \| A \|_2 \| x \| \leq \| A \|_F \| x \| \ , \forall x \in \mathbb{K}^n.$$

□

Ya estamos en condiciones de definir los números de condicionamiento usuales del Álgebra Lineal Numérica.

DEFINICIÓN 3. *Sea* $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$*. Dada una matriz* $A \in \mathcal{M}_{n\times m}(\mathbb{K})$*, se definen:*

i) El condicionamiento clásico de A, que denotaremos por $\kappa(A)$*:*

$$\kappa(A) := \| A \|_2 \cdot \| A^+ \|_2,$$

donde $\| \cdot \|_2$ *es la norma como operador lineal y* A^+ *es la pseudo-inversa de Moore-Penrose de A (ver Sección A.1 del Apéndice A donde se describe resumidamente la pseudo-inversa).*

ii) El condicionamiento de Demmel (también llamado de Smale en la literatura):

$$\mu(A) := \| A \|_F \cdot \| A^+ \|_2 .$$

El número de condicionamiento μ es usualmente asignado a J. Demmel por su trabajo [**Dem, 87**] y algunos autores lo denotan mediante $\kappa_D(A)$.

PROPOSICIÓN 1.2.6. *Si* $A \in GL(n, \mathbb{K})$ *es una matriz regular, los números de condicionamiento precedentes son los usuales para matrices regulares:*

$$\kappa(A) := \| A \|_2 \cdot \| A^{-1} \|_2,$$
$$\mu(A) := \| A \|_F \cdot \| A^{-1} \|_2 .$$

DEMOSTRACIÓN. Es inmediato porque $A^+ = A^{-1}$ cuando $A \in GL(n, \mathbb{K})$ (cf. Sección A.1 del Apéndice A). □

PROPOSICIÓN 1.2.7. *En el caso de matrices regulares* $A \in GL(n, \mathbb{K})$*, ambos números de condicionamiento sirven para acotar los errores relativos con respecto a las métricas canónicas en* $\mathbb{K}^n$ *y* $\mathbb{K}^m$ *respectivamente. Es decir, dada* $A \in GL(n, \mathbb{K}) \subseteq \mathcal{M}_n(\mathbb{K})$ *una matriz regular y dado* $b \in \mathbb{K}^n$*, supongamos* $x \in \mathbb{R}^n$ *una solución de la ecuación lineal:*

$$Ax = b.$$

Sean $\delta x, \delta b \in \mathbb{K}^n$ *dos vectores tales que* $A \cdot (x + \delta x) = b + \delta b$*. Entonces,*

$$\frac{\| \delta x \|}{\| x \|} \leq \kappa(A) \cdot \frac{\| \delta b \|}{\| b \|} \leq \mu(A) \cdot \frac{\| \delta b \|}{\| b \|}.$$

DEMOSTRACIÓN. Basta con recuperar los argumentos descritos en las motivaciones (ver Sección 1.1 para $\kappa(A)$). Como, por la Proposición 1.2.3 precedente, $\| A \|_2 \leq \| A \|_F$, concluimos también la segunda desigualdad. □

Ambos números de condicionamiento tienen una expresión natural en términos de los valores singulares de una matriz.

PROPOSICIÓN 1.2.8. *Con las notaciones precedentes, sea* $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ *una matriz y supongamos que* $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$ *son sus valores singulares con* $r = \operatorname{rank}(A)$*.*
Entonces, los números de condicionamiento de A introducidos satisfacen:

$$\kappa(A) = \frac{\sigma_1}{\sigma_r},$$
$$\mu(A) = \frac{(\sigma_1^2 + ... + \sigma_r^2)^{1/2}}{\sigma_r}.$$

DEMOSTRACIÓN. Es obvio por lo discutido en resultados precedentes, lo descrito en las Secciones A.3 y A.1 del Apéndice A, y las definiciones. □

Recordemos que si $\mathbb{K}$ es un cuerpo, el espacio proyectivo N-dimensional sobre $\mathbb{K}$ se define del modo siguiente:
Consideremos $A = \mathbb{K}^{N+1} \setminus \{0\}$ el conjunto de los elementos no nulos del $\mathbb{K}$-espacio vertorial $\mathbb{K}^{N+1}$. Sobre el conjunto A definimos la siguiente relación de equivalencia

$$x \sim y \iff \exists \lambda \in \mathbb{K} \setminus \{0\}, \lambda x = y.$$

Al conjunto cociente $A/\sim$ por esa relación de equivalencia se le donomina espacio proyectivo N-dimensional sobre $\mathbb{K}$ y lo denotaremos mediante $\mathbb{P}_N(\mathbb{K})$. Es obvio que con la construcción precedente, $\mathbb{P}_N(\mathbb{K})$ son los subespacios vectoriales de $\mathbb{K}^{N+1}$ de dimensión uno. Otra manera de interpretarlo es ver $\mathbb{K}^{N+1}$ como el espacio de las direcciones de las rectas (sobre $\mathbb{K}$) del espacio afin de dimension $N+1$.

PROPOSICIÓN 1.2.9. *Ambos números de condicionamiento son funciones naturalmente definidas en el espacio proyectivo:*

$$\kappa, \mu : \mathbb{P}_{nm-1}(\mathbb{K}) \to \mathbb{R}_+ \cup \{+\infty\},$$

donde $\mathbb{P}_{nm-1}(\mathbb{K})$ es el espacio proyectivo (de dimensión $n \cdot m - 1$ sobre $\mathbb{K}$) asociado al espacio de matrices $\mathcal{M}_{n\times m}(\mathbb{K})$. Es decir, dada $A \in \mathcal{M}_{n\times m}(\mathbb{K}) \setminus \{0\}$ y dado $\lambda \in \mathbb{K} \setminus \{0\}$, se tiene:

$$\kappa(A) = \kappa(\lambda \cdot A)\ ,\ \mu(A) = \mu(\lambda \cdot A). \tag{1.2.1}$$

La demostración es inmediata a partir de las definiciones de las normas involucradas.

Algunos autores prefieren utilizar la expresión *homogénea de grado 0* para referirse a las funciones que satisfacen la Identidad (1.2.1). Nosotros hemos preferido ver el condicionamiento de modo más geométrico, poniendo el énfasis en su condición de función proyectiva. Por eso podemos considerar ambos condicionamientos como funciones definidas en el espacio proyectivo $\mathbb{P}_{nm-1}(\mathbb{K}) = \mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{K}))$:

$$\begin{array}{rccc} \kappa\colon & \mathbb{P}_{nm-1}(\mathbb{K}) = \mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{K})) & \longrightarrow & \mathbb{R}_+ \\ & < A > & \longmapsto & \kappa(A), \end{array}$$

$$\begin{array}{rccc} \mu\colon & \mathbb{P}_{nm-1}(\mathbb{K}) = \mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{K})) & \longrightarrow & \mathbb{R}_+ \\ & < A > & \longmapsto & \mu(A), \end{array}$$

donde $< A >$ representa la clase de equivalencia (punto proyectivo) definido por el elemento $A \in \mathcal{M}_{n\times m}(\mathbb{K}) \setminus \{0\}$. El teorema de Eckart-Young, que especificaremos y probaremos en la sección siguiente, explicará que el condicionamiento de Demmel está fuertemente relacionado con propiedades métricas de espacios proyectivos.

1.3. Condicionamiento, Mínimos Cuadrados y Perturbaciones: Teorema de Wedin

Ya hemos indicado anteriormente que el condicionamiento está ligado al control de errores en el problema de resolución de ecuaciones lineales en el caso de matriz de coeficientes regular. En el caso en que una matriz no sea de rango máximo, el condicionamiento permite controlar el error relativo para el problema de mínimos cuadrados.

PROBLEMA 3 (**Problema de Mínimos Cuadrados Lineal**). *Sea $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$, $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ una matriz y $b \in \mathbb{K}^n$ un vector. Se define el problema de mínimos cuadrados con entrada (A, b) como el problema que consiste en:*
Hallar $x \in \mathbb{K}^m$ tal que el vector $b' = Ax$ minimiza la distancia entre b y la imagen $Im(A) \subseteq \mathbb{K}^n$ de la aplicación lineal definida por A, esto es, hallar $x \in \mathbb{K}^m$ tal que

$$\| Ax - b \|_2 = \min\{\| Az - b \|_2 : z \in \mathbb{K}^m\}.$$

La pseudo-inversa de Moore-Penrose responde a esta pregunta.

TEOREMA 1.3.1. *Dados $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y $b \in \mathbb{K}^n$, una solución del problema de mínimos cuadrados con datos (A, b) es el vector*

$$z = A^+ b \in \ker(A)^\perp \subseteq \mathbb{K}^m,$$

donde A^+ es la pseudo-inversa de Moore-Penrose y $\ker(A)^\perp$ es el complemento ortogonal del núcleo de la aplicación lineal definida por la matriz A.

DEMOSTRACIÓN. Basta considerar nuestra construcción de la pseudo-inversa de Moore-Penrose. Denotemos por ϑ la imagen de $A : \mathbb{K}^m \to \mathbb{K}^n$ (i.e. $\vartheta = Im(A)$). Consideremos la proyección ortogonal $\pi_\vartheta : \vartheta \cong \mathbb{K}^m \to \mathbb{K}^n$ y sea $b' = \pi_\vartheta(b)$ (cf. ver Sección A.1 del Apéndice A para mas detalles). Por lo visto en la Sección A.1 del Apéndice A, la proyección ortogonal $\pi_\vartheta(b)$ minimiza la distancia entre b y ϑ. Es decir,

$$\| \pi_\vartheta(b) - b \|_2 = \| b' - b \|_2 = \min\{\| \omega - b \|_2 : \omega \in \vartheta\} = \min\{\| Az - b \|_2 : z \in \mathbb{K}^m\}.$$

De otro lado, tenemos la restricción de A al complemento ortogonal de su núcleo, i.e.

$$A_{|_{\ker(A)^\perp}} : \ker(A)^\perp \to \vartheta = Im(A).$$

Esta apliación lineal es un isomorfismo, posee inversa y podemos considerar

$$z = \left(A_{|_{\ker(A)^\perp}}\right)^{-1} \circ \pi_\vartheta(b) = \left(A_{|_{\ker(A)^\perp}}\right)^{-1} b'.$$

Claramente se tiene, $Az = b'$ y, por tanto,

$$\| Az - b \|_2 = \| b' - b \|_2 = \min\{\| A\omega - b \|_2 : \omega \in \mathbb{K}^m\}.$$

Por tanto z es una solución al problema de mínimos cuadrados con datos (A, b). Además, como se indica en la Sección A.1 del Apéndice A, la pseudo-inversa de Moore-Penrose viene dada por la siguiente identidad:

$$A^+ = \left(A_{|_{\ker(A)^\perp}}\right)^{-1} \circ \pi_{Im(A)},$$

donde $\ker(A)^\perp$ es el complemento ortogonal del núcleo de A y $\pi_{Im(A)} = \pi_\vartheta$ es la proyección ortogonal sobre $Im(A)$ antes citada. Esto concluye la prueba del Teorema. □

Obsérvese que la solución al problema de mínimos cuadrados inicialmente propuesto admite varias soluciones si $\ker(A) \neq \{0\}$, pues si $z \in \mathbb{K}^m$ es una solución al problema de mínimos cuadrados y $\omega \in \ker(A)$, entonces $z + \omega$ también es una solución de nuestro enunciado del Problema 3.
Pero la solución destacada en el anterior teorema responde a un problema más exigente:

PROBLEMA 4 (**Solución de Norma Mínima al Problema de Mínimos Cuadrados**). *Dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y un vector $b \in \mathbb{K}^n$, el problema de hallar la solución de norma mínima al Problema de Mínimos Cuadrados con datos (A, b) consiste en hallar $x \in \mathbb{K}^m$ tal que:*

i) $\| Ax - b \|_2 = \min\{\| Az - b \|_2 : z \in \mathbb{K}^m\}$.
ii) El vector x minimiza la norma de todos los elementos de $\mathbb{K}^m$ que satisfacen la identidad i. Es decir, dado $y \in \mathbb{K}^m$ tal que

$$\| Ay - b \|_2 = \min\{\| Az - b \|_2 : z \in \mathbb{K}^m\},$$

entonces $\| x \|_2 \leq \| y \|_2$.

TEOREMA 1.3.2. *Dada una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y un vector $b \in \mathbb{K}^n$, el vector $x = A^+b$ es una solución de norma mínima al problema de mínimos cuadrados con datos (A, b). Además, tal solución es única.*

DEMOSTRACIÓN. Consideremos un elemento $y \in \mathbb{K}^m$ tal que $\| Ay - b \|_2$ minimiza la distancia de b con $Im(A)$. Denotemos por $\vartheta = Im(A)$ y se tendrá que verificar que $Ay = \pi_\vartheta(b) = b'$. El punto $x = A^+b$ minimiza la misma distancia y ya hemos visto que $Ax = b'$. Por tanto, $0 = b' - b' = Ax - Ay = A(x - y)$, con lo que concluimos que $x - y \in \ker(A)$. Tenemos, además, la descomposición como suma ortogonal $\mathbb{K}^m = \ker(A) \perp \ker(A)^\perp$, y sabemos que $x \in \ker(A)^\perp$. Por tanto, $y = x + (y - x)$ es una descomposición de y como suma de un elemento $x \in \ker(A)^\perp$ y otro elemento $y - x \in \ker(A)$. Por ser suma ortogonal, concluimos

$$\| y \|^2 = \| x \|^2 + \| y - x \|^2. \tag{1.3.1}$$

Por tanto, cualquier solución que minimiza la distancia a $Im(A)$ ha de satisfacer que

$$\| y \| \geq \| x \|,$$

con lo que se minimiza la norma de las soluciones del problema de mínimos cuadrados con datos (A, b). Más aun, por la identidad (1.3.1), cualquier otra solución $y \in \mathbb{K}^m$ de norma

mínima del problema de mínimos cuadrados con datos (A, b) ha de verificar $\| y - x \|^2 = 0$ o, equivalentemente, $x = y$ lo que prueba la unicidad de solución. □

A continuación probaremos un famoso Teorema de P.A. Wedin ([**Wed, 1973**]) que involucra la presencia de perturbaciones de los datos de un problema de mínimos cuadrados con el condicionamiento de la matriz de coeficientes. El resultado es destacado en el texto de N. Higham ([**Higham, 02**]), de donde hemos tomado la demostración, y, con alguna variación, aparece también en el texto de J. Demmel ([**Demmel, 1997**]).
Antes de introducir dicho Teorema de Wedin necesitamos una serie de Teoremas relacionados con los valores propios de matrices hermíticas, para ello comenzaremos definiciendo el cociente de Rayleigh-Fritz:

DEFINICIÓN 4. *Sea $A \in \mathcal{M}_n(\mathbb{C})$ una matriz hermítica (i.e. $A = A^*$). Se define el* cociente de Rayleigh-Fritz *como la función siguiente:*

$$\begin{array}{rccc} R_A\colon & \mathbb{C}^n \setminus \{0\} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & R_A(x) := \frac{\langle Ax, x\rangle}{\| x \|^2}. \end{array}$$

LEMA 1.3.3. *La función de Rayleigh-Fritz está bien definida, es decir, para toda matriz hermítica $A \in \mathcal{M}_n(\mathbb{C})$, se verifica que $\forall x \in \mathbb{C}^n \setminus \{0\}$, $R_A(x) \in \mathbb{R}$. Esta función R_A es una función homogénea de grado cero y se puede considerar como función definida en el espacio proyectivo complejo ($\mathbb{P}_n(\mathbb{C})$) o en la esfera unidad compleja ($S^{2n-1} := \{x \in \mathbb{C}^n : \| x \|^2 = 1\}$):*

$$R_A\colon \quad \mathbb{P}_n(\mathbb{C}) \quad \longrightarrow \quad \mathbb{R},$$

o bien

$$\begin{array}{rccc} R_A\colon & S^{2n-1} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \langle Ax, x\rangle. \end{array}$$

Más aún, como $\mathbb{P}_n(\mathbb{C})$ y S^{2n-1} son espacios topológicos compactos, supremos e ínfimos son, respectivamente, máximos y mínimos.

DEMOSTRACIÓN. Veamos que efectivamente está bien definida, para ello basta con comprobar, que para toda matriz hermítica $A \in \mathcal{M}_n(\mathbb{C})$, se verifica que $\forall x \in \mathbb{C}^n \setminus \{0\}$, $\langle Ax, x\rangle \in \mathbb{R}$, para ello comenzamos obteniendo la siguiente igualdad que se tiene para matrices hermíticas:

$$\langle Ax, x\rangle = x^* Ax = (Ax)^* x = \langle x, Ax\rangle,$$

si ahora relacionamos este resultado con la simetría conjugada del conjunto escalar,

$$\langle Ax, x\rangle = \langle x, Ax\rangle = \overline{\langle Ax, x\rangle},$$

y de esta forma $\langle Ax, x\rangle \in \mathbb{R}$ y el cociente de Rayleigh-Fritz está bien definido.
Veamos que es una función homogénea de grado cero, es decir, $R_A(x) = R_A(\alpha x)$ para un $\alpha \in \mathbb{C} \setminus \{0\}$,

$$R_A(\alpha x) = \frac{\langle A(\alpha x), \alpha x\rangle}{\| \alpha x \|^2} = \frac{\langle \alpha Ax, \alpha x\rangle}{|\alpha|^2 \| x \|^2} = \frac{\alpha \cdot \overline{\alpha}\langle Ax, x\rangle}{|\alpha|^2 \| x \|^2} = \frac{|\alpha|^2\langle Ax, x\rangle}{|\alpha|^2 \| x \|^2} = R_A(x).$$

El considerar la función desde el espacio proyectivo complejo ($\mathbb{P}_n(\mathbb{C})$) o en la esfera unidad compleja (S^{2n-1}) se realiza de forma natural.

□

TEOREMA 1.3.4 (**Principio del Mini-max de Courant–Fischer**). *Sea $A \in \mathcal{M}_n(\mathbb{C})$ una matriz hermítica. Supongamos que los valores propios de A, que son números reales, están ordenados del modo siguiente:*

$$\lambda_1(A) \leq \lambda_2(A) \leq ... \leq \lambda_n(A).$$

Entonces, se verifica la siguiente igualdad:

$$\lambda_k(A) = \min \left\{ \max_{x \in S^{2n-1}} \{R_A(x) : x \in U\} : U \subseteq \mathbb{C}^n \textit{subespacio, } \dim(U) = k \right\}.$$

También se verifica la siguiente igualdad:

$$\lambda_k(A) = \max \left\{ \min_{x \in S^{2n-1}} \{R_A(x) : x \in V\} : V \subseteq \mathbb{C}^n \textit{subespacio, } \dim(V) = n - k + 1 \right\}.$$

DEMOSTRACIÓN. Como la matriz A es hermítica, esta es diagonalizable, y podemos escoger una base ortonormal de vectores propios $\{u_1, ..., u_n\}$ esto es, u_i es el valor propio asociado al valor propio λ_i y tal que $\langle u_i, u_i\rangle = 1$ y $\langle u_i, u_j\rangle = 0$ para todo $i \neq j$.
Si U es un subespacio de dimension k entonces su intersección con el subespacio $< u_k, ..., u_n >$ es distinta de cero (basta con comprobar las dimensiones) y por lo tanto existe un vector $v \neq 0$ en dicha intersección y por lo tanto podremos describir como sigue

$$v = \sum_{i=k}^{n} \alpha_i u_i \in U,$$

y cuyo cociente de Rayleigh es

$$R_A(v) = \frac{\sum_{i=k}^{n} \lambda_i \alpha_i^2}{\sum_{i=k}^{n} \alpha_i^2} \geq \lambda_k,$$

ya que $\lambda_i \geq \lambda_k$ para todo $i : k, ..., n$, y por tanto

$$\max_{x \in S^{2n-1}} \{R_A(x) : x \in U\} \geq \lambda_k.$$

Como esto es cierto para todo U, podemos concluir que

$$\min \left\{ \max_{x \in S^{2n-1}} \{R_A(x) : x \in U\} : U \subseteq \mathbb{C}^n \textit{subespacio, } \dim(U) = k \right\} \geq \lambda_k.$$

Acabamos de probar la primera desigualdad, para la otra desigualdad, tomamos el subespacio $V =< u_1, ..., u_k >$, para el cual se tiene

$$\max_{x \in S^{2n-1}} \{R_A(x) : x \in V\} \leq \lambda_k,$$

ya que λ_k es el mayor valor propio asociado a dicho subespacio. Y por tanto, como la dimensión de V es k se tiene que

$$\min \left\{ \max_{x \in S^{2n-1}} \{R_A(x) : x \in U\} : U \subseteq \mathbb{C}^n \textit{subespacio, } \textit{dim}(U) = k \right\} \leq \lambda_k.$$

En el caso en que U sea un subespacio de dimensión $n - k + 1$, procedemos de forma similar a la anterior. Consideremos el subespacio $< u_1, ..., u_k >$, de dimensión k. Su intersección con el subespacio U es distinta de cero (de nuevo basta con comprobar las dimensiones) y por lo tanto existe un vector v en esta intersección que puede escribirse mediante

$$v = \sum_{i=1}^{k} \alpha_i u_i,$$

y cuyo cociente de Rayleigh es

$$R_A(v) = \frac{\sum_{i=1}^{k} \lambda_i \alpha_i^2}{\sum_{i=1}^{k} \alpha_i^2} \leq \lambda_k,$$

y por tanto

$$\min_{x \in S^{2n-1}} \{R_A(x) : x \in U\} \leq \lambda_k.$$

Como esto es cierto para cualquier subespacio U, podemos concluir que

$$\max \left\{ \min_{x \in S^{2n-1}} \{R_A(x) : x \in V\} : V \subseteq \mathbb{C}^n \textit{subespacio, } \dim(V) = n - k + 1 \right\} \leq \lambda_k.$$

Para probar la otra desigualdad tomamos nuevamente el subespacio $V =< u_k, ..., u_n >$, para el cual se tiene

$$\min_{x \in S^{2n-1}} \{R_A(x) : x \in V\} \geq \lambda_k,$$

ya que λ_k es el menor valor propio asociado a dicho subespacio. Y por tanto, como la dimensión de V es $n - k + 1$ se tiene que

$$\max \left\{ \min_{x \in S^{2n-1}} \{R_A(x) : x \in V\} : V \subseteq \mathbb{C}^n \textit{subespacio, } \dim(V) = n - k + 1 \right\} \geq \lambda_k.$$

□

En esencia, los valores propios de una matriz hermítica son valores extremos (min-max o max-min) del cociente de Rayleigh-Fritz sobre grassmannianas.
Los resultados que vienen a continuación son desigualdades de valores propios de matrices hermíticas conocidos como resultados variacionales. El principal ingrediente es el principio Minimax de Courant y Fischer descrito anteriormente.
Dada una matriz hermítica $M \in \mathcal{M}_n(\mathbb{C})$, denotaremos mediante

$$\lambda_1(M) \leq \cdots \leq \lambda_n(M) = \| M \|_2,$$

sus valores propios.

TEOREMA 1.3.5. *Sean $A, B \in \mathcal{M}_n(\mathbb{C})$ dos matrices hermíticas, entonces,*

$$|\lambda_i(A) - \lambda_i(B)| \leq \| A - B \|_2 .$$

DEMOSTRACIÓN. Escribamos $E = B - A$ y consideremos $B = A + E$.
Usemos el Minmax Principle de Courant-Fisher y la función de Rayleigh-Fritz. Tenemos que

$$\lambda_j(B) = \lambda_j(A + E) = \min\{\max\{R_B(x) : x \in U, \| x \|_2^2 = 1\} : U \subseteq \mathbb{C}^n, \dim(U) = j\}.$$

Ahora, como $B = A + E$ tenemos que para cualquier $U \subseteq \mathbb{C}^n$, subespacio de dimensión j, y $\forall x \in U$, $\| x \|_2 = 1$,

$$R_B(x) = \langle Bx, x\rangle = \langle (A + E)x, x\rangle = \langle Ax, x\rangle + \langle Ex, x\rangle.$$

Como $\langle Ax, x\rangle, \langle Ex, x\rangle \in \mathbb{R}$, observamos que

$$R_B(x) = \langle Ax, x\rangle + \langle Ex, x\rangle \leq \langle Ax, x\rangle + \| E \|_2 \| x \|_2^2 = \langle Ax, x\rangle + \| E \|_2,$$

donde hemos usado que $\langle Ex, x\rangle \leq \| Ex \|_2 \| x \|_2 \leq \| E \|_2 \| x \|_2^2$ y que $\| x \|_2^2 = 1$.
Entonces para cualquier subespacio $U \subseteq \mathbb{C}^n$, de cualquier dimensión, se tiene:

$$\{\max\{R_B(x) : x \in U, \| x \| = 1\} \leq \{\max\{R_A(x) : x \in U, \| x \| = 1\} + \| E \|_2 .$$

Finalmente, como $\| E \|_2$ es una constante tendremos

$$\begin{array}{rcl} \lambda_j(B) & = & \min\{\max\{R_B(x) : x \in U, \| x \| = 1\} : U \subseteq \mathbb{C}^n, \dim(U) = j\} \\ & \leq & \min\{\max\{R_A(x) : x \in U, \| x \| = 1\} : U \subseteq \mathbb{C}^n, \dim(U) = j\} + \| E \|_2 \\ & = & \lambda_j(A) + \| E \|_2 . \end{array}$$

Recordando que $E = B - A$, tenemos

$$\lambda_j(B) - \lambda_j(A) \leq \| B - A \|_2 .$$

Con el mismo argumento, cambiando E por $E' = A - B$ obtendremos que

$$\lambda_j(A) - \lambda_j(B) \leq \| E' \|_2 = \| A - B \|_2 = \| B - A \|_2 .$$

Juntando las dos expresiones tendremos que

$$|\lambda_j(A) - \lambda_j(B)| \leq \| A - B \|_2,$$

y la desigualdad enunciada queda probada.

□

Sea $A \in \mathcal{M}_{n\times m}(\mathbb{C})$ una matriz compleja. Llamaremos extensión "a la Mulmuley" (ver [**Mulmuley, 1986**]) de A a la matriz hermítica

$$Ext(A) := \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \in \mathcal{M}_{n+m}(\mathbb{C}).$$

Nótese que $Ext(A)$ es una matriz hermítica. Veamos algunas propiedades básicas de $Ext(A)$.

PROPOSICIÓN 1.3.6. *Con las notaciones precedentes:*

i) $\operatorname{rank}(Ext(A)) = 2 \cdot \operatorname{rank}(A)$.

ii) Dada una descomposición en valores singulares de A

$$PAQ^* = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_t & \\ \hline & 0 & & 0 \end{array}\right),$$

con $P \in \mathscr{U}(n)$, $Q \in \mathscr{U}(m)$, tendremos

$$\begin{bmatrix} 0 & P \\ Q & 0 \end{bmatrix} \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} Q^* & 0 \\ 0 & P^* \end{bmatrix} = \begin{pmatrix} 0 & PAQ^* \\ QA^*P^* & 0 \end{pmatrix}$$

y, por tanto, salvo matrices permutación que puedan cambiar los signos, los valores propios de $Ext(A)$ están en el conjunto

$$\{\pm\sigma_1, ..., \pm\sigma_r\} \cup \{0\}.$$

iii) $Ext(A+B) = Ext(A) + Ext(B)$.

Demostración. Todas las propiedades son de mera verificación. □

Corolario 1.3.7. *Sean $A, B \in \mathcal{M}_{n\times m}(\mathbb{C})$ dos matrices complejas. Supongamos* $\operatorname{rank}(A) = \operatorname{rank}(B)$ *y sean*

$$0 \leq \sigma_1(A) \leq ... \leq \sigma_n(A),$$

$$0 \leq \sigma_1(B) \leq ... \leq \sigma_n(B),$$

los valores singulares respectivamente de A y B, ordenados de menor a mayor, eventualmente con alguno igual a cero. Entonces, $\forall i, 1 \leq i \leq n$, se tiene

$$|\sigma_i(A) - \sigma_i(B)| \leq \| A - B \|_2,$$

Demostración. Consideremos las matrices $Ext(A)$, $Ext(B)$ y $Ext(A-B)$ con la construcción precedente. Nótese que $Ext(A-B) = Ext(A) - Ext(B)$.
Dado que $\operatorname{rank}(A) = \operatorname{rank}(B)$, los valores singulares i-ésimos de A y B ocuparán la misma posición entre los valores propios de $Ext(A)$ y $Ext(B)$ respectivamente.
Además, por construcción de $Ext(A-B)$ se ha de tener

$$\| Ext(A-B) \|_2 = |\sigma_n(A-B)| = \sigma_n(A-B) = \| A - B \|_2 .$$

Por el Teorema precedente (1.3.5) se tiene que

$$|\lambda_j(Ext(A)) - \lambda_j(Ext(B))| \leq \| Ext(A) - Ext(B) \|_2 = \| A - B \|_2 .$$

Además, las posiciones j (por tener el mismo rango) significan que para cada $i, 1 \leq i \leq n$, existe un único $j(i) \in \{1, ..., 2r\}$ tal que

$$\lambda_{j(i)}(Ext(A)) = \sigma_i(A),$$

$$\lambda_{j(i)}(Ext(E)) = \sigma_i(B).$$

Concluimos así que

$$|\sigma_i(A) - \sigma_i(B)| \leq \| A - B \|_2,$$

como se pretendía. □

Teorema 1.3.8 (**Teorema de Wedin**). *Sea $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ una matriz de rango máximo con $n \geq m$. Sea $b \in \mathbb{K}^n$ un vector. Sea $x \in \mathbb{K}^m$ un punto que resuelve el problema de mínimos cuadrados con datos (A, b), es decir,*

$$\| Ax - b \|_2 = \min\{\| Az - b \|_2 : z \in \mathbb{K}^m\}.$$

Denotamos por $r := b - Ax \in \mathbb{K}^n$, el residuo asociado a la solución x dada.
Consideramos una serie de perturbaciones $\delta A \in \mathcal{M}_{n\times m}(\mathbb{K})$, $\delta b \in \mathbb{K}^n$ y sea $\epsilon > 0$ un número real positivo. Supongamos que las perturbaciones satisfacen las siguientes propiedades:

- *El número real ϵ acota las perturbaciones relativas, esto es, se tiene:*

$$\frac{\| \delta A \|_2}{\| A \|_2} \leq \epsilon \ , \ \frac{\| \delta b \|_2}{\| b \|_2} \leq \epsilon$$

- *La matriz $A + \delta A \in \mathcal{M}_{n\times m}(\mathbb{K})$ es también una matriz de rango máximo.*

Sea $y \in \mathbb{K}^m$ una solución del problema de mínimos cuadrados con inputs (perturbados) $(A + \delta A, b + \delta b)$, es decir, supongamos

$$\|(A+\delta A)y - (b+\delta b)\|_2 = \min\{\|(A+\delta A)z - (b+\delta b)\|_2 : z \in \mathbb{K}^m\},$$

y denotemos por $\delta x = y - x$ y por $s = (b+\delta b) - (A+\delta A)y \in \mathbb{K}^m$ el resto asociado a estos datos perturbados y a la solución $y \in \mathbb{K}^m$.
Entonces, si $\kappa(A) \cdot \epsilon < 1$, se tienen las siguientes desigualdades:

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\|y-x\|_2}{\|x\|_2} \leq \frac{\epsilon \cdot \kappa(A)}{1 - \epsilon \cdot \kappa(A)} \left(2 + (\kappa(A)+1)\frac{\|r\|_2}{\|A\|_2 \|x\|_2}\right),$$

$$\frac{\|\delta r\|_2}{\|b\|_2} = \frac{\|r-s\|_2}{\|b\|_2} \leq (1 + 2\kappa(A)) \cdot \epsilon,$$

donde hemos usado $\delta r = s - r$.

Antes de meternos con la demostración, introduciremos un par de lemas que necesitaremos para la misma.

LEMA 1.3.9. *Sean $A, B \in \mathcal{M}_{n\times m}(\mathbb{R})$ del mismo rango, y sea $\eta = \|A^+\|_2 \|A - B\|_2 < 1$, entonces*

$$\|B^+\|_2 \leq \frac{1}{1-\eta}\|A^+\|_2.$$

DEMOSTRACIÓN. Sea $r = \operatorname{rank}(A)$. Por el Corolario 1.3.7 precedente, tenemos

$$\sigma_r(B) \geq \sigma_r(A) - \|A - B\|_2.$$

Como $\sigma_r(B) = \|B^+\|_2^{-1}$ y $\sigma_r(A) = \|A^+\|_2^{-1}$, se tiene

$$\|B^+\|_2^{-1} \geq \|A^+\|_2^{-1} - \|A - B\|_2,$$

por hipótesis tenemos que $\eta < 1$ lo que hace que $\|A^+\|_2^{-1} - \|A - B\|_2 > 0$, de esta forma, tomando las inversas en ambos lados de la desigualdad tenemos

$$\|B^+\|_2 \geq \frac{1}{\|A^+\|_2^{-1} - \|A - B\|_2},$$

ahora, multiplicando por $\frac{\|A^+\|_2}{\|A^+\|_2}$ a la derecha de la desigualdad obtenemos

$$\|B^+\|_2 \geq \frac{\|A^+\|_2}{1 - \|A - B\|_2 \|A^+\|_2},$$

y obtenemos la desigualdad que buscábamos.

□

LEMA 1.3.10. *Sean $A, B \in \mathcal{M}_{n\times m}(\mathbb{R})$ del mismo rango, entonces*

$$\|AA^+(I - BB^+)\|_2 = \|BB^+(I - AA^+)\|_2 \leq \|A - B\|_2 \min\{\|A^+\|_2, \|B^+\|_2\}.$$

DEMOSTRACIÓN. Dado que la norma de una matriz es igual a la de su traspuesta, tenemos

$$\|BB^+(I - AA^+)\|_2 = \|(BB^+(I - AA^+))^T\|_2.$$

Veamos que dada una matriz M, se tiene que $(MM^+)^T = MM^+$, para ello a partir de una descomposición en valores singulares de M, por la Proposición A.1.4 del Apéndice A, tenemos que

$$M = U \cdot \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) V^* \ , \ M^+ = V \cdot \left(\begin{array}{ccc|c} 1/\sigma_1 & & & \\ & \ddots & & 0 \\ & & 1/\sigma_r & \\ \hline & 0 & & 0 \end{array}\right) U^*,$$

de esta forma tenemos que

$$MM^+ = U \cdot \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) \cdot \left(\begin{array}{ccc|c} 1/\sigma_1 & & & \\ & \ddots & & 0 \\ & & 1/\sigma_r & \\ \hline & 0 & & 0 \end{array}\right) \cdot U^*,$$

no es dificil ver que MM^+ es simétrica, y acabamos teniendo

$$\| BB^+(I - AA^+) \|_2 = \|(BB^+(I - AA^+))^T \|_2 = \|(I - AA^+)BB^+ \|_2$$

de esta forma sumando y restando A obtenemos

$$\| BB^+(I - AA^+) \|_2 = \|(I - AA^+)(A + (B - A))B^+ \|_2,$$

operando, vemos que $(I - AA^+)A = 0$ luego,

$$\| BB^+(I - AA^+) \|_2 = \|(I - AA^+)(B - A)B^+ \|_2,$$

y empleando que $\| I - AA^+ \|_2 = \min\{1, m - n\}$ concluimos que

$$\| BB^+(I - AA^+) \|_2 \leq \| A - B \|_2 \| B^+ \|_2 .$$

Dado que A y B son del mismo rango, por el Teorema A.1.5 del Apéndice A tenemos que

$$\| AA^+(I - BB^+) \|_2 = \| BB^+(I - AA^+) \|_2,$$

De forma que podemos repetir este mismo proceso pero esta vez partiendo de $\| AA^+(I - BB^+) \|_2$, de esta forma obtenemos que

$$\| AA^+(I - BB^+) \|_2 \leq \| A - B \|_2 \| A^+ \|_2 .$$

Juntando ambos resultados podemos concluir que

$$\| AA^+(I - BB^+) \|_2 = \| BB^+(I - AA^+) \|_2 \leq \| A - B \|_2 \min\{\| A^+ \|_2, \| B^+ \|_2\}.$$

□

Ahora sí, comenzamos la demostración del Teorema de Wedin,

Demostración. Sea $B := A + \delta A$. Siendo $r = b - Ax$, además, por ser B una matriz de rango máximo, B^+B será la matriz identidad $\in \mathcal{M}_{m\times m}(\mathbb{R})$, de esta forma tenemos:

$$y - x = B^+(b + \delta b) - x = B^+(r + Ax + \delta b) - x,$$

dado que $B = A + \delta A$, tenemos que $Ax = Bx - \delta Ax$, luego sustituyendo tenemos

$$y - x = B^+(r + Bx - \delta Ax + \delta b) - x,$$

agrupando los elementos multiplicados por x

$$y - x = B^+(r - \delta Ax + \delta b) - (I - B^+B)x,$$

y utilizando que B^+B es la matriz identidad, finalmente obtenemos

$$y - x = B^+(r - \delta Ax + \delta b). \tag{1.3.2}$$

Ahora,

$$B^+r = B^+(BB^+)r = B^+BB^+r = B^+BB^+(I - AA^+)r. \tag{1.3.3}$$

Aplicando el Lema 1.3.10 que acabamos de desarrollar, obtenemos

$$\| B^+r \|_2 \leq \| B^+ \|_2(\| B - A \|_2 \| A^+ \|_2) \| r \|_2,$$

y dado que tenemos las hipótesis $\| \delta A \|_2 \leq \epsilon \cdot \| A \|_2$ y $\epsilon \cdot \kappa(A) < 1$, es decir, $\| A^+ \|_2 \| \delta A \|_2 < 1$ podemos aplicar el Lema 1.3.9 de forma que obtenemos

$$\leq \frac{\| A^+ \|_2}{1 - \| A^+ \|_2 \| \delta A \|_2} \| \delta A \|_2 \| A^+ \|_2 \| r \|_2,$$

y finalmente, utilizando que $\| \delta A \|_2 \leq \epsilon \| A \|_2$, junto a que $\kappa(A) \cdot \epsilon < 1$, tenemos

$$\| B^+r \|_2 \leq \frac{\epsilon \cdot \kappa(A)^2}{1 - \epsilon \cdot \kappa(A)} \frac{\| r \|_2}{\| A \|_2}. \tag{1.3.4}$$

De forma similar,

$$\begin{aligned} \| B^+(-\delta Ax + \delta b) \|_2 &\leq \frac{\| A^+ \|_2}{1 - \epsilon \cdot \kappa(A)} \epsilon(\| A \|_2 \| x \|_2 + \| b \|_2), \\ &= \frac{\epsilon \cdot \kappa(A)}{1 - \epsilon \cdot \kappa(A)} \left(1 + \frac{\| b \|_2}{\| A \|_2 \| x \|_2}\right) \| x \|_2, \\ &\leq \frac{\epsilon \cdot \kappa(A)}{1 - \epsilon \cdot \kappa(A)} \left(2 + \frac{\| r \|_2}{\| A \|_2 \| x \|_2}\right) \| x \|_2 . \end{aligned} \tag{1.3.5}$$

Sustituyendo las inecuaciones 1.3.4 y 1.3.5 en 1.3.2 obtenemos

$$\| y - x \| \leq \frac{\epsilon \cdot \kappa(A)}{1 - \epsilon \cdot \kappa(A)} \left(\kappa(A) \frac{\| r \|_2}{\| A \|_2} + \left(2 + \frac{\| r \|_2}{\| A \|_2 \| x \|_2} \right) \| x \|_2 \right),$$

que finalmente dividiendo entre $\| x \|_2$ obtenemos

$$\frac{\| y - x \|_2}{\| x \|_2} \leq \frac{\epsilon \cdot \kappa(A)}{1 - \epsilon \cdot \kappa(A)} \left(2 + (\kappa(A) + 1) \frac{\| r \|_2}{\| A \|_2 \| x \|_2} \right),$$

Pasando al residuo, usando la igualdad 1.3.2 tenemos que

$$\begin{aligned} s - r &= \delta b + B(x - y) - \delta A x \\ &= \delta b - BB^+(r - \delta A x + \delta b) - \delta A x \\ &= (I - BB^+)(\delta b - \delta A x) - BB^+ r. \end{aligned}$$

Como $\| I - BB^+ \|_2 = \min\{1, m - n\}$,

$$\| r - s \|_2 \leq \| \delta b \|_2 + \| \delta A x \|_2 + \| BB^+ r \|_2 \leq \epsilon(\| b \|_2 + \| A \|_2 \| x \|_2) + \| BB^+ r \|_2 . \tag{1.3.6}$$

Usando 1.3.3, el Lema 1.3.10, y $\| BB^+ \|_2 = 1$, obtenemos

$$\| BB^+ r \|_2 = \| BB^+ P_B (I - P_A) r \|_2 \leq \| \delta A \|_2 \| A^+ \|_2 \| r \|_2 \leq \kappa(A) \epsilon \| r \|_2 .$$

Por lo tanto, introduciendo este resultado en la desigualdad 1.3.6 obtenemos finalmente la cota sobre los residuos

$$\frac{\| r - s \|_2}{\| b \|_2} \leq \epsilon \left(1 + \frac{\| A \|_2 \| x \|_2}{\| b \|_2} + \kappa(A) \frac{\| r \|_2}{\| b \|_2} \right) \leq \epsilon(1 + 2\kappa(A)).$$

□

1.4. El teorema de Schmidt-Mirsky-Eckart-Young

En esta sección expandiremos un resultado varias veces redescubierto en la literatura del Álgebra Lineal. Al menos podemos citar a E. Schmidt ([**Schm, 1967**]), L Mirsky ([**Mrs, 1960**]) y G.Eckart con G. Young ([**E&Y, 1936**]). En esencia se trata de la "geometrización" de los números de condicionamiento: se trata, esencialmente, de invariantes métricos (inverso de la distancia a los problemas mal condicionados) en espacios proyectivos, usaremos el material precedente y la Sección A.2 del Apéndice A.

1.4.1. Métricas en el espacio proyectivo y en esferas.

Para comenzar comencemos recordando las esferas unitarias en los espacios de matrices en los casos real y complejo. Supongamos así $n, m \in \mathbb{N}$ con $n \geq m$ y sea $N := n \cdot m$. Tendremos

$$S^{N-1} := \{A \in \mathcal{M}_{n \times m}(\mathbb{R}) : \| A \|_F^2 = 1\},$$

$$S^1(\mathbb{C}^N) := \{A \in \mathcal{M}_{n \times m}(\mathbb{C}) : \| A \|_F^2 = 1\}.$$

Nótese la identificación natural entre $S^1(\mathbb{C}^N) = S^{2N-1}$, por lo que usaremos S^{2N-1} para referirnos al caso complejo. Las esferas unitarias nos permiten definir también unas proyecciones naturales que determinan los espacios proyectivos:

- En el caso real ($\mathbb{K} = \mathbb{R}$), denotemos por $\pi_1 : S^{N-1} \to \mathbb{P}_{N-1}(\mathbb{R})$, la proyección natural

$$\begin{array}{rccc} \pi_1 : & S^{N-1} & \longrightarrow & \mathbb{P}_{N-1}(\mathbb{R}) = \mathbb{P}(\mathcal{M}_{n \times m}(\mathbb{R})) \\ & A & \longmapsto & < A >, \end{array}$$

  donde $< A >$ es el $\mathbb{R}$-espacio vectorial de dimensión 1 generado por la matriz A. Nótese que la fibra sobre cada subespacio $r \in \mathbb{P}_{N-1}(\mathbb{R})$ es justamente la intersección de r con la esfera S^{N-1}, es decir, dado $r \in \mathbb{P}_{N-1}(\mathbb{R})$

$$\pi_1^{-1}(\{r\}) = r \cap S^{N-1} = \{A, -A\},$$

- En el caso complejo ($\mathbb{K} = \mathbb{C}$), denotemos también por π_1, la proyección natural sobre $S^1(\mathbb{C}^N)$

$$\begin{array}{rccc} \pi_1: & S^1(\mathbb{C}^N) & \longrightarrow & \mathbb{P}_{N-1}(\mathbb{C}) = \mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{C})) \\ & A & \longmapsto & < A >, \end{array}$$

donde $< A >$ es el $\mathbb{C}$-espacio vectorial de generado por la matriz A en $\mathcal{M}_{n\times m}(\mathbb{C})$. De nuevo, π_1 es suprayectiva y la fibra se determina de modo siguiente:
Dado $r \in \mathbb{P}_{N-1}(\mathbb{C})$, la fibra $\pi_1^{-1}(\{r\})$ de π_1 sobre ese punto es dada por la igualdad siguiente:

$$\pi_1^{-1}(\{r\}) =< A > \cap S^1(\mathbb{C}^N) = \{e^{i\theta} \cdot A : \theta \in \mathbb{R}\} \cong S^1(\mathbb{C}) = S^1.$$

Utilizando las topologías cociente, tenemos las siguientes identificaciones usuales, que le dan al proyectivo una naturaleza de espacio topológico compacto:

$$\mathbb{P}_{N-1}(\mathbb{R}) = \mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{R})) \cong S^{N-1}/\mathbb{Z}_2 \ , \ \mathbb{P}_{N-1}(\mathbb{C}) = \mathbb{P}(\mathcal{M}_{n\times m}(\mathbb{C})) \cong S^{2N-1}/S^1$$

Adicionalmente, los espacios proyectivos tienen una naturaleza métrica canónica, que vamos a definir tanto para el caso real (métrica Riemanniana) como para el caso complejo (métrica de Fubini-Study).

1.4.2. La métrica en $\mathbb{P}_{N-1}(\mathbb{R})$ y S^{N-1}. Consideremos el caso real y consideremos los subespacios vectoriales de dimensión 1 de $\mathbb{R}^N$, es decir, los puntos $v_1, v_2 \in \mathbb{P}_{N-1}(\mathbb{R})$. Se trata de dos rectas pasando por el origen $0 \in \mathbb{R}^N$. De manera clásica se define la distancia entre esas dos rectas como el ángulo entre ellas.
$dist(r_1, r_2) = \alpha \in [0, \pi/2]$, siendo α el ángulo entre las dos rectas. Supongamos que $r_1 =< A_1 >$ y $r_2 =< A_2 >$ son respectivamente, los $\mathbb{R}$-espacios vectoriales generados por las matrices no nulas $A_1, A_2 \in \mathcal{M}_{n\times m}(\mathbb{R})$.
En $\mathcal{M}_{n\times m}(\mathbb{R})$ consideremos la estructura de espacio de Hilbert real definida por el proudcto de Frobenius

$$\langle A, B\rangle_F = Tr(A \cdot B^*).$$

Entonces, tenemos el resultado clásico

$$\langle A, B\rangle_F = \| A \|_F \| B \|_F \cos(\alpha),$$

lo que nos permite definir

Definición 5. *Se denomina distancia Riemanniana en $\mathbb{P}_{N-1}(\mathbb{R})$ a la función*

$$\begin{array}{rccc} d_R: & \mathbb{P}_{N-1}(\mathbb{R}) \times \mathbb{P}_{N-1}(\mathbb{R}) & \longrightarrow & \mathbb{R}_+ \\ & (< A_1 >, < A_2 >) & \longmapsto & \arccos\left(\frac{|\langle A_1, A_2\rangle_F|}{\| A_1 \|_F \| A_2 \|_F}\right) \in [0, \pi/2]. \end{array}$$

Proposición 1.4.1. *La métrica Riemanniana d_R es una métrica, el espacio $\mathbb{P}_{N-1}(\mathbb{R})$ es compacto con esa métrica y la distancia máxima entre dos puntos de $\mathbb{P}_{N-1}(\mathbb{R})$ es $\pi/2$.*

Es fácil observar que esa distancia Riemanniana es la especialización a clases de equivalencia de la distancia Riemanniana en la esfera (i.e. la longitud del arco esférico)

$$\begin{array}{rccc} d_S: & S^{N-1} \times S^{N-1} & \longrightarrow & \mathbb{R}_+ \\ & (A_1, A_2) & \longmapsto & \alpha = \arccos\left(\langle A_1, A_2\rangle_F\right). \end{array}$$

En nuestro estudio resulta particularmente cómoda una variación de la distancia que algunos autores denominan "distancia del seno" y nosotros hemos denominado distancia proyectiva,

Definición 6. *Dados dos puntos proyectivos $r_1 =< A_1 >, r_2 =< A_2 >\in \mathbb{P}_{N-1}(\mathbb{R})$, llamaremos distancia proyectiva entre dos puntos r_1 y r_2 a*

$$d_{\mathbb{P}}(r_1, r_2) = \sin(d_R(r_1, r_2)) = \sqrt{1 - \frac{\langle A_1, A_2\rangle_F^2}{\| A_1 \|_F^2 \| A_2 \|_F^2}}.$$

Es fácil interpretar la distancia proyectiva del modo siguiente:

- Tenemos $r_1 =< A_1 >$ y supongamos $A_1 \in S^{N-1}$.
- Consideremos la recta $r_2 \subseteq \mathbb{R}^N$ (es subespacio vectorial de dimensión 1).
- Sea $B \in r_2$ la proyección ortogonal de A_1 en r_2 (i.e. el punto que minimiza la distancia entre A_1 y r_1.

Entonces

$$d_{\mathbb{P}}(r_1, r_2) = \| A_1 - B \|_F .$$

En la figura que esta a continuación, se describen las métricas que acabamos de mencionar, de izquiera a derecha:

i) Métrica Euclídea: $\| A - B \|_F$
ii) Métrica Riemanniana: $d_R(A, B)$
iii) Métrica Riemanniana: $d_{\mathbb{P}}(A, B)$

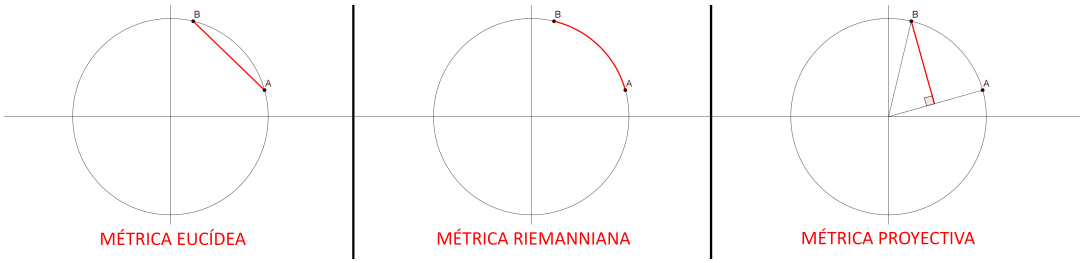

FIGURE 2. Métricas sobre el espacio proyectivo.

De hecho, tenemos

PROPOSICIÓN 1.4.2. *Dadas $A_1, A_2 \in S^{N-1}$ se tiene:*

$$d_{\mathbb{P}}(< A_1 >, < A_2 >) \leq \| A_1 - A_2 \|_F \leq d_R(< A_1 >, < A_2 >)$$

LEMA 1.4.3. *La función $d_{\mathbb{P}}$ es una métrica sobre $\mathbb{P}_{N-1}(\mathbb{R})$ que induce la topología cociente de $\mathbb{P}_{N-1}(\mathbb{R})$.*

DEMOSTRACIÓN. Veamos primero que efectivamente es una métrica,

i) $d_{\mathbb{P}}(r_1, r_2) = \sin(d_R(r_1, r_2)) \geq 0$, ya que $d_R(r_1, r_2) \in [0, \pi/2]$.
ii) $d_{\mathbb{P}}(r_1, r_2) = \sin(d_R(r_1, r_2)) = 0 \iff d_R(r_1, r_2) = 0 \iff r_1 = r_2$ por ser d_R métrica.
iii) $d_{\mathbb{P}}(r_1, r_2) = \sin(d_R(r_1, r_2)) = \sin(d_R(r_2, r_1)) = d_{\mathbb{P}}(r_2, r_1)$ por ser d_R métrica.
iv) $d_{\mathbb{P}}(r_1, r_3) \leq d_{\mathbb{P}}(r_1, r_2) + d_{\mathbb{P}}(r_2, r_3)$, como d_R es una métrica se tiene

$$d_R(r_1, r_3) \leq d_R(r_1, r_2) + d_R(r_2, r_3),$$

lo que significa que exiten $\alpha \leq d_R(r_1, r_2), \beta \leq d_R(r_2, r_3)$ no negativos tales que

$$d_R(r_1, r_3) = \alpha + \beta,$$

de esta forma tenemos que

$$\sin(d_R(r_1, r_3)) = \sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta),$$

aplicando que la función $\cos(x)$ vale como mucho uno, y que es positivo por ser $\alpha, \beta \in [0, \pi/2]$,

$$\sin(d_R(r_1, r_3)) = \sin(\alpha + \beta) \leq \sin(\alpha) + \sin(\beta),$$

y como $0 \leq \alpha \leq d_R(r_1, r_2) \leq \pi/2$, $0 \leq \beta \leq d_R(r_2, r_3) \leq \pi/2$ y la función $\sin(x)$ es creciente en $[0, \pi/2]$ concluimos que

$$\sin(d_R(r_1, r_3)) \leq \sin(d_R(r_1, r_2)) + \sin(d_R(r_2, r_3)).$$

Veamos ahora que $d_{\mathbb{P}}$ induce la topología cociente de $\mathbb{P}_{N-1}(\mathbb{R})$. Denotemos por τ_C la topología cociente y por $\tau_{d_{\mathbb{P}}}$ la topología inducida por la métrica proyectiva. Veamos que $\tau_C = \tau_{d_{\mathbb{P}}}$.
Probemos primero $\tau_{d_{\mathbb{P}}} \subseteq \tau_C$, para ello, es suficiente con ver que la siguiente aplicación es continua,

$$\begin{array}{rccc} \pi_1 : & S^{N-1} & \longrightarrow & (\mathbb{P}_{N-1}(\mathbb{R}), \tau_{d_{\mathbb{P}}}) \\ & A & \longmapsto & < A >, \end{array}$$

esto es, $\pi_1^{-1}(B(x_0, \alpha))$ es un abierto en S^{N-1} con la topología usual.
Tomando un vector x_0 de norma uno, y operando con el vector de la proyección de x sobre x_0,

$$\pi_1^{-1}(B(x_0, \alpha)) = \{x \in S^{N-1} : \| x - (x^T x_0)x_0 \|^2 < \alpha^2\} = \{x \in S^{N-1} : 1 - (x^T x_0)^2 < \alpha^2\}$$

este conjunto podemos descomponerlo de la siguiente forma,

$$\pi_1^{-1}(B(x_0,\alpha)) = \{x \in S^{N-1} : x_0^T x > \sqrt{1-\alpha^2}\} \cup \{x \in S^{N-1} : x_0^T x < -\sqrt{1-\alpha^2}\}.$$

Para ver que este conjunto es abierto definimos la siguiente aplicación,

$$\begin{array}{rccc} f\colon & S^{N-1} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & x_o^T x. \end{array}$$

Esta aplicación es continua, y se tiene la siguiente igualdad,

$$\pi_1^{-1}(B(x_0,\alpha)) = f^{-1}((\sqrt{1-\alpha^2},+\infty)) \cup f^{-1}((-\infty,-\sqrt{1-\alpha^2})),$$

y concluimos que $\pi_1^{-1}(B(x_0,\alpha))$ es abierto por ser la imagen inversa de un conjunto abierto por la aplicación continua f.

Probemos ahora que $\tau_C \subseteq \tau_{d_{\mathbb{P}}}$, para ello, tomemos $U \subseteq \mathbb{P}_{N-1}(\mathbb{R})$ tal que $\pi_1^{-1}(U)$ es abierto en S^{N-1}, es decir, $U \in \tau_C$. Sea $x \in U$ un punto de U, $\pi_1^{-1}(x) = \{z_1, z_2\} \subseteq \pi_1^{-1}(U)$, siendo $z_1 = -z_2$, de esta forma existirán dos bolas abiertas $B_1 = B(z_1, r_1), B_2 = B(z_2, r_2) \in \mathbb{R}^N$, tales que

$$B_1 \cap S^{N-1} \subseteq \pi_1^{-1}(U) \text{ , y } z_1 \in B_1,$$
$$B_2 \cap S^{N-1} \subseteq \pi_1^{-1}(U) \text{ , y } z_2 \in B_2.$$

Denotemos por $B_S = B(z,r)$ la bola que menor radio tenga entre B_1 y B_2, ahora tomando $y \in B_S \cap S^{N-1}$, con $y \neq z$, generamos la siguiente bola de $\mathbb{P}_{N-1}(\mathbb{R})$,

$$B_{\mathbb{P}} = B_{d_{\mathbb{P}}}(z,\alpha), \text{ donde } \alpha = \| y - P_{y,z} \| \text{ , siendo } P_{y,z} \text{ la proyección de } y \text{ sobre } z.$$

Veamos que $B_{\mathbb{P}} \subseteq U$, para ello, basta con probar que $\pi_1^{-1}(B_{\mathbb{P}}) \subseteq \pi_1^{-1}(U)$. Operando como en la otra contención, obtenemos la siguiente descomposición de $\pi_1^{-1}(B_{\mathbb{P}})$,

$$\pi_1^{-1}(B_{\mathbb{P}}) = \{x \in S^{N-1} : x^T z > |y^T z|\} \cup \{x \in S^{N-1} : x^T z < -|y^T z|\}.$$

Suponiendo que $B_S = B_1$, tenemos que

$$B_1 = \{x \in S^{N-1} : \| x - z \|^2 < r_1^2\} = \{x \in S^{N-1} : x^T z > |1 - r_1^2/2|\}.$$

De esta forma, esta claro que $\{x \in S^{N-1} : x^T z > |y^T z|\} \subseteq B_1$, ya que $y \in B_1$. De la misma forma tenemos que,

$$B_2 = \{x \in S^{N-1} : \| x - (-z) \|^2 < r_2^2\} = \{x \in S^{N-1} : x^T z < |r_2^2/2 - 1|\},$$

y al igual que antes, $\{x \in S^{N-1} : x^T z < -|y^T z|\} \subseteq B_2$, ya que $-y \in B_2$. Por lo tanto, $\pi_1^{-1}(B_{\mathbb{P}}) \subseteq B_1 \cup B_2 \subseteq \pi_1^{-1}(U)$, como queriamos ver. □

1.4.3. Las Métricas en $\mathbb{P}_{N-1}(\mathbb{C})$. En este caso, hablaremos de la métrica de Fubini-Study, observando que las rectas complejas (los subespacios complejos de dimensión 1), son planos desde la perspectiva real.

Definición 7. *Sean $r_1 =< A_1 >, r_2 =< A_2 > \in \mathbb{P}_{N-1}(\mathbb{C})$ dos puntos proyectivos complejos. Llamaremos distancia de Fubini-Study entre los puntos r_1, r_2 a la distancia*

$$d_{FS}(r_1, r_2) = \arccos\left(\frac{|\langle A_1, A_2\rangle_F|}{\| A_1 \|_F \| A_2 \|_F}\right).$$

Nótese que esta distancia está bien definida y no depende de los representantes A_1 y A_2 en $\mathbb{C}^N \setminus \{0\}$ elegidos. Nótese también que esta distancia coincide con el ángulo entre r_1 y r_2.

Análogamente al caso real, podemos definir la distancia proyectiva (o distancia del seno) del modo siguiente;

Definición 8. *Dados $r_1, r_2 \in \mathbb{P}_{N-1}(\mathbb{C})$, llamaremos distancia proyectiva entre r_1 y r_2 al seno de la distancia de Fubini-Study*

$$d_{\mathbb{P}}(r_1, r_2) = \sin(d_{FS}(r_1, r_2)) = +\sqrt{1 - \frac{\langle A_1, A_2\rangle_F^2}{\| A_1 \|_F^2 \| A_2 \|_F^2}},$$

donde $r_1 =< A_1 >$ y $r_2 =< A_2 >$.

Proposición 1.4.4. *Las funciones $d_{FS}, d_{\mathbb{P}} : \mathbb{P}_{N-1}(\mathbb{C}) \times \mathbb{P}_{N-1}(\mathbb{C}) \to \mathbb{R}_+$ son dos métricas topológicamente equivalentes (i.e. definen la misma topología en $\mathbb{P}_{N-1}(\mathbb{C})$ y hacen de $\mathbb{P}_{N-1}(\mathbb{C})$ un espacio métrico compacto.*

DEMOSTRACIÓN. Veamos que son dos métricas topológicamente equivalentes, es decir, existen dos constantes α, β tales que $0 < \alpha < \beta$ y para todo $r_1, r_2 \in \mathbb{P}_{N-1}(\mathbb{C})$ se cumple que

$$\alpha \cdot d_{FS}(r_1, r_2) \leq d_{\mathbb{P}}(r_1, r_2) \leq \beta \cdot d_{FS}(r_1, r_2),$$

para ello comenzaremos viendo el caso en que $d_{FS}(r_1, r_2) \neq 0$, y emplearemos las siguientes desiguialdades

$$\frac{2}{\pi} \leq \frac{\sin(x)}{x} \leq 1 \text{ , para } x \in (0, \pi/2],$$

de esta forma, se tiene

$$\frac{2}{\pi} \cdot d_{FS}(r_1, r_2) \leq \sin(d_{FS}(r_1, r_2)) \leq 1 \cdot d_{FS}(r_1, r_2),$$

o lo que es lo mismo,

$$\frac{2}{\pi} \cdot d_{FS}(r_1, r_2) \leq d_{\mathbb{P}}(r_1, r_2) \leq 1 \cdot d_{FS}(r_1, r_2).$$

Para finalizar, analizamos cuando $d_{FS}(r_1, r_2) = 0$, en este caso, $d_{\mathbb{P}}(r_1, r_2) = \sin(d_{FS}(r_1, r_2)) = \sin(0) = 0$, y por tanto también verifica

$$\frac{2}{\pi} \cdot d_{FS}(r_1, r_2) \leq d_{\mathbb{P}}(r_1, r_2) \leq 1 \cdot d_{FS}(r_1, r_2).$$

□

1.4.4. El Teorema de Schmidt-Mirsky-Eckart-Young.

Estamos ya en condiciones de abordar el enunciado y la demostración del Teorema de Schmidt-Mirsky-Eckart-Young. Comencemos con unas pocas notaciones.

De la Sección A.2 del Apéndice A, recordamos que $\sum_r' \subseteq \mathcal{M}_{n\times m}(\mathbb{K})$ es la variedad algebraica afín formada por todas las matrices de rango acotado por r. Denotaremos por $\sum_r \subseteq \mathbb{P}_{N-1}(\mathbb{K})$ a su imagen en el correspondiente espacio proyectivo.

Supondremos $n \geq m$ y para cada matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K}) \setminus \{0\}$ denotaremos por $< A >$ el punto proyectivo en $\mathbb{P}_{N-1}(\mathbb{K})$ determinado por A.

Para una matriz $B \in \mathcal{M}_{n\times m}(\mathbb{K})$ cualquiera denotaremos por $\sigma_i(B) \in \mathbb{R}$ a su i-ésimo valor singular (admitiendo $\sigma_i(B) = 0$ si $i > \text{rank}(B)$).

Haremos la descripción de los resultados para $\mathbb{K} = \mathbb{C}$. Los resultados del caso real son idénticos, reemplazando "grupo unitario" por "grupo ortogonal".

Sea $A \in \mathcal{M}_{n\times m}(\mathbb{C})$ una matriz de rango $t \leq m$ que fijaremos a lo largo de esta subsección. Sea $\alpha =< A >\in \mathbb{P}_{nm-1}(\mathbb{C})$ el punto proyectivo que define y sea $r < m$ un entero positivo, también fijado en esta subsección.

Supongamos dada una descomposición en valores singulares de A, también fija en esta subsección, de la forma siguiente:

$$A = U \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_t & \\ \hline & 0 & & 0 \end{array}\right) V^*, \tag{1.4.1}$$

donde $\sigma_1 = \sigma_1(A) \geq \sigma_2 = \sigma_2(A) \geq ... \geq \sigma_t = \sigma_t(A) > 0$ son los valores singulares (no nulos) de A y las matrices $U \in \mathscr{U}(n)$, $V \in \mathscr{U}(m)$ son unitarias.

DEFINICIÓN 9. *Con las notaciones precedentes, la matriz $A_r \in \sum_r'$ siguiente se denomina la aproximación de rango r de la matriz A:*

$$A_r = U \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & & 0 & \\ \hline & 0 & & 0 \end{array}\right) V^* \in \sum_r' .$$

El siguiente resultado prueba que el "término" aproximación por rango r de A está justificado, al menos con la norma como operador:

TEOREMA 1.4.5. *Con las notaciones precedentes, se verifica:*

$$\| A - A_r \|_2 = \sigma_{r+1} = \min\{\| A - C \|_2 : C \in \sum_r'\},$$

donde $\| \cdot \|_2$ *es la norma como operador.*

Demostración. Consideremos el subespacio $\vartheta = \mathbb{C} < \{v_1, ..., v_{r+1}\} > \subseteq \mathbb{C}^m$ generado por las $r+1$ primeras columnas de la matriz V, y sea $B \in \sum_r'$ una matriz cualquiera. Veamos que $\vartheta \cap \ker(B) \neq \{0\}$. Para ello, basta con considerar la restricción de B a ϑ

$$B_{|_\vartheta} : \vartheta \to \mathbb{K}^n.$$

Veamos que $B_{|_\vartheta}$ no puede ser un monomorfismo. Si fuera un monomorfismo, entonces el rango de la imagen de $B_{|_\vartheta}$ sería la dimensión de ϑ. Concluiríamos así

$$r + 1 = \operatorname{rank}(B_{|_\vartheta}) = \dim(\vartheta) \leq \operatorname{rank}(B) \leq r,$$

lo que nos llevaría a una contracicción. Por tanto, $B_{|_\vartheta}$ no es inyectiva y $\vartheta \cap \ker(B) \neq \{0\}$. Podemos suponer que existe $v \in \ker(B) \cap \vartheta$, con $\| v \| = 1$. Entonces,

$$\|(A - B)v\| \leq \sup\{\|(A - B) \cdot x\| : x \in \mathbb{K}^m, \|x\| = 1\} = \|A - B\|_2 .$$

De otro lado

$$\|(A - B)v\| = \|Av - Bv\| = \|Av\| = \|U \cdot Diag(\sigma_1, ..., \sigma_t, 0, ..., 0) \cdot V^* v\| .$$

Ahora bien v es una combinación lineal de las primeras $r+1$ columnas de V y la norma de $\mathbb{K}^n$ es invariante por la acción del gurpo unitario, tendremos

$$\|(A - B)v\| = \|Diag(\sigma_1, ..., \sigma_t, 0, ..., 0) \cdot (V^* v)\| = \|Diag(\sigma_1, ..., \sigma_t, 0, ..., 0) \cdot \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_{r+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}\|,$$

donde hemos observado que

$$V^* \cdot (\lambda_1 v_1 + ... + \lambda_{r+1} v_{r+1}) = \lambda_1 \cdot e_1 + ... + \lambda_{r+1} \cdot e_{r+1}, \text{ siendo}$$

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

el i-ésimo elemento de la base canónica usual de $\mathbb{K}^m$. En suma,

$$\|(A - B)v\| = \left(\sum_{i=1}^{r+1} \lambda_i^2 \sigma_i^2 \right)^{1/2},$$

sabiendo que $\sum_{i=1}^{r+1} \lambda_i^2 = 1$. Para concluir obsérvese que dados números reales positivos $a_1 \geq a_2 \geq ... \geq a_t > 0$ y la forma cuadrática que definen:

$$\begin{array}{rccc} l: & \mathbb{R}^t & \longrightarrow & \mathbb{R} \\ & (x_1, ..., x_t) & \longmapsto & x_1^2 a_1^2 + ... + x_t^2 a_t^2. \end{array}$$

El mínimo de l en la esfera unidad $S^{t-1} \subseteq \mathbb{R}^t$ se acanza en el punto $\omega_t = (0, ..., 0, 1)$ y el valor mínimo resulta ser

$$l(v_t) = a_t^2 = \min\{l(x) : x \in S^{t-1}\}.$$

En particular, $\|(A-B)v\|^2 \geq \sigma_{r+1}^2$, y, por tanto, habremos probado las desigualdades siguientes:

$$\sigma_{r+1} \leq \|A \cdot v\| \leq \|(A - B)v\| \leq \|A - B\|_2, \forall B \in \sum_r' .$$

De otro lado, es claro que el mayor valor singular de la matriz $A-A_r$ es σ_{r+1}, dado que tenemos

$$A - A_r = U \begin{pmatrix} \boxed{0} & & & \\ & \sigma_{r+1} & & \\ & & \ddots & \\ & & & \sigma_t \\ \hline & & & 0 \end{pmatrix} V^*,$$

donde $\sigma_{r+1} \geq ... \geq \sigma_t > 0$. Permutando esta identidad tendremos que los valores singulares de $A - A_r$ serán

$$\sigma_1(A - A_r) = \sigma_{r+1} \ , \ ... \ , \ \sigma_{t-r}(A - A_r) = \sigma_t.$$

Como $\| A - A_r \|_2 = \sigma_1(A - A_r)$, habremos concluido

$$\| A - A_r \| = \sigma_{r+1} \leq \| A - B \|_2, \forall B \in \textstyle\sum_r' .$$

Es decir, A_r es una de las mejores aproximaciones de A en $\sum_r'$ para $\| \cdot \|_2$ y el Teorema se sigue. □

Probemos también que A_r es la matriz de rango menor o igual que r más próxima a A también con respecto a la norma de Frobenius.

Teorema 1.4.6 (**Schmidt-Mirsky-Eckart-Young**). *Con las notaciones precedentes, se tiene:*

$$\| A - A_r \|_F = \left(\sigma_{r+1}^2 + ... + \sigma_n^2\right)^{1/2} = \min\{\| A - B \|_F : B \in \textstyle\sum_r'\},$$

Demostración. Para demostrarlo, escribamos

$$\mathcal{J} := \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_t & \\ \hline & 0 & & 0 \end{pmatrix} \in \mathcal{M}_{n\times m}(\mathbb{C})$$

e introduzcamos las matrices

$$\mathcal{J}_r := \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{pmatrix} \ , \ \mathcal{J}_{r+1} := \begin{pmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_{r+1} & & 0 \\ & & & & \ddots & \\ & & & & & \sigma_t \\ \hline & & & 0 & & 0 \end{pmatrix}.$$

ambas pertenecientes a $\mathcal{M}_{n\times m}(\mathbb{C})$.
Supongamos inicialmente que $A = \mathcal{J}$, es decir, que U y V son las respectivas matrices identidad en la Ecuación (1.4.1) anterior.
Probemos la siguiente afirmación:

Afirmación. *Con las notaciones precedentes, para cada $B \in \sum_r'$, matriz con valores singulares $\sigma_i(B) = \gamma_i, 1 \leq i \leq r$ verificando $\gamma_1 \geq ... \geq \gamma_r$.*
Entonces

$$\| \mathcal{J} - B \|_F \geq \| \mathcal{J}_{r+1} \|_F + \| \mathcal{J}_r - B \|_F^2 .$$

Demostración. Para probar el resultado, comencemos observando la siguiente igualdad

$$\| \mathcal{J} - B \|_F^2 = Tr((\mathcal{J} - B)^*(\mathcal{J} - B)) = Tr(\mathcal{J}^*\mathcal{J}) + Tr(B^*B) - Tr(\mathcal{J}^*B) - Tr(B^*\mathcal{J}). \quad (1.4.2)$$

Ahora observemos que $\mathcal{J} = \mathcal{J}_r + \mathcal{J}_{r+1}$ y que $\mathcal{J}_r \perp \mathcal{J}_{r+1}$. La primera igualdad es obvia mientras que la segunda se sigue de verificar que

$$Tr(\mathcal{J}_r^*\mathcal{J}_{r+1}) = Tr(\mathcal{J}_{r+1}^*\mathcal{J}_r) = Tr(0) = 0.$$

Por tanto

$$Tr(\mathcal{J}^*\mathcal{J}) = Tr((\mathcal{J}_r + \mathcal{J}_{r+1})^*(\mathcal{J}_r + \mathcal{J}_{r+1})) = Tr(\mathcal{J}_r^*\mathcal{J}_r) + Tr(\mathcal{J}_{r+1}^*\mathcal{J}_{r+1}).$$

De otro lado, por la desigualdad triangular para el valor absoluto en $\mathbb{C}$, si $a, b \in \mathbb{C}$ con $a+b \in \mathbb{C}$ se tendrá: $a+b \leq |a+b| \leq |a|+|b|$, con lo que $-(a+b) \geq -|a|-|b|$. Como $Tr(\mathcal{J}^*B)+Tr(B^*\mathcal{J}) \in \mathbb{R}$, concluiremos

(1.4.3)

$$\| \mathcal{J} - B \|_F^2 = Tr(\mathcal{J}_{r+1}^* \mathcal{J}_{r+1}) + \left(Tr(\mathcal{J}_r^* \mathcal{J}_r) + Tr(B^*B) - \left(Tr(\mathcal{J}^*B) + Tr(B^*\mathcal{J}) \right) \right)$$

$$\geq Tr(\mathcal{J}_{r+1}^* \mathcal{J}_{r+1}) + \left(Tr(\mathcal{J}_r^* \mathcal{J}_r) + Tr(B^*B) - |Tr(\mathcal{J}^*B)| - |Tr(B^*\mathcal{J})| \right).$$

Por la desigualdad de la Traza de von Neumann (ver Teorema A.3.4 en el Apéndice A) tenemos las siguientes desigualdades:

- $|Tr(\mathcal{J}^*B)| \leq \sum_{i=j}^m \sigma_i(\mathcal{J}) \cdot \sigma_i(B) = \sum_{i=1}^r \sigma_i \cdot \gamma_i$,
- $|Tr(B^*\mathcal{J})| \leq \sum_{i=j}^m \sigma_i(B) \cdot \sigma_i(\mathcal{J}) = \sum_{i=1}^r \sigma_i \cdot \gamma_i$,

puesto que $\sigma_i(B) = 0$ para $r+1 \leq i \leq m$. Por tanto, las desigualdades (1.4.3) se transforman en

$$\| \mathcal{J} - B \|_F^2 \geq Tr(\mathcal{J}_{r+1}^* \mathcal{J}_{r+1}) + \left(Tr(\mathcal{J}_r^* \mathcal{J}_r) + Tr(B^*B) - 2 \sum_{i=1}^r \sigma_i \cdot \gamma_i \right). \tag{1.4.4}$$

Ahora, definamos Γ a la matriz de valores singulares de B:

$$\Gamma = \left(\begin{array}{ccc|c} \gamma_1 & & & \\ & \ddots & & 0 \\ & & \gamma_r & \\ \hline & 0 & & 0 \end{array} \right),$$

y existen matrices unitarias $U_1 \in \mathscr{U}(n)$, $V_1 \in \mathscr{U}(m)$ tales que

$$B = U_1 \cdot \Gamma \cdot V_1^*.$$

Como la norma de Frobenius es un invariante por transformaciones unitarias, tendremos las siguientes igualdades:

- $Tr(B^*B) = Tr(\Gamma^*\Gamma) = \| \Gamma \|_F^2 = \sum_{i=1}^r \gamma_i^2$
- $Tr(\mathcal{J}_r^*\Gamma) = \langle \mathcal{J}_r, \Gamma \rangle_F = \sum_{i=1}^r \alpha_i \cdot \gamma_i$

Finalmente, es claro que

- $Tr(\mathcal{J}_r^* \mathcal{J}_r) = \| \mathcal{J}_r \|_F^2$
- $Tr(\mathcal{J}_{r+1}^* \mathcal{J}_{r+1}) = \| \mathcal{J}_{r+1} \|_F^2$

En conclusión, la desigualdad (1.4.4) resulta las siguientes

$$\| \mathcal{J} - B \|_F^2 \geq \| \mathcal{J}_{r+1} \|_F^2 + \left(\| \mathcal{J}_r \|_F^2 + \| \Gamma \|_F^2 - 2\langle \mathcal{J}_r, \Gamma \rangle_F \right). \tag{1.4.5}$$

Como se tiene que $\mathcal{J}_r$ y Γ son matrices con coordenadas reales, tenemos

$$\| \mathcal{J}_r - \Gamma \|_F^2 = \langle \mathcal{J}_r - \Gamma, \mathcal{J}_r - \Gamma \rangle_F = \| \mathcal{J}_r \|_F^2 + \| \Gamma \|_F^2 - 2\langle \mathcal{J}_r, \Gamma \rangle_F$$

con lo que la desigualdad (1.4.5) se transforma en

$$\| \mathcal{J} - B \|_F^2 \geq \| \mathcal{J}_{r+1} \|_F^2 + \| \mathcal{J}_r - \Gamma \|_F^2,$$

y se sigue la desigualdad señalada. □

Afirmación. *Sea $A \in \mathcal{M}_{n \times m}(\mathbb{K})$ una matriz de rango $t > r$ con valores singulares*

$$\sigma_1 \geq \sigma_2 ... \geq \sigma_t > 0.$$

Sea $B \in \sum_r'$ una matriz de rango a lo sumo r. Entonces,

$$\| A - B \|_F^2 \geq \left(\sum_{j=r+1}^t \sigma_j^2 \right).$$

Demostración. Usemos notaciones similares a las de la afirmación precedente. Así si $\gamma_1 \geq ... \geq \gamma_r > 0$ son los valores singulares de B, escribamos $\Gamma \in \mathcal{M}_{n\times m}(\mathbb{C})$ la matriz

$$\Gamma = \left(\begin{array}{ccc|c} \gamma_1 & & & \\ & \ddots & & 0 \\ & & \gamma_r & \\ \hline & 0 & & 0 \end{array}\right),$$

Del mismo modo supongamos $U_1 \in \mathscr{U}(n)$, $V_1 \in \mathscr{U}(m)$ matrices unitarias tales que

$$B = U_1 \cdot \Gamma \cdot V_1^*.$$

Denotemos las matrices

$$\mathcal{J} = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_t & \\ \hline & 0 & & 0 \end{array}\right) \in \mathcal{M}_{n\times m}(\mathbb{C}) \ , \ \mathcal{J}_r = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) \in \mathcal{M}_{n\times m}(\mathbb{C}),$$

y $\mathcal{J}_{r+1} = \mathcal{J} - \mathcal{J}_r \in \mathcal{M}_{n\times m}(\mathbb{C})$. Supongamos, como en la ecuación (A.3.1) que se tiene:

$$A = U \cdot \mathcal{J} \cdot V^*,$$

con $U \in \mathscr{U}(n)$, $V \in \mathscr{U}(m)$ matrices unitarias, como la norma de Frobenius es invariante por la acción de los grupos unitarios $\mathscr{U}(n)$ y $\mathscr{U}(m)$ tenemos la siguiente igualdad

$$\| A - B \|_F^2 = \| U\mathcal{J}V^* - B \|_F^2 = \| U(\mathcal{J} - U^*BV)V^* \|_F^2 = \| \mathcal{J} - B_1 \|_F^2,$$

donde $B_1 = U^*BV$. Ahora bien, tenemos que B_1 tiene los mismos valores singulares que B y el mismo rango. Más aún, se tiene que

$$B_1 = U_2 \cdot \Gamma \cdot V_2^*,$$

donde $U_2 = U^* \cdot U_1 \in \mathscr{U}(n)$ y $V_2 = V^* \cdot V_1$. Aplicando la desigualdad de la Afirmación 1.4.4, concluimos

$$\| A - B \|_F^2 = \| \mathcal{J} - B_1 \|_F^2 \geq \| \mathcal{J}_{r+1} \|_F^2 + \| \mathcal{J}_r - \Gamma \|_F^2 .$$

Ahora bien, como $\| \mathcal{J}_{r+1} \|_F^2 = \sum_{j=r+1}^t \sigma_j^2$, concluimos la desigualdad enunciada en la Afirmación:

$$\| A - B \|_F^2 \geq \sum_{j=r+1}^{t} \sigma_j^2.$$

□

Finalmente, estamos en condiciones de probar el enunciado del Teorema. De una parte, la aproximación de rango r de A satisface $A_r \in \sum_r'$ y, además, $\| A - A_r \|_F^2 = \sum_{j=r+1}^t \sigma_j^2 \geq \min\{\| A - C \|_F^2 : C \in \sum_r'\}$.
Pero, de otro lado, la última afirmación demostrada prueba que

$$\min\{\| A - C \|_F^2 : C \in \sum_r'\} \geq \sum_{j=r+1}^t \sigma_j^2 = \| A - A_r \|_F^2 .$$

Por tanto, $A_r \in \sum_r'$ minimiza la distancia de A a $\sum_r'$ con respecto a la norma de Frobenius. □

Finalmente, probemos la conexión entre el condicionamiento de Demmel y las métricas proyectivas.
Podemos definir la distancia de de una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ con respecto a un subconjunto $T \subseteq \mathcal{M}_{n\times m}(\mathbb{K})$ con respecto a la norma de Frobenius mediante:

$$d_F(A, T) := \inf\{\| A - B \|_F : B \in T\}.$$

En el caso de que T sea un cono (i.e. $\lambda T \subseteq T, \forall \lambda \in \mathbb{K}$) ese ínfimo es un mínimo.

Teorema 1.4.7 (**Schmidt-Mirsky-Eckart-Young y condicionamiento**). *Con las notaciones precedentes, sea $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ una matriz no nula de rango $t \leq m$, es decir, $A \in \sum_t'$. Sea $\alpha = \pi(A) \in \mathbb{P}_{nm-1}(\mathbb{K})$ el punto proyectivo que define A. Se tiene*

$$\mu(A) = \tfrac{1}{d_{\mathbb{P}}(\alpha, \sum_{t-1})}.$$

Además, si A es un representante de α que cumpla $\| A \|_F = 1$, para cada r, $1 \leq r \leq t$, se tiene la siguiente igualdad,

$$d_{\mathbb{P}}(\alpha, \sum\nolimits_r) = d_F(A, \sum\nolimits'_r).$$

Y en particular, se tendrá la siguiente igualdad

$$\mu(A) = \frac{1}{d_{\mathbb{P}}(\alpha, \sum_{t-1})} = \frac{1}{d_F(A, \sum'_{t-1})}.$$

Demostración. Haremos el caso complejo ($\mathbb{K} = \mathbb{C}$), dejando el caso real ($\mathbb{K} = \mathbb{R}$) por tener una demostración idéntica. Para nuestra prueba tomamos $\alpha =< A >$ siendo A un representante tal que $\| A \|_F^2 = 1$. Consideremos una descomposición en valores singulares de A:

$$A = U \cdot \mathcal{J} \cdot V^*, \tag{1.4.6}$$

donde $U \in \mathscr{U}(n)$, $V \in \mathscr{U}(m)$ y $\mathcal{J}$ es la matriz diagonal siguiente:

$$\mathcal{J} = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_t & \\ \hline & 0 & & 0 \end{array} \right) \in \mathcal{M}_{n \times m}(\mathbb{K}), \tag{1.4.7}$$

donde $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_t > 0$ son los valores singulares de A. Para cada $i, 1 \leq i \leq t-1$, sea A_i la aproximación con rango $\leq i$ de A descrita anteriormente. Esto es, definiremos

$$A_i = U \cdot \mathcal{J}_i \cdot V^*,$$

donde $\mathcal{J}_i \in \mathcal{M}_{n \times m}(\mathbb{C})$ es la matriz dada mediante:

$$\mathcal{J}_i = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_i & \\ \hline & 0 & & 0 \end{array} \right) \in \mathcal{M}_{n \times m}(\mathbb{K}).$$

Sea $\beta_i =< A_i >\in \mathbb{P}_{nm-1}(\mathbb{K})$ el punto proyectivo definido por A_i y tendremos que

$$d_{\mathbb{P}}(\alpha, \beta_i) = \sqrt{1 - \frac{|\langle A, A_i \rangle_F|^2}{\| A \|_F^2 \, \| A_i \|_F^2}} = \sqrt{1 - \frac{|\langle \mathcal{J}, \mathcal{J}_i \rangle_F|^2}{\| \mathcal{J}_i \|_F^2}},$$

donde hemos usado la invarianza de $\langle \cdot, \cdot \rangle_F$ por la acción de los grupos unitarios $\mathscr{U}(n)$ y $\mathscr{U}(m)$ (resp. a izquierda y derecha). Como ya se ha hecho en apartados precedentes, es fácil observar que

- $|\langle \mathcal{J}, \mathcal{J}_i \rangle_F|^2 = (\sigma_1^2 + ... + \sigma_i^2)^2$,
- $\| \mathcal{J}_i \|_F^2 = \sigma_1^2 + ... + \sigma_i^2$

Por tanto

$$d_{\mathbb{P}}(\alpha, \beta_i) = \sqrt{1 - \frac{(\sigma_1^2 + ... + \sigma_i^2)^2}{\sigma_1^2 + ... + \sigma_i^2}} = (1 - (\sigma_1^2 + ... + \sigma_i^2))^{1/2} = \left(\sum_{k=i+1}^{t} \sigma_k^2 \right)^{1/2}.$$

En particular, habremos concluido que

$$\left(\sum_{k=i+1}^{t} \sigma_k^2 \right)^{1/2} = d_{\mathbb{P}}(\alpha, \beta_i) \geq \min\{ d_{\mathbb{P}}(\alpha, \beta) : \beta \in \sum\nolimits_r \},$$

donde ya hemos usado el hecho de que $\sum_r$ es un compacto (y, por tanto, el ínfimo y el mínimo coinciden).

Consideremos ahora un punto $\beta =< B >\in \mathbb{P}_{nm-1}(\mathbb{C})$ de tal modo que $\beta \in \sum_r$ (i.e B tiene rango acotado por r). Consideremos los valores singulares de B dados por

$$\sigma_1(B) = \gamma_1 \geq \sigma_2(B) = \gamma_2 \geq ... \geq \sigma_r(B) = \gamma_r > 0.$$

Por la Desigualdad de la Traza de von Neumann (ver Teorema A.3.4 del Apéndice A), observando que los valores singulares de una matriz $B \in \mathcal{M}_{n\times m}(\mathbb{C})$ coinciden con los valores singulares de $B^* \in \mathcal{M}_{n\times m}(\mathbb{C})$, tendremos que

$$|Tr(B^*A)| \leq \sum_{i=1}^{r} \sigma_i \cdot \gamma_i$$

Por tanto

$$1 - \frac{|Tr(B^*A)|^2}{\| A \|_F^2 \| B \|_F^2} \geq 1 - \frac{\left(\sum_{i=1}^r \sigma_i \cdot \gamma_i\right)^2}{\sum_{i=1}^r \gamma_i^2},$$

donde hemos usado que $\| A \|_F^2 = 1$ y $\| B \|_F^2 = \sum_{i=1}^r \gamma_i$.
Por tanto, hemos probado que

$$d_{\mathbb{P}}(\beta, \alpha)^2 \geq \left(\sum_{i=1}^{t} \sigma_i^2\right) - \frac{\left(\sum_{i=1}^r \sigma_i \cdot \gamma_i\right)^2}{\sum_{i=1}^r \gamma_i^2}.$$

Por tanto, tenemos

$$d_{\mathbb{P}}(\beta, \alpha)^2 \geq \frac{\left(\sum_{i=1}^t \sigma_i^2\right)\left(\sum_{i=1}^r \gamma_i^2\right) - \left(\sum_{i=1}^r \sigma_i \cdot \gamma_i\right)^2}{\sum_{i=1}^r \gamma_i^2} =$$

$$= \frac{\left(\sum_{i=1}^r \sigma_i^2\right)\left(\sum_{i=1}^r \gamma_i^2\right) - \left(\sum_{i=1}^r \sigma_i \cdot \gamma_i\right)^2}{\sum_{i=1}^r \gamma_i^2} + \frac{\left(\sum_{i=r+1}^t \sigma_i^2\right)\left(\sum_{i=1}^r \gamma_i^2\right)}{\sum_{i=1}^r \gamma_i^2}. \tag{1.4.8}$$

Ahora bien. tomando $\underline{\sigma} = (\sigma_1, ...\sigma_r) \in \mathbb{R}^r$ y $\underline{\gamma} = (\gamma_1, ...\gamma_r) \in \mathbb{R}^r$, la desigualdad de Cauchy-Schwartz se escribe mediante

$$\langle \underline{\sigma}, \underline{\gamma} \rangle^2 = \left(\sum_{i=1}^{r} \sigma_i \cdot \gamma_i\right)^2 \leq \| \underline{\sigma} \|^2 \| \underline{\gamma} \|^2 = \left(\sum_{i=1}^{r} \sigma_i^2\right)\left(\sum_{i=1}^{r} \gamma_i^2\right).$$

Por tanto, tenemos que el primer término de la suma en (1.4.8) es un número real positivo, esto es,

$$\frac{\left(\sum_{i=1}^r \sigma_i^2\right)\left(\sum_{i=1}^r \gamma_i^2\right) - \left(\sum_{i=1}^r \sigma_i \cdot \gamma_i\right)^2}{\sum_{i=1}^r \gamma_i^2} \geq 0.$$

Por tanto, $d_{\mathbb{P}}(\beta, \alpha)$ es mayor o igual que el segundo término,

$$d_{\mathbb{P}}(\beta, \alpha)^2 \geq \frac{\left(\sum_{j=r+1}^t \sigma_j^2\right)\left(\sum_{i=1}^r \gamma_i^2\right)}{\sum_{i=1}^r \gamma_i^2} = \sum_{j=r+1}^{t} \sigma_j^2 = d_{\mathbb{P}}(\alpha, \beta_r)^2,$$

Siendo $\beta_r =< A_r >$. En conclusión

$$d_{\mathbb{P}}(\alpha, \textstyle\sum_r) = d_{\mathbb{P}}(\alpha, \beta_r) = \left(\sum_{j=r+1}^{t} \sigma_j^2\right),$$

lo que combinado con las igualdades obtenidas en el Teorema 1.4.6, implica:

$$d_{\mathbb{P}}(\alpha, \textstyle\sum_r) = d_F(A, \sum_r').$$

Para la última igualdad del enunciado, obsérvese que si A satisface la Identidad (1.4.6) donde $\mathcal{J}$ es la matriz diagonal definida en (1.4.7), entonces, por ser A tal que $\| A \|_F = 1$ se tiene que:

$$\mu(A) = \| A \|_F \| A^+ \|_2 = \frac{1}{\sigma_t}.$$

Por el Teorema 1.4.6 precedente, junto al resultado que acabamos de ver, tenemos que:

$$d_{\mathbb{P}}(\alpha, \textstyle\sum_{t-1}) = d_F(A, \sum_{t-1}') = \sigma_t.$$

Luego se tiene la igualdad final del enunciado. Basta con darse cuenta de la naturaleza proyectiva del condicionamiento para dado cualquier representante A de $\alpha = \pi(A)$ obtener:

$$\mu(A) = \tfrac{1}{d_{\mathbb{P}}(\alpha, \sum_{t-1})},$$

con lo que finaliza la demostración del Teorema. □

CAPíTULO 2

Elementos de la geometría Integral en algunas variedades de matrices.

Índice

2.1.	**Introducción y Resumen del Contenido del Capítulo**	**29**
2.2.	**La esfera como variedad de Riemann**	**32**
2.3.	**Estructura Riemanniana del grupo ortogonal**	**33**
2.4.	**El jacobiano normal de la parametrización desde la SVD**	**42**
2.5.	**El volumen de la fibra de la parametrización SVD de una matriz**	**45**
2.6.	**El uso de la fórmula de la Co-área para la parametrización SVD**	**46**

2.1. Introducción y Resumen del Contenido del Capítulo

Como ya se señaló en el Teorema de Wedin o en la Subsección 1.1.1, el condicionamiento lineal es un parámetro muy necesario para controlar la resolución de problemas (ecuaciones lineales, mínimos cuadrados, etc.) tras perturbar los datos del mismo. Es, por tanto, esencial, en la aplicación de algoritmos numéricos, el disponer "a priori" de buenas razones (tranquilizadoras) que nos garanticen que el problema particular que se trata está bien condicionado. Sin embargo, no resulta muy eficiente exigir que cada aplicación de nuestro algoritmo calcule el condicionamiento del input, lo que puede resultar en un alto aumento del tiempo de ejecución. Surge así la pregunta de conocer el comportamiento en promedio de los números de condicionamiento lineales. Algunas referencias que citan el problema (y son citadas en la tesis de A. Edelman) son [**vN, 1963**], [**Birk&Gul, 1963**], [**Smale, 1985**], [**Demmel, 1988**]. De hecho, será A. Edelman quien presente los resultados más "ajustados" (en especial en el caso complejo) del comportamiento en promedio del condicionamiento lineal en [**Edelman, 1988**] y en su Ph. D. Thesis [**Edelman, 1989**].
Estos resultados vienen a garantizar un resultado del tipo siguiente:

> "La probabilidad de que una matriz $A \in \mathcal{M}_n(\mathbb{K})$ tenga un condicionamiento $\mu(A) > \epsilon^{-1}$ para $\epsilon > 0, \epsilon \in \mathbb{R}$, está acotada para la distribución gaussiana por una cantidad de la forma
>
> $$\varphi(n) \cdot \epsilon^k \text{ "}$$

donde $k \in \{1, 2\}$ según sea el caso real o el complejo y $\varphi(n)$ es una cantidad "razonable".
Este tipo de resultados (y muchos otros que se generaron a posteriori), tienen la siguiente pseudo-interpretación algorítmica: dado un algoritmo que recibe como input una matriz A aleatoriamente elegida, si su comportamiento depende del condicionamiento de A, este comportamiento será "bueno" con alta probabilidad. No es pues necesario obligar a nuestros algoritmos a una pre-computación del condicionamiento del input que empeora nuestro rendimiento.
De otro lado, el resultado meramente como resultado matemático es interesantísimo por su propia naturaleza y los "elementos" que lo envuelven. Y sólo ese argumento estético bastaría para ser interesante.
Sin embargo hay un error de principio (o, si se prefiere, una sutileza) que debe ponernos en estado de atención y cuidado. En el mundo real, los problemas para los que se diseñan los algoritmos no suelen aplicarse a inputs genéricos sino, por ejemplo, a subclases de matrices y problemas. Pongamos tres ejemplos:

- Nuestro algoritmo trata solamente con matrices A estocásticas en $\mathcal{M}_n(\mathbb{R})$.
- Nuestro algoritmo trata solamente con matrices A tri-diagonales en $\mathcal{M}_n(\mathbb{R})$.
- Nuestro algoritmo trata solamente con matrices singulares A en $\mathcal{M}_n(\mathbb{R}) \setminus GL(n, \mathbb{R})$.

Estos tres ejemplos, y muchos otros fáciles de entender, transportan una característica común:

> Todos ellos tienen inputs en un conjunto $\mathscr{L} \in \mathcal{M}_n(\mathbb{R})$ que tiene medida de Lebesgue nula en el espacio afín $\mathbb{R}^{n^2} = \mathcal{M}_n(\mathbb{R})$.

Por tanto, los resultados de probabilidad genérica (i.e. en dimensión n^2) de Edelman, Demmel o Smale no se les aplican: no se puede heredar un resultado de distribución de probabilidad sobre un espacio probabilístico (X, ν) a un subespacio $(Y, \widetilde{\nu})$ cuando $\nu(Y) = 0$.
Esto supone, de facto, la necesidad de estudios "ad hoc" del comportamiento en promedio del condicicionamiento lineal restringido a "clases especiales" de matrices. La dificultad es más que evidente: Tomemos como ejemplo el comprender el condicionamiento de matrices $A \in \mathcal{M}_n(\mathbb{C})$, con rango a lo sumo $n-1$. Es decir, sea $\sum'_{n-1} \subseteq \mathcal{M}_n(\mathbb{C})$ la clase de matrices con determinante nulo. Se trata de una matriz A que satisface una ecuación polinomial homogénea de grado n:

$$\sum'_{n-1} := \{A \in \mathcal{M}_n(\mathbb{C}) : \det(A) = 0\}.$$

De hecho ni siquiera es una subvariedad diferenciable de $\mathbb{C}^{n^2}$ dado que es un conjunto con abundantes puntos singulares ($\sum'_{n-1}$ son los puntos singulares de $\sum'_n$ (ver [**AGV, 1985**] o Sección A.2 del Apédice A)). Si ahora supongo una distribución gaussiana $(\mathcal{M}_n(\mathbb{C}), \gamma)$ y un resultado sobre la probabilidad, es inaplicable al comportamiento de matrices aleatoriamente elegidas $A \in \sum'_n$.
Ha habido varias aproximaciones a entender este problema. Lo que se desarrolla en este Capítulo es la elegante estrategia de [**Be, 09**]. Consideremos la estratificación por rango ya citada

$$\sum'_1 \subsetneq \sum'_2 \subsetneq ... \subsetneq \sum'_t \subseteq \mathcal{M}_{n\times m}(\mathbb{K}),$$

con $t = \min\{n, m\}$. Nuestra noción de condicionamiento (tanto κ como μ) se adaptan bien a esta estratificación por usar la pseudo-inversa de Moore-Penrose. Consideremos las proyectivizaciones de esas variedades donde los condicionamientos tienen, además, el significado geométrico que les confiere el Teorema de Schmidt-Mirsky-Eckart-Young.

$$\sum_1 \subsetneq \sum_2 \subsetneq ... \subsetneq \sum_t \subseteq \mathbb{P}_{nm-1}(\mathbb{K}).$$

Tanto en el caso $\mathbb{K} = \mathbb{R}$ como en el caso $\mathbb{K} = \mathbb{C}$, $\sum_r \setminus \sum_{r-1}$ es una subvariedad de Riemann compacta de $P_{nm-1}(\mathbb{K})$ de volumen finito. Luego hay una distribución de probabilidad natural que da sentido a la cuestión:

PROBLEMA 5. *"Estimar la Probabilidad de elegir aleatoriamente una matriz (un punto proyectivo) $A \in \sum_r \setminus \sum_{r-1}$ tal que $\mu(A) > \epsilon^{-1}$ ".*

De eso trata este Capítulo. Además, la idea de [**Be, 09**] es muy interesante por calcular (objetivo esencial de este Trabajo de Fin de Grado) la siguiente estrategia: La describiremos para el caso $\mathbb{K} = \mathbb{R}$. Consideremos la esfera unidad $S^{r-1} \subseteq \mathbb{R}^r$ y definamos el conjunto de puntos de un "octante" estrictamente positivo:

$$S_r^+ := \{(\sigma_1, ..., \sigma_r) \in S^{r-1} : \sigma_1 > \sigma_2 > ... > \sigma_r > 0\}.$$

Denotemos también por S_r^+ al conjunto de matrices en $\mathcal{M}_n(\mathbb{R})$ cuya diagonal está en S_r^+. Es decir,

$$S_r^+ := \left\{ \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array} \right) \in \mathcal{M}_n(\mathbb{R}) : \sigma_1^2 + ... + \sigma_r^2 = 1, \sigma_1 > ... > \sigma_r > 0 \right\}.$$

Ahora consideremos la parametrización "natural" de $\left(\sum'_r \setminus \sum'_{r-1}\right) \cap S^{nm-1}$ dada mediante la Descomposición en Valores Singulares representada por S_r^+:

$$\begin{array}{rccc} \varphi: & \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m) & \longrightarrow & \left(\sum'_r \setminus \sum'_{r-1}\right) \cap S^{nm-1} \\ & (U, D, V) & \longmapsto & U \cdot D \cdot V^*. \end{array}$$

La clásica fórmula de la Co-áera (ver Subsección A.6.3 del Apéndice A), nos da una regla natural para trasladar esperanzas y probabilidades de la compleja variedad de Riemann $\sum_r \setminus \sum_{r-1}$ a la (más simple) variedad producto $\mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m)$.
Dicha fórmula se resume del modo siguiente:

TEOREMA 2.1.1. [**Fórmula de la Co-área**] *Sean M y N dos variedades Riemannianas. Sean m, n sus dimensiones respectivas, con $m \geq n$. Sea $F : M \to N$ una aplicación diferenciable sobreyectiva tal que la diferencial $d_x F$ es sobreyectiva para casi todo punto $x \in M$. Sea $\phi : M \to [0, \infty]$ una función integrable. Entonces, se tienen las dos siguientes igualdades:*

$$\int_{x\in M} \phi(x) dM = \int_{y\in N} \int_{x\in F^{-1}(y)} \phi(x) \frac{1}{NJ_x F} dF^{-1}(y) dN,$$

$$\int_{x\in M} \phi(x) NJ_x F dM = \int_{y\in N} \int_{x\in F^{-1}(y)} \phi(x) dF^{-1}(y) dN.$$

Es una forma general del Teorema de Cambio de Variable en integración donde el usual jacobiano es reemplazado por el jacobiano normal $NJ_x F$. (ver la Definición 19 de la Subsección A.6.3 del Apéndice A).
En nuestro caso, supongamos $\phi : \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m) \longrightarrow \mathbb{R}$ es una función y $\varphi : \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m) \longrightarrow \sum_r \setminus \sum_{r-1}$ es la parametrización basada en la descomposición en valores singulares. Entonces, el comportamiento de ϕ se puede estudiar mediante la identidad siguiente:

$$I = \int_{\mathscr{O}(n)\times S_r^+ \times \mathscr{O}(m)} \phi(U, D, V) d(\mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m)) = \tag{2.1.1}$$

$$= \int_{A\in \sum_r \setminus \sum_{r-1}} \int_{B\in\varphi^{-1}(A)} \frac{\phi(B)}{NJ_B\varphi} d\varphi^{-1}(A) d(\textstyle\sum_r \setminus \sum_{r-1})$$

A primera vista si ϕ sólo depende de D (como es el caso del condicionamiento) la expresión se simplifica más:

$$I = \int_{\mathscr{O}(n)\times\mathscr{O}(m)} \left(\int_{S_r^+} \phi(D) dS_r^+ \right) d(\mathscr{O}(n) \times \mathscr{O}(m)).$$

Pero además, los valores singulares asociados a una matriz A son únicos por lo que $\phi(U, D, V) = \phi(D)$ es constante en $\varphi^{-1}(A)$ para cualquier A. Así, vamos reduciendo la expresión a

$$I = \int_{A\in \sum_r \setminus \sum_{r-1}} \phi(A) \left(\int_{B\in\varphi^{-1}(A)} \frac{1}{NJ_B\varphi} d\varphi^{-1}(A) \right) d(\textstyle\sum_r \setminus \sum_{r-1}).$$

Resulta, por tanto, crucial comprender el **jacobiano normal de la parametrización por la SVD**, esto es, $NJ_{(U,D,V)}\varphi$ con la notación precedente.
No es extraño observar que el jacobiano normal $NJ_{(U,D,V)}\varphi$ es invariante por la acción de ambos grupos ortogonales o, si se prefiere, es constante en la fibra $\varphi^{-1}(A)$ para cada $A \in \sum_r \setminus \sum_{r-1}$. Así, la identidad toma la forma (para un ϕ dependiente solamente de los valores singulares):

$$I = \int_{\sum_r \setminus \sum_{r-1}} \frac{\phi(D)}{NJ_{(U,D,V)}\varphi} \left(\int_{B\in\varphi^{-1}(A)} d\varphi^{-1}(A) \right) d(\textstyle\sum_r \setminus \sum_{r-1}).$$

En otras palabras, la identidad 2.1.1 para funciones ϕ que sólo dependen de los valores singulares resulta

$$I = \mathrm{Vol}[\mathscr{O}(n)] \cdot \mathrm{Vol}[\mathscr{O}(m)] \int_{S_r^+} \phi(D) d\nu_S \tag{2.1.2}$$

$$= \int_{\sum_r \setminus \sum_{r-1}} \frac{\phi(D)}{NJ_A\varphi} \mathrm{Vol}[\varphi^{-1}(A)] d(\textstyle\sum_r \setminus \sum_{r-1}).$$

donde $NJ_A\varphi$ sería el jacobiano normal en cualquier punto de la fibra $\varphi^{-1}(A)$.
Dado que los volúmenes de los grupos ortogonales $\mathscr{O}(n)$ y $\mathscr{O}(m)$ son conocidos, si elegimos $\phi(D) = \frac{1}{\sigma_r}$ (que es el condicionamiento que nos ocupa) hay solamente dos ingredientes técnicos a controlar y calcular:

- El jacobiano normal $NJ_{(U,D,V)}\varphi$ en cualquier punto $(U, D, V) \in \varphi^{-1}(A)$.
- El volumen de la fibra $\varphi^{-1}(A)$.

Ambos son los resultados fundamentales que se describen en [**Be, 09**] y que demostramos en este Trabajo Fin de Grado. El primero es el Teorema 2.4.5 de la Sección 2.4 y que reproducimos aquí:

TEOREMA 2.1.2. *Sean $U \in \mathscr{O}(n)$, $V \in \mathscr{O}(m)$, $D \in S_r^+$ y sea $A = UDV^* \in \sum_r \setminus \sum_{r-1}$. Sea $NJ_{(U,D,V)}\varphi$ el jacobiano normal de la parametrización φ de la Descomposición en Valores Singulares. Entonces,*

$$NJ_{(U,D,V)}\varphi = \frac{\det(D)^{n+m-2r}\Delta(D)}{\left(\sqrt{2}\right)^{r(n+m-r-1)}}.$$

donde $\det(D) = \prod_{i=1}^r \sigma_i$ es el determinante y $\Delta(D) = \prod_{i<j}(\sigma_i^2 - \sigma_j^2)$ es el discriminante.

El segundo resultado fundamental de [**Be, 09**] es el cálculo del volumen de la fibra $\varphi^{-1}(A)$ para $A \in \sum_r \setminus \sum_{r-1}$. Es el Teorema 2.5.1 de la Sección 2.5 y que reproducimos a continuación:

TEOREMA 2.1.3. *La cantidad $\mathrm{Vol}[\varphi^{-1}(A)]$ es independiente de $A \in \sum_r \setminus \sum_{r-1}$, excepto para un conjunto de medida cero. En efecto, sea $A \in \sum_r \setminus \sum_{r-1}$ tal que todos sus r valores singulares distintos de cero sean diferentes. Entonces tenemos que*

$$\mathrm{Vol}[\varphi^{-1}(A)] = 2^r \mathrm{Vol}[\mathscr{O}(n-r)]\mathrm{Vol}[\mathscr{O}(m-r)].$$

Con estos resultados ya podemos establecer la identidad final fundamental para cualquier función $\phi(A)$ que sólo dependa de los valores singulares de A (escribiremos $\phi(\sigma(A))$ y que, despejando en (2.1.2) se transforma en el siguiente resultado correspondiente al Teorema 2.6.1:

TEOREMA 2.1.4. *Sea $A \in \sum_r \setminus \sum_{r-1}$ el punto proyectivo asociado a una matriz de rango r. Sea $\sigma(A) = (\sigma_1(A), ..., \sigma_r(A)) \in S_r^+$ la lista de sus valores singulares. Sea $\phi : \mathbb{R}^r \to \mathbb{R}$ una función medible. Entonces, se tiene:*

$$\int_{A \in \sum_r \setminus \sum_{r-1}} \phi(\sigma(A))\mathrm{d}(\sum_r \setminus \sum_{r-1}) = H(n,m,r) \cdot \int_{D \in S_r^+} \phi(D)(\det(D)^{n+m-2r}\Delta(D))\mathrm{d}S_r^+,$$

donde

$$H(n,m,r) = \frac{\mathrm{Vol}[\mathscr{O}(n)]\mathrm{Vol}[\mathscr{O}(m)]}{\mathrm{Vol}[\mathscr{O}(n-r)]\mathrm{Vol}[\mathscr{O}(m-r)]\left(\sqrt{2}\right)^{r(n+m-r)+r}}.$$

Por último, aplicaremos este resultado que junto a los resultados de Chen y Dongarra [**Chen&Dongarra, 2005**] y estableceremos unas cotas para la función de distribución de κ para matrices pertenecientes a $\sum_r \setminus \sum_{r-1}$ con la distribución uniforme, esto se realiza en el Teorema 2.6.2 que se enuncia como sigue:

TEOREMA 2.1.5. *Con la distribución uniforme en $\sum_r \setminus \sum_{r-1}$ se tiene la siguiente igualdad para cada t:*

$$P_U\left[\kappa(A) > t : A \in \sum_r \setminus \sum_{r-1}\right] = P_\gamma[\kappa(A) > t : A \in GL(r \times (n+m-r), \mathbb{R})].$$

donde $GL(r \times (n+m-r), \mathbb{R})$ es el abierto Zariski en $\mathcal{M}_{r\times(n+m-r)}(\mathbb{R})$ formado por las matrices de rango máximo y γ es la distribución Gaussiana.
Más aún, se verifican las siguientes cotas

$$\frac{1}{(2\pi)^{1/2}}\left(\frac{0.245}{t}\right)^{(n+m-2r+1)} \leq P_U\left[\frac{\kappa(A)}{\frac{n+m-r}{n+m-2r+1}} > t : A \in \sum_r \setminus \sum_{r-1}\right],$$

$$P_U\left[\frac{\kappa(A)}{\frac{n+m-r}{n+m-2r+1}} > t : A \in \sum_r \setminus \sum_{r-1}\right] \leq \frac{1}{(2\pi)^{1/2}}\left(\frac{6.414}{t}\right)^{(n+m-2r+1)}.$$

2.2. La esfera como variedad de Riemann

Consideremos $\mathbb{R}^n$ y el producto escalar canónico

$$\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}.$$

Denotaremos por S^{n-1} a la esfera real usual dada a partir del producto escalar canónico anterior

$$S^{n-1} = \{x \in \mathbb{R}^n : \langle x, x \rangle = 1\} = \{x \in \mathbb{R}^n : \| x \|_2^2 = 1\}.$$

Sea $p \in S^{n-1}$ un punto y sea $x : (-\epsilon, \epsilon) \to S^{n-1}$ una curva diferenciable cualquiera tal que $x(0) = p$. Denotemos por $\dot{x}(t) = (\dot{x}_1(t), ..., \dot{x}_n(t))$ su derivada y observamos que

$$1 = \| x(t) \|^2 = \sum_{i=1}^{n} x_i(t)^2 \ , \ \forall t \in (-\epsilon, \epsilon).$$

Por tanto, se tiene:

$$0 = 2 \cdot \sum_{i=1}^{n} x_i(t) \cdot \dot{x}_i(t) \ , \ \forall t \in (-\epsilon, \epsilon).$$

En particular, el vector tangente $\dot{x}(0) \in < p >^{\perp}$, donde $< p >^{\perp}$ es el complemento ortogonal del subespacio generado por p en $\mathbb{R}^n$. Recíprocamente, sea $v \in < p >^{\perp}$ y definamos

$$\begin{array}{rccc} \varphi\colon & \mathbb{R} & \longrightarrow & S^{n-1} \\ & t & \longmapsto & \frac{p+tv}{\| p+tv \|} \end{array}$$

Calculemos $\dot{\varphi}(t)$ y observaremos que

$$\dot{\varphi}(0) = v,$$

para verlo, basta con aplicar la definición de la derivada para φ en $t = 0$

$$\dot{\varphi}(0) = \lim_{h \to 0} \frac{\varphi(0+h) - \varphi(0)}{h} = \lim_{h \to 0} \frac{v}{\| p + hv \|} = \frac{v}{\| p \|} = v.$$

Hemos probado así que:

PROPOSICIÓN 2.2.1. *El conjunto $(S^{n-1}, \cdot)$ es una subvariedad de Riemann de $\mathbb{R}^n$. Para cada punto $p \in S^{n-1}$, el espacio tangente T_pS^{n-1} es el complemento ortogonal $< p >^{\perp}$ del subespacio generado por p en $\mathbb{R}^n$. Más aún, la métrica de T_pS^{n-1} es la inducida por la métrica canónica $\langle \cdot, \cdot \rangle$ de $\mathbb{R}^n$, es decir,*

$$\begin{array}{rccc} \langle \cdot, \cdot \rangle_{T_pS^{n-1}}\colon & T_pS^{n-1} \times T_pS^{n-1} & \longrightarrow & \mathbb{R} \\ & (\dot{v}, \dot{w}) & \longmapsto & \langle \dot{v}, \dot{w} \rangle = \sum_{i=1}^{n} \dot{v} \cdot \dot{w}. \end{array}$$

DEMOSTRACIÓN. Esencialmente es la idea que acabamos de discutir. □

PROPOSICIÓN 2.2.2. *Sean $p, q \in S^{n-1}$ dos puntos cualesquiera de la esfera unidad y sea $U \in \mathscr{O}(n)$ una matriz ortogonal tal que $U \cdot p = q$.*

i) La matriz U define una isometría lineal sobre S^{n-1} con respecto al producto escalar canónico de $\mathbb{R}^n$.

ii) U define también una isometría entre los espacios tangentes T_pS^{n-1} y T_qS^{n-1} antes descritos.

$$\begin{array}{ccccc} < p >^{\perp} = T_pS^{n-1} & \longrightarrow & T_qS^{n-1} = < q >^{\perp} \\ \dot{V} & \longmapsto & U \cdot \dot{V}. \end{array}$$

En particular, U es una isometría de S^{n-1} en sí misma como subvariedad del Riemann de $\mathbb{R}^n$.

DEMOSTRACIÓN. Basta recordar que si $U \in \mathscr{O}(n)$, $\forall u, v \in \mathbb{R}^n$

$$\langle u, v \rangle = \langle Uu, Uv \rangle.$$

□

2.3. Estructura Riemanniana del grupo ortogonal

Consideremos el grupo de matrices ortogonales reales $(\mathscr{O}(n), \cdot)$ donde

$$\mathscr{O}(n) = \{A \in \mathcal{M}_n(\mathbb{R}) : AA^t = Id_n\}.$$

Escribiremos A^* en lugar de A^t. El grupo ortogonal $(\mathscr{O}(n), \cdot)$ es un subgrupo del grupo $(\mathscr{U}(n), \cdot)$ de las matrices unitarias. Recordemos, además, que las matrices unitarias definen una isometría con respecto al producto hermítico complejo $\langle \cdot, \cdot \rangle_{\mathbb{C}}$ canónico. Es decir, consideramos el producto hermítico en $\mathbb{C}^n$

$$\begin{array}{rccc} \langle \cdot, \cdot \rangle_{\mathbb{C}}\colon & \mathbb{C}^n \times \mathbb{C}^n & \longrightarrow & \mathbb{C} \\ & (\underline{a}, \underline{b}) & \longmapsto & \sum_{i=1}^{n} a_i \overline{b_i}, \end{array}$$

donde $\overline{b_i}$ quiere decir conjugado de b_i y $\underline{a} = (a_1, ..., a_n$, $\underline{b} = (b_1, ..., b_n)$. De esta forma, dada $U \in \mathscr{U}(n)$ matriz unitaria y dados $u, v \in \mathbb{C}^n$, se tiene $\langle Uu, Uv \rangle_{\mathbb{C}} = \langle u, v \rangle_{\mathbb{C}}$.

En particular, toda matriz ortogonal $U \in \mathscr{O}(n)$ también define una isometría sobre $\mathbb{C}^n$ con respecto a $\langle \cdot, \cdot \rangle_{\mathbb{C}}$ y, en particular, también define una isometría sobre el espacio vectorial real $\mathbb{R}^n$ con respecto al producto escalar canónico.

LEMA 2.3.1. *Sea $U \in \mathscr{U}(n)$ una matriz unitaria. Sea $\lambda \in \mathbb{C}$ uno de sus valores propios, entonces $|\lambda|^2 = 1$, donde $|\cdot|$ quiere decir valor absoluto en $\mathbb{C}$.*

DEMOSTRACIÓN. Si λ es un valor propio de U, consideremos un vector propio no nulo $v \in \mathbb{C}^n \setminus \{0\}$ tal que $U \cdot v = \lambda \cdot v$. Por ser U una isometría

$$\| v \|^2 = \langle v, v \rangle_{\mathbb{C}} = \langle Uv, Uv \rangle_{\mathbb{C}} = \lambda \cdot \overline{\lambda} \langle v, v \rangle_{\mathbb{C}} = |\lambda|^2 \| v \|^2 .$$

En consecuencia $|\lambda|^2 = 1$. □

En particular, si $U \in \mathscr{O}(n)$ sus valores propios son números reales o complejos de valor absoluto 1. Observemos, además, que si U es una matriz ortogonal, en particular es una matriz normal y se le aplica el Teorema Espectral Real (cf. Teorema A.4.3 del Apéndice A). En particular,

COROLARIO 2.3.2. *Sea $U \in \mathscr{O}(n)$ una matriz ortogonal, entonces, existe una matriz ortogonal $P \in \mathscr{O}(n)$ y una matriz "diagonal real" D de tal modo que*

i) $PUP^* = D$

ii) La matriz D tiene la forma siguiente:

$$D = \begin{pmatrix} Id_r & & & & \\ & -Id_s & & & \\ & & M_1 & & \\ & & & \ddots & \\ & & & & M_t \end{pmatrix}, \tag{2.3.1}$$

donde $r + s + 2t = n$ y para cada i, $1 \leq i \leq t$,

$$M_i = \begin{pmatrix} a_i & b_i \\ -b_i & a_i \end{pmatrix} = \begin{pmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{pmatrix}$$

para $\theta_i \in [0, 2\pi]$.

DEMOSTRACIÓN. Es consecuencia inmediata del Teorema Espectral Real citado y del hecho de que los valores propios de U tienen módulo 1 en $\mathbb{C}$. El número r representa la multiplicidad de 1 como valor propio, s es la multiplicidad de -1 como valor propio y las matrices $_{2\times 2}$ representan los valores propios en $S^1 \subseteq \mathbb{C} = \mathbb{R}^2$ que no son números reales. □

PROPOSICIÓN 2.3.3. *EL siguiente es un subgrupo normal de $(\mathscr{O}(n), \cdot)$ conocido como el grupo especial ortogonal $SO(n)$,*

$$SO(n) := \{U \in \mathscr{O}(n) : \det(U) = 1\}.$$

Más aún, para cada matriz $U \in SO(n)$, existe $P \in \mathscr{O}(n)$ tal que

$$PUP^* = \begin{pmatrix} Id_r & & & & \\ & -Id_s & & & \\ & & M_1 & & \\ & & & \ddots & \\ & & & & M_t \end{pmatrix} = D,$$

como la identidad (2.3.1) precedente, con s un número par.

DEMOSTRACIÓN. Recordemos que el determinante es un morfismo de grupos (cf. [**Pardo, 2021**]):

$$\det : (GL(n, \mathbb{R}), \cdot) \longrightarrow (\mathbb{R}^\times, \cdot),$$

donde $(\mathbb{R}^\times, \cdot)$ es el grupo de las unidades del cuerpo de los números reales. También será morfismo de grupos su restricción:

$$\det\nolimits_{|\mathscr{O}(n)} : (\mathscr{O}(n), \cdot) \longrightarrow (\mathbb{R}^\times, \cdot).$$

En particular, el núcleo de $\det_{|_{\mathscr{O}(n)}}$, como morfismo de grupos que es, es un subgrupo normal de $\mathscr{O}(n)$. Más aún, como $1 \in \mathbb{R}^\times$ es el elemento neutro de $(\mathbb{R}^\times, \cdot)$ tenemos que

$$SO(n) = \ker\left(\det_{|_{\mathscr{O}(n)}}\right) = \{U \in \mathscr{O}(n) : \det(n) = 1\}.$$

Por el Colorario 2.3.2 precedente, admite una descomposición del tipo (2.3.1) salvo $P \in \mathscr{O}(n)$, es decir,

$$PUP^* = \begin{pmatrix} Id_r & & & & \\ & -Id_s & & & \\ & & M_1 & & \\ & & & \ddots & \\ & & & & M_t \end{pmatrix}, \tag{2.3.2}$$

donde M_i es la realización como matriz $_{2\times 2}$ de un elemento $z \in S^1 \subseteq \mathbb{C}$. Entonces,

$$\det(PUP^*) = \det(U)\det(PP^*) = \det(U) = 1.$$

Pero, usando (2.3.2) tendremos

$$1 = \det(PUP^*) = \det(Id_r) \cdot \det(Id_s) \cdot \prod_{i=1}^{t} \det(M_i) = (-1)^s.$$

Por tanto, s ha de ser un número par. □

Consideremos el conjunto de las matrices anti-simétricas reales, esto es,

$$A_n(\mathbb{R}) = \{A \in \mathcal{M}_n(\mathbb{R}) : A + A^* = 0\}.$$

Proposición 2.3.4. *El conjunto $A_n(\mathbb{R})$ es un subespacio vectorial de $\mathcal{M}_n(\mathbb{R})$, de dimensión $\frac{n(n-1)}{n}$. Una base de $A_n(\mathbb{R})$ como espacio vectorial son las matrices*

$$\{A_{ij} : i < j, 1 \leq i \leq n\},$$

donde $A_{ij} = \left(a_{r,s}^{(i,j)}\right)_{1\leq r,s\leq n}$ es una matriz cuyas coordenadas tienen la forma siguiente:

$$a_{r,s}^{(i,j)} = \begin{cases} 0 & si \quad (r,s) \neq (i,j),(j,i) \\ 1 & si \quad (r,s) = (i,j) \\ -1 & si \quad (r,s) = (j,i). \end{cases}$$

Demostración. Veamos primero que $A_n(\mathbb{R})$ es un subespacio vectorial de $\mathcal{M}_n(\mathbb{R})$, para ello basta con tomar dos matrices cualquiera $A, B \in A_n(\mathbb{R})$ y dos constantes cualesquiera $\lambda, \mu \in \mathbb{R}$, y simplemente operando vemos como $\lambda A + \mu B \in A_n(\mathbb{R})$:

$$\lambda A + \mu B + (\lambda A + \mu B)^* = \lambda A + \mu B + \mu B^* + \lambda A^* = \lambda(A + A^*) + \mu(B + B^*) = 0,$$

y concluimos así que es un subespacio vectorial de $\mathcal{M}_n(\mathbb{R})$.
Veamos ahora que el conjunto $\{A_{ij} : i < j, 1 \leq i \leq n\}$ forma una base de $A_n(\mathbb{R})$, es evidente que sus elementos son todos linealmente independientes, ya que ninguna matriz de la base comparte coordenadas no nulas, por lo que no podremos expresar un elemento de la base mediante la combinación lineal del resto. Veamos ahora que es un sistema generador de $A_n(\mathbb{R})$.
Sea $A \in A_n(\mathbb{R})$, esto significa que $A = -A^*$, es decir, A será una matriz de la siguiente forma

$$A = \begin{pmatrix} 0 & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ -a_{1,2} & 0 & a_{2,3} & \dots & a_{2,n} \\ -a_{1,3} & -a_{2,3} & 0 & & \vdots \\ \vdots & \vdots & & \ddots & a_{n-1,n} \\ -a_{1,n} & -a_{2,n} & \dots & -a_{n-1,n} & 0 \end{pmatrix},$$

de esta forma, podremos expresar A de mediante

$$A = \sum_{\{i<j, 1\leq i\leq n\}} a_{i,j} \cdot A_{i,j}$$

y concluimos que $\{A_{ij} : i < j, 1 \leq i \leq n\}$ es una base de $A_n(\mathbb{R})$. De esta forma, para ver la dimensión de $A_n(\mathbb{R})$ basta con calcular el cardinal de la base, que mediante un simple ejercicio de combinatoria vemos que es $\frac{n(n-1)}{2}$. □

TEOREMA 2.3.5. *Con las notaciones precedentes, la siguiente es epimorfismo de grupos:*

$$\begin{array}{rccc} \exp\colon & A_n(\mathbb{R}) & \longrightarrow & SO(n) \\ & A & \longmapsto & \sum_{k\geq 0} \frac{A^k}{k!}. \end{array}$$

Más aún, $(SO(n), \cdot)$ será un subgrupo normal y conexo de $(\mathscr{O}(n), \cdot)$.

DEMOSTRACIÓN. Para empezar nótese que si $A \in \mathcal{M}_n(\mathbb{R})$ es una matriz real, se tiene:

$$\exp(A^*) = \sum_{k\geq 0} \frac{(A^*)^k}{k!} = \left(\sum_{k\geq 0} \frac{A^k}{k!}\right)^* = (\exp(A))^*,$$

donde usamos implícitamente el Teorema de Hamilton-Cayley y la forma de Jordan para calcular la exponencial de matriz.
Por tanto, dada $A \in \mathcal{M}_n(\mathbb{R})$ una matriz anti-simétrica y $U = \exp(A)$, tenemos que $U^* = \exp(A^*)$ y, por las propiedades de la exponencial de matriz,

$$U \cdot U^* = \exp(A) \cdot \exp(A^*) = \exp(A + A^*) = \exp(0) = Id_n.$$

De esta forma, $\det(U)\cdot\det(U^*) = 1$, es decir, $\det(U) = \det(U^*) = 1$ y se tiene que $\det(\exp(A)) = 1$. Por tanto, $\exp(A) \in SO(n)$ y $\exp : A_n \to SO(n)$ está bien definida. Veamos que exp es suprayectiva. Para ello, consideremos una presentación de U como en la Proposición (2.3.3) precedente:

$$PUP^* = \begin{pmatrix} Id_r & & & & & \\ & -Id_{2s} & & & & \\ & & M_1 & & & \\ & & & \ddots & & \\ & & & & M_t \end{pmatrix},$$

donde $r + 2s + 2t = n$. Ahora, observemos que

- $-I_{2s}$ es la suma diagonal de s matrices ${}_{2\times 2}$ de la forma

$$B_i = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

observemos que estas matrices B_i son, en realidad,

$$B_i = \begin{pmatrix} \cos\pi & \sin\pi \\ -\sin\pi & \cos\pi \end{pmatrix}$$

y podemos expresarla mediante

$$B_i = \exp\left(\pi \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right).$$

- Por su parte, las matrices $M_1, ..., M_t$ tienen la forma

$$M_i = \begin{pmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{pmatrix} = \exp\left(\theta_i \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right).$$

En particular, podemos definir la matriz dada como la diagonal de matrices

$$A = \begin{pmatrix} O_{r\times r} & & & & & & \\ & Q_1 & & & & & \\ & & \ddots & & & & \\ & & & Q_s & & & \\ & & & & N_1 & & \\ & & & & & \ddots & \\ & & & & & & N_t \end{pmatrix},$$

donde

$$Q_i = \pi \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

$$N_i = \theta_i \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

observamos que

$$\exp(A) = \begin{pmatrix} \exp(O_{r\times r}) & & & & & & & \\ & \exp(Q_1) & & & & & & \\ & & \ddots & & & & & \\ & & & \exp(Q_s) & & & & \\ & & & & \exp(N_1) & & & \\ & & & & & \ddots & & \\ & & & & & & \exp(N_t) & \end{pmatrix} = D.$$

Finalmente, tendremos que

$$U = \exp(P^*AP) = P^* \exp(A)P = P^*DP.$$

Veamos, finalmente que $P^*AP \in A_n(\mathbb{R})$ es una matriz anti-simétrica real. Nótese que $A \in A_n(\mathbb{R})$ ya es anti-simétrica (es decir, tal y como hemos definido A se tiene que $A + A^* = 0$). Entonces, observamos

$$P^*AP + (P^*AP)^* = P^*AP + P^*A^*P = P^*(A + A^*)P = P^*0P = 0.$$

En conclusión, exp es un morfismo de grupos cuya imagen es $SO(n)$, es decir,

$$\exp(A_n(\mathbb{R})) = SO(n).$$

Además, $A_n(\mathbb{R}) \subseteq \mathbb{R}^{n^2} = \mathcal{M}_n(\mathbb{R})$ es un subespacio vectorial de $\mathbb{R}^n$, por tanto, como exp es continua su imagen es conexo, y $SO(n)$ es un subgrupo normal conexo de $\mathscr{O}(n)$. □

Como $SO(n)$ es un subgrupo normal de $(\mathscr{O}(n), \cdot)$, tenemos un isomorfismo de grupos

$$\mathscr{O}(n)/SO(n) \cong (\mathbb{Z}^\times, \cdot) = (\{\pm 1\}, \cdot).$$

En particular, tenemos la descomposición como unión disjunta de clases módulo $SO(n)$:

$$\mathscr{O}(n) = SO(n) \,\dot\cup\, U_0 \cdot SO(n),$$

donde U_0 es cualquier matriz ortogonal tal que $\det(U_0) = -1$. Podemos elegir, por ejemplo

$$U_o = \begin{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \\ & Id_{n-2} \end{pmatrix} \in \mathscr{O}(n).$$

COROLARIO 2.3.6. *El conjunto $\mathscr{O}(n) \setminus SO(n)$ es un conjunto conexo, homeomorfo a $SO(n)$ mediante el homeomorfismo:*

$$\begin{array}{ccc} SO(n) & \longrightarrow & \mathscr{O}(n) \setminus SO(n) \\ P & \longmapsto & U_0 P. \end{array}$$

En particular, $\mathscr{O}(n)$ tiene dos componentes conexas que son, simultaneamente abiertos y cerrados en $\mathscr{O}(n)$.

La estructura diferenciable de $\mathscr{O}(n)$ puede estudiarse de diversas formas. La más sencilla pasa por retomar las escuaciones que definen $\mathscr{O}(n)$ como subvariedad algebraica real del espacio afín real $\mathcal{M}_n(\mathbb{R})$ (de hecho, una subvariedad del producto de n esferas $(S^{n-1})^n = S^{n-1} \times ... \times S^{n-1}$).

LEMA 2.3.7. *Sea* $\mathrm{Sym}_n(\mathbb{R}) = \{A \in \mathcal{M}_n(\mathbb{R}) : A = A^*\}$ *el subespacio vectorial de* $\mathcal{M}_n(\mathbb{R})$ *de las matrices simétricas* $_{n\times n}$ *con coordenadas reales. Entonces,* $\mathrm{Sym}_n(\mathbb{R})$ *es un subespacio vectorial de dimensión* $\frac{n(n+1)}{2}$*. En particular, es una subvariedad algebraica, analítica y diferenciable y su espacio tangente puede identificarse con* $\mathrm{Sym}_n(\mathbb{R})$*.*

DEMOSTRACIÓN. La demostración se realiza de forma practicamente idéntica a la realizada para las matrices $A_n(\mathbb{R})$, en este caso tomamos las siguientes matrices como base de $\mathrm{Sym}_n(\mathbb{R})$ como espacio vectorial

$$\{A_{ij} : i \leq j, 1 \leq i \leq n\},$$

donde $A_{ij} = \left(a_{r,s}^{(i,j)}\right)_{1\leq r,s\leq n}$ es una matriz cuyas coordenadas tienen la forma siguiente:

$$a_{r,s}^{(i,j)} = \begin{cases} 0 & si \quad (r,s) \neq (i,j),(j,i) \\ 1 & si \qquad (r,s) = (i,j) \\ 1 & si \qquad (r,s) = (j,i). \end{cases}$$

De esta forma es claro que $\mathrm{Sym}_n(\mathbb{R})$ es un subespacio vectorial de $\mathcal{M}_n(\mathbb{R})$ de dimensión $\frac{n(n+1)}{2}$ y su espacio tangente puede identificarse con $\mathrm{Sym}_n(\mathbb{R})$. □

Consideremos ahora la siguiente aplicación

$$\begin{array}{rccc} w\colon & GL(n,\mathbb{R}) & \longrightarrow & \mathrm{Sym}_n(\mathbb{R}) \\ & A & \longmapsto & A^* \cdot A. \end{array}$$

PROPOSICIÓN 2.3.8. *w es una aplicación de clase $\mathscr{C}^\infty$ entre un abierto de $\mathcal{M}_n(\mathbb{R})$ y la variedad $\mathrm{Sym}_n(\mathbb{R})$. Además, para cada $P \in GL(n,\mathbb{R})$ el rango de la aplicación diferenciable verifica:*

$$\operatorname{rank} D_p w = \frac{n(n+1)}{2}.$$

Más aún, $w(GL(n,\mathbb{R}))$ es la subvariedad de dimensión $\frac{n(n+1)}{2}$ de $\mathrm{Sym}_n(\mathbb{R})$ formada por las matrices regulares simétricas definidas positivas.

DEMOSTRACIÓN. En primer lugar, es claro que w es una función de clase $\mathscr{C}^\infty$ y su imagen está contenida en $\mathrm{Sym}_n(\mathbb{R})$. Más aún, por el Teorema de Existencia de Descomposición en Valores Singulares, si $A \in GL(n,\mathbb{R})$, los valores propios de A^*A son, precisamente, los cuadrados de los valores singulares de A. Por tanto, $w(GL(n,\mathbb{R}))$ está contenido en el conjunto de las matrices simétricas semi-definidas positivas. Ahora bien, si $A \in GL(n,\mathbb{R})$, entonces A^*A es también una matriz regular. Por tanto, $w(GL(n,\mathbb{R}))$ está contenido en las matrices simétricas definidas positivas. Es decir,

$$w(GL(n,\mathbb{R})) \subseteq \mathrm{Sym}_n^+(\mathbb{R}) = \{A \in \mathrm{Sym}_n(\mathbb{R}) : \forall x \in \mathbb{R}^n, x \neq 0, \langle Ax, x\rangle > 0\}.$$

Probemos que $\mathrm{Sym}_n^+(\mathbb{R})$ es un abierto en $\mathrm{Sym}_n(\mathbb{R})$ y, por tanto, una subvariedad diferenciable de dimensión $\frac{n(n+1)}{2}$.
Por ello, obsérvese que una matriz $A \in \mathrm{Sym}_n(\mathbb{R})$ es definida positiva si y solamente si

$$\langle Ax, x\rangle > 0, \forall x \in S^{n-1}.$$

Consideremos la función continua sobre el compacto S^{n-1} siguiente:

$$\begin{array}{rccc} \langle A\cdot,\cdot\rangle\colon & S^{n-1} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \langle Ax, x\rangle. \end{array}$$

Si $A \in \mathrm{Sym}_n^+(\mathbb{R})$ entonces, $\langle A\cdot,\cdot\rangle$ alcanza un mínimo en un punto $\delta \in (0,+\infty)$. En otras palabras, existe $\delta > 0$ tal que

$$\langle Ax, x\rangle \geq \delta, \forall x \in S^{n-1}.$$

Recordemos la norma de A como operador $\| A \|_2$, que es equivalente a la norma de Frobenius (véase Proposición 1.2.3). Sea $B \in \mathrm{Sym}_n(\mathbb{R})$ una matriz cualquiera tal que

$$\| B - A \|_2 < \epsilon = \delta/2.$$

Entonces, para cada $x \in S^{n-1}$

$$|\langle Bx, x\rangle - \langle Ax, x\rangle| = |\langle (B-A)x, x\rangle| \leq \| B - A \|_2 \| x \|^2 < \epsilon.$$

Por tanto, $\forall x \in S^{n-1}$

$$\langle Bx, x\rangle \geq \langle Ax, x\rangle - \epsilon \geq \delta - \epsilon \geq \delta/2 > 0,$$

y B será una matriz simétrica real definida positiva. Esto prueba que $\mathrm{Sym}_n^+(\mathbb{R})$ es un abierto en $\mathrm{Sym}_n(\mathbb{R})$ y tenemos

$$w : GL(n,\mathbb{R}) \longrightarrow \mathrm{Sym}_n^+(\mathbb{R}).$$

Veamos, además, que w es suprayectiva. Para ello, usaremos el hecho de que una matriz simétrica admite una diagonalización salvo congruencia ortogonal. Esto es, si $S \in \mathrm{Sym}_n(\mathbb{R})$ existe $P \in \mathscr{O}(n)$ y D una matriz diagonal tal que

$$S = PDP^*.$$

Supongamos $D = Diag(d_1, ..., d_n)$. Si $S \in \mathrm{Sym}_n^+(\mathbb{R})$ es una matriz regular definida positiva, entonces, sus valores propios son todos positivos. Es decir, $d_i > 0, 1 \leq i \leq n$. Definamos la matriz raíz cuadrada de D como la matriz diagonal $\sqrt{D} = Diag(\sqrt{d_1}, ..., \sqrt{d_n})$ y definamos

$$A = P^*\sqrt{D}P.$$

Es claro que $A \in GL(n,\mathbb{R})$ y, por tanto,

$$w(A) = (P^*\sqrt{D}P)^* \cdot (P^*\sqrt{D}P) = PDP^* = S,$$

con lo que hemos probado que w es suprayectiva en un abierto del espacio afin $\mathrm{Sym}_n(\mathbb{R}) = \mathbb{R}^{\frac{n(n+1)}{2}}$. Ahora observemos que la diferencial de w es una matriz A_0 toma la forma

$$\begin{array}{rccc} D_{A_0}w\colon & T_{A_0}GL(n,\mathbb{R}) = T_{A_0}\mathcal{M}_n(\mathbb{R}) & \longrightarrow & T_{w(A_0)}\mathrm{Sym}_n^+(\mathbb{R}) \\ & \dot{A} & \longmapsto & \dot{A}^*A_0 + A_0^*\dot{A}. \end{array}$$

Nótese que si A_0 es una matriz regular se tiene que

$$D_{A_0}w(\dot{A}) = D_{Id_n}w(A_0^*\dot{A}).$$

Por tanto, es fácil verificar que el rango de $D_{A_0}w$ es independiente de A_0 y coincide con el rango de la imagen de $D_{Id_n}w$.
Observemos la forma particular de $D_{Id_n}w$

$$\begin{array}{rccc} D_{Id_n}w\colon & T_{Id_n}GL(n,\mathbb{R}) = \mathcal{M}_n(\mathbb{R}) & \longrightarrow & \mathrm{Sym}_n(\mathbb{R}) = T_{Id_n}\mathrm{Sym}_n(\mathbb{R}) \\ & \dot{A} & \longmapsto & \dot{A}^* + \dot{A}. \end{array}$$

Es claro que el núcleo de $D_{Id_n}w$ es, precisamente el conjunto de las matrices anti-simétricas:

$$\ker(D_{Id_n}w) = \{\dot{A} \in \mathcal{M}_n(\mathbb{R}) : \dot{A}^* + \dot{A} = 0\} = A_n.$$

Como $\dim(A_n) = \frac{n(n-1)}{2}$, por el primer Teorema de isomorfía, tendremos que

$$\dim Im(D_{Id_n}w) = \dim(\mathcal{M}_n(\mathbb{R})) - \dim(A_n) = n^2 - \frac{n(n-1)}{2} = \frac{n(n+1)}{2}.$$

□

Corolario 2.3.9. *El conjunto de las matrices ortogonales $\mathscr{O}(n) \subseteq \mathcal{M}_n(\mathbb{R})$ es una subvariedad diferenciable de dimensión $\frac{n(n-1)}{2}$. Además, para cada $P \in \mathscr{O}(n)$, podemos identificar su espacio tangente con:*

$$T_P\mathscr{O}(n) = \ker(D_pw) = \{\dot{A} \in \mathcal{M}_n(\mathbb{R}) : \dot{A}^*P + P^*\dot{A} = 0\}.$$

Demostración. Por la Proposición precedente, la aplicación $w : GL(n,\mathbb{R}) \to \mathrm{Sym}_n^+(\mathbb{R})$ es una aplicación de clase $\mathscr{C}^\infty$ suprayectiva y para cada $P \in GL(n,\mathbb{R})$ se tiene que

$$\mathrm{rank}(D_pw) = \frac{n(n+1)}{2} = \dim(\mathrm{Sym}_n(\mathbb{R})) = \dim T_{w(P)}\mathrm{Sym}_n^+(\mathbb{R}).$$

Por tanto, w es una submersión en todo punto $P \in GL(n,\mathbb{R})$, que es un abierto en $\mathcal{M}_n(\mathbb{R})$. En particular, por la Proposición A.5.2 del Apéndice A, para cada $B \in \mathrm{Sym}_n^+(\mathbb{R})$ la fibra $w^{-1}(\{B\})$ es una subvariedad diferenciable de $GL(n,\mathbb{R})$ (y de $\mathcal{M}_n(\mathbb{R})$) cuya dimensión satisface

$$\dim w^{-1}(\{B\}) = \dim \mathcal{M}_n(\mathbb{R}) - \dim T_B\mathrm{Sym}_n(\mathbb{R}) = n^2 - \left(\frac{n(n+1)}{2}\right) = \frac{n(n-1)}{2}.$$

Además, el espacio tangente a un punto $P \in w^{-1}(\{B\})$ es precisamente el núcleo $\ker(D_pw)$.
Tomemos $B = Id_n$ y observamos que la fibra $w^{-1}(\{Id_n\})$ es justamente el grupo ortogonal

$$\mathscr{O}(n) = w^{-1}(\{Id_n\}) = \{A \in GL(n,\mathbb{R}) : A^*A = Id_n\},$$

con $\dim \mathscr{O}(n) = \frac{n(n-1)}{2}$ como subvariedad diferenciable de $\mathcal{M}_n(\mathbb{R})$.
Además, para cada $P \in \mathscr{O}(n)$, el espacio tangente verifica:

$$T_P\mathscr{O}(n) = \ker(D_p w) = \{\dot{A} \in \mathcal{M}_n(\mathbb{R}) : \dot{A}^* P + P^* \dot{A} = 0\}.$$

□

Podemos caracterizar el espacio tangente $T_P\mathscr{O}(n)$ de forma alternativa usando exp.

PROPOSICIÓN 2.3.10. *Con las notaciones precedentes, para cada $P \in \mathscr{O}(n)$, el espacio tangente puede identificarse con*

$$T_P\mathscr{O}(n) = \{A \cdot P : A \in A_n\}.$$

DEMOSTRACIÓN. Probémoslo en primer lugar para matrices en $SO(n)$. Es claro que el conjunto $\{\dot{A}P : \dot{A} \in A_n\} \subseteq T_P\mathscr{O}(n)$ verificando la identidad del Colorario precedente:

$$(\dot{A}P)^* P + P^*(\dot{A}P) = P^*\dot{A}^* P + P^*\dot{A}P = P^*(\dot{A}^* + \dot{A})P = 0,$$

para cada matriz anti-simétrica $\dot{A} \in A_n$. De otro lado, supongamos que $P \in SO(n)$. Por el Teorema 2.3.5 precedente,

$$\exp\colon A_n(\mathbb{R}) \longrightarrow SO(n)$$

es suprayectiva por lo que existe $A_0 \in A_n(\mathbb{R})$ tal que $P = \exp(A_0)$. Consideremos ahora la siguiente función diferenciable para una matriz $A \in A_n(\mathbb{R})$

$$\begin{array}{rccc} f\colon & \mathbb{R} & \longrightarrow & SO(n) \\ & t & \longmapsto & \exp(A_0 + tA). \end{array}$$

Diferenciando con respecto a t obtendremos

$$f'(t) = A \cdot \exp(A_0 + tA),$$

con lo que

$$f'(0) = A \cdot \exp(A_0) = A \cdot P \in T_P\mathscr{O}(n).$$

Falta un caso, si $P \in \mathscr{O}(n) \setminus SO(n)$ es una matriz de determinante -1, tenemos el homeomorfismo entre dos abiertos de $\mathscr{O}(n)$

$$\begin{array}{rccc} \varphi_0\colon & SO(n) & \longrightarrow & \mathscr{O}(n) \setminus SO(n) \\ & Q & \longmapsto & Q \cdot U_0, \end{array}$$

donde $U_0 = \left(\begin{array}{cc|c} 0 & 1 & \\ 1 & 0 & \\ \hline & & Id_{n-2} \end{array}\right) \in \mathscr{O}(n) \setminus SO(n)$ es la matriz antes definida.
Se observa que φ_0 es también un difeomorfismo entre los respectivos abiertos, por lo que tenemos un isomorfismo entre los respectivos espacios tangentes. Es decir dado $P \in \mathscr{O}(n) \setminus SO(n)$ y dado $Q \in SO(n)$ tal que $P = Q \cdot U_0$, tendremos que

$$T_P\mathscr{O}(n) = D\varphi_0(Q)(T_Q\mathscr{O}(n)) = D\varphi_0(Q)(\{\dot{A}Q : \dot{A} \in A_n(\mathbb{R})\}).$$

Pero $D\varphi_0(Q)(\dot{B}) = \dot{B} \cdot U_0$ para cada $\dot{B} \in T_Q(\mathscr{O}(n))$. Por tanto,

$$T_Q\mathscr{O}(n) = \{\dot{A}QU_0 : \dot{A} \in A_n(\mathbb{R})\} = \{\dot{A}P : \dot{A} \in A_n(\mathbb{R})\}.$$

□

Finalmente, fijemos la estructura Riemanniana del grupo ortogonal. En $\mathcal{M}_n(\mathbb{R})$ disponemos del producto escalar de Frobenius.

$$\begin{array}{rccc} \langle \cdot, \cdot \rangle_F\colon & \mathcal{M}_n(\mathbb{R}) \times \mathcal{M}_n(\mathbb{R}) & \longrightarrow & \mathbb{R} \\ & (A, B) & \longmapsto & Tr(AB^*). \end{array}$$

Consideraremos en $T_P\mathscr{O}(n)$ el producto escalar inducido por este producto de Frobenius

$$\begin{array}{rccc} \langle \cdot, \cdot \rangle_F\colon & T_P\mathscr{O}(n) \times T_P\mathscr{O}(n) & \longrightarrow & \mathbb{R} \\ & (\dot{A}P, \dot{B}P) & \longmapsto & Tr((\dot{A}PP^*\dot{B}^*) = Tr(\dot{A}\dot{B}^*). \end{array}$$

PROPOSICIÓN 2.3.11. *El grupo ortogonal $(\mathscr{O}(n),\cdot)$ es una subvariedad de Riemann de $\mathcal{M}_n(\mathbb{R})$. Los espacios tangentes $T_P\mathscr{O}(n)$ están dotados de la estructura de espacio euclídeo inducido por el producto escalar de Frobenius, $\langle\cdot,\cdot\rangle_F$. Más aún, todos los espacios tangentes $T_P\mathscr{O}(n)$ son isométricos entre sí mediante la identificación*

$$\begin{array}{ccc} T_P\mathscr{O}(n) & \longrightarrow & T_Q\mathscr{O}(n) \\ \dot{A} & \longmapsto & \dot{A}P^*Q. \end{array}$$

DEMOSTRACIÓN. Basta con observar que multiplicar por P^*Q es una isometría. □

OBSERVACIÓN 2.3.12. De hecho, todos los espacios tangentes serán isométricos al espacio vectorial $A_n(\mathbb{R}) = T_{Id_n}\mathscr{O}(n)$ con la métrica inducida por el producto escalar de Frobenius $\langle\cdot,\cdot\rangle_F$.

TEOREMA 2.3.13. *Para cada matriz ortogonal $P \in \mathscr{O}(n)$ una base ortonormal del espacio tangente $T_P\mathscr{O}(n)$ es dada por la siguiente descripción:*

$$\mathscr{B}_P = \left\{ \frac{1}{\sqrt{2}} E_{i,j} \cdot P : 1 \leq i < j \leq n \right\},$$

donde $E_{ij} = \left(a_{r,s}^{(i,j)}\right)_{1\leq r,s\leq n}$ es una matriz cuyas coordenadas tienen la forma siguiente:

$$a_{r,s}^{(i,j)} = \begin{cases} 0 & si \quad (r,s) \neq (i,j),(j,i) \\ 1 & si \quad (r,s) = (i,j) \\ -1 & si \quad (r,s) = (j,i). \end{cases}$$

DEMOSTRACIÓN. Es claro que todas las matrices $\frac{1}{\sqrt{2}}E_{i,j}$ son matrices antisimétricas, por lo que $\mathscr{B}_P \in T_P\mathscr{O}(n)$. Más aun, es claro que la norma de Frobenius satisface:

$$\| \frac{1}{\sqrt{2}} E_{i,j} \|_F = \frac{1}{\sqrt{2}} \| E_{i,j} \|_F = \frac{1}{\sqrt{2}} \cdot \sqrt{2} = 1.$$

Por tanto, queda por demostrar simplemente que $Tr(E_{i,j}\cdot(E_{l,m})^*) = 0$ para todo $(i,j) \neq (l,m)$. Escribamos el producto

$$(E_{i,j} \cdot (E_{l,m})^*) = (C_{r,s})_{1\leq r,s\leq n}. \tag{2.3.3}$$

Y veamos que todos los elementos de la diagonal de estas matrices son 0 si $(i,j) \neq (l,m)$. Es decir, veamos que

$$c_{r,r} = 0 \ , \ 1 \leq r \leq n.$$

Ahora bien, tenemos

$$c_{r,r} = \sum_{k=1}^{n} a_{r,k}^{(i,j)} \cdot a_{r,k}^{(l,m)},$$

donde ya hemos tenido en cuenta que es el producto de $E_{i,j}$ por la traspuesta de $E_{l,m}$.
Ahora, consideremos todos los casos posibles: Si $r \neq i,j$, entonces $a_{r,k}^{(i,j)} = 0$ y, en ese caso, $c_{r,r} = 0$.
Supongamos $r = i$. En ese caso $a_{r,k}^{(i,j)} = a_{i,k}^{(i,j)} = 0$ para todo $k \neq j$ con lo que

$$c_{r,r} = c_{i,i} = a_{i,j}^{(i,j)} \cdot a_{i,j}^{(l,m)}.$$

Ahora bien $a_{i,j}^{(l,m)} = 0$ si $(l,m) \neq (i,j),(j,i)$. Pero, además, como $l < m$ e $i < j$, no es posible que $(l,m) = (j,i)$. Por tanto, si $(l,m) \neq (i,j)$, necesariamente $(l,m) \neq (j,i)$ y $a_{i,j}^{(l,m)} = 0$. Es decir,

$$\text{si } r = i, c_{rr} = c_{i,i} = 0.$$

Supongamos finalmente, que $r = j$. En este caso $a_{r,k}^{(i,j)} = a_{j,k}^{(i,j)} \neq 0$ para todo $k \neq i$. Así pues

$$c_{r,r} = c_{j,j} = a_{j,i}^{(i,j)} \cdot a_{j,i}^{(l,m)} = -a_{j,i}^{(l,m)}.$$

Ahora bien, $(l,m) \neq (i,j)$ luego $(j,i) \neq (m,l)$ y, en conclusión, $-a_{j,i}^{(l,m)} = 0$ (porque $(j,i) \neq (l,m)$ y $(l,m) \neq (i,j)$). En conclusión

$$\text{si } r = j, c_{rr} = c_{j,j} = 0.$$

Hemos concluido así que todos los elementos de la diagonal principal del producto $(E_{i,j}\cdot(E_{l,m})^*)$ son ceros. En particular, si $(i,j)\neq(l,m)$,

$$\langle E_{i,j},E_{l,m}\rangle_F = Tr(E_{i,j}\cdot(E_{l,m})^*) = \sum_{k=1}^{n} 0 = 0.$$

En suma, las matrices de la familia $\mathscr{B}_P$ son dos a dos ortogonales, cada una tiene norma de Frobenius 1 y, finalmente,

$$\#\mathscr{B}_P = \#\{(i,j) : i<j, 1\leq i,j\leq n\} = \frac{n(n-1)}{2}.$$

Por tanto, $\mathscr{B}_P$ es una base ortonormal de $T_P\mathscr{O}(n)$ y queda probado el Teorema. □

2.4. El jacobiano normal de la parametrización desde la SVD

Sea $r\in\mathbb{N}$ y sea $S_r^+ = \{\mathrm{Diag}(\sigma_1 > ... > \sigma_r) : \sigma_r > 0, \sigma_1^2 + ... + \sigma_r^2 = 1\}$ la esfera unidad en el conjunto de las matrices diagonales con elementos reales, positivos y ordenados en orden decreciente. Sea

$$\begin{array}{rccc} \varphi\colon & \mathscr{O}(n)\times S_r^+\times\mathscr{O}(m) & \longrightarrow & \sum_r\setminus\sum_{r-1} \\ & (U,D,V) & \longmapsto & U\begin{pmatrix} D & 0\\ 0 & 0\end{pmatrix}V^*. \end{array} \tag{2.4.1}$$

Queremos usar la fórmula de la Co-área (A.6.7) para φ, así que debemos calcular el jacobiano normal $NJ_{(U,D,V)}\varphi$. Comenzaremos viendo que dicho jacobiano normal solo depende de D.

LEMA 2.4.1. *Sea $D\in S_r^+$, el valor de $NJ_{(U_0U,D,VV_0)}\varphi$ es invariante para cualesquiera $U_0\in\mathscr{O}(n)$ y $V_0\in\mathscr{O}(m)$.*

DEMOSTRACIÓN. Sean $U_0\in\mathscr{O}(n)$ y $V_0\in\mathscr{O}(m)$. Notemos que las siguientes dos aplicaciones son isometrías:

$$\begin{array}{rccc} \varphi_M\colon & \mathscr{O}(n)\times S_r^+\times\mathscr{O}(m) & \longrightarrow & \mathscr{O}(n)\times S_r^+\times\mathscr{O}(m) \\ & (U,D,V) & \longmapsto & (U_0U,D,V_0V), \end{array}$$

$$\begin{array}{rccc} \varphi_N\colon & \sum_r\setminus\sum_{r-1} & \longrightarrow & \sum_r\setminus\sum_{r-1} \\ & A & \longmapsto & U_0\cdot A\cdot V_0^*. \end{array}$$

Además, conmutan en el sentido que $\varphi_N\circ\varphi = \varphi\circ\varphi_M$. Por el Corolario A.6.6 del Apéndice A, esto implica que $NJ_{(U,D,V)}\varphi = NJ_{(U_0U,D,VV_0)}\varphi$ para cualesquiera (U,D,V) y (U_0,V_0). □

De esta forma $NJ_{(U,D,V)}\varphi$ solo depende de D y podemos tomar U,V matrices identidad y calcular $NJ_{(Id_n,D,Id_m)}\varphi$.

LEMA 2.4.2. *El siguiente conjunto es una base del espacio tangente $T_{(Id_n,D,Id_m)}(\mathscr{O}(n)\times S_r^+\times\mathscr{O}(m))$:*

$$\{(E_{ij},0,0) : 1\leq i<j<n\}\cup\{(0,0,E_{ij}) : 1\leq i<j<m\}\cup\{(0,R_j,0) : 1\leq j<r-1\}$$

donde $\{R_j\}_{1\leq j<r-1}$ es una base de $<D>^\perp$ y las matrices E_{ij} son definidas como en el Teorema (2.3.13) y su dimensión dependerá del contexto.

DEMOSTRACIÓN. El espacio tangente

$$T_{(Id_n,D,Id_m)}(\mathscr{O}(n)\times S_r^+\times\mathscr{O}(m)) = T_{Id_n}\mathscr{O}(n)\times T_DS_r^+\times T_{Id_m}\mathscr{O}(m),$$

es decir,

$$T_{(Id_n,D,Id_m)}(\mathscr{O}(n)\times S_r^+\times\mathscr{O}(m)) = \{(A,R,B) : A\in T_{Id_n}\mathscr{O}(n), B\in T_{Id_m}\mathscr{O}(m), R\in T_DS_r^+\}.$$

De las Proposiciones (2.2.1) y (2.3.10) tenemos que

$$T_{(Id_n,D,Id_m)}(\mathscr{O}(n)\times S_r^+\times\mathscr{O}(m)) = \{(A,R,B) : A\in A_n, B\in A_m, R\in<D>^\perp\}.$$

De esta forma, nombrando por $\mathscr{A}$ una base de A_n, por $\mathscr{B}$ una base de A_m y por $\mathscr{C}$ una base de $<D>^\perp$, tendremos que una base de $T_{(Id_n,D,Id_m)}(\mathscr{O}(n)\times S_r^+\times\mathscr{O}(m))$ estará formada por la unión de los tres conjuntos siguientes:

$$\{(a,0,0) : a\in\mathscr{A}\}\cup\{(0,0,b) : b\in\mathscr{B}\}\cup\{(0,c,0) : c\in\mathscr{C}\}$$

Por lo que tomando $\{R_j\}_{1\leq j<r-1}$ como base de $< D >^{\perp}$, y tomando las matrices E_{ij} como se describen en el Teorema (2.3.13), donde su dimensión dependerá del contexto, tendremos finalmente que el siguiente conjunto es una base de $T_{(Id_n,D,Id_m)}(\mathscr{O}(n)\times S_r^+\times\mathscr{O}(m))$

$$\{(E_{ij},0,0):1\leq i<j<n\}\cup\{(0,0,E_{ij}):1\leq i<j<m\}\cup\{(0,R_j,0):1\leq j<r-1\}$$

□

Observación 2.4.3. Podemos expresar la aplicación diferencial en el punto (Id_n,D,Id_m) mediante

$$d_{(Id_n,D,Id_m)}\varphi(A,R,B)=A\begin{pmatrix}D&0\\0&0\end{pmatrix}+\begin{pmatrix}R&0\\0&0\end{pmatrix}+\begin{pmatrix}D&0\\0&0\end{pmatrix}B^*.$$

Veamos las posibles imágenes de nuestra base del espacio tangente por la aplicación diferencial. En lo que sigue para cualquier par de enteros positivos i,j, la matriz $\delta_{i,j}$ será de n filas y m columnas con todas sus entradas nulas excepto la posición (i,j) donde tendrá valor 1.

- Si $1\leq i<j\leq r$, $d_{(Id_n,D,Id_m)}\varphi(E_{ij},0,0)=E_{ij}\begin{pmatrix}D&0\\0&0\end{pmatrix}=\delta_{ij}\sigma_j-\delta_{ji}\sigma_i$.
- Si $1\leq i\leq r<j\leq n$, $d_{(Id_n,D,Id_m)}\varphi(E_{ij},0,0)=E_{ij}\begin{pmatrix}D&0\\0&0\end{pmatrix}=-\delta_{ji}\sigma_i$.
- Si $1\leq r<i<j\leq n$, $d_{(Id_n,D,Id_m)}\varphi(E_{ij},0,0)=E_{ij}\begin{pmatrix}D&0\\0&0\end{pmatrix}=0$.
- Si $1\leq i<j\leq r$, $d_{(Id_n,D,Id_m)}\varphi(0,0,E_{ij})=\begin{pmatrix}D&0\\0&0\end{pmatrix}E_{ij}^*=-\delta_{ij}\sigma_i+\delta_{ji}\sigma_j$.
- Si $1\leq i\leq r<j\leq m$, $d_{(Id_n,D,Id_m)}\varphi(0,0,E_{ij})=\begin{pmatrix}D&0\\0&0\end{pmatrix}E_{ij}^*=-\delta_{ij}\sigma_i$.
- Si $1\leq r<i<j\leq m$, $d_{(Id_n,D,Id_m)}\varphi(0,0,E_{ij})=\begin{pmatrix}D&0\\0&0\end{pmatrix}E_{ij}^*=0$.
- $d_{(Id_n,D,Id_m)}\varphi(0,R_j,0)=\begin{pmatrix}R_j&0\\0&0\end{pmatrix}$.

De esta forma, el siguiente resultado nos ofrece una base ortonormal de $\ker(d_{(Id_n,D,Id_m)}\varphi)^{\perp}$.

Lema 2.4.4. *Una base ortonormal de* $\ker(d_{(Id_n,D,Id_m)}\varphi)^{\perp}$ *es la formada por los siguientes elementos:*

i) $\frac{1}{2}(E_{ij},0,E_{ij}):1\leq i<j\leq r$,
ii) $\frac{1}{2}(E_{ij},0,-E_{ij}):1\leq i<j\leq r$,
iii) $\frac{1}{\sqrt{2}}(E_{ij},0,0):1\leq i\leq r<j\leq n$,
iv) $\frac{1}{\sqrt{2}}(0,0,E_{ij}):1\leq i\leq r<j\leq m$,
v) $(0,R_j,0)$ *donde* $\{R_j\}_{1\leq j<r-1}$ *es una base ortonormal de* $< D >^{\perp}$

Demostración. Notemos, a partir de las imágenes de $d_{(Id_n,D,Id_m)}\varphi$ de nuestra base del espacio tangente del Lema 2.4.2 precedente, junto con las bases ortonormales del espacio tangente $T_P\mathscr{O}(n)$ para cualquier matriz ortogonal P que vimos en el Teorema (2.3.13), la siguiente es una base ortonormal de $\ker(d_{(Id_n,D,Id_m)}\varphi)^{\perp}$

i) $\frac{1}{\sqrt{2}}(E_{ij},0,0):1\leq i<j\leq r$,
ii) $\frac{1}{\sqrt{2}}(0,0,E_{ij}):1\leq i<j\leq r$,
iii) $\frac{1}{\sqrt{2}}(E_{ij},0,0):1\leq i\leq r<j\leq n$,
iv) $\frac{1}{\sqrt{2}}(0,0,E_{ij}):1\leq i\leq r<j\leq m$,
v) $(0,R_j,0)$ con $\{R_j\}_{1\leq j<r-1}$ una base ortonormal de $< D >^{\perp}$.

Es evidente que dicha base es equivalente a la citada en este Lema, luego solo nos falta comprobar que sea ortonormal. Por tanto, basta con comprobar que los siguientes elementos de la base son ortonormales

i) $\frac{1}{2}(E_{ij},0,E_{ij}):1\leq i<j\leq r$,
ii) $\frac{1}{2}(E_{ij},0,-E_{ij}):1\leq i<j\leq r$.

Es evidente que dados dos elementos de la base cuyos índices i,j difieran, dichos elementos son ortogonales, veamos que pasa cuando coinciden. Es decir, si un elemento es del tipo i) y otro es de tipo ii), esto es, tenemos los elementos $\frac{1}{2}(E_{ij},0,E_{ij})$ y $\frac{1}{2}(E_{ij},0,-E_{ij})$, su producto escalar

es cero ya que se anulan los productos de las matrices E_{ij} respectivamente. Para finalizar, ver que estos elementos son de norma uno es mera comprobación. □

TEOREMA 2.4.5. *Sean* $U \in \mathscr{O}(n)$, $V \in \mathscr{O}(m)$, $D \in S_r^+$ *y sea* $A = UDV^* \in \sum_r \setminus \sum_{r-1}$. *Sea* $NJ_{(U,D,V)}\varphi$ *el jacobiano normal de la parametrización* φ *de la Descomposición en Valores Singulares. Entonces,*

$$NJ_{(U,D,V)}\varphi = \frac{\det(D)^{n+m-2r}\Delta(D)}{(\sqrt{2})^{r(n+m-r-1)}}.$$

donde $\det(D) = \prod_{i=1}^{r} \sigma_i$ *es el determinante y* $\Delta(D) = \prod_{i<j}(\sigma_i^2 - \sigma_j^2)$ *es el discriminante.*

DEMOSTRACIÓN. Tomamos la siguiente base ortonormal para la variedad de la imagen de la aplicación diferencial:

i) $\delta_{ij} : 1 \le i < j \le r$,
ii) $\delta_{ji} : 1 \le i < j \le r$,
iii) $\delta_{ij} : 1 \le i \le r < j \le m$,
iv) $\delta_{ji} : 1 \le i \le r < j \le n$,
v) $\begin{pmatrix} R_j & 0 \\ 0 & 0 \end{pmatrix}$ siendo R_j perteneciente a la base ortonormal antes descrita.

Donde al igual que en la Observación 2.4.3, para cualquier par de enteros positivos i, j, la matriz $\delta_{i,j}$ es de n filas y m columnas con todas sus entradas nulas excepto la posición (i, j) donde tiene valor 1.

Ahora, podemos expresar la aplicación diferencial $d_{(Id_n,D,Id_m)}\varphi$ restringida a $\ker(d_{(Id_n,D,Id_m)}\varphi)^\perp$ a través de las imágenes de la base calculada en el Lema 2.4.4, expresadas en la base de la imagen que acabamos de desarrollar para un cierto orden de la misma, mediante la siguiente matriz:

$$\Psi = \left(\begin{array}{cc|c} A & 0 & \\ 0 & B & 0 \\ \hline \multicolumn{2}{c|}{0} & Id \end{array}\right),$$

siendo Id la matriz identidad de tamaño $r-1$ y las matrices A y B son las que siguen:

$$A = \frac{1}{2} \cdot \begin{pmatrix} \sigma_1+\sigma_2 & \sigma_2-\sigma_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\sigma_1-\sigma_2 & \sigma_2-\sigma_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_i+\sigma_j & \sigma_j-\sigma_i & 0 & 0 & 0 \\ 0 & 0 & 0 & -\sigma_i-\sigma_j & \sigma_j-\sigma_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{r-1}+\sigma_r & \sigma_r-\sigma_{r-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & -\sigma_{r-1}-\sigma_r & \sigma_r-\sigma_{r-1} \end{pmatrix},$$

$$B = -\frac{1}{\sqrt{2}} \cdot \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_r & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_r \end{pmatrix},$$

$$\underbrace{\qquad\qquad}_{m+n-2r} \qquad\qquad \underbrace{\qquad\qquad}_{m+n-2r}$$

donde la matriz A se corresponde a las imágenes de $\{\frac{1}{2}(E_{ij}, 0, E_{ij}) : 1 \le i < j \le r\}$, y $\{\frac{1}{2}(E_{ij}, 0, -E_{ij}) : 1 \le i < j \le r\}$, la matriz B a las imágenes de $\{\frac{1}{\sqrt{2}}(E_{ij}, 0, 0) : 1 \le i \le r < j \le n\}$, y $\{\frac{1}{\sqrt{2}}(0, 0, E_{ij}) : 1 \le i \le r < j \le m\}$, y la matriz Id a las imágenes de $(0, R_j, 0)$ con

$R_j \in \{R_j\}_{1\leq j<r-1}$.
De esta forma, $NJ_{(Id_n,D,Id_m)}\varphi$ sera el valor absoluto del determinante de la matriz Ψ, es decir,

$$NJ_{(Id_n,D,Id_m)}\varphi = |\det(\Psi)| = |\det(A)| \cdot |\det(B)|.$$

es fácil ver que $|\det(B)| = \frac{1}{\sqrt{2}^{r(m+n-2r)}} \prod_{i=1}^{r} \sigma_i^{m+n-2r}$, por otro lado, tenemos que $|\det(A)| = |\det(\overline{A})|$, siendo $\overline{A}$ la matriz resultante de sumar las filas de A dos a dos, es decir

$$\overline{A} = \frac{1}{2} \cdot \begin{pmatrix} \sigma_1+\sigma_2 & \sigma_2-\sigma_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2(\sigma_2-\sigma_1) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_i+\sigma_j & \sigma_j-\sigma_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2(\sigma_j-\sigma_i) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{r-1}+\sigma_r & \sigma_r-\sigma_{r-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2(\sigma_r-\sigma_{r-1}) \end{pmatrix},$$

y es fácil ver que,

$$|\det(\overline{A})| = \frac{1}{2^{\frac{r(r-1)}{2}}} \cdot \prod_{i<j\leq r} (\sigma_i^2 - \sigma_j^2).$$

Por lo tanto,

$$NJ_{(Id_n,D,Id_m)}\varphi = \frac{1}{\left(\sqrt{2}\right)^{r(m+n-r+1)}} \prod_{i=1}^{r} \sigma_i^{m+n-2r} \prod_{i<j\leq r} (\sigma_i^2 - \sigma_j^2).$$

□

2.5. El volumen de la fibra de la parametrización SVD de una matriz

Sea

$$\begin{array}{rccl} \varphi\colon & \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m) & \longrightarrow & \sum_r \setminus \sum_{r-1} \\ & (U,D,V) & \longmapsto & U \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} V^*, \end{array}$$

tal y como la definimos en la sección anterior, nuevamente con nuestro objetivo de usar la fórmula de la Co-área (A.6.7) para φ, ahora, dada una matriz $A \in \sum_r \setminus \sum_{r-1}$, debemos calcular el volumen de la fibra $\varphi^{-1}(A)$, es decir, $\mathrm{Vol}[\varphi^{-1}(A)]$.

Teorema 2.5.1. *La cantidad* $\mathrm{Vol}[\varphi^{-1}(A)]$ *es independiente de* $A \in \sum_r \setminus \sum_{r-1}$*, excepto para un conjunto de medida cero. En efecto, sea* $A \in \sum_r \setminus \sum_{r-1}$ *tal que todos sus* r *valores singulares distintos de cero sean diferentes. Entonces tenemos que*

$$\mathrm{Vol}[\varphi^{-1}(A)] = 2^r \mathrm{Vol}[\mathscr{O}(n-r)]\mathrm{Vol}[\mathscr{O}(m-r)].$$

Demostración. Notemos que los elementos de $\sum_r \setminus \sum_{r-1}$ son matrices de la forma $A = U \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} V^*$ con U, V ortogonales, sin embargo, dado que el producto de matrices ortogonales es una matriz ortogonal, bastará con probar esta fórmula para el caso $A = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$, con $D \in S_r^+$ y tal que todas sus entradas son diferentes. Ahora notemos que

$$\varphi^{-1}(A) = \{(U,D,V) \in \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m) : UAV^* = A\}.$$

Sea $(U,D,V) \in \varphi^{-1}(A)$. Escribimos $U = \begin{pmatrix} U_1 & U_2 \\ U_3 & U_4 \end{pmatrix}$, donde los bloques se ajustan a los de A. Entonces tenemos que

$$UA^2U^* = (UAV^*)(VA^*U^*) = AA^* = A^2,$$
$$UA^4U^* = (UA^2U^*)(UA^2U^*) = A^2A^2 = A^4.$$

Más generalmente, $UA^{2k}U^* = A^{2k}$ para todo entero positivo k, esto es,

$$\begin{pmatrix} U_1D^{2k}U_1^* & U_1D^{2k}U_3^* \\ U_3D^{2k}U_1^* & U_3D^{2k}U_3^* \end{pmatrix} = \begin{pmatrix} D^{2k} & 0 \\ 0 & 0 \end{pmatrix}.$$

En particular, $U_1 D^{2k} U_1^* = D^{2k}$ para $k \geq 1$, y por lo tanto, U_1 es invertible y

$$(U_1 D^2 U_1^*)(U_1 D^2 U_1^*) = D^4 = U_1 D^4 U_1^*.$$

Después de multiplizar a izquierda por U_1^{-1} y a derecha por $(U_1^*)^{-1}$, tenemos que $D^2 U_1^* U_1 D^2 = D^4$, es decir, $U_1^* U_1 = I_r$ y por tanto U_1 es ortogonal. Además U_3 ha de ser la matriz nula. Si ahora desarrollamos $U^* A^{2k} U = A^{2k}$ obtenemos

$$\begin{pmatrix} U_1^* D^{2k} U_1 & U_1^* D^{2k} U_2 \\ U_2^* D^{2k} U_1 & U_2 D^{2k} U_2 \end{pmatrix} = \begin{pmatrix} D^{2k} & 0 \\ 0 & 0 \end{pmatrix}.$$

Y obtenemos que U_2 también tiene que ser la matriz nula. Además, como U en sí misma una matriz ortogonal, esto implica que U_4 ha de ser ortogonal. Un argumento similar para V hace que concluyamos que

$$U = \begin{pmatrix} U_1 & 0 \\ 0 & U_4 \end{pmatrix},\ V = \begin{pmatrix} V_1 & 0 \\ 0 & V_4 \end{pmatrix},$$

donde $U_1, V_1 \in \mathscr{O}(r)$, $U_4 \in \mathscr{O}(n-r)$ y $V_4 \in \mathscr{O}(m-r)$. Finalmente, esto implica que $U_1 D V_1^* = D$ es una descomposición en valores singulares de D. Como todas las entradas de D son diferentes, la descomposición en valores singulares es única a excepción de constantes multiplicativas 1 y -1. Como $U_1 D V_1^* = D$ implica que $U_1 = \mathrm{Diag}(\alpha_1, ..., \alpha_r)$ y $V_1 = \mathrm{Diag}(\alpha_1, ..., \alpha_r)$ con $\alpha_j \in \{1, -1\}$, hemos probado que $\varphi^{-1}(A)$ es igual al siguiente conjunto:

$$\left\{ \left(\begin{pmatrix} \mathrm{Diag}(\alpha_j) & 0 \\ 0 & U_4 \end{pmatrix}, D, \begin{pmatrix} \mathrm{Diag}(\alpha_j) & 0 \\ 0 & V_4 \end{pmatrix} \right) : \alpha_j \in \{1, -1\}, U_4 \in \mathscr{O}(m-r), V_4 \in \mathscr{O}(n-r). \right\}.$$

Notemos que para los distintos valores que tomen los α_j generamos conjuntos independientes, por lo que podemos expresar el volumen de $\varphi^{-1}(A)$ mediante

$$\mathrm{Vol}[\varphi^{-1}(A)] = \mathrm{Vol}\left[\bigcup_{\epsilon \in \{\pm 1\}^r} \{\epsilon\} \times \mathscr{O}(n-r) \times \mathscr{O}(m-r) \right],$$

es decir, tenemos que

$$\mathrm{Vol}[\varphi^{-1}(A)] = 2^r \mathrm{Vol}[\mathscr{O}(n-r)] \mathrm{Vol}[\mathscr{O}(m-r)].$$

□

2.6. El uso de la fórmula de la Co-área para la parametrización SVD

En las dos secciones anteriores hemos ido calculando los ingredientes necesarios para poder usar la fórmula de la Co-área para nuestra parametrización SVD descrita en la ecuación (2.4.1).

TEOREMA 2.6.1. *Sea $A \in \sum_r \setminus \sum_{r-1}$ el punto proyectivo asociado a una matriz de rango r. Sea $\sigma(A) = (\sigma_1(A), ..., \sigma_r(A)) \in S_r^+$ la lista de sus valores singulares. Sea $\phi : \mathbb{R}^r \to \mathbb{R}$ una función medible. Entonces, se tiene:*

$$\int_{A \in \sum_r \setminus \sum_{r-1}} \phi(\sigma(A)) \mathrm{d}(\textstyle\sum_r \setminus \sum_{r-1}) = H(n,m,r) \cdot \displaystyle\int_{D \in S_r^+} \phi(D) (\det(D)^{n+m-2r} \Delta(D)) \mathrm{d}S_r^+,$$

donde

$$H(n,m,r) = \frac{\mathrm{Vol}[\mathscr{O}(n)] \mathrm{Vol}[\mathscr{O}(m)]}{\mathrm{Vol}[\mathscr{O}(n-r)] \mathrm{Vol}[\mathscr{O}(m-r)] \left(\sqrt{2}\right)^{r(n+m-r)+r}}.$$

DEMOSTRACIÓN. A partir de la Fórmula de la Co-área (Teorema A.6.7) tenemos que

$$\int_{(U,D,V) \in \mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m)} \phi(D) N J_{(U,D,V)} \varphi \mathrm{d}(\mathscr{O}(n) \times S_r^+ \times \mathscr{O}(m)) =$$

$$= \int_{A \in \sum_r \setminus \sum_{r-1}} \mathrm{Vol}[\varphi^{-1}(A)] \phi(\sigma(A)) \mathrm{d}(\textstyle\sum_r \setminus \sum_{r-1}).$$

Donde $N J_{(U,D,V)} \varphi$ ya lo hemos calculado en el Teorema 2.4.5, y a su vez, $\mathrm{Vol}[\varphi^{-1}(A)]$ lo hemos calculado en la Proposición 2.5.1. Ahora, aplicando el Teorema de Fubini a la parte izquierda de la igualdad y sustituyendo los valores que ya conocemos obtenemos

$$\mathrm{Vol}[\mathscr{O}(n)] \mathrm{Vol}[\mathscr{O}(m)] \int_{D \in S_r^+} \phi(D) \left(\frac{\det(D)^{n+m-2r} \Delta(D)}{\left(\sqrt{2}\right)^{r(n+m-r-1)}} \right) \mathrm{d}S_r^+ =$$

$$= 2^r \text{Vol}[\mathscr{O}(n-r)]\text{Vol}[\mathscr{O}(m-r)] \int_{A\in\sum_r \setminus \sum_{r-1}} \phi(\sigma(A))\text{d}(\sum_r \setminus \sum_{r-1}),$$

o lo que es lo mismo

$$\int_{A\in\sum_r \setminus \sum_{r-1}} \phi(\sigma(A))\text{d}(\sum_r \setminus \sum_{r-1}) = H(n,m,r)\cdot \int_{D\in S_r^+} \phi(D)(\det(D)^{n+m-2r}\Delta(D))\text{d}S_r^+,$$

□

Notemos que ya estamos en disposición de poder realizar estimaciones probabilísticas sobre $\sum_r \setminus \sum_{r-1}$ para funciones que dependan de los valores singulares. En este caso, usando el resultado de [**Chen&Dongarra, 2005**] obtenemos una acotación tanto superior como inferior para función de distribución de κ.

Teorema 2.6.2. *Con la distribución uniforme en $\sum_r \setminus \sum_{r-1}$ se tiene la siguiente igualdad para cada t:*

$$P_U\left[\kappa(A) > t : A \in \sum_r \setminus \sum_{r-1}\right] = P_\gamma[\kappa(A) > t : A \in GL(r\times(n+m-r),\mathbb{R})].$$

donde $GL(r\times(n+m-r),\mathbb{R})$ es el abierto Zariski en $\mathcal{M}_{r\times(n+m-r)}(\mathbb{R})$ formado por las matrices de rango máximo. "U es la distribución uniforme en $\sum_r \setminus \sum_{r-1}$ y γ es la distribución Gaussiana en $\mathcal{M}_{r\times(n+m-r)}(\mathbb{R})$.
Más aún, se verifican las siguientes cotas

$$\frac{1}{(2\pi)^{1/2}}\left(\frac{0.245}{t}\right)^{(n+m-2r+1)} \leq P_U\left[\frac{\kappa(A)}{\frac{n+m-r}{n+m-2r+1}} > t : A \in \sum_r \setminus \sum_{r-1}\right],$$

$$P_U\left[\frac{\kappa(A)}{\frac{n+m-r}{n+m-2r+1}} > t : A \in \sum_r \setminus \sum_{r-1}\right] \leq \frac{1}{(2\pi)^{1/2}}\left(\frac{6.414}{t}\right)^{(n+m-2r+1)}.$$

Demostración. Veamos en primer lugar que efectivamente

$$P_U\left[\kappa(A) > t : A \in \sum_r \setminus \sum_{r-1}\right] = P_\gamma[\kappa(A) > t : A \in GL(r\times(n+m-r),\mathbb{R})].$$

Sea

$$\phi_1(x_1,...,x_r) = 1,\ \phi_2(x_1,...,x_r) = \begin{cases} 1 & \text{si } \frac{x_1}{x_r} \geq t \\ \\ 0 & \text{en otro caso} \end{cases}$$

A partir del resultado del Teorema 2.6.1, aplicando la función ϕ_1 tenemos que

$$\text{Vol}\left[\sum_r \setminus \sum_{r-1}\right] = H(n,m,r)\int_{D\in S_r^+}(\det(D)^{n+m-2r}\Delta(D))\text{d}S_r^+.$$

Ahora, aplicando el mismo resultado del Teorema 2.6.1 a la función ϕ_2 obtenemos que

$$\text{Vol}\left[A \in \sum_r \setminus \sum_{r-1} : \kappa(A) > t\right] = H(n,m,r)\int_{D\in S_r^+,\kappa(A)>t}(\det(D)^{n+m-2r}\Delta(D))\text{d}S_r^+.$$

De esta forma tenemos que

$$\text{P}_U\left[\kappa(A) > t : A \in \sum_r \setminus \sum_{r-1}\right] = \frac{\text{Vol}\left[A \in \sum_r \setminus \sum_{r-1} : \kappa(A) > t\right]}{\text{Vol}\left[\sum_r \setminus \sum_{r-1}\right]},$$

es decir, se tiene:

$$\text{P}_U\left[\kappa(A) > t : A \in \sum_r \setminus \sum_{r-1}\right] = \frac{\int_{D\in S_r^+,\kappa(A)>t}(\det(D)^{n+m-2r}\Delta(D))\text{d}S_r^+}{\int_{D\in S_r^+}(\det(D)^{n+m-2r}\Delta(D))\text{d}S_r^+}.$$

Notemos la relación natural entre $\sum_r \setminus \sum_{r-1}$ y $\left(\sum'_r \setminus \sum'_{r-1}\right)\cap S^{N-1}$ donde $N = n\cdot m$, es decir, las matrices de rango r con norma de Frobenius igual a 1. De forma que considerando ahora φ como sigue

$$\begin{array}{rccc} \varphi\colon & \mathscr{O}(n)\times S_r^+\times\mathscr{O}(m) & \longrightarrow & \left(\sum'_r \setminus \sum'_{r-1}\right)\cap S^{N-1} \\ & (U,D,V) & \longmapsto & U\begin{pmatrix} D & 0 \\ 0 & 0\end{pmatrix}V^*, \end{array}$$

tenemos la siguiente interpretación de la fórmula de la Co-área para φ que en el fondo es equivalente a la obtenida en el Teorema 2.6.1

(2.6.1)

$$\int_{A\in\left(\sum'_r\setminus\sum'_{r-1}\right)\cap S^{N-1}}\phi(\sigma(A))\mathrm{d}((\sum'_r\setminus\sum'_{r-1})\cap S^{N-1})=H\int_{D\in S_r^+}\phi(D)(\det(D)^{n+m-2r}\Delta(D))\mathrm{d}S_r^+$$

donde $H=H(n,m,r)$ de la misma forma que en dicho Teorema.

De esta forma, repitiendo el proceso de aplicar ϕ_1 y ϕ_2 esta vez en (2.6.1) para $r'=n'=r$, $m'=n+m-r$, es decir, el caso de las matrices de rango máximo con norma de Frobenius uno, esto es $S\left(GL(r\times(n+m-r),\mathbb{R})\right)$, la esfera del grupo lineal de matrices de r filas y $n+m-r$ columnas.

$$\mathrm{P}_U[\kappa(A)>t:A\in S(GL(r\times(n+m-r),\mathbb{R}))]=\frac{\int_{D\in S_r^+,\kappa(A)>t}(\det(D)^{n'+m'-2r'}\Delta(D))\mathrm{d}S_r^+}{\int_{D\in S_r^+}(\det(D)^{n'+m'-2r'}\Delta(D))\mathrm{d}S_r^+},$$

que sustituyendo los valores n',m' y r' acabamos obteniendo

$$\mathrm{P}_U[\kappa(A)>t:A\in S(GL(r\times(n+m-r),\mathbb{R}))]=\frac{\int_{D\in S_r^+,\kappa(A)>t}(\det(D)^{n+m-2r}\Delta(D))\mathrm{d}S_r^+}{\int_{D\in S_r^+}(\det(D)^{n+m-2r}\Delta(D))\mathrm{d}S_r^+}.$$

Es decir, hemos obtenido que las ambas probabilidades coinciden, es decir

$$\mathrm{P}_U\left[\kappa(A)>t:A\in\textstyle\sum_r\setminus\sum_{r-1}\right]=\mathrm{P}_U[\kappa(A)>t:A\in S\left(GL(r\times(n+m-r),\mathbb{R})\right)]$$

Para finalizar la primera parte de la demostración, notemos que $\kappa(A)$ es invariante por multiplicación escalar sobre A, es decir, es una función homogénea de grado cero. De este modo, siguiendo el resultado de la Proposición A.6.11 del Apéndice A, la función de distribución de κ sobre las matrices de rango máximo con la distribución Gaussiana es igual a la función de distribución de κ sobre las matrices de rango máximo en la esfera unidad $S\left(GL(r\times(n+m-r),\mathbb{R})\right)$ con la distribución uniforme. Es decir, acabamos obteniendo

$$\mathrm{P}_U\left[\kappa(A)>t:A\in\textstyle\sum_r\setminus\sum_{r-1}\right]=P_\gamma[\kappa(A)>t:A\in GL(r\times(n+m-r),\mathbb{R})].$$

Por otro lado, Chen y Dongarra en [**Chen&Dongarra, 2005**] obtienen un resultado de acotación de la función de distribución de κ para matrices de rango máximo con la distribución Gaussiana. Dicho resultado es refinado en [**Be, 09**] mediante:

PROPOSICIÓN 2.6.3. *Sea $GL(m\times n,\mathbb{K})$ el espacio de matrices de m filas y n columnas de rango máximo, con $m\leq n$, con coeficientes en $\mathbb{K}$. Entonces se tiene*

$$\frac{1}{(2\pi)^{\beta/2}}\left(\frac{c}{t}\right)^{\beta(n-m+1)}\leq P_\gamma\left[\frac{\kappa(A)}{\frac{n}{n-m+1}}>t:A\in GL(m\times n,\mathbb{K})\right]\leq\frac{1}{(2\pi)^{\beta/2}}\left(\frac{C}{t}\right)^{\beta(n-m+1)}.$$

Donde los coeficientes β,c,C se recogen en la siguiente tabla dependiendo de $\mathbb{K}$:

$\mathbb{K}$	β	c	C
$\mathbb{R}$	1	0.245	6.414
$\mathbb{C}$	2	0.319	6.298

Notemos que $GL(r\times(n+m-r),\mathbb{R})$ es un caso particular de este resultado, y obtenemos de forma natural las desigualdades de nuestro Teorema. □

APÉNDICE A

Algunos Conocimientos previos al condicionamiento de matrices

Índice

A.1.	**Pseudo-inversa de Moore-Penrose**	**49**
A.2.	**Estratificación por co-rango del espacio del matrices**	**52**
A.3.	**Existencia de Descomposición en Valores Singulares**	**53**
A.4.	**El Teorema Espectral**	**54**
A.5.	**Teoremas de Transversalidad**	**55**
A.6.	**Algunas propiedades destacables de la Fórmula de la Co-área**	**56**
A.6.1.	Integración en espacios euclídeos.	56
A.6.2.	Integración en variedades diferenciables.	57
A.6.3.	La Fórmula de la Co-área.	58
A.6.4.	Probabilidad e integrales.	59

A.1. Pseudo-inversa de Moore-Penrose

Como en el resto del manuscrito, supondremos $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$. En ambos casos, tenemos un espacio de Hilbert de dimensión finita (real o complejo, según el caso) en $\mathbb{K}^n$. Así, para un subespacio vectorial $\vartheta \subseteq \mathbb{K}^n$, tenemos su *complemento ortogonal*:

$$\vartheta^{\perp} := \{u \in \mathbb{K}^n : \langle u, v\rangle = 0, \forall v \in \vartheta\}.$$

Como estamos en dimensión finita tenemos una descomposición de $\mathbb{K}^n$ como suma ortogonal (i.e. suma directa de subespacios ortogonales) para cualquier subespacio $\vartheta \subseteq \mathbb{K}^n$

$$\mathbb{K}^n := \vartheta \oplus \vartheta^{\perp}$$

Definición 10. *Con las notaciones precedentes, dado $\vartheta \subseteq \mathbb{K}^n$, llamaremos proyección ortogonal sobre ϑ a la aplicación lineal:*

$$\begin{array}{rccc} \pi_{\vartheta}: & \mathbb{K}^n & \longrightarrow & \vartheta \\ & x & \longmapsto & y \end{array}$$

donde $y \in \vartheta$ es el elemento de ϑ que minimiza la distancia

$$\| x - y \| = \min\{\| x - \omega \| : \omega \in \vartheta\}.$$

Proposición A.1.1. *La anterior definición es una buena definición, es decir, π_{ϑ} está bien definida como aplicación y es una aplicación lineal.*

Demostración. Probaremos el resultado por su simplicidad. Gracias a las técnicas de J.P. Gram y E. Schimidt (aunque parece que eran técnicas parcialmente conocidas por P. Laplace), podemos garantizar que existen:

i) Una base $\mathcal{B}_1 = \{u_1, ..., u_r\}$ de ϑ que es ortonormal, con $r = \dim(\vartheta)$.
ii) Una base $\mathcal{B}_2 = \{u_{r+1}, ..., u_n\}$ de $\vartheta^{\perp}$ que es también ortonormal.

de tal modo que $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ es una base ortonormal de $\mathbb{K}^n$. Definamos la aplicación lineal:

$$\begin{array}{rccc} \varphi: & \mathbb{K}^n & \longrightarrow & \vartheta \\ & v & \longmapsto & \sum_{i=1}^{r} \langle v, u_i\rangle \cdot u_i. \end{array}$$

claramente, φ es una aplicación lineal (de fácil verificación), suprayectiva y bien definida. Veamos que $\varphi = \pi_{\vartheta}$ es, precisamente, la proyección ortogonal citada en la definición precedente.

Para ello, consideremos $x \in \mathbb{K}^n$. Tendremos una descripción de x en la base $\mathcal{B}$ dada por la igualdad siguiente:

$$x = \sum_{i=1}^{n} \lambda_i \cdot u_i.$$

Por ser $\mathcal{B}$ base ortonormal, tenemos que por ser bilineal (o sesquilineal) se tiene,

$$\langle x, u_i \rangle = \sum_{j=1}^{n} \lambda_j \langle u_j, u_i \rangle = \sum_{j=1}^{n} \lambda_j \cdot \delta_{ji} = \lambda_i,$$

donde δ_{ji} es la delta de Kronecker.
Por tanto, $\varphi(x) = \sum_{i=1}^{r} \lambda_i \cdot u_i$, mientras que $\sum_{j=r+1}^{n} \lambda_j \cdot u_j \in \vartheta^{\perp}$. En suma, tendremos que

$$\| x \|^2 = \sum_{i=1}^{n} |\lambda_i|^2 = \left(\sum_{i=1}^{r} |\lambda_i|^2 \right) + \left(\sum_{j=r+1}^{n} |\lambda_j|^2 \right) = \| \varphi(x) \|^2 + \| x - \varphi(x) \|^2,$$

verificandose, además, que $x - \varphi(x) \in \vartheta^{\perp}$. Supongamos además, que $y \in \vartheta$ es un elemento cualquiera tal que

$$\| x - y \|^2 = \min\{ \| x - \omega \|^2 : \omega \in \vartheta \}.$$

Dado que $y \in \vartheta$, tendremos que

$$\| x - y \|^2 = \| \varphi(x) + (x - \varphi(x)) - y \|^2 = \| \varphi(x) - y \|^2 + \| x - \varphi(x) \|^2 \tag{A.1.1}$$

Por tanto, si y minimiza la distancia de x a ϑ, tendremos que, necesariamente,

$$\| x - y \|^2 \geq \| x - \varphi(x) \|^2$$

y, tomando $y = \varphi(x)$ concluiremos que $\| x - \varphi(x) \|^2 \leq \min\{ \| x - \omega \|^2 : \omega \in \vartheta \}$.
Por tanto, $\varphi(x)$ satisface la definición precedente y es uno de los elementos de ϑ que minimiza la distancia de x a ϑ. Pero, además, hemos visto que

$$\| x - \varphi(x) \| = \min\{ \| x - \omega \| : \omega \in \vartheta \},$$

con lo que si y es un punto que también minimiza esta distancia, por la Ecuación (A.1.1) anterior, concluiremos

$$\| x - y \|^2 = \| x - \varphi(x) \|^2 \Longrightarrow \| y - \varphi(x) \|^2 = 0 \Longrightarrow y = \varphi(x).$$

En suma, φ satisface las condiciones de la definición precedente y, además, $\varphi(x)$ es el único punto de ϑ que minimiza la distancia de x a ϑ. Es decir,

$$\varphi = \pi_{\vartheta}$$

y π_{ϑ} está bien definida. □

Algunas propiedades inmediatas de la proyección ortogonal son las siguientes:

PROPOSICIÓN A.1.2. *Con las notaciones precedentes, dado $\vartheta \subseteq \mathbb{K}^n$, se tiene:*

i) La proyección ortogonal $\pi_{\vartheta} : \mathbb{K}^n \to \vartheta$ es un epimorfismo y, además, su restricción a ϑ es la identidad:

$$\pi_{\vartheta}|_{\vartheta} = Id_{\vartheta}.$$

ii) Para cada $x \in \mathbb{K}^n$, $\pi_{\vartheta}(\pi_{\vartheta}(x)) = \pi_{\vartheta}(x)$, lo que suele denotarse mediante

$$\pi_{\vartheta}^2 = \pi_{\vartheta}.$$

Consideremos ahora $A \in \mathcal{M}_{n \times m}(\mathbb{K})$ una matriz y la aplicación lineal que define $A : \mathbb{K}^m \to \mathbb{K}^n$, consideremos $K = \ker(A)$ el núcleo de A y su complemento ortogonal $K^{\perp}$. Consideremos, asimismo, $\vartheta = Im(A) \subseteq \mathbb{K}^n$ el subespacio imagen de A y la proyección ortogonal

$$\pi_{\vartheta} : \mathbb{K}^n \to \vartheta.$$

Ahora, consideremos la restricción de A en $K^{\perp}$, que será un isomorfismo con $\vartheta = Im(A)$

$$\begin{array}{rccc} A_{red} := A|_{K^{\perp}} : & K^{\perp} & \longrightarrow & \vartheta \\ & x & \longmapsto & Ax. \end{array}$$

Por el primer Teorema de Isomorfía, A_{red} es un isomorfismo de espacios vectoriales y tiene, por tanto, aplicación inversa $(A_{red})^{-1}$. Finalmente, denotemos por A^+ la aplicación lineal, componiendo con la proyección ortogonal π_ϑ:

$$\begin{array}{rccc} A^+: & \mathbb{K}^n & \longrightarrow & \mathbb{K}^m \\ & y & \longmapsto & \left(\left(A_{|_{\ker(A)^\perp}}\right)^{-1} \circ \pi_\vartheta\right)(y). \end{array}$$

Esta es la famosa pseudo-inversa de Moore-Penrose, que fue introducida, de modo aparentemente independientemente, por E.H.Moore (en 1927), A. Bjerhammer (en 1931) y R. Penrose (en 1958). Una definición alternativa, usualmente descrita en la literatura del tema, suele ser el siguiente enunciado:

TEOREMA A.1.3. *Sea $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$ y $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ una matriz con n filas y m columnas. Entonces, existe una matriz $A^+ \in \mathcal{M}_{m\times n}(\mathbb{K})$ que verifica las siguientes propiedades:*

i) $AA^+A = A$.
ii) $A^+AA^+ = A^+$.
iii) $(AA^+)^* = AA^+$, *i.e. AA^+ es simétrica (si $\mathbb{K} = \mathbb{R}$) o hermítica (si $\mathbb{K} = \mathbb{C}$).*
iv) $(A^+A)^* = A^+A$, *es también una matriz hermítica.*

Además, la matriz A^+ es única para cada $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y se denomina la pseudo-inversa de Moore-Penrose de A.

Es fácil verificar que la matriz A^+ descrita en la construcción que precede a este teorema es una pseudo-inversa de Moore-Penrose. La unicidad nos garantiza que tenemos una forma de construir la pseudo-inversa a partir del producto canónico en $\mathbb{K}^n$ y $\mathbb{K}^m$; así como el cálculo de $\ker(A)$ e $Im(A)$ y sus complementos ortogonales.
Es también evidente que si $A \in GL(n, \mathbb{K})$ es una matriz regular, $A^+ = A^{-1}$ es la inversa de A.

PROPOSICIÓN A.1.4. *Sea $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y sea $A^+ \in \mathcal{M}_{m\times n}(\mathbb{K})$ su pseudo-inversa de Moore-Penrose. Supongamos que A admite una descomposición en valores singulares de la forma:*

$$A = U \cdot \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) V^*$$

con U y V matrices ortogonales si $\mathbb{K} = \mathbb{R}$ (unitarias si $\mathbb{K} = \mathbb{C}$) de tamaños apropiados y $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$ son los valores singulares (en $\mathbb{R}$) de A. Se tiene:

i) La pseudo-inversa de Moore-Penrose de A viene dada por:

$$A^+ = V \cdot \left(\begin{array}{ccc|c} 1/\sigma_1 & & & \\ & \ddots & & 0 \\ & & 1/\sigma_r & \\ \hline & 0 & & 0 \end{array}\right) U^*$$

ii) Las normas de A^+ satisfacen:

$$\| A^+ \|_2 = \frac{1}{\sigma_r}\ ,\ \| A^+ \|_F = \left(\sum_{i=1}^{r} \frac{1}{\sigma_i^2}\right)^{1/2},$$

DEMOSTRACIÓN. De nuevo basta con verificar que esa matriz satisface las propiedades *i)* a *iv)* del Teorema precedente. □

El siguiente resultado puede seguirse en [**Stw, 1977**] y nos será de utilidad en la demostración del Teorema de Wedin.

TEOREMA A.1.5. *Para cualquier par de matrices $A, B \in \mathcal{M}_{n\times m}(\mathbb{K})$ se tienen los siguientes resultados:*

- *Si el rango* $\operatorname{rank}(A) = \operatorname{rank}(B)$, *entonces los valores de $AA^+(I - BB^+)$ y $BB^+(I - AA^+)$ son los mismos, por lo que*

$$\| AA^+(I - BB^+) \| = \| BB^+(I - AA^+) \|.$$

Es más, los valores singulares σ de $AA^+(I-BB^+)$ se corresponden con pares $\pm\sigma$ de valores propios de $BB^+ - AA^+$, por lo que se tiene

$$\| BB^+ - AA^+ \|_2 = \| AA^+(I-BB^+) \|_2 = \| BB^+(I-AA^+) \|_2 .$$

- *Si* $\| BB^+ - AA^+ \|_2 < 1$, *entonces A y B son del mismo rango.*
- *Si* $\operatorname{rank}(B) \geq \operatorname{rank}(A)$, *entonces*

$$\| BB^+(I-AA^+) \| \geq \|(I-BB^+)AA^+ \| .$$

A.2. Estratificación por co-rango del espacio del matrices

Lo que sigue resume un material clásico del Álgebra Lineal, que consiste en mostrar el espacio de matrices $\mathcal{M}_n(\mathbb{K})$, $V \in \mathcal{M}_m(\mathbb{K})$ como una cadena ascendente de variedades algebraicas homogéneas determinadas por el corango, donde

$$corank(A) = \min\{n,m\} - \operatorname{rank}(A).$$

Seguiremos los resultados que pueden consultarse en [**AGV, 1985**], aunque es una clasificación conocida.

Definición 11. *Sean $n, m \in \mathbb{N}$ dos números naturales, supongamos $n \geq m$ y sea $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$. Definiremos la variedad de matrices de rango acotado por r como la variedad*

$$\textstyle\sum'_r(n,m) := \{A \in \mathcal{M}_{n\times m}(\mathbb{K}) : \operatorname{rank}(A) \leq r\}$$

Fijemos $n, m \in \mathbb{N}$, con $n \geq m$ a lo largo de esta Sección y denotemos por $\sum'_r = \sum'_r(n,m)$. Nótese que $\sum'_m = \mathcal{M}_{n\times m}(\mathbb{K})$ y que $\sum'_0 = \{0\}$.
Consideremos $\pi : \mathcal{M}_{n\times m}(\mathbb{K}) \setminus \{0\} \to \mathbb{P}_{mn-1}(\mathbb{K})$ la proyección canónica y denotemos mediante $\sum_r := \pi\left(\sum'_r\right)$ para $1 \leq r \leq m$. Denotaremos por $\sum'_o = \phi$.

Proposición A.2.1. *Los subconjuntos $\sum'_r(n,m) \subseteq \mathcal{M}_{n\times m}(\mathbb{K}) = \mathbb{K}^{n\cdot m}$ son variedades algebraicas afines y $\mathbb{Q}$-definibles (i.e. definidas por ecuaciones polinomiales multi-variadas con coeficientes en $\mathbb{Q}$ de dimensión*

$$n \cdot m - (n-r)(m-r).$$

No haremos una demostración formal, pero demos una idea. Nótese que una matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ está en $\sum'_r$ si y solamente si todos los menores de A de orden mayor o igual que $r+1$ son idénticamente nulos. Gracias a la famosa fórmula de Laplace, basta con que sean nulos los menores de orden $r+1$. Los menores de una matriz de orden t son polinomios multivariados en las coordenadas de la matriz A de grado t con coeficientes en $\mathbb{Q}$ (los coeficientes son los índices (en $\{\pm 1\}$) de las permultaciones involucradas). Por tanto $\sum'_r$ es una variedad algebraica $\mathbb{Q}$-definible que vive en el espacio afín $\mathbb{K}^{n\cdot m}$.
La prueba de la dimensión es ligeramente más compleja y puede seguirse en [**AGV, 1985**].
De hecho, se tiene:

Proposición A.2.2. *Con las notaciones precedentes, el conjunto $\sum'_r \setminus \sum'_{r-1}$ es una variedad diferenciable (analítica, holomorfa en el caso $\mathbb{K} = \mathbb{C}$) cuya dimensión coincide con la dimensión de $\sum'_r$, i.e.*

$$dim\left(\textstyle\sum'_r \setminus \sum'_{r-1}\right) = dim\left(\sum'_r\right).$$

De hecho, $\sum'_r \setminus \sum'_{r-1}$ es la variedad de puntos lisos (o regulares) de la variedad $\sum'_r$ y $\sum'_{r-1}$ es, justamente, la subvariedad de puntos singulares de $\sum'_r$ (i.e. los puntos donde el espacio tangente a $\sum'_r$ tienen la dimensión inapropiada).
Tenemos así una cadena de variedades algebraicas

$$\textstyle\sum'_0 \subseteq \sum'_1 \subseteq \sum'_2 \subseteq \ldots \subseteq \sum'_m = \mathcal{M}_{n\times m}(\mathbb{K}),$$

que define, propiamente, una estratificación de $\mathcal{M}_{n\times m}(\mathbb{K})$ como unión de subvariedades

$$\mathcal{M}_{n\times m}(\mathbb{K}) = \left(\textstyle\sum'_m \setminus \sum'_{m-1}\right) \cup \left(\sum'_{m-1} \setminus \sum'_{m-2}\right) \cup \ldots \cup \sum'_0 .$$

Obsérvese que las ecuaciones que definen las variedades algebraicas $\sum'_r$ son polinomios homogéneos (de grado $r+1$) porque los menores de una matriz, que son determinantes, son polinomios homogéneos, cuyas variables son las coordenadas de una matriz. Por tanto, para cada matriz $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y cada $\lambda \in \mathbb{K} \setminus \{0\}$, se tiene que

$$A \in \sum_r' \iff \lambda \cdot A \in \sum_r' .$$

Esto nos permite interpretar de modo natural los subonjuntos de $\sum_r \subseteq \mathbb{P}_{nm-1}(\mathbb{K})$ y ver que $\sum_r'$ es el cono proyectante sobre $\sum_r$, i.e.

$$\sum_r' = \{B \in \mathcal{M}_{n\times m}(\mathbb{K}) : \exists A \in \mathcal{M}_{n\times m}(\mathbb{K}), \pi(A) \in \sum_r \text{, y } \exists \lambda \in \mathbb{K} \setminus \{0\} \text{ t.q } B = \lambda \cdot A\} .$$

La definición es redundante pero ayuda a interpretar $\sum_r'$ como cono proyectante. Más aún, tenemos el siguiente corolario que usaremos en la memoria.

COROLARIO A.2.3. *Con las notaciones precedentes, tenemos:*

i) $\sum_r = \pi\left(\sum_r'\right)$ *es una variedad algebraica proyectiva definida por polinomios homogéneos de grado* $r+1$ *con coeficientes en* $\mathbb{Q}$*, de dimensión*

$$dim\left(\sum_r\right) = dim\left(\sum_r'\right) - 1.$$

ii) $\sum_r \setminus \sum_{r-1} \subseteq \mathbb{P}_{nm-1}(\mathbb{K})$ *es una variedad analítica de dimensión*

$$dim\left(\sum_r \setminus \sum_{r-1}\right) = dim\left(\sum_r'\right) - 1.$$

iii) La siguiente es una descomposición de $\mathbb{P}_{nm-1}(\mathbb{K})$ *como unión disjunta de subvariedades anaíticas*

$$\mathbb{P}_{nm-1}(\mathbb{K}) = \left(\sum_m \setminus \sum_{m-1}\right) \cup ... \cup \sum_1 .$$

A.3. Existencia de Descomposición en Valores Singulares

La descomposición en valores singulares de una matriz con coordenadas reales o complejas es un resultado clásico descubierto independientemente por E. Beltrami y C. Jordan en el siglo XIX. Posteriormente, fue redescubierto por J.J. Sylvester, E. Schmidt y H. Weyl, un recuento histórico de estas contribuciones puede seguirse en [**Stw, 1993**]. Aquí nos limitaremos a seguir la presentación, en forma de Problemas, hecha en las notas del curso [**Pardo, 2021**]. No entraremos a discutir las muchísimas aplicaciones de esta descomposición.

LEMA A.3.1. *Sea* $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$*. Para cada matriz* $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ *existen matrices* $U \in \mathcal{M}_n(\mathbb{K})$*,* $V \in \mathcal{M}_m(\mathbb{K})$*, y* $A_1 \in \mathcal{M}_{n-1\times m-1}(\mathbb{K})$ *tales que se verifica:*

i) Si $\mathbb{K} = \mathbb{R}$*, las matrices* U *y* V *son ortogonales, i. e.* $U \in \mathscr{O}(n)$*,* $V \in \mathscr{O}(m)$*.*

ii) Si $\mathbb{K} = \mathbb{C}$*, las matrices* U *y* V *son unitarias, i. e.* $U \in \mathscr{U}(n)$*,* $V \in \mathscr{U}(m)$*.*

iii) Se tiene:

$$UAV^* = \begin{pmatrix} \| A \|_2 & 0 \\ 0 & A_1 \end{pmatrix},$$

donde $\| A \|_2$ *es la norma de* A *como operador lineal y los* 0*'s representan vectores nulos de longitud apropiada.*

Usando inductiavmente este Lema, se obtiene:

TEOREMA A.3.2 (**Existencia de Descomposición en Valores Singulares**)**.** *Sea* $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$*. Para cada matriz* $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ *existen matrices* $U \in \mathcal{M}_n(\mathbb{K})$*,* $V \in \mathcal{M}_m(\mathbb{K})$ *y números reales* $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$ *de tal modo que:*

i) Si $\mathbb{K} = \mathbb{R}$*,* $U \in \mathscr{O}(n)$*,* $V \in \mathscr{O}(m)$*.*

ii) Si $\mathbb{K} = \mathbb{C}$*,* $U \in \mathscr{U}(n)$*,* $V \in \mathscr{U}(m)$*.*

iii) Se tiene:

$$UAV^* = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right), \tag{A.3.1}$$

donde $r = \text{rank}(A)$ *es el rango de la matriz* A*,* $\sigma_1 = \| A \|_2$*, y los* 0*'s son submatrices de tamaños apropiados para que se de la igualdad.*

iv) Los números reales $\sigma_1 \geq ... \geq \sigma_r > 0$*,* $\sigma_i \in \mathbb{R}$*, que satisfacen (A.3.1) son únicos satisfaciendo tal ecuación.*

DEFINICIÓN 12. *A los elementos* $\sigma_1 \geq ... \geq \sigma_r > 0$*, únicos, se les denomina valores singulares de la matriz* A*.*

La unicidad de los valores singulares se sigue de la siguiente propiedad, que será de utilidad en nuestro estudio.

PROPOSICIÓN A.3.3. *Con las notaciones precendentes, si $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y los valores singulares de A son dados por $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$, entonces $\sigma_1^2 \geq \sigma_2^2 \geq ... \geq \sigma_r^2 > 0$ son los valores propios no nulos de la cualquiera de las dos matrices hermíticas siguientes:*

$$AA^* \ , \ A^*A$$

DEMOSTRACIÓN. Es obvio desde la SVD, pero si suponemos $\mathbb{K} = \mathbb{C}$ y tenemos

$$A = U \cdot \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) V^*, U \in \mathscr{U}(n), V \in \mathscr{U}(m),$$

tendremos que

$$AA^* = U \cdot \left(\begin{array}{ccc|c} \sigma_1^2 & & & \\ & \ddots & & 0 \\ & & \sigma_r^2 & \\ \hline & 0 & & 0 \end{array}\right) U^* = U \cdot \left(\begin{array}{ccc|c} \sigma_1^2 & & & \\ & \ddots & & 0 \\ & & \sigma_r^2 & \\ \hline & 0 & & 0 \end{array}\right) U^{-1},$$

por lo que concluimos que la forma canónica de Jordan de $A \cdot A^* \in \mathcal{M}_n(\mathbb{C})$ es la matriz diagonal

$$\text{Diag}(\sigma_1^2, ..., \sigma_r^2, \underbrace{0, ..., 0}_{n-1}).$$

Con lo que los valores propios no nulos de AA^* son $\sigma_1^2 \geq ... \geq \sigma_r^2$.
El argumento es el mismo si consideramos la matriz cuadrada hermítica $A^* \cdot A \in \mathcal{M}_m(\mathbb{C})$ y su forma canónica de Jordan será

$$\text{Diag}(\sigma_1^2, ..., \sigma_r^2, \underbrace{0, ..., 0}_{m-r}).$$

□

De entre las muchas y muy relevantes propiedades que satisfacen los valores singulares de una matriz, una de las más útiles es la famosa Desigualdad de la Traza de von Neumann que podemos condibir como sigue:

TEOREMA A.3.4 (**von Neumann Trace Inequality**). *Sean $A \in \mathcal{M}_{n\times m}(\mathbb{K})$ y $B \in \mathcal{M}_{m\times t}(\mathbb{K})$ dos matrices con coordenadas complejas, de tamaños respectivos $n \times m$ y $m \times t$. Consideramos la siguiente familia de números reales:*

- $\sigma_1(A) \geq \sigma_2(A) \geq ... \geq \sigma_m(A) \geq 0$, *donde*
  - $\sigma_i(A)$ *es el i-ésimo valor singular de* A *para* $1 \leq i \leq \text{rank}(A)$.
  - $\sigma_j(A) = 0$, *para cada* $j \geq \text{rank}(A) + 1$.
- $\sigma_1(B) \geq \sigma_2(B) \geq ... \geq \sigma_m(B) \geq 0$, *donde*
  - $\sigma_i(B)$ *es el i-ésimo valor singular de* A *para* $1 \leq i \leq \text{rank}(B)$.
  - $\sigma_j(B) = 0$, *para cada* $j \geq \text{rank}(B) + 1$.

Entonces, se da la siguiente desigualdad:

$$|Tr(A \cdot B)| \leq \sum_{i=1}^{m} \sigma_i(A) \cdot \sigma_i(B)$$

A.4. El Teorema Espectral

El material de esta sección se ha tomado del texto básico [**Olazábal, 1988**], aunque son resultados clásicos ampliamente descritos en los textos básicos de Álgebra Lineal. Como notación básica, para una matriz $A \in \mathcal{M}_n(\mathbb{C})$, escribiremos $A^* \in \mathcal{M}_n(\mathbb{C})$ para denotar la traspuesta de la matriz obtenida de A mediante la conjugación de sus coordenadas. Si $A \in \mathcal{M}_n(\mathbb{R})$, la matriz A^* también denotará la matriz traspuesta de la matriz A.

DEFINICIÓN 13 (**Matriz Normal**). *Una matriz $A \in \mathcal{M}_n(\mathbb{C})$ se dice normal si $AA^* = A^*A$.*

Un clásico resultado de gran relevancia es el siguiente resultado.

TEOREMA A.4.1. *Para una matriz compleja $A \in \mathcal{M}_n(\mathbb{C})$, las siguientes propiedades son equivalentes:*

i) A es una matriz normal.

ii) Existe $P \in \mathscr{U}(n)$ una matriz unitaria tal que PAP^* es una matriz diagonal, esto es

$$PAP^* = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} = \mathrm{Diag}(\lambda_1, ..., \lambda_n),$$

donde $\{\lambda_1, ..., \lambda_n\} \subseteq \mathbb{C}$ son los valores propios de la matriz A.

Nótese que, en particular, las matrices normales son diagonalizables.

COROLARIO A.4.2. *Las matrices ortogonales (en $\mathscr{O}(n)$) y las matrices unitarias (en $\mathscr{U}(n)$) son matrices diagonalizables sobre $\mathbb{C}$ salvo congruencia unitaria.*

Se puede hacer algo en el caso real, considerando, con cierto cuidado, la representación de sus valores propios complejos.

TEOREMA A.4.3. *Sea $A \in \mathcal{M}_n(\mathbb{R})$ una matriz con coordenadas reales. Son equivalentes:*

i) A es una matriz normal.

ii) Existen $\lambda_1, ..., \lambda_t \in \mathbb{R}$, $z_1, ..., z_s \in \mathbb{C} \setminus \mathbb{R}$ tales que $t + 2s = n$ y una matriz ortogonal real P de tal modo que

$$PAP^* = \begin{pmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & & \\ & & \lambda_n & & & & \\ & & & M(z_1) & & & \\ & & & & \ddots & & \\ & & & & & M(z_s) & \end{pmatrix},$$

donde

$$M(z_j) = \begin{pmatrix} a_j & b_j \\ -b_j & a_j \end{pmatrix} \text{ cuando } z_j = a_j + ib_j \in \mathbb{C} \setminus \mathbb{R}.$$

El caso de matrices simétricas reales es aún más significativo.

TEOREMA A.4.4 (**Diagonalización ortogonal de matrices simétricas reales**). . *Dada una matriz simétrica real $S \in \mathrm{Sym}_n(\mathbb{R})$, entonces existen*

i) Una matriz ortogonal $P \in \mathscr{O}(n)$

ii) Una matriz diagonal $D = \mathrm{Diag}(\lambda_1, ..., \lambda_n)$ donde $\lambda_1, ..., \lambda_n \in \mathbb{R}$

tales que

$$S = P \cdot D \cdot P^*.$$

En particular, las matrices simétricas reales son diagonalizables, sus valores propios son todos reales y existe una base ortonormal formada por vectores propios de la matriz.

A.5. Teoremas de Transversalidad

La referencia para estos resultados será el texto de M. Demazure [**Demazure, 1989**].
Denotemos por μ_n (o simplemente por μ) la medida de Lebesgue de un espacio afín real n-dimensional. Un subconjunto $S \subseteq \mathbb{R}^n$ se dice de medida nula si su volumen $\mu(S) = 0$.

DEFINICIÓN 14. *Sea U un abierto de $\mathbb{R}^n$ y $f : U \to \mathbb{R}^m$ una función de clase $\mathscr{C}^\infty$. Un punto $x \in U$ se denomina punto crítico si la diferencial*

$$Df(x) : T_xU = \mathbb{R}^n \longrightarrow T_{f(x)}\mathbb{R}^m = \mathbb{R}^m$$

no es suprayectiva. Si $Df(x)$ es suprayectiva, decimos que $x \in U$ es un punto regular. Un punto $y = f(x) \in f(U)$ se denomina valor crítico si $Df(x)$ no es suprayectiva.

TEOREMA A.5.1 (**Morse-Sard**). *Sea $U \subseteq \mathbb{R}^n$ un abierto y $f : U \to \mathbb{R}^m$ una función de clase $\mathscr{C}^\infty$. Entonces, el conjunto de los valores críticos de f es un conjunto de medida nula.*

Nótese que es fundamental poner cuidado en $f(U)$. Si la dimensión de $f(U)$ es menor estricto que m (por ejemplo, si $n < m$) es perfectamente posible que todos los puntos de la imagen sean valores críticos como puntos en $\mathbb{R}^m$.

DEFINICIÓN 15. *Sea $U \subseteq \mathbb{R}^n$ un abierto y $f : U \to \mathbb{R}^m$ una función de clase $\mathscr{C}^\infty$. Sea $a \in U$, diremos que f es una submersión en $a \in U$ si $T_a f : \mathbb{R}^n \to \mathbb{R}^m$ es una aplicación suprayectiva. Diremos que f es una submersión en U si es una submersión en todos los puntos de U.*

PROPOSICIÓN A.5.2. *Sea $U \subseteq \mathbb{R}^n$ un abierto y $f : U \to \mathbb{R}^m$ una submersión en el punto $a \in U$. Entonces, si $b = f(a)$*

i) La fibra $f^{-1}(\{b\})$ es una subvariedad diferenciable de U de dimensión $\dim(U) - m$.
ii) El espacio tangente a $f^{-1}(\{b\})$ es precisamente $\ker(T_a f)$.
iii) Existe un abierto $\Omega \subseteq \mathbb{R}^m$ y una aplicación $\mathscr{C}^\infty$ $h : \Omega \to \mathbb{R}^n$ tal que $h(b) = 0$, $h(y) \in U, \forall y \in \Omega$ y $f(h(y)) = y$, para todo $y \in \Omega$.

A.6. Algunas propiedades destacables de la Fórmula de la Co-área

En esta sección del apéndice tomaremos como referencia la Tesis de C. Beltrán [**Be, 06**].

A.6.1. Integración en espacios euclídeos. Sea W un espacio vectorial real de dimensión n. Diremos que una forma de volumen ω en W es el valor absoluto de una forma n–multilineal alternada $\omega' : W^n \to \mathbb{R}$. Esto es, $\omega(v_1, ..., v_n) := |\omega'(v_1, ..., v_n)|$, donde ω' es n–multilineal y

$$\omega'(v_1, ..., v_i, ..., v_j, ..., v_n) = -\omega'(v_1, ..., v_j, ..., v_i, ..., v_n), \forall i \neq j.$$

Es bien sabido que una forma de volumen ω en W queda determinada a partir del valor $\omega(v_1, ..., v_n)$ de ω en una base cualquiera $[v_1, ..., v_n]$ de W.
Para cualquier tupla de n vectores de W, $[w_1, ..., w_n]$, tal que $w_j = \sum_{i=1}^n a_{ij} v_i$ donde $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$ es una matriz real, se tiene que

$$\omega(w_1, ..., w_n) = |\det(A)|\omega(v_1, ..., v_n).$$

Fijada una forma de volumen ω en un espacio vectorial W de dimensión n, y dados un conjunto de vectores $v_1, ..., v_n$ de W, llamamos volumen del paralelepípedo formado por $v_1, ..., v_n$ al número real $\omega(v_1, ..., v_n)$.

DEFINICIÓN 16. *Sea W un espacio vectorial real de dimensión n con producto interior $\langle \cdot, \cdot \rangle$. Entonces, W tiene una forma de volumen natural ω dada por*

$$\omega(b_1, ..., b_n) = 1,$$

para toda base ortonormal $\{b_1, ..., b_n\}$ de W. Entonces, el volumen (en W) del paralelpípedo formado por $v_1, ..., v_n$ con respecto a ω es

$$\omega(v_1, ..., v_n) = |\det(v_{ij})|,$$

donde (v_{ij}) es la matriz formada por las coordenadas de los vectores v_i en cualquier base ortonormal de W.

Dicha definición no depende de la base ortonormal elegida para W.

LEMA A.6.1. *Sea W un espacio vectorial de dimensión n con producto interior $\langle \cdot, \cdot \rangle$. Sea $\{v_1, ..., v_s\}$ un conjunto de vectores en W, con $s \leq n$, y sea $W' :=< v_1, ..., v_s >$ el espacio vectorial generado por esa familia de vectores. Finalmente, sea $M := (v_{ij}) \in \mathcal{M}_{n \times s}(\mathbb{R})$ la matriz cuyas columnas son las coordenadas de los vectores v_i en cualquier base ortonormal de W. Entonces, el volumen en W' del paralelepípedo formado por $v_1, ..., v_s$ es igual a:*

$$\det(M^* M)^{1/2},$$

donde M^* es la matriz traspuesta conjugada de M, si estamos en un espacio vectorial complejo, o la matriz traspuesta estándar en caso contrario.

A.6.2. Integración en variedades diferenciables. Para nosotros, "diferenciable" significará siempre "infinitamente diferenciable", tanto al referirnos a variedades como a aplicaciones entre variedades. Para cada variedad diferenciable M supondremos que el espacio topológico subyacente satisface las siguientes hipótesis:

- El conjunto M es un espacio topológico Hausdorff.
- M es localmente compacto (esto es, para todo elemento $p \in M$, existe un abierto $U \subseteq M$ que contiene a p y un compacto $K \subseteq M$ tal que $U \subseteq K$).
- Se puede escribir M como unión contable de compactos.
- M es paracompacto (esto es, además de Hausdorff, todo recubrimiento abierto $\mathcal{C}$ de M tiene un refinamiento que es localmente finito, en el sentido de que todo punto $p \in M$ tiene un entorno U tal que el número de elementos de $\mathcal{C}$ que intersecan a U es finito).
- La topología de M tiene una base contable.

Para cada $p \in M$, denotaremos por T_pM el espacio tangente a M en p. Para cada aplicación diferenciable entre variedades $F : M \to M'$ y para cada $p \in M$, denotaremos por $d_pF : T_pM \to T_{F(p)}M'$ la aplicación diferencial entre los respectivos espacios tangentes. También denotaremos por $\dim(M)$ la dimensión (real o compleja, según el contexto) de M.
Una variedad Riemanniana real (compleja) es una variedad diferenciable real (compleja) en la que además hay una forma bilinear simétrica y definida positiva (Hermitiana) $\langle \cdot, \cdot \rangle_p$ en cada espacio tangente T_pM, que varía de forma diferenciable. Esto permite definir de modo natural conceptos de ortogonalidad y ortonormalidad de vectores tangentes en un punto p. También queda asignada así una norma $\| \cdot \|_p$ en cada espacio tangente T_pM, del modo natural: $\| v \|_p = \langle v, v \rangle_p^{1/2}$. También diremos que una aplicación diferenciable entre variedades $F : M \to M'$ es isometría en un punto $p \in M$ si la diferencial d_pF es una isometría de espacios vectoriales, también diremos que F es una isometría si lo es para todo punto $p \in M$.
Una *función de densidad* en una variedad diferenciable M de dimensión m viene dada por una medida positiva (absolutamente continua respecto a la medida de Lebesgue) en cada abierto $\mathbb{R}^m$ asociado a un atlas de M, de forma que se verifique la ley de composición de cartas. Así, se definen una σ-Álgebra y una medida positiva en la variedad M a través de las cartas, lo que permite definir volúmenes e integrales en M siguiendo el esquema clásico de Teoría de la Medida. Una variedad Riemanniana tiene asociada de modo natural una función de densidad, y por tanto definiciones de volumen e integral. También quedan así establecidos los conceptos de conjuntos medibles, funciones medibles e integrables, conjuntos de medida cero y demás nociones relacionadas. Con los axiomas topológicos que hemos asumido, se tiene además que la variedad Riemanniana M es σ-finita con respecto a la medida asociada, esto es, M puede escribirse como unión numerable de conjuntos de medida finita. Por ello, podemos aplicar cuando lo deseemos el Teorema de Fubini para integración en producto de variedades. Diremos que una propiedad se satisface en casi todo punto, si se satisface en el complementario de un conjunto de medida cero. En general, para una función integrable $f : M \to [-\infty, +\infty]$ definida en una variedad Riemanniana M, denotaremos su integral en M mediante

$$\int f dM, \text{ o bien } \int_{x \in M} f(x) dM.$$

Para un subconjunto medible $A \subseteq M$, denotaremos por $\nu_M[A]$ el volumen de A. En el caso de que M sea $\mathbb{R}^k$ con la estructura usual, $\nu_{\mathbb{R}^k}$ es la medida de Lebesgue.
Si $N \subseteq M$ es una subvariedad diferenciable de M, entonces N hereda de M una estructura de variedad Riemanniana, y tiene por tanto una medida positiva asociada a ella de modo natural.

Definición 17 (**Transversalidad**). *Sea $F : M \to N$ una aplicación diferenciable entre variedades diferenciables. Sea $W \subseteq N$ una subvariedad diferenciable de N. Decimos que F es transversal a W si para todo par de puntos $x \in M, y \in W$ tales que $F(x) = y$ se tiene:*

$$T_yW + d_xF(t_xM) = T_yN.$$

Es decir, si el espacio tangente en y a W y la imagen de la diferencial d_xF generan todo el espacio tangente en y a N.

Corolario A.6.2. *Sea $F : M \to N$ una aplicación diferenciable entre variedades diferenciables. Sea $W \subseteq N$ una subvariedad diferenciable de N, tal que F es transversal a W. Entonces, el*

conjunto $F^{-1}(W)$ es una subvariedad diferenciable de M, y la codimensión de $F^{-1}(W)$ en M es igual a la codimensión de W en N. Esto es,

$$\dim(M) - \dim(F^{-1}(W)) = \dim(N) - \dim(W).$$

TEOREMA A.6.3 (**Thom**). *Sean M, N, T variedades diferenciables. Sea también $W \subseteq N$ una subvariedad diferenciable de N. Sea $j : T \to \mathcal{C}^\infty(M, N)$ una aplicación cualquiera, tal que la aplicación*

$$\begin{array}{rccc} \Theta\colon & T \times M & \longrightarrow & N \\ & (t, x) & \longmapsto & j(t)(x). \end{array}$$

es diferenciable. Supongamos además que Θ es transversal a W. Entonces, para casi todo $t \in T$ se tiene que $j(t)$ es transversal a W.

A.6.3. La Fórmula de la Co-área. En esta sección introduciremos la Fórmula de la Co-área, que generaliza el Teorema de Fubini y el Teorema del Cambio de Variable simultaneamente.

DEFINICIÓN 18. *Sean M y N dos variedades Riemannianas, de dimensiones respectivas $m \geq n$. Sea $x \in M$ un punto cualquiera y sea $F : M \to N$ una aplicación diferenciable tal que la diferencial d_xF es sobreyectiva.*
Definimos el espacio horizontal en x, $H_x \subseteq T_xH$, como el complemento ortogonal de $\ker(d_xF)$ *en el espacio tangente T_xM. Esto es,*

$$H_x := (\ker(d_xF))^\perp \subseteq T_xM.$$

El hecho de suponer que d_F es sobreyectiva hace que la dimensión de H_x sea igual a $n = \dim(N)$. Esto permite definir el concepto de jacobiano normal, como sigue.

DEFINICIÓN 19. *Sean M y N dos variedades Riemannianas, de dimensiones respectivas $m \geq n$. Sea $x \in M$ un punto cualquiera y sea $F : M \to N$ una aplicación diferenciable tal que la diferencial d_xF es sobreyectiva. Sea $[v_1, ..., v_n]$ una base ortonormal del espacio horizontal en x, H_x. Definimos el jacobiano normal de F en x, denotado NJ_xF, como el volumen en $T_{F(x)}N$ del paralelepípedo definido por el conjunto de vectores*

$$d_xF(v_1), ..., d_xF(v_n).$$

En otras palabras, tenemos:

$$NJ_xF := |\det(d_xF|_{H_x})|,$$

donde $d_xF|_{H_x}$ se identifica con cualquiera de sus matrices coordenadas, en bases ortonormales. Si d_xF no es sobreyectiva, o si $m < n$, se define $NJ_xF := 0$.

El siguiente resultado permite calcular el jacobiano normal explícitamente en multitud de ocaciones.

LEMA A.6.4. *Sea $\mathbb{K} = \mathbb{R} \vee \mathbb{C}$ y $F : \mathbb{K}^m \to \mathbb{K}^n$ una aplicación diferenciable, y supongamos que $m \geq n$. Sea $x \in \mathbb{K}^m$ un punto cualquiera y sea $D \in \mathcal{M}_{n\times m}(\mathbb{K})$ la matriz de la diferencial de F en x en cualesquiera bases ortonormales de $\mathbb{R}^m$ y $\mathbb{R}^n$ (respectivamente en bases unitarias de $\mathbb{C}^m$ y $\mathbb{C}^n$). Entonces, se tiene la siguiente igualdad:*

$$NJ_xF = \det(DD^*)^{1/2}.$$

Donde si D^* denotará la matriz traspuesta de la matriz D en caso que $\mathbb{K} = \mathbb{R}$, y D^* denotará la traspuesta de la matriz obtenida de D mediante la conjugación de sus coordenadas en caso que $\mathbb{K} = \mathbb{C}$.

PROPOSICIÓN A.6.5. *Sean M_1, M_2, M_3 variedades Riemannianas, y sean $F : M_1 \to M_2$ y $G : M_2 \to M_3$ dos aplicaciones diferenciables. Sea $x \in M_1$ un punto y supongamos además que se satisface al menos una de las dos condiciones siguientes:*

- *La aplicación F es una isometría en x.*
- $\dim(M_2) = \dim(M_3)$.

Entonces, también se satisface la siguiente igualdad:

$$NJ(G \circ F) = NJ_{F(x)}G \cdot NJ_xF.$$

El siguiente Corolario indica que podemos mantener el jacobiano normal de una aplicación diferencial a través de isometrías.

COROLARIO A.6.6. *Sean M, N dos variedades Riemannianas cualesquiera, y sea $F : M \to N$ una aplicación diferenciable. Sean $x_1, x_2 \in M$ dos puntos regulares de F. Supongamos que existen isometrías $\varphi_M : M \to M$ y $\varphi_N : N \to N$ tales que $\varphi_M(x_1) = x_2$ y*

$$F \circ \varphi_M = \varphi_N \circ F.$$

Entonces, tenemos:

$$NJ_{x_1}F = NJ_{x_2}F.$$

Además, si existe una aplicación inversa $G : N \to M$ localmente definida alrededor de $F(x) \in N$, entonces

$$NJ_xF = \frac{1}{NJ_{F(x)}G}.$$

Con las notaciones de la definición de jacobiano normal, sea $y \in N$ un valor regular de la aplicación F (esto es, F es transversal a $\{y\}$). Entonces, el Corolario A.6.2 nos garantiza que $F^{-1}(y)$ es una variedad Riemanniana de dimensión $m - n$.
Más aún el Teorema A.5.1 nos garantiza que casi todo punto y de N es un punto regular de F. Por tanto, para casi todo punto y de N, podemos considerar integrales de funciones de la fibra $F^{-1}(y)$. Esta observación dota de sentido al siguiente resultado.

TEOREMA A.6.7 (**Fórmula de la Co-área**). *Sean M y N dos variedades Riemannianas. Sean m, n sus dimensiones respectivas, con $m \geq n$. Sea $F : M \to N$ una aplicación diferenciable sobreyectiva tal que la diferencial d_xF es sobreyectiva para casi todo punto $x \in M$. Sea $\phi : M \to [0, \infty]$ una función integrable. Entonces, se tienen las dos siguientes igualdades:*

$$\int_{x\in M} \phi(x)dM = \int_{y\in N}\int_{x\in F^{-1}(y)} \phi(x)\frac{1}{NJ_xF}dF^{-1}(y)dN,$$

$$\int_{x\in M} \phi(x)NJ_xFdM = \int_{y\in N}\int_{x\in F^{-1}(y)} \phi(x)dF^{-1}(y)dN.$$

Dicho Teorema es una forma del Teorema de Cambio de Variable en integración donde el usual jacobiano es reemplazado por el jacobiano normal NJ_xF.

A.6.4. Probabilidad e integrales. En esta sección siguiendo los resultados de [**BCSS, 1998**], introduciremos conceptos básicos de la teoría de probabilidad. Todas las probabilidades que vamos a considerar estan dadas por funciones de densidad.

DEFINICIÓN 20. *Sea M una variedad diferenciable con una forma de volumen $d\omega$. Sea $f : M \to R_+$ definida en casi todo punto de M, digamos en el complemento de una unión finita de subvariedades de menor dimensión de M, y supongamos que f es suficientemente regular para ser integrable con $\int_M f d\omega = 1$. Entonces, f es una función de densidad y definimos para $A \subset M$ (un dominio sobre el que podemos integrar) la probabilidad de A será*

$$m(A) = \int_M f d\omega.$$

Hemos escrito una integral simple representando la integral múltiple de f en lugar de repetir el signo de la integral tantas veces como sea la dimensión de M.

Veamos algunos ejemplos de probabilidades dadas por su función de densidad

LEMA A.6.8. *La distribución uniforme en la esfera unidad S^{n-1} es dada por*

$$m(A) = \frac{1}{\text{Vol}(S^{n-1})}\int_A 1 d\Theta,$$

donde $d\Theta$ es el volumen en la esfera.

LEMA A.6.9. *La distribución Gaussiana en $\mathbb{R}$ es dada por*

$$m(A) = \frac{1}{(2\pi)^{1/2}}\int_A e^{-(x^2/2)}dx, \ A \subset \mathbb{R}.$$

Este resultado podemos generalizarlo a R^k

LEMA A.6.10. *La distribución Gaussiana en* $\mathbb{R}^k$ *es dada por*

$$m(A) = \frac{1}{(2\pi)^{k/2}} \int \int \cdots \int_A \prod_{i=1}^{k} e^{-(x_i^2/2)} dx_1 \cdots dx_k, \ A \subset \mathbb{R}^k.$$

Sin embargo, debido a que la probabilidad de A *es el producto de las probabilidades de los* A_i*, donde* A_i *son los* k *lados de una sola dimensión de* A*, es decir los lados* A_i *son independientes. Por este motivo a través del Teorema de Fubini acabamos obteniendo*

$$m(A) = \frac{1}{(2\pi)^{k/2}} \int_A e^{-(\| x \|^2 /2)} dx_1 \cdots dx_k, \ A \subset \mathbb{R}^k.$$

El siguiente resultado muestra que si $\phi : \mathbb{R}^n \to \mathbb{R}$ es una función homogénea de grado cero (es decir, $\phi(\lambda x) = \phi(x)$ para todo $\lambda \in \mathbb{R} \setminus \{0\}$), el valor medio de ϕ respecto a la distribución Gaussiana en $\mathbb{R}^n$ es el mismo que el valor medio de ϕ respecto a la distribución uniforme en S^{n-1}.

PROPOSICIÓN A.6.11. *Sea* $\phi(x)$ *una función homogénea de grando cero, entonces*

$$\frac{1}{(2\pi)^{k/2}} \int_{\mathbb{R}^n} \phi(x) e^{-(\| x \|^2 /2)} dx = \frac{1}{\text{Vol}(S^{n-1})} \int_{S^{n-1}} \phi(\theta) d\theta,$$

asumiendo que las integrales existen.

Bibliografía

[BCSS, 1998] L. Blum, F. Cucker, M. Shub, and S. Smale, *"Complexity and real computation"*, Springer-Verlag, New York, 1998.

[CMP, 08] M. Callau, J.L. Montaña, Luis M. Pardo, *VC Dimension Bounds for Analytic Algebraic Computations*. In *"Computing and Combinatorics"* , Lecture Notes in Computer Science **5092**, X. Hu & J. Wang eds., Springer Verlag, 2008, 62-71.

[BePa, 07] C. Beltrán, Luis M. Pardo, *Estimates on the distribution of the condition number of singular matrices*. Foundations of Computational Mathematics **7** (2007), 87-134.

[High, 02] N. Highman, *"Accuracy and Stability of Numerical Algorithms"*. SIAM, second edition, 2002.

[Mal, 1933] R.R Mallock, *An electrical calculating machine*. Proc. of the Royal Soc. A, **140** (1933), 457-483.

[Wilk, 1963] J. H. Wilkinson, *"Rounding Errors in Algebraic Process"*. Prentice-Hall, Inc., Englewood Cliffs, 1963.

[Wilk, 1965] J. H. Wilkinson, *"The Algebraic Eugenvalue Problem"*. Monographs on Numerical Analysis (1st.edition), Oxford University Press/ Clarendon Press, 1965.

[Turing, 1948] A.M. Turing, *Rounding-off errors in matrix process* The Quarterly Journal of Mechanics and Applied Mathematics **1** (1948), 287-308.

[vNG, 1947] J. von Neumann, H.H. Goldstine, *Numerical inverting of matrices of high order* Bull. Amer. Math. Soc. **53** (1947), 1021-1099.

[Blum, 04] L. Blum, *Computability over the reals: Where Turing meets Newton*, Notices Amer. Math Soc. **51** (2004), 1024-1034.

[Blum, 12] L. Blum, *Alan Turing and the other theory of computation*. In S.Barry Cooper and Jan van Leeuwen, *"The Selected Works of A.M. Turing: His Work and Impact"*, Elsevier, 2012.

[Dem, 87] J.W. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*. Numer. Mathematic **51** (1987), 251-289.

[Higham, 02] N. Higham, *"Accuracy and Stability of Numerical Algorithms"*. SIAM, 2002.

[Demmel, 1997] J. Demmel, *"Applied Numerical Linear Algebra"*. SIAM, 1997.

[Wed, 1973] P.A. Wedin, *Perturbation Theory for pseudo-inverses*. BIT **13** (1973), 217-232.

[E&Y, 1936] C. Eckart, G. Young, *The approximation of one matrix by another of lower rank*. Psychometrika **1** (1936), 211-218.

[Mrs, 1960] L.Mirsky, *Symmetric Gauge Functions and Unitarily Invariant Norms*. Quarterly Journal of Mathematics Oxford **11** (1960), 50-59.

[Schm, 1967] E. Schmidt, *Zur Theorie der Linearen und Nichtlinearen Integralgleichungen. I Teil. Entwicklung Willkürlichen Funktionen nach System Vorgeschriebener*. Mathematische Annalen **63** (1907), 433-476.

[Stw, 1993] G. W. Stewart, *On the early history of the singular value decomposition*. SIAM Review **35** (1993), 551-566.

[Stw, 1977] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear least squares problems*. SIAM Rev., 19(4) (1977), 634-662.

[Pardo, 2021] L. M. Pardo, *"Apuntes para un curso de Álgebra Conmutativa"*, Manuscrito, Universidad de Cantabria, 2021.

[AGV, 1985] V.I. Arnold, S.M. Gusein-Zade, Alexander N. Varchenko, *" Singularities of Differentiable Maps, Volume 1: Classification of Critical Points, Caustics and Wave Fronts"*. Birkhäuser, 1985.

[Olazábal, 1988] J.M. Olazábal, *"Procedimientos Simbólicos en Álgebra Lineal"*. Servicio de Publicaciones, Universidad de Cantabria, 1988.

[Demazure, 1989] M. Demazure, *"Catastrophes et Bifurcations"*. Ellipses, École Polytechnique, 1989.

[Be, 09] C. Beltrán, *"Estimates on the condition number of random rank-deficient matrices"*. IMA Journal of Numerical Analysis **31** (2011),25-39.

[Be, 06] C. Beltrán, *"Sobre el Problema 17 de Smale: Teoría de la Intersección y Geometría Integral"*. Universidad de Cantabria, 2006.

[vN, 1963] J. von Neumann, *"Collected Works Volume V: Design of Computers, Theory of Automata and Numerical Analysis"* A.H. Taub, ed., pergamon, New York (1963).

[Birk&Gul, 1963] G. Birkhoff, S. Gulati, *Isotropic distributions of test matrices*. Journal of Applied Mathematics and Physics **30** (1979), 148–158.

[Smale, 1985] S. Smale *On the efficiency of algorithms of analysis*. Bull. Amer. Math. Soc. **13** (1985), 87-121.

[Demmel, 1988] J. Demmel *The probability that a numerical analysis problem is difficult*. Math. of Comput. **50** (1989), 449-480.

[Edelman, 1988] A. Edelman *"Eigenvalues and Condition Numbers of Random Matrices"*. SIAM J. Matrix Analysis Appl. **9** (1989), 553-560.

[Edelman, 1989] A. Edelman *"Eigenvalues and Condition Numbers of Random Matrices"*. Ph. D. Thesis, MIT, (1989).

[Chen&Dongarra, 2005] Z. Chen, J. J. Dongarra *"Condition Numbers of Gaussian Random Matrices"*. SIAM J. Matrix Analysis Appl. **27** (2005), 603-620.

[Mulmuley, 1986] K. Mulmuley. *A fast parallel algorithm to compute the rank of a matrix over an arbitrary field.* En Proceedings of the eighteenth annual ACM symposium on Theory of computing. (1986), 338-339.