Descomposición en valores singulares y análisis de factores en ciencias humans y sociales

Sergio A. Pernice[1]

Universidad del CEMA
Av. Córdoba 374, Buenos Aires, 1054, Argentina

5 de octubre de 2022

Abstract

La descomposición en valores singulares sistematiza y generaliza el análisis de factores, una metodología de investigación muy usada en ciencias humans y sociales, pero no forma parte de la formación estándar de investigadores en estas disciplinas. En este documento presentamos de manera didáctica dicha técnica. Además, la descomposición en valores singulares y sus numerosas aplicaciones recientes, invitan a ciertas reflexiones respecto del análisis de factores como metodología de investigación.

Keywords: descomposición en valores singulares, álgebra lineal, regresiones, Big Data, machine learning, aprendizaje automático, ciencias humans y sociales.

1 Introducción

Los objetos de estudio de las ciencias humanas y sociales son intrínsecamente complejos. Porque es filosóficamente atractiva, y además porque ayuda en la práctica a manejar dicha complejidad, una de las ideas fuerza más influyente a lo largo de la historia de dichas disciplinas, es la noción de que la gran cantidad de manifestaciones empíricas que caracterizan sus objetos de estudio son en realidad expresiones de unos pocos factores que influyen sobre todas las demás variables.

La correspondiente metodología estadística para implementar esas ideas tienen diferente nombre en distintas disciplinas, y la implementación concreta también difiere en los detalles en diferentes disciplinas, pero un nombre que puede ser claramente reconocido en muchas de ellas es el *análisis de factores* [Harman (1976)].

El método surgió hace más de un siglo, originalmente en psicología, en los trabajos de Charles Spearman [Spearman (1904), Spearman (1927)], Cyril Burt [Burt (1909)], Karl Pearson [Pearson (1901)],

[1] sp@ucema.edu.ar
Los puntos de vista del autor no representan necesariamente la posición de la Universidad del CEMA.

Godfrey H. Thomson [Thomson (1938)], J. C. Maxwell Garnett [Garnett-Whitehead (1919)], Karl Holzinger [Holzinger (1930)] y otros. Esos trabajos giraban alrededor de las ideas de Spearman, quien notando que diferentes tests que medían distintos tipo de habilidades relacionadas a la inteligencia tenían correlación positiva entre sí, concluyó que debía haber un factor central que influye en esas diversas capacidades cognitivas. A dicho factor lo llamó *inteligencia general*.

En estudios subsiguientes, para explicar mejor los datos, Spearman propuso su influyente "Teoría de los dos factores", a los que llamó "inteligencia general" y "habilidad especial". Además desarrolló métodos matemáticos que le permitían encontrar relaciones lineales entre los resultados de los diferentes tipos de tests y esos factores subyacentes, que explicaban gran parte de la variabilidad de los mismos.

Eventualmente, con datos empíricos más detallados, comenzó a acumularse evidencia que sugería que hay en realidad más factores detrás de las habilidades que genéricamente llamamos inteligencia. Fue entonces natural que, con los resultados de varios tipos diferentes de tests para muchos individuos, la correspondiente matriz de correlaciones se analizara como objeto matemático a ser aproximado por matrices más sencillas, con un número relativamente pequeño de "factores" que explicaran la mayor parte posible de la variabilidad de los datos.

Fue Louis Leon Thurstone quien propuso el criterio del *rango de la matriz* de correlación como base para determinar el número de factores comunes [Thurstone (1931), Thurstone (1947)] y formuló el problema en términos matriciales, lo que simplificó mucho el análisis posterior.

Concretamente, supongamos que d_{ij} de es el resultado del test tipo j del $i-$ésimo individuo, donde hay n tipos de tests que se le toman a m individuos elegidos al azar en una población mucho mayor que m. Podemos ordenar dichos resultados en la matriz

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} \tag{1.1}$$

donde los elementos de la columna j se pueden pensar como muestras de una variable aleatoria representativa de las habilidades poblacionales medidas en el test tipo j. Restándole a d_{ij} la media de los resultados del tests j

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} d_{ij}, \quad j = 1, \cdots, n \tag{1.2}$$

obtenemos la matriz "centrada"

$$\tilde{D} = \begin{bmatrix} d_{11} - \mu_1 & d_{12} - \mu_2 & \cdots & d_{1n} - \mu_n \\ d_{21} - \mu_1 & d_{22} - \mu_2 & \cdots & d_{2n} - \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} - \mu_1 & d_{m2} - \mu_2 & \cdots & d_{mn} - \mu_n \end{bmatrix} \tag{1.3}$$

La estimación empírica de la desviación estándar σ_j es

$$\sigma_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(d_{ij}-\mu_j\right)^2, \quad j = 1, \cdots, n \tag{1.4}$$

Las variables aleatorias $z_{ij} = (d_{ij} - \mu_j)/\sigma_j$ están normalizadas a varianza 1 con media cero.

Pasamos entonces de la matriz D en (1.1), o $\tilde{D}$ en (1.3), a la matriz normalizada Z:

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix} \tag{1.5}$$

Spearman propone un modelo del tipo

$$z_{ij} = a_{i1}I_{1j} + a_{i2}I_{2j} + \epsilon_{ij} \tag{1.6}$$

e interpreta a I_{1j} como el impacto que tiene la "inteligencia general" en las habilidades medidas en el test j, a I_{2j} como el impacto que tienen las "habilidades especiales" en las capacidades medidas en el test j, a_{i1} y a_{i2}, respectivamente, como la inteligencia general y habilidad especial del individuo i, y ϵ_{ij} es el error asociado a la medición del test j para el individuo i. La expectativa de estos modelos es que ϵ_{ij} tiendan a ser pequeños.

El modelo (1.6) se puede escribir matricialmente como

$$\underset{m\times n}{Z} = \underset{m\times 2}{A}\ \underset{2\times n}{I} + \underset{m\times n}{E} \tag{1.7}$$

donde la primera fila de I corresponde a la inteligencia general y la segunda a las habilidades especiales. La matriz AI tiene rango 2, y representa la mejor aproximación de rango 2 de la matriz Z.

En un modelo de k factores, la matriz A tiene dimensiones $m \times k$ y la matriz I dimensiones $k \times n$, con las k filas de A linealmente independientes y lo mismo para las k columnas de I. En ese caso la matriz AI tiene rango k, y representa la mejor aproximación de rango k de la matriz Z.

Muchos años más tarde, en finanzas, para entender los mercados de capitales, se replicó el mismo patrón: primero se propuso una teoría de un factor, para luego concluir que varios factores explican mejor los datos. El celebrado CAPM (Capital Assets Pricing Model) es un modelo de un factor [Sharpe (1964), Lintner (1965), Mossin (1966)] para predecir el retorno esperado de activos de riesgo en un mercado de capitales en equilibrio. Ameritó el premio Nobel de Economía para William Forsyth Sharpe en 1990 (compartido con Harry Markowitz y Merton Miller).

El modelo se generalizó a modelos de varios factores, por ejemplo en la popular "Teoría de arbitraje de precios" (Arbitrage pricing theory) de Stephen A. Ross [Ross (1976)], y otros modelos multifactoriales de Riesgo y Retorno. Tanto el CAPM como los modelos de varios factores forman parte de la formación estándar en maestrías y doctorados con especialidad en finanzas.

En macroeconomía, por ejemplo en la literatura sobre el ciclo económico real (RBC) y el equilibrio general estocástico dinámico (DSGE), típicamente se modelan unos pocos tipos de perturbaciones (factores) que afectan a todas las variables (productividad, demanda, oferta, etc.), ver por ejemplo [Stock and Watson (2015), Stock and Watson (2011)], [Bai-Ng (2002), Bai-Ng (2008)].

No es una exageración decir que esencialmente todas las ciencias humanas y sociales han utilizado, y siguen utilizando, modelos de factores. Las razones se mencionaron antes, pero vale la pena enfatizarlas. Por un lado, la idea de que unos pocos factores explican muchas datos empíricos es filosóficamente atractiva. Por el otro, es computacionalmente mucho más manejable: mientras que la matriz Z tiene $m \times n$ elementos, el producto AI tiene $(m + n) \times k$ elementos, lo cual es una gran ventaja si $k \ll n$.

A la luz de la importancia que tiene el análisis de factores en ciencias humanas y sociales, es un poco sorprendente que un método clásico de álgebra lineal conocido como "Descomposición en valores singulares" (SVD), que sistematiza y generaliza la descomposición en factores de *cualquier* matriz, y que constituye un tema estándar en cursos de posgrado en matemática aplicada, no sea parte de la formación estándar en las mencionadas disciplinas. Si bien su uso en estas disciplinas muestra una pendiente fuertemente creciente, ver [Athey y otros (2017), Athey-Imbens (2019)] y referencias en dichos trabajos, en la opinión del autor, entenderlo y manejarlo con destreza sería de gran importancia para cualquier investigador en estas áreas por al menos dos razones.

Por un lado, como se mencionó, este método automatiza y generaliza en muchas direcciones el análisis de factores.

Si el interés pasa por una matriz con un claro significado estadístico, como lo tienen la matriz D en (1.1), o $\tilde{D}$ en (1.3), o la matriz Z en (1.5), veremos que la matriz varianza-covarianza es $\tilde{D}^\top \tilde{D}/(m-1)$, y la matriz de correlaciones es $Z^\top Z/(n-1)$. Aplicando SVD a estas matrices, que por características específicas de las mismas que se explicarán más adelante, lleva el nombre de "análisis de componentes principales", uno obtiene automáticamente el análisis completo de factores de las correspondientes distribuciones.

Como la matriz original D en (1.1) fue generada a partir de n datos para cada uno de m individuos, es natural pensar en dichos datos como m vectores en un espacio vectorial de n dimensiones. Entonces, en esos casos, al análisis estadístico se le suma uno geométrico: el análisis de factores se reduce a aproximar esa "nube" de datos en $\mathbb{R}^n$ por sus proyecciones en el "mejor" hiperplano de k dimensiones.

Pero hay muchos ejemplos de datos en los que no hay una diferenciación clara de significado entre filas y columnas, como sí lo hay en la matriz D, ni hay un correspondiente significado estadístico de los datos. Por ejemplo, imaginemos en una economía una matriz en la que tanto las filas como las columnas representan personas humanas y jurídicas, y el elemento ij de la matriz representa alguna medida de cuánto comercian entre sí la persona i con la persona j. Como SVD funciona para *cualquier* matriz, con SVD tiene sentido pensar en factores aún para matrices como esta.

Más aún, la existencia misma de la macroeconomía como disciplina, con sus modelos basados en relativamente pocas variables como trabajo, capital, productividad, inflación, etc., depende de que dicha matriz tenga, y mantenga a lo largo del tiempo, un muy bajo "rango efectivo".

Una segunda razón para familiarizarse con SVD es que, debido en parte a algoritmos extremadamente optimizados para implementarlo, el método es de enorme importancia en la era de big data y machine learning, que influyen de manera creciente en la investigación en todas las áreas de estudio, y las ciencias humanas y sociales no son la excepción.

Por ejemplo, en aplicaciones de procesamiento de imágenes, como en el cálculo de "autocaras" (Eigenfaces) para proporcionar una representación eficiente de imágenes faciales en el reconocimiento facial [Muller-Magaia-Herbst (2004)] [Turk-Pentland (1991a), Turk-Pentland (1991b)].

En genómica [Alter-Brown-Botstein (2000)] [Holter y otros (2000)]. En un sorprendente trabajo, Novembre, Johnson, Bryc y otros, literalmente reproducen el mapa de Europa a partir de los dos factores principales en los vectores genéticos caracterizando las mutaciones de 3000 individuos europeos. Esto es especialmente sorprendente cuando uno observa que dichos vectores "viven" en más de medio millón de dimensiones [Novembre y otros (2008)].

Tal vez más sorprendente son los trabajos de procesamiento de lenguaje natural [Deerwester y otros (1990)]. En trabajos recientes basados en matrices de co-ocurrencia de palabras, donde el elemento ij de la matriz cuenta la cantidad de veces que las palabras i y j ocurren en un texto a menos de d palabras entre sí (por lo tanto la matriz es simétrica), siendo esos textos, por ejemplo, artículos de Google News, SVD descubre sorprendente estructura semántica y sintáctica en subespacios de relativamente baja dimension (en dimension 100 por ejemplo, cuando las palabras "viven" en $\sim$ 10.000 dimensiones). Además ha servido para detectar sesgos en dichos textos, y desarrollar algoritmos para quitarle dichos sesgos (debias). Esto es especialmente importante dado que tales algoritmos son usados por empresas para preseleccionar curriculum vitae de manera automática [Bolukbasi y otros (2016)].

SVD también muestra su poder para completar bases con datos faltantes, donde el objetivo es proporcionar la mejor predicción posible de lo que deberían ser esas entradas, ver por ejemplo [Athey y otros (2017), Athey-Imbens (2019)].

¿Qué es SVD?

El teorema fundamental establece que *cualquier* matriz A de dimensiones $m \times n$ se puede expresar como el producto de tres matrices:

$$\underset{m\times n}{A} = \underset{m\times m}{U}\ \underset{m\times n}{\Lambda}\ \underset{n\times n}{V^\top} \tag{1.8}$$

donde las matrices U y V son ortogonales (matrices cuadradas cuyas filas y columnas son vectores ortonormales), el superíndice "$\top$" indica transpuesta, y la matriz Λ es diagonal y semidefinida positiva. Es decir, los elementos no diagonales de Λ son todos cero, y los elementos de la diagonal principal λ_i son, o bien positivos, o cero, con tantos valores no nulos como rango k tenga la matriz A. Los elementos diagonales de Λ generalmente se ordenan de mayor a

menor $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Es decir, asumiendo que $n \leq m$, en cuyo caso necesariamente $k \leq n$,

$$\underset{m\times n}{A} = U\Lambda V^\top = \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u_1} & \cdots & \mathbf{u_m} \\ | & & | \end{bmatrix}}_{m\times m} \underbrace{\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}}_{m\times n} \underbrace{\begin{bmatrix} -\mathbf{v_1}^\top - \\ \vdots \\ -\mathbf{v_n}^\top - \end{bmatrix}}_{n\times n} \tag{1.9}$$

$$= \underbrace{\begin{bmatrix} | & & | \\ \mathbf{u_1} & \cdots & \mathbf{u_k} \\ | & & | \end{bmatrix}}_{m\times k} \underbrace{\begin{bmatrix} -\lambda_1\mathbf{v_1}^\top - \\ \vdots \\ -\lambda_k\mathbf{v_k}^\top - \end{bmatrix}}_{k\times n} \tag{1.10}$$

donde la segunda igualdad se deriva trivialmente de la primera asumiendo que $\lambda_i = 0$ para $i = k+1, \ldots, n$.

La forma (1.10) de la descomposición SVD es idéndtica a la forma (1.7) utilizada en análisis de factores y descartando la matriz de errores E, pero la forma (1.9) es la fundamental. De hecho, (1.10) no es la única manera de darle a A la forma (1.7). Por ejemplo, los valores singulares λ_i pueden multiplicar a las respectivas columnas de U en vez de las filas de $V^\top$, y cualquier combinación entre ambas también funciona.

La implementación de SVD en las librerías de software es estándar y tremendamente optimizada. Por ejemplo, en Matlab, escribiendo

```
[U, L, V] = svd(A)
```

Matlab devuelve las matrices U, L y V correspondientes a (1.9). Lo mismo ocurre con la librería linalg de numpy en Python, donde escribiendo

```
U, s, VT = np.linalg.svd(A, full_matrices=False)
```

Python devuelve las matrices U, $V^\top$, y s, que es un vector con los elementos diagonales de Λ. Por más detalles sobre el significado de "full_matrices=False" ver el link más arriba.

Si uno quiere aproximar una matriz A de rango k con una matriz de rango $\ell < k$, la mejor aproximación de rango ℓ consiste simplemente reemplazar por cero en (1.9) o (1.10) los valores singulares λ_i, para $i = \ell+1, \ldots, k$. "Mejor aproximación" en el sentido de la norma ℓ_2 de Frobenius:

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} \tag{1.11}$$

Es decir, para toda otra matriz B de rango k,

$$\left\|A - U\Lambda_k V^\top\right\|_F \leq \|A - B\|_F \tag{1.12}$$

donde Λ_k es Λ en (1.9) con los valores singulares más chicos, λ_i, $i = \ell + 1, \ldots, k$, reemplazados por cero.

La elección del rango ℓ de la aproximación deseada depende del usuario, es decir, ℓ es un parámetro exógeno en el método SVD.

El tiempo de ejecución de SVD escala rápido, como el menor de $O(m^2n)$ y $O(n^2m)$, por lo que una típica computadora portátil puede calcular el SVD de una matriz de 5000×5000 sin problemas, pero se le complica con una matriz de 10.000×10.000. Si uno se contenta con los mayores k valores singulares, como es usualmente el caso cuando uno busca la mejor aproximación de la matriz en k factores en lugar la descomposición completa, el tiempo de ejecución de SVD escala como $O(mnk)$, lo cual es una enorme ventaja si $k \ll m, n$. Además, para matrices especiales (como es el caso en la mayoría de las aplicaciones) hay algoritmos mucho más rápidos [Liberty y otros (2007)].

Este trabajo tiene dos objetivos. Por un lado, presentar a la comunidad de ciencias humanas y sociales un método sistemático, conocido como "Descomposición en valores singulares" (SVD), para descomponer *cualquier* matriz en factores. Aunque un mínimo de conocimiento de álgebra lineal se asume, el camino elegido tiene como sub-objetivo familiarizar aún más a la comunidad mencionada con métodos de álgebra lineal, de enrome y creciente importancia en la era de Big Data, y además permite introducir SVD de manera natural, donde el lector casi que va adivinando la forma antes de leerla.

El segundo objetivo, que desarrollaremos en la conclusión cuando se haya entendido el método, es invitar a cuestionar ciertas hipótesis subyacentes en el uso tradicional del análisis de factores en estas disciplinas. En la opinión del autor, la naturaleza misma de SVD, y los sorprendentes resultados de numerosas aplicaciones recientes, una minúscula parte de las cuales fueron mencionadas más arriba, ameritan una revisión crítica de tales hipótesis.

El resto del trabajo se organiza de siguiente manera: en la siguiente sección definimos la notación que vamos a utilizar y recordamos algunos elementos básicos de algebra lineal que se asumen conocidos, en la sección 3 recordamos las relación biyectiva entre matrices y funciones lineales y derivamos de manera simple la igualdad entre en "rango fila" y "rango columna" de cualquier matriz, en la sección 4 definimos el producto "outer" entre vectores, que suele no cubrirse en cursos básicos de algebra lineal y resulta fundamental para entender SVD de manera intuitiva, en la sección 5 derivamos SVD, en la sección 6 clarificamos la relación entre SVD, el análisis de factores y el análisis de componentes principales, finalmente, en la sección 7 resumimos el trabajo y, como mencionamos antes, analizamos ciertas hipótesis subyacentes en el uso tradicional del análisis de factores en estas disciplinas, invitando a una revisión crítica de tales hipótesis.

2 Notación

Usamos letras minúsculas en negrita "$\mathbf{v}$" para vectores, letras minúsculas "a" para escalares, y mayúsculas "A" para matrices. A menos que se especifique lo contrario los vectores son

"columna", y si queremos referirnos a vectores "fila" lo hacemos explícitamente con el símbolo de traspuesto $\mathbf{v}^\top$.

A los vectores de la base "canónica" la denotamos $\hat{\mathbf{e}}_{\mathbf{i,n}}$, donde $\mathbf{n}$ es la dimensión del espacio vectorial subyacente. Por ejemplo, en $\mathbb{R}^3$ tenemos:

$$\hat{\mathbf{e}}_{\mathbf{1,3}} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \hat{\mathbf{e}}_{\mathbf{2,3}} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \hat{\mathbf{e}}_{\mathbf{3,3}} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{2.1}$$

Vamos a usar la notación matricial $\mathbf{v}^\top\mathbf{w}$ para el producto escalar entre vectores, a los que pensamos como matrices $n \times 1$. Por ejemplo, para vectores en $\mathbb{R}^2$:

$$\mathbf{v}^\top\mathbf{w} = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = v_1 w_1 + v_2 w_2 \tag{2.2}$$

Con el producto escalar definimos la **"norma", o magnitud, o longitud de vectores** como

$$\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^\top\mathbf{v}} \tag{2.3}$$

y la **distancia entre dos vectores** $\mathbf{v}$ e $\mathbf{w}$ como $\|\mathbf{v} - \mathbf{w}\|$.

Con el producto escalar (2.2) el **ángulo** θ entre dos vectores $\mathbf{v}$ y $\mathbf{w} \in \mathbb{R}^n$ es

$$\cos\theta = \frac{\mathbf{v}^\top\mathbf{w}}{\|\mathbf{v}\|_2\|\mathbf{w}\|_2} \tag{2.4}$$

De (2.3) vemos que $\mathbf{v}/\|\mathbf{v}\|_2$ y $\mathbf{w}/\|\mathbf{w}\|_2$ tienen norma 1. Si el producto escalar entre dos vectores es cero, $\cos\theta = 0$, es decir θ es $\pi/2$ o $3\pi/2$, y los vectores son **ortogonales** (o perpendiculares).

Si $\mathbf{v}$ tiene norma 1, $\mathbf{v}^\top\mathbf{w}$ es la **proyección** de $\mathbf{w}$ en la dirección de $\mathbf{v}$, y $(\mathbf{v}^\top\mathbf{w})\,\mathbf{v}$ es un vector que apunta en la dirección de $\mathbf{v}$ y su norma es la proyección de $\mathbf{w}$ en la dirección de $\mathbf{v}$.

Dada una matriz real A "cuadrada", es decir, $A \in \mathbb{R}^{n\times n}$, la misma es simétrica si el elemento $a_{ij} = a_{ji}$, es decir, si $A^\top = A$. Una matriz cuadrada genérica de $n \times n$ tiene n^2 elementos independientes. Si es simétrica, tenemos los n elementos de la diagonal por un lado, y de los restantes $n^2 - n = n(n-1)$ elementos, solo la mitad son independientes. Entonces hay $n + n(n-1)/2 = n(n+1)/2$ elementos independientes.

Dada una matriz cuadrada A, un vector $\mathbf{v_i}$ no nulo se llama "autovector", y un escalar λ_i es el correspondiente "autovalor", si satisfacen la siguiente ecuación

$$A\mathbf{v_i} = \lambda_i\mathbf{v_i} \tag{2.5}$$

Todas las librerías de software numérico tienen funciones que calculan autovalores y autovectores.

Para una matriz real cuadrada genérica A, en general, tanto los autovalores como las coordenadas de los autovectores son números complejos. Pero si la matriz es simétrica resulta que varias propiedades se satisfacen:

1. Tanto los autovalores como las coordenadas de los autovectores son números reales.

2. Si la matriz es de $n \times n$, hay exactamente n autovalores y n autovectores (si dos o más autovalores son iguales esto sigue valiendo con ciertas aclaraciones, pero son irrelevantes para nuestros propósitos.)

3. Los autovectores son ortogonales entre sí, y se los puede elegir de norma 1. Es decir, si $\mathbf{v_i}$ y $\mathbf{v_j}$ son dos autovectores diferentes, entonces $\mathbf{v_i}^\top \mathbf{v_j} = 0$, y $\mathbf{v_i}^\top \mathbf{v_i} = 1$.

3 Matrices y funciones lineales entre espacios vectoriales

Hay una relación biyectiva (uno a uno) entre matrices $A \in \mathbb{R}^{m \times n}$ y **funciones lineales** $f_A : \mathbb{R}^n \to \mathbb{R}^m$.

Si $\mathbf{v_1}$ y $\mathbf{v_2} \in \mathbb{R}^n$, $\mathbf{w_1} = A\mathbf{v_1}$ y $\mathbf{w_2} = A\mathbf{v_2} \in \mathbb{R}^m$, entonces $A(a_1\mathbf{v_1} + a_2\mathbf{v_2}) = a_1 A\mathbf{v_1} + a_2 A\mathbf{v_2} = a_1\mathbf{w_1} + a_2\mathbf{w_2}$. Mirado como funciones lineales, en general, si V y W son espacios vectoriales reales de dimensiones n y m respectivamente, una función $f : V \to W$ es **lineal** si

$$\mathbf{f}(a\mathbf{v_1} + b\mathbf{v_2}) = a\mathbf{f}(\mathbf{v_1}) + b\mathbf{f}(\mathbf{v_2}) \tag{3.1}$$

donde $\mathbf{v_1}, \mathbf{v_2} \in V$, $\mathbf{f}(\mathbf{v_1}), \mathbf{f}(\mathbf{v_2}) \in W$, y $a, b \in \mathbb{R}$.

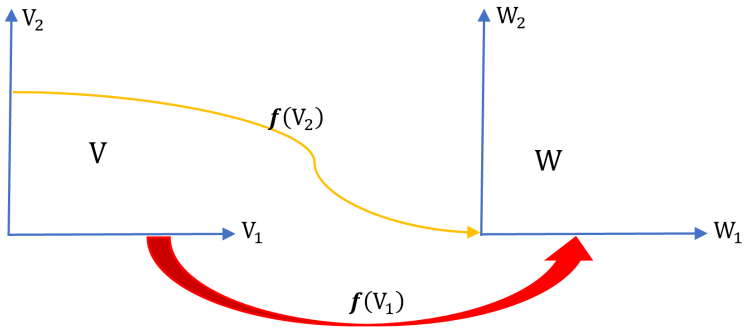

Figure 1: $\dim V = \dim V_1 + \dim V_2$, and $\dim V_1 = \dim W_1$.

Consideremos el conjunto $V_2 \subset V$ tal que si $\mathbf{v_2} \in V_2$, $\mathbf{f}(\mathbf{v_2}) = \mathbf{0}$, ver figura 1. V_2 es un subespacio vectorial de V (si $\mathbf{v}_{2,a}$ y $\mathbf{v}_{2,b} \in V_2$, la linealidad de $\mathbf{f}$ implica $\mathbf{f}\left(\alpha_a \mathbf{v}_{2,a} + \alpha_b \mathbf{v}_{2,b}\right) = \mathbf{0}$, por lo que $\alpha_a \mathbf{v}_{2,a} + \alpha_b \mathbf{v}_{2,b} \in V_2$) que se conoce como **espacio nulo** de $\mathbf{f}$, o $N(\mathbf{f})$.

Todo elemento $\mathbf{v} \in V$ se puede expresar de manera única de la forma $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, donde $\mathbf{v}_1$ pertenece al complemento ortogonal de V_2, al que llamamos V_1, y $\mathbf{v}_2 \in V_2$. En general, en un espacio con producto escalar, el complemento ortogonal de un subespacio es también un subespacio. Se dice que V es la "suma directa" de V_1 y V_2, $V = V_1 \oplus V_2$, $V_1 \cap V_2 = \{\mathbf{0}\}$), y $\mathbf{v}_1^\top \mathbf{v}_2 = 0$.

Ya sabemos que la imagen de V_2 es $\{\mathbf{0}\}$, consideremos ahora la imagen de V_1, a la que llamamos $W_1 \subset W$, ver figura 1, es decir,

$$W_1 = \{\mathbf{w} \in W \ / \ \exists\ \mathbf{v} \in V_1, \mathbf{f}(\mathbf{v}) = \mathbf{w}\} \tag{3.2}$$

Resulta que $\dim W_1 = \dim V_1$.

Para ver esto simplemente consideremos una base de vectores linealmente independientes de V_1, $\{\mathbf{v_{1,1}}, \ldots, \mathbf{v_{d,1}}\}$, $d \leq n$, y elijamos los dos primeros, $\mathbf{v_{1,1}}$ y $\mathbf{v_{2,1}}$. $\mathbf{f}(\mathbf{v_{1,1}})$ y $\mathbf{f}(\mathbf{v_{2,1}})$ son linealmente independientes. Si no lo fueran, existirían dos escalares a_1 y a_2 no nulos, tales que $a_1\mathbf{f}(\mathbf{v_{1,1}}) + a_2\mathbf{f}(\mathbf{v_{2,1}}) = \mathbf{0}$. Pero $a_1\mathbf{f}(\mathbf{v_{1,1}}) + a_2\mathbf{f}(\mathbf{v_{2,1}}) = \mathbf{f}(a_1\mathbf{v_{1,1}} + a_2\mathbf{v_{2,1}})$, por lo que, o bien $a_1\mathbf{v_{1,1}} + a_2\mathbf{v_{2,1}} = \mathbf{0}$, o bien $\in V_2$. Pero ninguna de esas alternativas es posible, ya que por hipótesis $\mathbf{v_{1,1}}$ y $\mathbf{v_{2,1}}$ son linealmente independientes, y pertenecen a V_1, por lo que cualquier combinación lineal de ellos también pertenece a V_1.

Consideremos ahora cualquier vector linealmente dependiente de $\{\mathbf{v_{1,1}}, \mathbf{v_{2,1}}\}$, al que llamamos $\mathbf{v_{12,1}}$, y $\mathbf{v_{3,1}}$, el tercer elemento de la base de V_1. Por el mismo razonamiento que antes, $\mathbf{f}(\mathbf{v_{12,1}})$ y $\mathbf{f}(\mathbf{v_{3,1}})$ son linealmente independientes. Iterando hasta completar la base $\{\mathbf{v_{1,1}}, \ldots, \mathbf{v_{d,1}}\}$ vemos que el conjunto de d vectores de W_1 $\{\mathbf{f}(\mathbf{v_{1,1}}), \ldots, \mathbf{f}(\mathbf{v_{d,1}})\}$ es linealmente independiente. Por otro lado, vimos que cualquier vector $\mathbf{v}$ que no sea linealmente dependiente de la base $\{\mathbf{v_{1,1}}, \ldots, \mathbf{v_{d,1}}\}$, se puede escribir como $\mathbf{v} = \mathbf{v_1} + \mathbf{v_2}$, con $\mathbf{v_1} \in V_1$ y $\mathbf{v_2} \in V_2$, por lo que $\mathbf{f}(\mathbf{v}) = \mathbf{f}(\mathbf{v_1})$, entonces $\mathbf{f}(\mathbf{v})$ también es linealmente dependiente de $\{\mathbf{f}(\mathbf{v_{1,1}}), \ldots, \mathbf{f}(\mathbf{v_{d,1}})\}$. Por lo tanto $\dim W_1 = \dim V_1$.

Para traducir este resultado al lenguaje de matrices, conviene recordar dos maneras diferentes de mirar el producto matriz por vector. La primera es mirando a la matriz como un conjunto de filas:

$$\mathbf{w} = A\mathbf{v} = \begin{bmatrix} -\mathbf{f}_1^\top - \\ \vdots \\ -\mathbf{f}_m^\top - \end{bmatrix} \mathbf{v} = \begin{pmatrix} \mathbf{f}_1^\top \mathbf{v} \\ \vdots \\ \mathbf{f}_m^\top \mathbf{v} \end{pmatrix} \tag{3.3}$$

cada componente es el producto escalar del correspondiente vector fila de A con $\mathbf{v}$.

Vimos que todo elemento $\mathbf{v} \in V = \mathbb{R}^n$ se puede expresar de manera única de la forma $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, donde $\mathbf{v}_1 \in V_1$, $\mathbf{v}_2 \in V_2 = N(A)$ es el espacio nulo de A, y $\mathbf{v}_1^\top \mathbf{v}_2 = 0$. La expresión (3.3) hace explícito que para todo elemento $\mathbf{v}_2 \in V_2 = N(A)$, $A\mathbf{v_2} = \mathbf{0}$, si y solo si $\mathbf{f}_i^\top \mathbf{v_2} = \mathbf{0}$ para todo $i = 1, \ldots, m$. Es decir, $\mathbf{v_2}$ es ortogonal a los vectores cuyos transpuestos son las filas de A, por lo que V_1 es el subespacio de $\mathbb{R}^n$ linealmente dependiente de dichos vectores y V_2 es el subespacio de $\mathbb{R}^n$ ortogonal a V_1. Llamamos **rango fila** de A a la dimension de V_1, es decir, al numero de filas linealmente independientes.

La segunda manera de mirar al producto matriz por vector es mirando a la matriz como un conjunto de columnas:

$$\mathbf{w} = A\mathbf{v} = \begin{bmatrix} | & & | \\ \mathbf{c_1} & \cdots & \mathbf{c_n} \\ | & & | \end{bmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = v_1\mathbf{c_1} + \cdots + v_n\mathbf{c_n} \tag{3.4}$$

El resultado es una combinación lineal de los vectores columna de A con coeficientes iguales a

las componentes de $\mathbf{v}$. Llamamos **rango columna** de A a la dimension de W_1 (la imagen de V_1.) El rango columna, es entonces el numero de columnas de A linealmente independientes.

Pero vimos que $\dim W_1 = \dim V_1$, por lo que **rango columna = rango fila = rango de** A. Es decir, el subespacio de todas las combinaciones lineales de las filas de A tiene igual dimension que el subespacio de todas las combinaciones lineales de las columnas de A. O, equivalentemente, el número de filas linealmente independiente es igual al número de columnas linealmente independientes, independientemente de cuánto sea m y n. Este resultado no es del todo intuitivo a partir de lo visto hasta este punto. Se volverá intuitivo cuando veamos una tercer manera de mirar a las matrices: como suma de productos "outer", el tema de la siguiente sección.

4 Producto "outer"

Llamamos producto "outer"[2] entre un vector $\mathbf{v} \in \mathbb{R}^m$ y otro vector $\mathbf{w} \in \mathbb{R}^n$ a

$$\mathbf{v}\mathbf{w}^\top = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \begin{pmatrix} w_1 & \cdots & w_n \end{pmatrix} = \begin{bmatrix} v_1 w_1 & v_1 w_2 & \cdots & v_1 w_n \\ v_2 w_1 & v_2 w_2 & \cdots & v_2 w_n \\ \vdots & \vdots & \ddots & \vdots \\ v_m w_1 & v_m w_2 & \cdots & v_m w_n \end{bmatrix} \in \mathbb{R}^{m \times n} \tag{4.1}$$

Es decir, el producto outer es simplemente el producto matricial entre la matriz $m \times 1$ asociada al vector $\mathbf{v}$ y la matriz $1 \times n$ asociada al vector $\mathbf{w}^\top$, siendo el resultado una matriz de $m \times n$.

Por ejemplo,

$$\hat{\mathbf{e}}_{\mathbf{2},2}\hat{\mathbf{e}}_{\mathbf{2},3}^\top = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{4.2}$$

Siguiendo con este ejemplo, queda claro que cualquier matriz A de 2×3 puede escribirse como una combinación lineal de productos outer de las correspondientes bases canónicas:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} = \sum_{i=1}^{2} \sum_{j=1}^{3} a_{ij}\, \hat{\mathbf{e}}_{\mathbf{i},2}\hat{\mathbf{e}}_{\mathbf{j},3}^\top \tag{4.3}$$

En general, si $\hat{\mathbf{e}}_{\mathbf{i},\mathbf{m}}$, $i = 1, \ldots, m$, denotan a los vectores de la base canónica en $\mathbb{R}^m$ y $\hat{\mathbf{e}}_{\mathbf{j},\mathbf{n}}$, $j = 1, \ldots, n$, a los vectores de la base canónica en $\mathbb{R}^n$, cualquier matriz A de $m \times n$ puede escribirse como una combinación lineal de productos outer de las correspondientes bases canónicas:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}\, \hat{\mathbf{e}}_{\mathbf{i},\mathbf{m}}\hat{\mathbf{e}}_{\mathbf{j},\mathbf{n}}^\top \tag{4.4}$$

[2] El producto "outer" no tiene un nombre estándar en español, por lo que adoptamos para este trabajo su nombre en inglés.

La matriz A se puede escribir de infinitas maneras diferentes como combinación lineal de productos outer. Nos enfocamos en dos.

Mirando a la matriz A como un conjunto de n vectores columna, cada uno en $\mathbb{R}^m$:

$$\begin{aligned} A &= \begin{bmatrix} | & & | \\ \mathbf{c_1} & \cdots & \mathbf{c_n} \\ | & & | \end{bmatrix} = \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} + \cdots + \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} \begin{pmatrix} 0 & \dots & 0 & 1 \end{pmatrix} \\ &= \sum_{j=1}^{n} \mathbf{c_j}\, \hat{\mathbf{e}}_{\mathbf{j,n}}^{\top}, \end{aligned} \tag{4.5}$$

donde $(\mathbf{c_j})_i = a_{ij}$. Es decir, el elemento i del vector columna j es $(\mathbf{c_j})_i = a_{ij}$.

De manera análoga, mirándola como un conjunto m vectores, cada uno en $\mathbb{R}^n$, cuyos transpuestos corresponden a las filas de A:

$$\begin{aligned} A &= \begin{bmatrix} -\mathbf{f_1}^{\top}- \\ \vdots \\ -\mathbf{f_m}^{\top}- \end{bmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} a_{11} & \dots & a_{1n} \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} a_{m1} & \dots & a_{mn} \end{pmatrix} \\ &= \sum_{i=1}^{m} \hat{\mathbf{e}}_{\mathbf{i,m}}\, \mathbf{f_i}^{\top} \end{aligned} \tag{4.6}$$

donde $(\mathbf{f_i}^{\top})_j = a_{ij}$. Es decir, el elemento j del vector fila i es $(\mathbf{f_i}^{\top})_j = a_{ij}$.

Veamos el efecto de multiplicar por derecha un vector columna por el producto outer entre dos vectores (asumimos que las dimensiones son correctas):

$$(\mathbf{x}\,\mathbf{y}^{\top})\,\mathbf{z} = \mathbf{x}\,(\mathbf{y}^{\top}\mathbf{z}) \tag{4.7}$$

Como todo producto matricial, el producto outer satisface la propiedad asociativa. La última igualdad, sin embargo, tiene gran utilidad, porque el producto escalar entre $\mathbf{y}$ y $\mathbf{z}$ va a ser cero si esos vectores son ortogonales. Es decir que el efecto de la matriz $\mathbf{x}\ \mathbf{y}^{\top}$ sobre el vector $\mathbf{z}$ es multiplicar *escalarmente* a $\mathbf{z}$ por $\mathbf{y}$ y al escalar resultante multiplicarlo por el vector $\mathbf{x}$. El resultado tiene la dirección de $\mathbf{x}$. Por lo tanto si $\mathbf{y} = \mathbf{x}$ y $|\mathbf{x}| = 1$, $\mathbf{x}\ \mathbf{x}^{\top}$ es un *proyector* en el subespacio generado por $\mathbf{x}$.

De la misma manera, multiplicando por izquierda un vector fila por el producto outer entre dos vectores (nuevamente asumimos que las dimensiones son correctas) es:

$$\mathbf{z}^{\top}\,(\mathbf{x}\,\mathbf{y}^{\top}) = (\mathbf{z}^{\top}\mathbf{x})\,\mathbf{y}^{\top} \tag{4.8}$$

Además de la propiedad asociativa por derecha y por izquierda, es fácil ver que el producto outer satisface las siguientes propiedades:

$$\begin{aligned} (\mathbf{x}\mathbf{y}^{\top})^{\top} &= (\mathbf{y}\mathbf{x}^{\top}) \\ (\mathbf{y}+\mathbf{z})\mathbf{x}^{\top} &= \mathbf{y}\mathbf{x}^{\top} + \mathbf{z}\mathbf{x}^{\top} \\ \mathbf{x}(\mathbf{y}^{\top}+\mathbf{z}^{\top}) &= \mathbf{x}\mathbf{y}^{\top} + \mathbf{x}\mathbf{z}^{\top} \\ c(\mathbf{x}\mathbf{y}^{\top}) &= (c\mathbf{x})\mathbf{y}^{\top} = \mathbf{x}(c\mathbf{y}^{\top}) \end{aligned} \tag{4.9}$$

Prometimos en la sección anterior que el hecho de hay igual número de filas y columnas linealmente independientes, o, equivalentemente, que $\dim V_1 = \dim W_1$, se volvería intuitivo con productos outer. Veremos ahora que las propiedades (4.9) es todo lo que necesitamos para convencernos de este resultado.

Supongamos que el rango fila de la matriz $A \in \mathbb{R}^{m\times n}$ es k. Claramente $k \leq n$ porque las filas de A son vectores en $\mathbb{R}^n$ transpuestos, y hay como máximo n vectores linealmente independientes en $\mathbb{R}^n$. Además $k \leq m$ porque hay solo m filas.

Mirando a la matriz A en la forma (4.6), supongamos, sin perdida de generalidad, que las primeras k filas son linealmente independientes y las filas $k+1, \ldots, n$ son linealmente dependientes de las primeras k, es decir

$$\mathbf{f_j} = \sum_{i=1}^{k} f_j^i \, \mathbf{f_i}, \qquad j = k+1, \ldots, m \tag{4.10}$$

donde los escalares $f_j^i,\ i = 1, \ldots, k,\ j = k+1, \ldots, m$ son únicos. Entonces, de (4.6)

$$A = \sum_{i=1}^{m} \hat{\mathbf{e}}_{\mathbf{i,m}} \, \mathbf{f_i}^\top = \sum_{i=1}^{k} \hat{\mathbf{e}}_{\mathbf{i,m}} \, \mathbf{f_i}^\top + \sum_{j=k+1}^{m} \hat{\mathbf{e}}_{\mathbf{j,m}} \, \mathbf{f_j}^\top \tag{4.11}$$

Insertando (4.10) en el último término de (4.11), y usando las propiedades (4.9) del producto outer

$$\sum_{j=k+1}^{m} \hat{\mathbf{e}}_{\mathbf{j,m}} \, \mathbf{f_j}^\top = \sum_{i=1}^{k} \left(\sum_{j=k+1}^{m} f_j^i \, \hat{\mathbf{e}}_{\mathbf{j,m}} \right) \mathbf{f_i}^\top \tag{4.12}$$

Reemplazando esta expresión en (4.11), usando nuevamente las propiedades (4.9),

$$A = \sum_{i=1}^{k} \hat{\mathbf{e}}_{\mathbf{i,m}} \, \mathbf{f_i}^\top + \sum_{i=1}^{k} \left(\sum_{j=k+1}^{m} f_j^i \hat{\mathbf{e}}_{\mathbf{j,m}} \right) \mathbf{f_i}^\top = \sum_{i=1}^{k} \left(\hat{\mathbf{e}}_{\mathbf{i,m}} + \sum_{j=k+1}^{m} f_j^i \hat{\mathbf{e}}_{\mathbf{j,m}} \right) \mathbf{f_i}^\top \tag{4.13}$$

En esta última expresión es una suma de k productos outer. Por hipótesis, los vectores $\mathbf{f_i}$ son linealmente independientes, y los vectores

$$\mathbf{p_i} = \hat{\mathbf{e}}_{\mathbf{i,m}} + \sum_{j=k+1}^{m} f_j^i \hat{\mathbf{e}}_{\mathbf{j,m}}, \qquad i = 1, \cdots, k \tag{4.14}$$

obviamente también lo son, ya que los $\hat{\mathbf{e}}_{\mathbf{i,m}}$ son los primeros k vectores de la base canónica de $\mathbb{R}^m$. Entonces, en (4.13) logramos expresar a la matriz A como una suma de k productos outer de la forma

$$A = \sum_{i=1}^{k} \mathbf{p_i} \mathbf{f_i}^\top \tag{4.15}$$

donde los k vectores $\mathbf{p_i} \in \mathbb{R}^m$ son linealmente independientes y los k vectores $\mathbf{f_i} \in \mathbb{R}^n$ también lo son. (4.15) inmediatamente implica que $\dim V_1 = \dim W_1$. Por hipótesis sabemos que $\dim V_1 = k$, y para todo vector $\mathbf{v} \in \mathbb{R}^n$,

$$A\mathbf{v} = \sum_{i=1}^{k} \mathbf{p_i} (\mathbf{f_i}^\top \mathbf{v}) \tag{4.16}$$

es una combinación lineal de los k vectores $\mathbf{p_i}$. Variando $\mathbf{v}$ en V_1 podemos hacer que los k productos escalares $\mathbf{f_i}^\top \mathbf{v}$, $i = 1, \ldots, k$, adquieran cualquier valor deseado, por lo que cubrimos toda posible combinación lineal de los k vectores $\mathbf{p_i} \in \mathbb{R}^m$. Como los $\mathbf{p_i}$ son linealmente independientes, se sigue que dim W_1 también es k.

Un argumento análogo muestra que si tenemos k columnas de A linealmente independientes, entonces tiene que haber k filas linealmente independientes, o, que si dim $W_1 = k$, entonces dim $V_1 = k$.

Finalizamos esta sección expresando el producto de matrices en términos de productos outer. Supongamos que la matriz $C \in \mathbb{R}^{m\times n}$ es el producto de la matriz $A \in \mathbb{R}^{m\times p}$ por la matrix $B \in \mathbb{R}^{p\times n}$. Entonces, el elemento ij de C es

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj} \tag{4.17}$$

Consideremos, por ejemplo, el elemento $k = 2$ en la suma en (4.17): $a_{i2}b_{2j}$. Es el elemento $i-$ésimo de la columna 2 de la matriz A, multiplicado por el elemento $j-$ésimo de la fila 2 de la matriz B. Esto, a su vez, se puede pensar como el elemento ij del producto outer entre el vector $\mathbf{a_2} \in \mathbb{R}^{m\times 1}$, correspondiente a la segunda columna de A y el vector $\mathbf{b_2} \in \mathbb{R}^{n\times 1}$, cuyo transpuesto es la fila 2 de la matriz B. En otras palabras, $(\mathbf{a_2}\mathbf{b_2}^\top)_{ij} = a_{i2}b_{2j}$, o

$$C = AB = \begin{bmatrix} | & & | \\ \mathbf{a_1} & \cdots & \mathbf{a_p} \\ | & & | \end{bmatrix} \begin{bmatrix} -\mathbf{b_1}^\top - \\ \vdots \\ -\mathbf{b_p}^\top - \end{bmatrix} = \sum_{k=1}^{p} \mathbf{a_k}\, \mathbf{b_k}^\top \tag{4.18}$$

5 Descomposición en valores singulares

Como vimos, la expresión (4.15), válida para cualquier matriz A de rango k, inmediatamente implica dim V_1 = dim W_1. Como veremos, la descomposición en valores singulares simplemente requiere probar que A se puede escribir como la suma k productos outer donde, además, los k vectores $\{\mathbf{p_i}\}$ y los k vectores $\{\mathbf{f_i}\}$ se pueden elegir simultáneamente ortonormales. Empecemos el analisis con matrices de 2×2.

En la figura 2 vemos que generando una matriz D al azar, las matrices simétricas y semidefinidas positivas $DD^\top$ y $D^\top D$ tienen obviamente autovectores ortogonales, pero no es obvia la relación entre ellos, como se ve en las diferentes figuras. Sin embargo, los autovalores de ambas matrices siempre son iguales. Tratemos de entender por qué esto es así.

$DD^\top$ es simétrica, ya que $(DD^\top)^\top = (D^\top)^\top D^\top = DD^\top$, por lo que sus autovalores son reales y sus autovectores son reales ortogonales. Además, dado cualquier vector $\mathbf{v}$, la norma al cuadrado de $\mathbf{w} = D^\top \mathbf{v}$ es $\mathbf{w}^\top \mathbf{w} = \mathbf{v}^\top DD^\top \mathbf{v} \geq 0$, por lo que $DD^\top$ es semidefinida positiva. Mismas conclusiones valen para $D^\top D$.

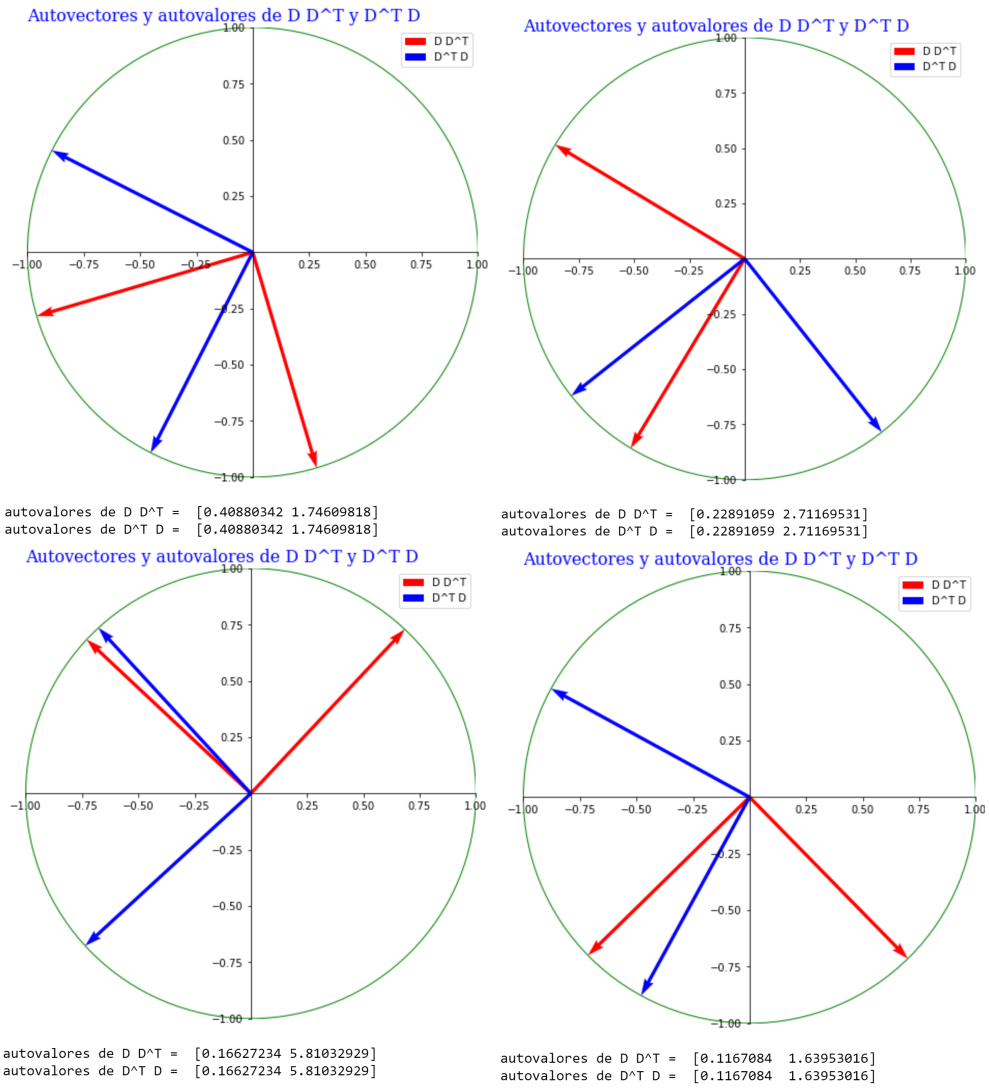

Figure 2: La matriz D es generada al azar con valores surgidos de una normal $N(0,1)$ (**D = np.random.randn(2,2)** en Python numpy.) Las matrices simétricas y semidefinidas positivas $DD^\top$ y $D^\top D$ tienen autovectores ortogonales, pero no obvia relación entre ellos, como se ve en las diferentes figuras. Sin embargo, los autovalores de ambas matrices siempre son iguales.

Los autovectores y autovalores de $DD^\top$ y $D^\top D$ satisfacen las siguientes ecuaciones

$$D^\top D\mathbf{v}_i = \lambda_i^2 \mathbf{v}_i, \quad i = 0, 1 \tag{5.1}$$

$$DD^\top \mathbf{w}_j = (\lambda')_j^2 \mathbf{w}_j, \quad j = 0, 1 \tag{5.2}$$

Escribimos los autovalores como el cuadrado de un número real porque sabemos que son no negativos. La figura 2 indica que, aunque generamos las matrices D al azar, podemos ordenar los autovalores de $D^\top D$ y $DD^\top$ de modo que $\lambda_i^2 = (\lambda')_i^2$. Cómo podemos entender esto? Consid-

eremos por ejemplo (5.2), y multipliquemos por izquierda por $D^\top$:

$$D^\top (DD^\top)\,\mathbf{w}_i = (D^\top D)\,(D^\top \mathbf{w}_i) = \lambda_i^2\,(D^\top \mathbf{w}_i) \tag{5.3}$$

comparando (5.3) con (5.1) e identificando a $\mathbf{v}_i$ con $D^\top \mathbf{w}_i$, explicamos por qué $\lambda_i^2 = (\lambda')_i^2$ y además descubrimos una relación entre los autovectores de ambas matrices que no era obvia desde la figura 2:

$$\mathbf{v}_i \propto D^\top \mathbf{w}_i, \quad \lambda_i^2 = (\lambda')_i^2 \tag{5.4}$$

Ponemos el signo de proporcionalidad y no el de igualdad porque nada asegura que $D^\top \mathbf{w}_i$ esté normalizado a uno si $\mathbf{w}_i$ lo está.

De la misma manera, multiplicando con D por izquierda a (5.1) llegamos a la conclusión de que

$$\mathbf{w}_i \propto D^\top \mathbf{v}_i, \quad (\lambda')_i^2 = \lambda_i^2 \tag{5.5}$$

Acabamos de encontrar una relación interesante entre los autovectores de $DD^\top$ y $D^\top D$. Como mencionamos, estas matrices son simétricas, por lo que sus autovectores normalizados a uno forman bases ortonormales de $\mathbb{R}^2$. Además son semidefinidas positivas, es decir, sus autovalores son positivos o cero.

$$\{\mathbf{v}_i\} \text{ base de } \mathbb{R}^2, \quad \mathbf{v}_i^\top \mathbf{v}_j = \delta_{ij}, \quad D^\top D\mathbf{v}_i = \lambda_i^2 \mathbf{v}_i, \quad \lambda_i^2 \geq 0, \quad i = 0, 1, \tag{5.6}$$

$$\{\mathbf{w}_i\} \text{ base de } \mathbb{R}^2, \quad \mathbf{w}_i^\top \mathbf{w}_j = \delta_{ij}, \quad DD^\top \mathbf{w}_i = \lambda_i^2 \mathbf{w}_i, \quad \lambda_i^2 \geq 0, \quad i = 0, 1, \tag{5.7}$$

esto significa que admiten la siguiente descripción en términos de productos outer:

$$D^\top D = \sum_{i=0}^{1} \lambda_i^2 \mathbf{v}_i \mathbf{v}_i^\top \tag{5.8}$$

$$DD^\top = \sum_{i=0}^{1} \lambda_i^2 \mathbf{w}_i \mathbf{w}_i^\top \tag{5.9}$$

Por ejemplo, los autovectores y autovalores determinan unívocamente a la matriz $D^\top D$, y $D^\top D\mathbf{v}_j = \sum_{i=0}^{1} \lambda_i^2 \mathbf{v}_i(\mathbf{v}_i^\top \mathbf{v}_j) = \lambda_j^2 \mathbf{v}_j$.

Mirando las expresiones (5.8-5.9) resulta casi obvio que D y $D^\top$ se pueden escribir así:

$$D = \sum_{i=0}^{1} \lambda_i\,(\pm\mathbf{w}_i)(\pm\mathbf{v}_i^\top) \tag{5.10}$$

$$D^\top = \sum_{j=0}^{1} \lambda_j\,(\pm\mathbf{v}_j)(\pm\mathbf{w}_j^\top) \tag{5.11}$$

donde λ_i es la raíz real no negativa del número real no negativo λ_i^2. La restricción de que los vectores de las bases $\{\mathbf{v}_i\}$ y $\{\mathbf{w}_i\}$ en (5.6-5.7) estén normalizados a 1 deja ambiguo un signo ± 1,

pero alguna de las $2^2 = 4$ combinaciones tiene que ser correcta. Sean cuales sean los signos correctos,

$$\begin{aligned} DD^\top &= \sum_{i=0}^{1}\sum_{j=0}^{1} \lambda_i\lambda_j\ (\pm\mathbf{w}_i)(\pm\mathbf{v}_i^\top)(\pm\mathbf{v}_j)(\pm\mathbf{w}_j^\top) \\ &= \sum_{i=0}^{1} \lambda_i^2\ \mathbf{w}_i\mathbf{w}_i^\top \end{aligned} \tag{5.12}$$

recuperamos (5.9). En (5.12) usamos la propiedad asociativa y $(\pm\mathbf{v}_i^\top)(\pm\mathbf{v}_j) = 1$ si $i = j$, o 0 si $i \neq j$. Lo mismo vale para $D^\top D$.

A partir las expresiones (5.8-5.9) obtuvimos (5.10-5.11) con una ambigüedad de signo, pero, independientemente del signo, a partir de (5.10-5.11) recuperamos (5.8-5.9). Esto indica que podemos elegir el signo de los autovectores $\{\mathbf{w}_i\}$ de $DD^\top$, y $\{\mathbf{v}_i\}$ de $D^\top D$ de modo que

$$D = \sum_{i=0}^{1} \lambda_i\ \mathbf{w}_i\mathbf{v}_i^\top \tag{5.13}$$

$$D^\top = \sum_{j=0}^{1} \lambda_j\ \mathbf{v}_j\mathbf{w}_j^\top \tag{5.14}$$

Como indicamos en el comienzo de esta sección, expresiones (5.13-5.14) era lo que buscábamos, son análogas a (4.15), pero con $\{\mathbf{p}_i\}$ y $\{\mathbf{f}_i\}$ ortogonales. (5.13) y (5.14) son la descomposición en valores singulares de las matrices D y $D^\top$ respectivamente.

A pesar de que fue derivado para matrices de 2×2, todas las operaciones realizadas valen para matrices reales de cualquier dimension finita $D \in \mathbb{R}^{m\times n}$. En ese caso $DD^\top \in \mathbb{R}^{m\times m}$ mientras que $D^\top D \in \mathbb{R}^{n\times n}$. Pero como vimos en la sección 3 en general, y en la sección 5 en términos de productos outer, $\dim W_1 = \dim V_1$, o el número de filas linealmente independiente es igual al número de columnas linealmente independientes, independientemente de cuánto sea m y n. Esto implica que si,por ejemplo, $n \leq m$, $DD^\top$ va a tener el menos $m - n$ autovalores nulos y las expresiones (5.13-5.14) siguen valiendo incluyendo en la suma solo los "valores singulares" λ_i no nulos.

En libros de texto, y en librerías de software como linalg de numpy, la descomposición en valores singulares de una matriz $D \in \mathbb{R}^{m\times n}$ suele presentarse indicando que la matriz D se puede escribir como

$$D = USV^\top \tag{5.15}$$

donde $V \in \mathbb{R}^{n\times n}$, con sus columnas dadas por los vectores ortonormales $\mathbf{v_i}$, de modo que las filas de $V^\top$ son $\mathbf{v_i}^\top$, $U \in \mathbb{R}^{m\times m}$ con sus columnas dadas por los vectores ortonormales $\mathbf{w_i}$, y $D \in \mathbb{R}^{m\times n}$ con los elementos de la diagonal principal iguales a los valores singulares positivos λ_i y todos los demás elementos iguales a cero.

Es fácil ver que la expresión (5.15) es equivalente a (5.13). Recordando la ecuación (4.18) para

productos de matrices, asumiendo que el rango de D es p, (5.15) implica:

$$D = USV^\top = \begin{bmatrix} | & & | \\ \mathbf{w_1} & \cdots & \mathbf{w_p} \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{bmatrix} \begin{bmatrix} -\mathbf{v_1}^\top - \\ \vdots \\ -\mathbf{v_p}^\top - \end{bmatrix} = \sum_{k=1}^{p} \lambda_k \mathbf{w_k}\, \mathbf{v_k}^\top \tag{5.16}$$

Finalizamos esta sección notando que si bien la descomposición en valores singulares vale para cualquier matriz[3] $V \in \mathbb{R}^{n\times n}$, en la práctica es especialmente útil cuando relativamente pocos valores singulares λ_i son mucho mayores que el resto. Supongamos por ejemplo que, como ocurre en (5.16), el rango de D es p pero los primeros $s \ll p$ valores principales son mucho mayores el resto, en ese caso, para muchas aplicaciones,

$$D = \sum_{k=1}^{p} \lambda_k \mathbf{w_k}\, \mathbf{v_k}^\top \sim \sum_{k=1}^{s} \lambda_k \mathbf{w_k}\, \mathbf{v_k}^\top, \quad s \ll p \tag{5.17}$$

suele ser una excelente aproximación de D y mucho mas fácil de analizar.

Como fue adelantado en la introducción, se puede probar que (5.16) es la mejor aproximación de la matriz $D \in \mathbb{R}^{m\times n}$ de rango p con matrices de rango s, en el sentido de que minimiza la "distancia Euclidiana" en el espacio de las matrices de $\mathbb{R}^{m\times n}$ (o norma ℓ_2 de Frobenius):

$$\|D\|_F = \sqrt{\sum_{i,j} d_{ij}^2} \tag{5.18}$$

Es decir, para toda otra matriz B de rango s,

$$\left\| D - \sum_{k=1}^{s} \lambda_k \mathbf{w_k}\, \mathbf{v_k}^\top \right\|_F \leq \|D - B\|_F \tag{5.19}$$

6 Relación entre SVD, análisis de componentes principales y análisis de factores

Antes de atacar de manera directa el tema de esta sección, es conveniente entender qué pasa con SVD cuando se aplica a matrices simétricas.

Recordemos que en la introducción vimos que para una matriz real cuadrada y simétrica A, los autovectores, y autovalores

$$A\mathbf{v_i} = \lambda_i \mathbf{v_i} \tag{6.1}$$

son reales, hay n de ellos, y se pueden elegir ortonormales:

$$\mathbf{v_i}^\top \mathbf{v_j} = 0 \quad \text{si} \quad i \neq j \tag{6.2}$$

$$\mathbf{v_i}^\top \mathbf{v_i} = 1 \quad \text{si} \quad i = 1, \cdots, n \tag{6.3}$$

[3] También vale para matrices complejas.

Comparando esto con la última igualdad en (5.16), notamos que para el caso de matrices reales y simétricas, los vectores singulares por derecha $\mathbf{v_i}$ y por izquierda $\mathbf{w_i}$ tienen que ser iguales entre sí e iguales a los autovectores $\mathbf{v_i}$, y los valores singulares λ_i tienen que ser los autovalores de A. Es decir, A se tiene que poder escribir como

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{v_i}\, \mathbf{v_i}^\top \tag{6.4}$$

de modo que cuando actúa sobre un autovector $\mathbf{v_j}$ tenemos

$$A\mathbf{v_j} = \sum_{i=1}^{n} \lambda_i \mathbf{v_i} \left(\mathbf{v_i}^\top \mathbf{v_j}\right) = \lambda_j \mathbf{v_j} \tag{6.5}$$

donde en la última igualdad usamos (6.2-6.3). Como el producto matriz por vector es lineal, la validez de (6.4) para los n autovectores linealmente independientes $\mathbf{v_i}$ asegura que dicha expresión vale en general.

La expresión (6.4) es muy parecida a las ecuaciones (5.8-5.9), pero en estas últimas expresiones λ_i^2 nunca es negativo, mientras que los autovalores de A en (6.4) pueden ser negativos: que A sea simétrica asegura que los λ_i sean reales, pero pueden ser negativos.

Pero parte de la definición de SVD en (5.16) es que los λ_i sean no negativos. Esto se arregla muy fácilmente: supongamos que en la expresion (6.4) el sumando $k-$ésimo tiene λ_k negativo, entonces, simplemente cambiándoles el signo a λ_k y a un $\mathbf{v_k}$ tenemos:

$$\lambda_k \mathbf{v_k}\, \mathbf{v_k}^\top = (-\lambda_k)\,(-\mathbf{v_k})\; \mathbf{v_k}^\top \tag{6.6}$$

donde $-\lambda_k$ es positivo. Si hacemos esto con todos los autovalores negativos recuperamos la expresión SVD en (5.16), que si bien demandaba que los λ_k sean positivos, no demandaba que los vectores singulares por derecha y por izquierda sean iguales.

En cualquier caso, si $\mathbf{v_i}$ es un autovector de A, $-\mathbf{v_i}$ también es un autovector de A con el mismo autovalor, linealmente dependiente de $\mathbf{v_i}$, es decir, esta en el mismo subespacio lineal que $\mathbf{v_i}$. Como se verá en unas líneas, esto habilita una interpretación geométrica muy útil de SVD para el caso en que la matriz A es cuadrada $n \times n$ y simétrica: la mejor aproximación de rango ℓ de la matriz A es la proyección ortogonal de los vectores columna (o fila, es indistinto) que viven en $\mathbb{R}^n$ sobre el "mejor" subespacio de dimensión ℓ.

Resumiendo, para matrices simétricas, en la expresión general (5.16) de SVD, los vectores singulares por derecha e izquierda correspondientes a autovalores no negativos son iguales entre sí e iguales a los autovectores de la matriz, y los vectores singulares por derecha e izquierda correspondientes a autovalores negativos también son autovectores de la matriz pero difieren entre sí en un signo menos.

Como vimos en la introducción, en análisis de factores en las ciencias humanas y sociales, típicamente tenemos una matriz D de datos empíricos donde, retomando el ejemplo usado en esa sección, el elemento d_{ij} es el resultado del test tipo j del $i-$ésimo individuo, hay n tipos de tests

que se le toman a m individuos elegidos al azar en una población mucho mayor que m. A partir de D obtenemos la matriz "centrada" $\tilde{D}$ (1.3) restándole a cada elemento d_{ij} la media μ_j de los resultados del tests j. Y, dependiendo de la aplicación, normalizábamos las columnas a varianza 1 obteniendo la matriz Z en (1.5), para finalmente realizar el análisis de factores de dicha matriz.

Este último paso no siempre se realiza, porque los diferentes valores de las varianzas de las variables aleatorias muchas veces es parte importante de la señal que queremos capturar, y en lo que sigue suponemos que *no* normalizamos las columnas a varianza 1. En cualquier caso, todo lo que sigue continúa valiendo si decidiéramos normalizarlas.

Recordemos (1.4) para la estimación empírica de la desviación estándar σ_j

$$\mathrm{Var}_j = \sigma_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(d_{ij}-\mu_j\right)^2 = \frac{1}{m-1}\left(\tilde{D}^\top\tilde{D}\right)_{jj} \tag{6.7}$$

y la estimación empírica de la covarianza es

$$\mathrm{CoVar}_{jk} = \frac{1}{m-1}\sum_{i=1}^{m}\left(d_{ij}-\mu_j\right)\left(d_{ik}-\mu_k\right) = \frac{1}{m-1}\left(\tilde{D}^\top\tilde{D}\right)_{jk} \tag{6.8}$$

Es decir, toda la información estadística de segundo orden está encapsulada en la matriz $\tilde{D}^\top\tilde{D}$. Pero vimos en la sección 5 que una matriz de la forma $\tilde{D}^\top\tilde{D}$: a) es cuadrada, b) es simétrica, y c) es semi-definida positiva. Es decir, sus autovalores nunca son negativos.

Más aún, la manera como la derivamos hizo explícito que si la SVD de la matriz D es

$$D = \sum_{i=0}^{n}\lambda_i\ \mathbf{w}_i\mathbf{v}_i^\top \tag{6.9}$$

la SVD de la matriz $\tilde{D}^\top\tilde{D}$ es

$$\tilde{D}^\top\tilde{D} = \sum_{i=0}^{n}\lambda_i^2\ \mathbf{v}_i\mathbf{v}_i^\top \tag{6.10}$$

Es decir, los vectores singulares por derecha de D son los autovectores de la matriz varianza-covarianza, y si λ_i es un valor singular de D, λ_i^2 es el correspondiente autovalor de la matriz varianza-covarianza. Dado que λ_i^2 nunca es negativo, para $\tilde{D}^\top\tilde{D}$ los vectores singulares por derecha y por izquierda coinciden incluso en los signos, y son ortonormales entre sí como en (6.2-6.3).

Expresada la SVD de D en la forma (5.15), recordando que la transpuesta de un producto de matrices es el producto en orden inverso de las transpuestas y que $(V^\top)^\top = V$, tenemos

$$\tilde{D}^\top D = VS\underbrace{U^\top U}_{=I}SV^\top = VS^2V^\top \tag{6.11}$$

donde $U^\top U$ es igual a la matriz identidad porque las columnas de U (filas de $U^\top$) son vectores ortonormales. Dado que la matriz S es diagonal con elementos λ_i, ver (5.16), la matriz S^2

también es diagonal con elementos λ_i^2. Es decir, obtenemos el mismo resultado que en (6.10) utilizando la forma (5.15) de SVD.

La expresión (6.11) de la matriz varianza-covarianza (proporcional a $\tilde{D}^\top D$) de la matriz de datos D se conoce como "descomposición en valores principales" (PCA) de D. Como vemos, no es otra cosa que SVD aplicado a $\tilde{D}^\top D$.

Pero dado que los vectores singulares por derecha y por izquierda de $\tilde{D}^\top D$ son idénticos, además de la interpretación estadística que acabamos de ver, PCA tiene una interesante interpretación geométrica: recordando que por ser V ortogonal, $V^\top V = V^\top V = I$, multiplicando $\tilde{D}^\top D$ en (6.11) por izquierda por $V^\top$ y por derecha por V obtenemos:

$$S^2 = V^\top \left(\tilde{D}^\top D\right) V \tag{6.12}$$

Siendo V ortogonal, se puede mostrar que V implementa en $\mathbb{R}^n$ una suerte de "rotación" (pueden también ser reflexiones, en general se llaman "transformaciones ortogonales"), y que el producto del lado derecho de la igualdad en (6.12) es exactamente cómo se "ve" la matriz $\tilde{D}^\top D$ en el sistema de coordenadas rotado por V, que transforma la base canónica en la base de autovectores de $\tilde{D}^\top D$. En ese sistema de coordenadas, dicha matriz es la matriz diagonal S^2 en el lado izquierdo de la igualdad en (6.12).

Estadísticamente significa que la matriz V transforma las variables aleatorias muestreada en las columnas de D, que en general tienen covarianzas (6.8) (o correlaciones) no nulas entre sí, en variables aleatorias independientes (por lo menos hasta segundo orden).

Por supuesto podemos hacer el mismo tipo de aproximación de bajo rango de la matriz varianza-covarianza (6.10), o (6.11), simplemente reemplazando los $n - \ell$ valores singulares λ_i^2 más pequeños por cero, obteniendo de esa manera la mejor aproximación de rango ℓ de la matriz varianza-covarianza, ver (5.18-5.19). Vemos entonces que en el hecho de que esta aproximación tiene sentido, se basa la idea de análisis de factores, tan influyente en las ciencias humans y sociales.

La "mejor aproximación de rango ℓ" de la matriz varianza-covarianza, que ya fue explicada en general en la sección 5, para esta matriz cuadrada, simétrica y semidefinida positiva, en la que, por lo tanto, el espacio fila y columna (ver sección 3) son en realidad el mismo espacio $\mathbb{R}^n$ donde viven los datos, implica dos nuevas interpretations complementarias y consistentes entre sí (recordemos que las filas de D son m puntos-dato que viven en $\mathbb{R}^n$.)

Mirada desde $\mathbb{R}^n$ que, de nuevo, es a la vez el espacio fila y columna de D, la mejor aproximación de rango ℓ es la mejor proyección de los m puntos-dato dados por las filas de D en el "mejor" subespacio de ℓ dimensiones. Mejor en el doble sentido de ser el subespacio de ℓ dimensiones que captura la mayor varianza de los datos, explicando la mayor variabilidad de los mismos, y mejor en el sentido de ser el subespacio de ℓ dimensiones que minimiza la distancia al cuadrado de los puntos a dicho subespacio.

7 Conclusiones

En este trabajo nos planteamos dos objetivos. Ante el hecho de que en las ciencias humanas y sociales el análisis de factores es de gran importancia, el primer objetivo es presentar de manera accesible para cualquier persona en estas disciplinas con una base elemental de álgebra lineal, la técnica conocida como análisis de valores singulares.

Dicha técnica sistematiza y generaliza el análisis de factores. Además tiene implementaciones algorítmicas sumamente optimizadas, lo que la hace apta para el análisis de factores incluso en la era de "Big Data", donde las bases de datos son mucho más voluminosas de lo que solían ser, tanto en número de puntos-dato como en la dimensión de los mismos. Presentamos varios ejemplos de dichas aplicaciones.

Enfatizo lo de "generaliza" el análisis de factores porque, como vimos, SVD hace posible dicho análisis incluso para matrices de datos mucho más generales, que no tienen la estructura típica estadística de m datos de n dimensiones, donde el interés pasa por el análisis de factores de la matriz varianza-covarianza (o, alternativamente, de la matriz de correlaciones).

Para presentar SVD se eligió un camino que conduzca a una comprensión intuitiva del método. Primero fijamos la notación y recordamos ciertos aspectos básicos de álgebra lineal. Luego enfatizamos la relación biyectiva entre matrices y funciones lineales entre espacios vectoriales, lo que condujo de manera natural a entender la igualdad entre el rango fila y el rango columna de cualquier matriz. Luego presentamos el producto "outer", que por un lado no es usualmente aprendido en cursos estándar de álgebra lineal, y por el otro es fundamental para conectar la igualdad entre el rango fila y el rango columna con SVD. Finalmente, con todas esas herramientas, presentamos SVD de manera tal que el lector casi que adivina su estructura. También vimos cómo se implementa en Matlab y Python.

Por último, volviendo al análisis de factores tradicional con la estructura estadística de m datos de n dimensiones, vimos que SVD es la base de dicho análisis, y que para el caso particular de su aplicación a la matriz varianza-covarianza, que es cuadrada, simétrica, y semidefinida positiva, los vectores singulares por derecha y por izquierda son idénticos, son los autovectores de dicha matriz, y los valores singulares son los correspondientes autovalores.

Mirado desde la perspectiva que brinda SVD, si en el espacio $\mathbb{R}^n$ donde viven los "m" datos nos centramos en la media de los mismos (en el centro de la "nube" de datos), el general, hay tantos "factores", en el sentido del análisis de factores, como rango $k \leq n$ tenga la matriz varianza-covarianza, y corresponden a las direcciones de los autovectores de dicha matriz.

Si uno mira la matriz varianza covarianza desde tales direcciones (desde la base de autovectores), la misma es diagonal. Es decir, los "factores" corresponden a las combinaciones lineales de variables que hacen las covarianzas de las nuevas variables sean iguales a cero. El hecho de que la matriz varianza-covarianza es simétrica asegura que siempre existan tales combinaciones y que forman una base ortogonal. (no necesariamente asegura que los momentos de orden 2 son suficientes para caracterizar las distribuciones)

Una teoría de ℓ factores simplemente corresponde a la decision del investigador de elegir los ℓ

factores (entre los n) con mayores valores singulares. Esas son las ℓ dimensiones que capturan la mayor varianza de los datos, explicando la mayor variabilidad posible de los mismos condicional a que uno eligió trabajar con ℓ factores.

El segundo objetivo de este trabajo era hacer ciertas observaciones sobre el análisis de factores como idea fuerza de investigación en ciencias humans y sociales en general. Las mismas requieren sólo unas pocas líneas.

Si bien la idea de que las diversas manifestaciones de campos de estudio complejos como el de tales disciplinas son en realidad manifestaciones de unos pocos factores que influyen sobre todas las demás variables es filosóficamente atractiva y ayuda en principio a ordenar semejante complejidad, el análisis matemático general de dichas ideas indica que hay que ser cuidadoso en su aplicación y en la derivación de las consecuencias a las aparentemente conduce.

Vimos que, si bien SVD sistematiza y generaliza el análisis de factores, también le quita cierta mística. No hay una gran revelación al plantear una teoría de factores porque los factores siempre existen. No importa el objeto de estudio, si los datos se pueden ordenar en una matriz, la descomposición en factores está asegurada.

En todo caso, recordando que los valores singulares son siempre positivos o cero, y que su orden natural es de mayor a menor, donde sí puede haber contenido genuino, es en rapidez con la que los valores singulares decaen. Si se da el caso que unos pocos valores singulares son mucho mayores que el resto, entonces sí hay contenido genuino en el análisis de factores.

Pero los desarrollos y numerosas aplicaciones de SVD con bases de datos enormes de los últimos 15 o 20 años, algunos de los cuales mencionamos en la introducción, parecen conducir a un patrón que se repite independientemente de la aplicación. Cuando una disciplina logra compilar grandes bases de datos, el número de factores relevantes tiende a ser muy grande como para identificar en ellos modelos efectivos basados en unas pocas variables, en unas pocas causas que uno puede adivinar a priori.

Volviendo a los ejemplos de la introducción, que haya unas pocos factores en la inteligencia, o que haya modelos macroeconómicos basados en unas pocas variables con real poder de predicción y validez duradera en el tiempo, son ideas que, si bien nos pueden sorprender para bien, no se ajustan a lo que se va observando en otras áreas en ciencia de datos en la era del Big Data. Quizás el gran esfuerzo haya que ponerlo en generar en estas disciplinas esas masivas bases de datos entonces.

Por el otro lado, también se da el caso que cuando una disciplina logra compilar grandes bases de datos, el número de factores relevantes tienden a ser mucho *menor* que el rango original de la matriz de datos.

Es decir, en número de factores relevantes tiende a ser demasiado grande como para habilitar modelos teóricos basados en unas pocas variables, pero mucho más pequeño de que lo que a priori hubiera podido ser. Quizás el misterio genuinamente interesante pase por entender el por qué de esa complejidad intermedia.

References

[Alter-Brown-Botstein (2000)] O. Alter, P. O. Brown, D. Botstein. "Singular value decomposition for genome-wide expression data processing and modeling." Proceedings of the National Academy of Sciences 97: 10101-10106 (2000).

[Athey y otros (2017)] Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. "Matrix Completion Methods for Causal Panel Data Models." Cornell University Library, 2017.

[Athey (2017)] Athey, S. "The Impact of Machine Learning on Economics," en The Economics of Artificial Intelligence: An Agenda, University of Chicago Press, 507-547 (2017).

[Athey-Imbens (2019)] "Machine learning methods that economists should know about." Annual Review of Economics 11: 685-725 (2019).

[Bai-Ng (2002)] Bai, J. and Ng, S. "Determining the Number of Factors in Approximate Factor Models". Econometrica, 70: 191-221 (2002).

[Bai-Ng (2008)] Bai, J, and Ng, S. "Large dimensional factor analysis." Foundations and Trends in Econometrics 3, no. 2: 89-163 (2008).

[Bolukbasi y otros (2016)] Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29 (2016).

[Burt (1909)] Burt, C., "Experimental Tests of General Intelligence." British Journal of Psychology, 1904-1920, 3: 94-177, (1909).

[Deerwester y otros (1990)] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. "Indexing by latent semantic analysis." J. Am. Soc. Inf. Sci., 41: 391-407 (1990).

[Garnett-Whitehead (1919)] Maxwell Garnett, J. C., Whitehead, A. N. "On certain independent factors in mental measurements." Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 96 (675): 91-111 (1919).

[Harman (1976)] Harman, H.H., "Modern factor analysis", University of Chicago press (1976).

[Holter y otros (2000)] N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, N. V. Fedoroff. "Fundamental patterns underlying gene expression profiles: Simplicity from complexity." Proceedings of the National Academy of Sciences 97: 8409-8414 (2000).

[Holzinger (1930)] Holzinger, K. J. "Statistical résumé of the Spearman two-factor theory". 75. University of Chicago Press (1930).

[Liberty y otros (2007)] E. Liberty, F. Woolfe, P. G. Martinsson, V. Rokhlin, M. Tygert, "Randomized algorithms for the low-rank approximation of matrices." Proceedings of the National Academy of Sciences 104: 20167-20172 (2007).

[Lintner (1965)] Lintner J., "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." The Review of Economics and Statistics, Vol. 47, No. 1: 13-37 (1965).

[Mossin (1966)] Mossin, J. “Equilibrium in a Capital Asset Market.” Econometrica 34, no. 4: 768–83 (1966).

[Muller-Magaia-Herbst (2004)] N. Muller, L. Magaia, B. M. Herbst. “Singular value decomposition, eigenfaces, and 3D reconstructions.” SIAM Review 46: 518-545 (2004).

[Novembre y otros (2008)] Novembre, J., Johnson, T., Bryc, K. et al. “Genes mirror geography within Europe.” Nature 456, 98–101 (2008).

[Pearson (1901)] Pearson, K. “On lines and planes of closest fit to systems of points in space.” The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11):559-572 (1901).

[Ross (1976)] Ross, S.A. “The arbitrage theory of capital asset pricing.” Journal of Economic Theory, v.13, (3), 341-360 (1976)

[Sharpe (1964)] Sharpe, W.F. (1964), “Capital Asset Prices: A Theory of Market Equilibrium under conditions of risk.” The Journal of Finance, 19: 425-442 (1964).

[Spearman (1904)] Spearman, C., ” ‘General Intelligence,’ Objectively Determined and Measured.” The American Journal of Psychology 15, no. 2 (1904): 201–92. https://doi.org/10.2307/1412107.

[Spearman (1927)] Spearman, C. “The Abilities of Man: Their Nature and Measurement.” Journal of Philosophical Studies 2 (8):557-560 (1927).

[Stock and Watson (2011)] “Dynamic factor models.” Oxford Handbooks Online (2011).

[Stock and Watson (2015)] Stock, J. and Watson, M. W. “Factor Models for Macroeconomics.” J. B. Taylor and H. Uhlig (eds), Handbook of Macroeconomics, Vol. 2, North Holland (2015).

[Thomson (1938)] Thomson, G. “Methods of Estimating Mental Factors.” Nature 141, 246 (1938).

[Thurstone (1931)] Thurstone, L. L. “Multiple factor analysis.” Psychological review 38, no. 5: 406 (1931).

[Thurstone (1947)] Thurstone, L. L. “Multiple-factor analysis; a development and expansion of The Vectors of Mind.” (1947).

[Turk-Pentland (1991a)] M. Turk, A. Pentland. “Eigenfaces for recognition.” Journal of Cognitive Neuroscience **3** (1991a).

[Turk-Pentland (1991b)] M. Turk, A. Pentland. “Face recognition using Eigenfaces.” Proc. of Computer Vision and Pattern Recognition 3: 586-591 (1991b).