UNIVERSIDAD AUTONOMA DE MADRID

Facultad de Ciencias

19|20

Trabajo fin de grado

Aplicación de la descomposición en valores singulares al análisis de datos

David Moreno Maldonado

Campus Internacional excelencia UAM CSIC

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
FACULTAD DE CIENCIAS

Grado en Matemáticas

TRABAJO FIN DE GRADO

Aplicación de la descomposición en valores singulares al análisis de datos

Autor: David Moreno Maldonado
Tutor: Amparo Baíllo Moreno

junio 2020

David Moreno Maldonado
Aplicación de la descomposición en valores singulares al análisis de datos

David Moreno Maldonado

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

Agradecimientos

Me gustaría agradecer a Amparo Baíllo todo el esfuerzo y empeño que ha puesto en ayudarme durante la realización de este trabajo. Junto a tí he aprendido mucho y tus consejos han sido de gran ayuda para desarrollar y pulir esta memoria, pese a las dificultades que hemos tenido a lo largo de este año.

Gracias a todos los profesores del Departamento de Matemáticas que han aportado en mi desarrollo como estudiante durante el Grado. Así como a todo el personal de la Escuela sin el que esta no podría funcionar.

A mis compañeros de clase, por todo el apoyo mutuo que nos hemos dado para superar las diferentes asignaturas. Ha sido un placer trabajar junto a vosotros y os deseo mucha suerte de aquí en adelante.

Por último, agradecer a mi familia y amigos que siempre han estado apoyándome desde antes de esta etapa universitaria. En especial, quería agradecer a mis padres por sus sabios consejos a lo largo de mi vida y su comprensión.

Muchas gracias a todos.

Resumen

En las últimas décadas, nuestra capacidad de recolectar y almacenar grandes cantidades de datos observados ha aumentado enormemente. Esto hace que, una vez que se va a proceder a analizar y comprender estos datos, la cantidad de información de la que se dispone pueda ser abrumadora. En algunos casos, puede ser interesante realizar una reducción de la dimensión, esto es, descartar aquella información que es menos relevante y quedarnos con la verdaderamente significativa. La manera en la que realizamos esta selección es uno de los objetivos de la Estadística Multivariante.

La Estadística Multivariante es la rama de la Estadística que se ocupa de los datos multivariados, aquellos con dimensión mayor que 1. Una muestra de observaciones multivariadas se recoge siempre en una matriz. La descomposición en valores singulares (SVD por sus siglas en inglés) es una técnica del Álgebra Matricial que, en este contexto, permite resumir y comprimir la información contenida en la matriz de datos.

Una ventaja inmediata de aplicar la descomposición en valores singulares es facilitar el análisis de la información muestral, haciendo más sencilla la comprensión de los datos que se están manejando. El objetivo de este Trabajo Fin de Grado es estudiar cómo se utiliza concretamente la descomposición en valores singulares en diferentes técnicas multivariadas: componentes principales, correlaciones canónicas y aproximación y compleción de matrices.

La base teórica y los procedimientos estudiados y demostrados se prueban después sobre un conjuntos de datos reales con librerías del programa R. La aplicación satisfactoria de las herramientas estudiadas pretende demostrar la potencia y utilidad de la descomposición en valores singulares en datos actuales de diversas fuentes y temáticas.

Palabras clave

Descomposición en valores singulares, Componentes principales, Correlaciones canónicas, Aproximación y compleción de matrices

Abstract

In the last decades, our capability for gathering and storing large amounts of data has increased a lot. Once we want to analyze and understand this information, it might be difficult due to its size. In some cases, it could be interesting to reduce the dimension, that is, discard the information that is not highly relevant to keep only the pieces of information which are meaningful. The way we proceed with this selection is one of objectives of Multivariant Statistics.

Multivariate Statistics is the branch of Statistics which handles multivariate data, that with dimension greater than one. We store multivariate observations in the data matrix. Singular Value Decomposition (SVD) is a technique from Matrix Algebra which can help to summarize end compress the information in the data matrix.

One straightforward advantage of SVD application is to facilitate the analysis of sample information. The objective of this project is to study how SVD is used in several multivariate techniques: principal components analysis, canonical correlation analysis and aproximation and completrion of matrices.

The theory and the procedures studied and proved are later tested with real data using R. The adequate application of these tools shows the power and utility of SVD in current data of a great variety of topics.

Keywords

Singular value decomposition; Principal components; Canonical correlation; Aproximation and completion of matrices

ÍNDICE

Listas

Lista de figuras

Lista de tablas

1

Introducción

En este capítulo se describirán la motivación y objetivos que se persiguen con la realización de este Trabajo de Fin de Grado. Además, se expondrá la estructura que sigue este texto especificando el contenido de cada capítulo. Por último, se incluye una breve explicación relativa a la notación que se seguirá a lo largo de todo el trabajo.

1.1. Motivación y objetivos

La Estadística Multivariante es una de las ramas de las Matemáticas que más aplicación tiene en el mundo actual. Debido a la cantidad de datos que se recogen en multitud de ámbitos, el estudio de estos puede ayudar a una mejor comprensión de nuestro entorno y a una mejora de la calidad de la sociedad, así como a su desarrollo. Dado que la Estadística Multivariante ofrece un campo muy amplio de herramientas y procedimientos para el análisis de datos, es necesario centrarse en alguna de estas técnicas en particular para poder profundizar lo suficiente y poder valorar su potencial.

La principal motivación del trabajo es estudiar y comprender las utilidades que tiene la descomposición en valores singulares en el contexto de la Estadística Multivariante. Para ello, se estudiarán diferentes técnicas estadísticas que hacen uso de dicha descomposición y se aplicarán a casos reales con datos de diferentes fuentes y temáticas. Las técnicas seleccionadas para estudiar son: componentes principales, correlaciones canónicas y la aproximación y compleción de matrices.

El objetivo más importante de este texto es, tras analizar y demostrar la base teórica detrás de cada una de las técnicas mencionadas, poder aplicar satisfactoriamente estas herramientas en datos reales y actuales. Para ello, se irá haciendo una exposición de los conceptos y teoremas claves en los que nos apoyaremos para sacar conclusiones fundamentadas sobre datos de diferentes temáticas y ámbitos.

Con la aplicación a estos conjuntos de datos concretos, lo que se pretende es demostrar la potencia y utilidad de la descomposición en valores singulares en datos actuales. Las aplicaciones que puede tener son enormes y no se reducen solo a los ámbitos estudiados sino que pueden ser muy diversos.

1.2. Estructura de la memoria

El texto esta dividido en cinco capítulos, donde los principales y en aquellos que se desarrolla toda la teoría y aplicaciones de técnicas son el segundo, tercero y cuarto. En este primer capítulo de introducción hemos establecido las motivaciones que propiciaron la escritura de este trabajo y los objetivos que pretendemos alcanzar. Además, se incluye un pequeño apartado notacional para establecer los convenios que se seguirán en todo el texto con respecto a la descomposición en valores singulares.

Los siguientes tres capítulos tratan acerca de componentes principales, correlaciones canónicas y aproximación y compleción de matrices, respectivamente. Tienen una estructura análoga: primero se expone toda la parte teórica que demuestra como se aplican la descomposición en valores singulares en cada caso, y después se aplican estos conceptos a datos reales. En cada capítulo se ofrece una descripción detallada de las demostraciones matemáticas utilizadas y las conclusiones extraídas de los datos.

Por último, tenemos un capítulo donde se recogen las principales conclusiones a las que se ha llegado tras la realización del trabajo. Se resumen los principales resultados obtenidos al investigar y aplicar cada una de las técnicas estadísticas propuestas.

1.3. Descomposición en valores singulares

A lo largo de todo el trabajo, se utilizará recurrentemente la descomposición en valores singulares de una matriz. Dado que existen diferentes definiciones de la misma con pequeñas variaciones se especifican en esta sección las convenciones y notación que se utilizarán a lo largo de todo el texto.

La descomposición en valores singulares (SVD) de una matriz $\mathbf{A}$ de dimensiones $m \times n$ es una factorización que definimos de la siguiente manera

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

donde $\mathbf{U}$ es una matriz $m \times n$ ortogonal ($\mathbf{U}'\mathbf{U} = \mathbf{I}_n$) cuyas columnas $\mathbf{u}_j \in \mathbb{R}^m$ son los *vectores singulares izquierdos*. De manera similar, $\mathbf{V}$ es una matriz $n \times n$ ortogonal ($\mathbf{V}'\mathbf{V} = \mathbf{I}_n$) cuyas columnas $\mathbf{v}_j \in \mathbb{R}^n$ son los *vectores singulares derechos*. La matriz $\mathbf{D}$ es $n \times n$ y diagonal, cuyos elementos $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ son conocidos como los valores singulares o propios de la matriz.

Los valores singulares de $\mathbf{A}$ que son no nulos coinciden con la raíz cuadrada de los autovalores no nulos de $\mathbf{A}'\mathbf{A}$ o $\mathbf{A}\mathbf{A}'$.

2

COMPONENTES PRINCIPALES

En este capítulo se describirá la teoría investigada acerca de las componentes principales. Es una técnica de reducción de la dimensión que sustituye las variables originales de una muestra por combinaciones lineales de las mismas, es decir, proyecciones de los datos sobre rectas. De esta manera, podemos obtener la información más relevante que nos ofrecían las variables originales, pero con menor dimensión.

2.1. Valores y vectores propios de una matriz. Componentes principales

Las componentes principales de una matriz de datos multivariantes con n observaciones y p variables $\mathbf{X}$ son unas variables incorreladas tales que unas pocas explican la mayor parte de la variabilidad de $\mathbf{X}$.

Definición 2.1 *Las componentes principales son las combinaciones lineales*

$$Y_1 = \mathbf{X}'\mathbf{t}_1, Y_2 = \mathbf{X}'\mathbf{t}_2, ..., Y_p = \mathbf{X}'\mathbf{t}_p$$

tales que:

1.– La varianza muestral de Y_1 (var(Y_1)) es máxima condicionado a $\mathbf{t}_1'\mathbf{t}_1 = 1$.

2.– Entre todas las combinaciones lineales de $\mathbf{X}$ incorreladas con Y_1, la variable Y_2 es tal que la varianza muestral de Y_2 (var(Y_2)) es máxima condicionado a $\mathbf{t}_2'\mathbf{t}_2 = 1$.

3.– Y_3 es una variable incorrelada con Y_1 e Y_2 y con varianza máxima. Análogamente, definimos las demás componentes principales.

El siguiente teorema nos permite relacionar las componentes principales definidas con la descomposición en valores singulares de una matriz de covarianzas.

Teorema 2.1 *Sean $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_p$ los p vectores propios ortonormalizados de la matriz de covarianzas*

muestral $\mathbf{S}$ *y* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ *los correspondientes valores propios, es decir,*

$$\mathbf{S}\mathbf{t}_i = \lambda_i \mathbf{t}_i, \;\; \mathbf{t}_i'\mathbf{t}_i = 1, \;\; i = 1, ..., p.$$

Entonces, para las combinaciones lineales $Y_i = \mathbf{X}'\mathbf{t}_i$, $i = 1, ..., p$ *se cumple que:*

1.– Las varianzas son los valores propios de $\mathbf{S}$

$$\text{var}(Y_i) = \lambda_i, \;\; i = 1, ..., p.$$

2.– Son variables incorreladas:

$$\text{cov}(Y_i, Y_j) = 0, \; i \neq j = 1, ..., p.$$

3.– Las variables $Y_i = \mathbf{X}'\mathbf{t}_i, \; i = 1, ..., p$ son las componentes principales.

Demostración:

Tenemos que $\text{cov}(Y_i, Y_j) = \mathbf{t}_i'\mathbf{S}\mathbf{t}_j = \mathbf{t}_i'\lambda_j\mathbf{t}_j = \lambda_j\mathbf{t}_i'\mathbf{t}_j = 0$ si $i \neq j$. Además, $\text{cov}(Y_i, Y_i) = \text{var}(Y_i) = \lambda_i\mathbf{t}_i'\mathbf{t}_i = \lambda_i$ y, como $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$, entonces $\text{var}(Y_1)$ es máxima entre todas las Y_i.

Si suponemos ahora $Y = \sum_{i=1}^p a_i X_i = \sum_{i=1}^p \alpha_i Y_i$, variable compuesta tal que $\sum_{i=1}^p \alpha_i^2 = 1$ tendremos que $\text{var}(Y) = \text{var}(\sum_{i=1}^p \alpha_i Y_i) = \sum_{i=1}^p \text{var}(\alpha_i Y_i)$ ya que $\text{var}(Z_1 + Z_2) = \text{var}(Z_1) + \text{var}(Z_2) + 2\text{cov}(Z_1, Z_2)$ y hemos visto que $\text{cov}(Y_i, Y_j) = 0$ si $i \neq j$.

Tenemos entonces,

$$\text{var}(\mathbf{Y}) = \sum_{i=1}^p \text{var}(\alpha_i Y_i) = \sum_{i=1}^p \alpha_i^2 \text{var}(Y_i) = \sum_{i=1}^p \alpha_i^2 \lambda_i \leq \sum_{i=1}^p \alpha_i^2 \lambda_1 = \lambda_1 = \text{var}(Y_1).$$

Es decir, Y_1 es la primera componente principal. Comprobemos ahora que Y_2 es la segunda componente.

Todas las variables Y_i son incorrelacionadas con Y_1. De todas ellas miramos las $Y = \sum_{i=1}^p b_i X_i = \sum_{i=1}^p \beta_i Y_i$ tal que $\sum_{i=1}^p \beta_i^2 = 1$. Tenemos que,

$$\sum_{i=1}^p \text{var}(\beta_i Y_i) = \sum_{i=1}^p \beta_i^2 \text{var}(Y_i) = \sum_{i=1}^p \beta_i^2 \lambda_i \leq \sum_{i=1}^p \beta_i^2 \lambda_2 = \lambda_2 = \text{var}(Y_2).$$

Por lo tanto, la varianza de $\mathbf{Y}$ es máxima cuando $\beta_2 = 1$ y $\beta_i = 0 \; \forall i \neq 2$. Tenemos que Y_2 está incorrelacionada con Y_1 y tiene varianza máxima de entre todas las posibles. El procedimiento para demostrar que $Y_3, Y_4, ..., Y_p$ $p \geq 3$ son las restantes componentes principales es análogo.

□

2.2. Variabilidad explicada por las componentes principales.

Una vez que hemos calculado las componentes principales de una matriz de covarianzas muestrales $\mathbf{S}$ queremos saber qué *cantidad* de información de la matriz explica cada componente. Esto permitirá más adelante tomar decisiones en cuanto a que componentes principales son las más importantes y crear criterios para seleccionar las mismas.

Por el momento necesitamos tener alguna medida que nos permita entender mejor la variabilidad de unos datos multivariantes a partir de sus componentes principales. Para ello vamos a definir una serie de conceptos.

Definición 2.2 *Sea $\mathbf{X}$ una matriz $n \times p$ de datos multivariantes con sus filas $\mathbf{x}_1', \mathbf{x}_2', ..., \mathbf{x}_n'$, la distancia euclídea (al cuadrado) entre dos filas de $\mathbf{X}$, $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ y $\mathbf{x}_j = (x_{j1}, ..., x_{jp})'$ es*

$$\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = \sum_{h=1}^{p}(x_{ih} - x_{jh})^2.$$

La matriz $\mathbf{\Delta} = (\delta_{ij})$ es la matriz $n \times n$ de distancias entre las filas.

Definición 2.3 *La variabilidad geométrica de la matriz de distancias Δ es la media de sus elementos al cuadrado*

$$V_\delta(\mathbf{X}) = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2.$$

Entonces $\mathbf{Y} = \mathbf{XT}$ es una transformación lineal de $\mathbf{X}$, donde $\mathbf{T}$ es una matriz $p \times q$ de constantes,

$$\delta_{ij}^2(q) = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = \sum_{h=1}^{q}(y_{ih} - y_{jh})^2$$

es la distancia euclídea entre las filas i y j de $\mathbf{Y}$. La variabilidad geométrica en dimensión $q \leq p$ es

$$V_\delta(\mathbf{Y})_q = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2(q).$$

Observemos que $V_\delta(\mathbf{X})_q$ es una medida de la dispersión de la muestra contenida en la matriz $\mathbf{X}$, como puntos del espacio euclídeo $\mathbb{R}^p$

Ahora, lo que hacemos es relacionar esta medida de variabilidad que hemos definido con la matriz de covarianzas muestrales $\mathbf{S}$ de la matriz de datos multivariantes $\mathbf{X}$ y, posteriormente, concluir que la transformación que maximiza esta variabilidad es la transformación por componentes principales.

Teorema 2.2 *La variabilidad geométrica de la distancia euclídea es la traza de la matriz de covarianzas*

$$V_\delta(\mathbf{X}) = tr(\mathbf{S}) = \sum_{h=1}^{p} \lambda_h.$$

Demostración: Dada una muestra univariante $x_1, \ldots, x_n$ con varianza s^2 vamos a ver que:

$$\frac{1}{2n^2} \sum_{i,j=1}^{n} (x_i - x_j)^2 = s^2.$$

Desarrollando, obtenemos

$$\begin{aligned}
\frac{1}{2n^2} \sum_{i,j=1}^{n} (x_i - x_j)^2 &= \frac{1}{2n^2} \sum_{i,j=1}^{n} (x_i - \bar{x} - (x_j - \bar{x}))^2 \\
&= \frac{1}{2n^2} (\sum_{i,j=1}^{n} (x_i - \bar{x})^2 + \sum_{i,j=1}^{n} (x_j - \bar{x})^2 - 2 \sum_{i,j=1}^{n} (x_i - \bar{x})(x_j - \bar{x})) \\
&= \frac{1}{2n^2} (n^2 s^2 + n^2 s^2 - 0) \\
&= s^2
\end{aligned}$$

ya que

$$\begin{aligned}
\sum_{i,j=1}^{n} (x_i - \bar{x})(x_j - \bar{x})) &= \sum_{i=1}^{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_j - \bar{x}) \\
&= n \sum_{i=1}^{n} (x_i - \bar{x}) \sum_{i=1}^{n} (x_j - \bar{x}) \\
&= n \sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - n\bar{x}) \\
&= n \sum_{i=1}^{n} (x_i \bar{x} - \bar{x}^2 - n x_i \bar{x} + n\bar{x}^2) \\
&= n\bar{x}^2 - n\bar{x}^2 + n^2 \bar{x}^2 - n^2 \bar{x}^2 \\
&= 0.
\end{aligned}$$

Por lo tanto, si aplicamos esto a cada columna de $\mathbf{X}$ obtenemos que

$$V_\delta(\mathbf{X}) = \sum_{j=1}^{p} s_{jj} = \text{tr}(\mathbf{S})$$

□

Si ahora quisiéramos una buena representación, pero reduciendo la dimensión de la matriz de datos de p a q, tendríamos que encontrar la que tenga una variabilidad geométrica máxima. De esta manera los puntos estarán lo más separados posible.

Teorema 2.3 *La transformación lineal $\mathbf{T}$ que maximiza la variabilidad geométrica en dimensión q es la transformación por componentes principales, es decir, $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_q]$ que contiene los q primeros*

vectores propios ortonormalizados de $\mathbf{S}$.

Demostración: Sabemos que la variabilidad geométrica de $\mathbf{Y} = \mathbf{XT}$ (siendo $\mathbf{T}$ cualquier transformación) es

$$V_\delta(\mathbf{Y})_q = \sum_{j=1}^{p} \text{var}^2(Y_j).$$

Dado que la transformación por componentes principales es aquella que maximiza var(Y_i) en todo caso, resulta que

$$\text{máx}\, V_\delta(\mathbf{Y})_q = \sum_{j=1}^{p} \lambda_j.$$

□

De esta manera, lo que estamos expresando es que, dada una matriz de datos $\mathbf{X}$ y su matriz de covarianzas $\mathbf{S}$, la primera componente principal de $\mathbf{S}$ nos da los coeficientes de la transformación lineal de $\mathbf{X}$ que más explica la variabilidad de los datos. La segunda componente principal los coeficientes de la transformación lineal que más explica, teniendo en cuenta que esta segunda componente principal esta incorrelada con la primera. Sucesivamente, las componentes principales nos quedan ordenadas de la que más explica a la que menos.

La varianza de la componente principal Y_i, como ya hemos visto, es var$(Y_i) = \lambda_i$ y la variabilidad geométrica total es tr$(\mathbf{S}) = \sum_{i=1}^{p} \lambda_i = V_\delta(\mathbf{X})$, como acabamos de ver. Por lo tanto:

1.– Y_i contribuye con la cantidad λ_i a la variabilidad total tr$(\mathbf{S})$.

2.– Si $q \leq p$, $Y_1, \ldots, Y_q$ explican la cantidad $\sum_{i=1}^{q} \lambda_i$ de la variabilidad total.

3.– El porcentaje de variabilidad explicada por las m primeras componentes principales es

$$P_m = 100\frac{V_\delta(\mathbf{Y})_m}{V_\delta(\mathbf{X})_p} = 100\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}.$$

2.3. Lema de Perron

Para ciertos tipos de datos, en los que todos los valores de la matriz de covarianzas sean positivos, nos puede ser de especial utilidad el teorema de Perron, enunciado en [3]. El teorema dice lo siguiente:

Teorema 2.4 *Sea* $\mathbf{X}$ *una matriz positiva* $n \times n$ *entonces:*

1.– $\mathbf{X}\mathbf{t}_1 = \lambda_1 \mathbf{t}_1$ para algún $\lambda_1 > 0,\ \mathbf{t}_1 > 0$.

2.– Para todo λ autovalor de $\mathbf{X}$, tenemos que $\lambda \leq \lambda_1$.

3.– El autovalor λ_1 tiene multiplicidad 1.

4.– Para cualquier otro autovalor de $\mathbf{X}$ diferente de λ_1, se tiene que no todas las componentes del correspondiente autovector son van a ser positivas simultáneamente.

Este teorema tiene como consecuencia que, si tenemos datos cuyas variables tengan todas las correlaciones positivas entre sí, la primera de las componentes principales tendrá todas sus componentes positivas. Esta situación se da frecuentemente en alometría. En este contexto la primera componente principal se puede interpretar como la componente de *tamaño*. Además, ninguna de las componentes principales restantes tendrá todas sus componentes positivas. En particular, la segunda componente principal se considera la componente de forma.

2.4. Criterios de selección de componentes principales

Ahora ya sabemos que las componentes principales nos dan información acerca de una matriz de datos multivariantes y nos permiten reducir la dimensión de los mismos a través de combinaciones lineales. Además, tenemos un mecanismo, el porcentaje de variabilidad, que nos indica en cada caso lo que aporta cada componente principal a la variabilidad total de los datos.

Se plantea ahora el problema de que número de componentes principales tomar en cada caso. Para decidirlo tenemos una serie de criterios que podemos seguir. Para cada conjunto de datos el criterio puede variar y es necesario realizar un estudio de cual es el más coherente.

2.4.1. Criterio del porcentaje

Este criterio es el más sencillo e intuitivo. Dado que, por las características de los autovalores tenemos que $P_1 > P_2 > \cdots > P_p$, se puede especificar un porcentaje r (dado por el usuario por ejemplo) y tomar las m primeras componentes principales tal que $P_m > r$ o muy próximo a él.

También puede resultar interesante ver a partir de que $k < p$ se estabiliza el porcentaje de variabilidad y tomar las $m = k$ primeras componentes principales.

2.4.2. Criterio de Kaiser

Aunque en general este criterio se suele utilizar con la matriz de correlaciones $\mathbf{R}$, se puede extender a la matriz de covarianzas $\mathbf{S}$. La premisa seguida es la de quedarnos únicamente con las componentes principales que expliquen más que una variable observada.

De esta manera, si obtenemos las componentes principales en base a la matriz de correlaciones, equivale a suponer que las variables observadas tienen varianza 1. Descartaremos entonces las que tengan una varianza menor que 1, quedándonos con las m primeras componentes principales tales que $\lambda_m \geq 1$. Experimentos de Monte Carlo demuestran que en la práctica es mejor tomar $\lambda_m \geq 0.7$

La manera de extenderlo a la matriz de covarianzas es tomar las componentes principales tales

que $\lambda_m \geq v$ siendo $v = \text{tr}(\mathbf{S})/p$ que es la media de las varianzas. De la misma manera, experimentos de Monte Carlo indican que se debe de tomar $\lambda_m \geq 0.7v$.

2.5. Aplicación de componentes principales a datos reales

Para comprobar la utilidad de las componentes principales de una manera práctica hemos escogido dos conjuntos de datos obtenidos de *Kaggle.com*: uno de medidas biométricas de huesos de diferentes tipos de aves (`www.kaggle.com/zhangjuefei/birds-bones-and-living-habits`) y otro de imágenes de células analizadas para determinar si son benignas o malignas con respecto al cáncer de mama (`www.kaggle.com/uciml/breast-cancer-wisconsin-data`).

En ambos casos, se describirán los datos, se explicará el procedimiento seguido para su análisis y se expondrán las conclusiones a las que se ha llegado.

2.5.1. Medidas biométricas de diferentes tipos de aves

Estos datos contienen medidas de la longitud y el diámetro de diferentes huesos de aves. En total se disponen de 413 observaciones de diferentes tipos de aves. Los tipos de aves se dividen en 8 grupos ecológicos dependiendo de sus habitats y estilos de vida. De estos, 6 están cubiertos en esta base de datos.

Para llevar a cabo un análisis más detallado se separan las variables en el grupo de diámetros y en el de longitudes de los 5 huesos de los que se tienen medidas. De esta manera, se podrán realizar dos análisis de componentes principales: una para las longitudes y otra para los diámetros de los huesos (húmero, cúbito, fémur, tibiotarso y tarsometatarso). En todo caso, se realizará un análisis sobre la matriz de covarianzas $\mathbf{S}$. Las varianzas escaladas utilizadas para determinar la selección de las componentes principales se obtienen escalando cada variable para que tengan varianza 1

En primer lugar, se muestran los resultados obtenidos al hallar las 5 componentes principales de las longitudes de los huesos de las aves en la Tabla 2.1. Utilizando los criterios descritos en la Sección 2.4 tomaremos las dos primeras componentes principales. Se puede observar como el porcentaje acumulado llega a prácticamente el 98 % en la segunda componente principal y se estabiliza ahí, el resto de componentes apenas explican nada más ni aportan información relevante. Además, utilizando el criterio de Kaiser vemos que las únicas componentes que superan el umbral de 0.7 son las dos primeras.

Cabe aclarar que los dos primeros huesos tomados como variables (el húmero y el cúbito) pertenecen a las alas de las aves, mientras que los otros tres (fémur, tibiotarso y tarsometatarso) pertenecen a las patas. Esto es especialmente relevante a la hora de analizar el significado de las componentes principales seleccionadas.

Variable	CP1	CP2	CP3	CP4	CP5
Húmero	0.602	-0.200	0.472	-0.266	-0.551
Cúbito	0.649	-0.434	-0.369	0.304	0.403
Fémur	0.187	0.257	-0.674	-0.649	-0.154
Tibiotarso	0.375	0.668	0.351	-0.117	0.525
Tarsometatarso	0.202	0.509	-0.250	0.635	-0.484
Porcentaje de variación	89.81 %	7.98 %	1.19 %	0.68 %	0.31 %
Porcentaje acumulado	89.81 %	97.79 %	98.99 %	99.68 %	100 %
Varianza escalada	2.056	0.723	0.416	0.255	0.111

Tabla 2.1: Componentes principales para las longitudes de los huesos de aves.

La primera componente principal corresponde a un autovector con componentes positivas. Esto ocurre ya que las medidas tomadas son biométricas y, por tanto, la matriz de covarianzas tiene todas sus componentes positivas (cuanto más grande sea un ave, más grande serán en general la longitud de sus huesos). Por lo tanto, se cumple el lema de Perron y la primera componente principal tiene todas sus componentes positivas y será la única que tenga estas características. Por las características de las variables analizadas podemos entender esta primera componente principal como la componente de *tamaño*. Observando los valores que indican lo que aporta cada variable, vemos que en la clasificación de aves tiene especial importancia el tamaño de las alas, que se complementa con el tamaño que puedan tener las patas.

La segunda componente principal pone en contraste la longitud de las alas con la de las patas de las aves. Podemos entender esta componente como la componente de *forma*, distinguiendo entre aves con alas mucho más largas que las alas y aquellas en las que ambas extremidades tienen una longitud similar.

Se realiza un análisis de componentes principales también de las medidas de los diámetros de los huesos. Los resultados obtenidos se muestran en la Tabla 2.2. Como se puede observar, los resultados son similares. Comparando las componentes principales halladas con respecto a los diámetros con las de las longitudes podemos ver que la primera componente principal es la que llamamos de *tamaño*. El autovector correspondiente tiene todas sus componentes positivas, una vez más consecuencia del lema de Perron. Será la única componente principal con estas características.

Los pesos de cada variable son, sin embargo, diferentes de los de las longitudes de huesos (Tabla 2.1). En el análisis por diámetro los huesos correspondientes a las alas pierden importancia. Esto lo podemos justificar con la forma que tienen los huesos, no aumentan en la misma medida su longitud y su diámetro cuando tratamos aves más grandes, sino que aumenta más en longitud que en diámetro. Aún así, siguen siendo las alas las que más información aportan a la hora de clasificar las aves.

Variable	CP1	CP2	CP3	CP4	CP5
Húmero	0.570	0.445	0.119	-0.637	0.239
Cúbito	0.427	0.540	0.040	0.705	-0.165
Fémur	0.399	-0.362	0.191	-0.162	-0.805
Tibiotarso	0.400	-0.538	0.481	0.260	0.502
Tarsometatarso	0.417	-0.299	-0.847	0.055	0.129
Porcentaje de variación	92.77 %	3.06 %	2.73 %	0.95 %	0.48 %
Porcentaje acumulado	92.77 %	95.83 %	98.56 %	99.51 %	100 %
Varianza escalada	2.149	0.407	0.383	0.207	0.167

Tabla 2.2: Componentes principales para los diámetros de los huesos de aves.

Las diferencias con respecto a las longitudes las podemos comprobar también en el porcentaje de variación de cada componente principal, que es mayor en la primera, y en la varianza escalada. En el caso de los diámetros, según el criterio de Kaiser no deberíamos tener en cuenta la segunda componente, dado que tiene una varianza que indica que explica menos que una variable individual de las originales (en promedio). Podemos recalcar que la segunda componente principal vuelve a contraponer las medidas de los huesos de las alas con los de las patas.

2.5.2. Medidas de células relacionadas con el cáncer de mama

En esta base de datos encontramos variables correspondientes a imágenes digitalizadas de biopsias de senos. Las medidas tomadas se corresponden con el núcleo celular y son las siguientes:

- Radio: Media de las distancias del centro al perímetro.
- Textura: Desviación estándar de la escala de grises de la imagen.
- Perímetro.
- Área.
- Suavidad: Variación local de la longitud del radio.
- Compacidad: $\frac{\text{perimetro}^2}{\text{area}-1}$.
- Concavidad: Agudeza de las porciones cóncavas del contorno.
- Puntos de concavidad: Número de porciones cóncavas del contorno.
- Simetría.
- Dimensión fractal.

Además de esto, se dispone de una clasificación de cada imagen en células benignas (63 % de la muestra total) y malignas (37 % de la muestra total). Esto nos será especialmente útil para ver como

afecta cada variable en cada caso y poder hacer un análisis de componentes principales diferenciando entre cada tipo de imagen.

Vamos a realizar primero el análisis centrándonos en las imágenes clasificadas como malignas. Los resultados de realizar el análisis por componentes principales sobre la matriz de covarianzas $\mathbf{S}$, se muestran en la Tabla 2.3. De las 10 componentes principales se muestran solo 7 ya que el resto no aportan información relevante.

Variable	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**
Radio	0.299	-0.410	0.054	-0.05	3 0.000	-0.233	0.071
Textura	0.021	-0.093	-0.980	0.021	-0.163	-0.048	-0.024
Perímetro	0.328	-0.379	0.041	-0.052	0.049	-0.192	0.083
Área	0.302	-0.403	0.057	-0.003	-0.023	-0.297	0.232
Suavidad	0.271	0.353	0.093	0.513	-0.607	-0.189	0.211
Compacidad	0.381	0.254	-0.101	-0.080	0.526	0.103	0.050
Concavidad	0.440	0.074	-0.056	0.125	0.079	0.636	0.412
Puntos de concavidad	0.450	-0.044	0.047	0.121	-0.118	0.173	-0.845
Simetría	0.253	0.323	0.021	-0.823	-0.379	-0.071	0.047
Dimensión fractal	0.181	0.465	-0.082	0.131	0.401	-0.577	-0.039
Porcentaje de variación	46.36 %	34.02 %	10.00 %	4.548 %	2.641 %	1.319 %	0.528 %
Porcentaje acumulado	46.36 %	80.38 %	90.39 %	94.93 %	97.57 %	98.89 %	99.42 %
Varianza escalada	2.153	1.845	1.000	0.674	0.514	0.363	0.230

Tabla 2.3: Componentes principales para las imágenes de células malignas.

Observando los porcentajes de variación que aporta cada componente y aplicando el criterio de Kaiser decidimos quedarnos con las tres primeras componentes principales. Con ellas, explicaríamos un 90 % de la variabilidad de la muestra, lo que es un porcentaje aceptable teniendo en cuenta la reducción de dimensión que estamos consiguiendo.

La primera de las componentes principales tiene todos los valores de su autovector positivo. Como en el caso de las aves esto indica que podemos tratar a esta componente como la componente de *tamaño*. Además, no hay ninguna variable que tenga un peso considerablemente mayor al resto. Únicamente la textura destaca por la poca información que aporta a esta componente de *tamaño*.

La segunda componente principal tiene características que requieren de un análisis más detallado. En primer lugar, podemos descartar la aportación que hacen la textura, la concavidad y los puntos de concavidad dado que es muy reducida con respecto del resto de variables. No quiere decir que a la hora de, por ejemplo, intentar clasificar células en benignas o malignas, no utilicemos estas variables, sino que en esta componente en específico no son de gran utilidad.

Por otra parte, vemos que en esta componente principal se contraponen los valores de radio, perímetro y área con los de suavidad, compacidad, simetría y dimensión fractal. Esto significa que está diferenciando células que son más *redondeadas* y sin muchas irregularidades en su forma de aquellas que tienen formas más onduladas (que suelen corresponder a células tumorales). Es por esto que consideramos esta componente principal la de *forma*.

La tercera componente principal es significativa a nivel de porcentaje de variación y ha sido seleccionada por tener una varianza escalada de 1. Si nos fijamos en los valores del autovector, vemos que la textura es la única variable que tiene un peso significativo (aunque la compacidad aporta también). La conclusión a la que llegamos con esta variable es que aporta prácticamente la misma información que la textura por si sola. Muestra de ello es que tiene una varianza escalada de 1 (como cualquier variable individualmente) y explica el 10 % de la variabilidad de una muestra de 10 variables. Por lo tanto, aunque no la descartamos continuaremos el análisis centrándonos en las dos primeras componentes principales.

En el gráfico de la Figura 2.1 se ha aplicado la transformación de las componentes principales de las células malignas al conjunto de todos los datos. En verde podemos ver los puntos correspondientes a las células benignas y en rojo a las células malignas. Se puede apreciar que existe una tendencia que puede permitir separar unas de otras y detectarlas en función de las variables estudiadas, que únicamente tienen que ver con la forma y el tamaño de cada célula.

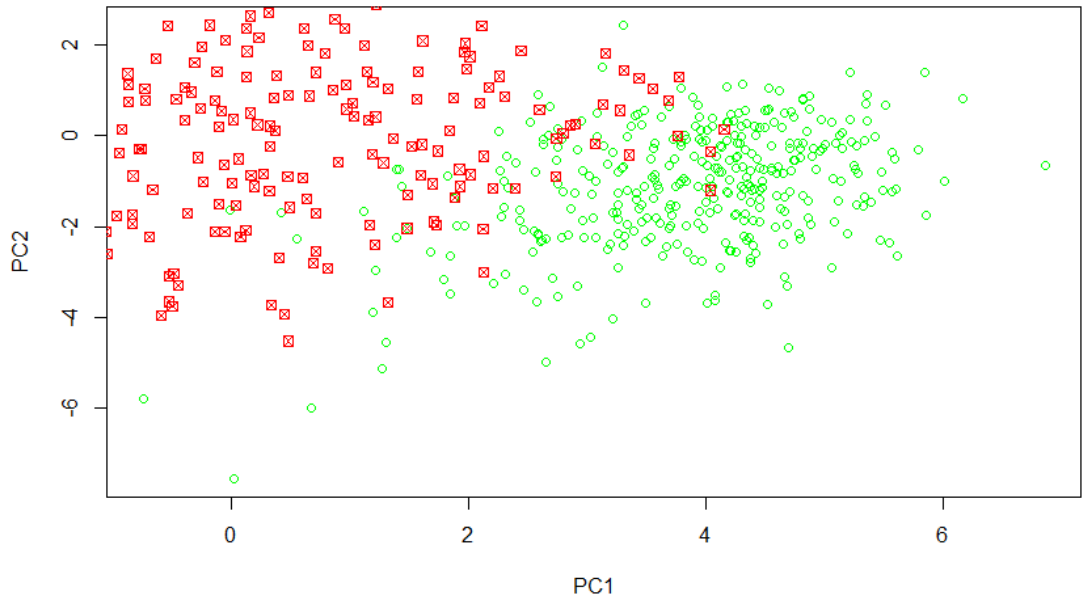

Figura 2.1: Gráfico de la aplicación de la transformación de las componentes principales de células malignas a todos los datos

Vamos ahora a realizar un análisis similar al realizado con las células malignas, pero con las be-

nignas. Realizando un análisis por componentes principales obtenemos los resultados mostrados en la Tabla 2.4

Variable	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**
Radio	0.425	-0.302	-0.044	-0.101	0.026	-0.170	0.099
Textura	0.040	0.091	0.853	-0.414	0.290	0.044	0.026
Perímetro	0.404	-0.333	-0.028	-0.086	0.027	-0.190	0.072
Área	0.426	-0.299	-0.039	-0.087	0.000	-0.194	0.119
Suavidad	-0.305	-0.237	-0.347	-0.306	0.643	0.192	0.429
Compacidad	-0.246	-0.435	0.146	0.090	0.108	-0.377	-0.303
Concavidad	-0.173	-0.397	0.320	0.347	-0.380	0.290	0.599
Puntos de concavidad	-0.048	-0.500	0.020	0.044	0.067	0.557	-0.579
Simetría	-0.286	-0.149	-0.130	-0.736	-0.577	-0.061	-0.020
Dimensión fractal	-0.453	-0.159	0.085	0.194	0.086	-0.564	-0.007
Porcentaje de variación	36.83 %	34.33 %	10.80 %	7.68 %	5.60 %	2.51 %	1.23 %
Porcentaje acumulado	36.83 %	71.15 %	81.95 %	89.63 %	95.24 %	97.74 %	98.98 %
Varianza escalada	1.919	1.853	1.039	0.877	0.748	0.501	0.352

Tabla 2.4: Componentes principales para las imágenes de células benignas.

Aunque, a priori, estos resultados pueden parecer diferentes a los de las células malignas, existen muchas similitudes. En primer lugar, siguiendo los mismos criterios de selección podemos quedarnos con las tres primeras componentes principales. Los datos muestran que la cuarta y quinta componente principal pueden ser de utilidad también, pero esto nos llevaría a no disminuir tanto la dimensión como sería deseable y dado el poco porcentaje de variación que explican nos quedamos únicamente con las tres primeras componentes.

Observando las dos primeras componentes principales parecen diferentes a las halladas en las células malignas, pero simplemente están intercambiadas. La primera componente principal en las malignas tiene su equivalente en la segunda componente de las benignas y al revés. Además, los signos de las componentes de ambas componentes están invertidos también, en este caso la matriz de varianzas no es enteramente positiva con lo que no se tiene por que cumplir el lema de Perron. Esto hace recalcar la importancia de ambas ya que explican una cantidad similar de la información, reforzando el argumento de centrarnos en ellas.

El análisis de estas dos componentes es análogo al de las células malignas. En la componente de *tamaño* (la segunda) la información que aporta la textura es despreciable, pero la concavidad y los puntos de concavidad si que tienen una mayor importancia. La componente de *forma* contrapone, de la misma forma que antes, variables de tamaño con las de forma. Diferencia así entre células con más o menos irregularidades.

Cabe destacar que la tercera componente en este caso explica más que en las células malignas. A la importancia de la textura (con muy poco peso en las dos primeras componentes) se suman otras variables como la suavidad o la concavidad.

Se ha realizado como en el otro caso un gráfico que se puede observar en la Figura 2.2. Los resultados son, una vez más, similares. Existen dos grupos diferenciados de células: las benignas en verde y las malignas en rojo. Potencialmente, estos resultados podrían servir para realizar una clasificación de las células en benignas o malignas en función de su tamaño y forma.

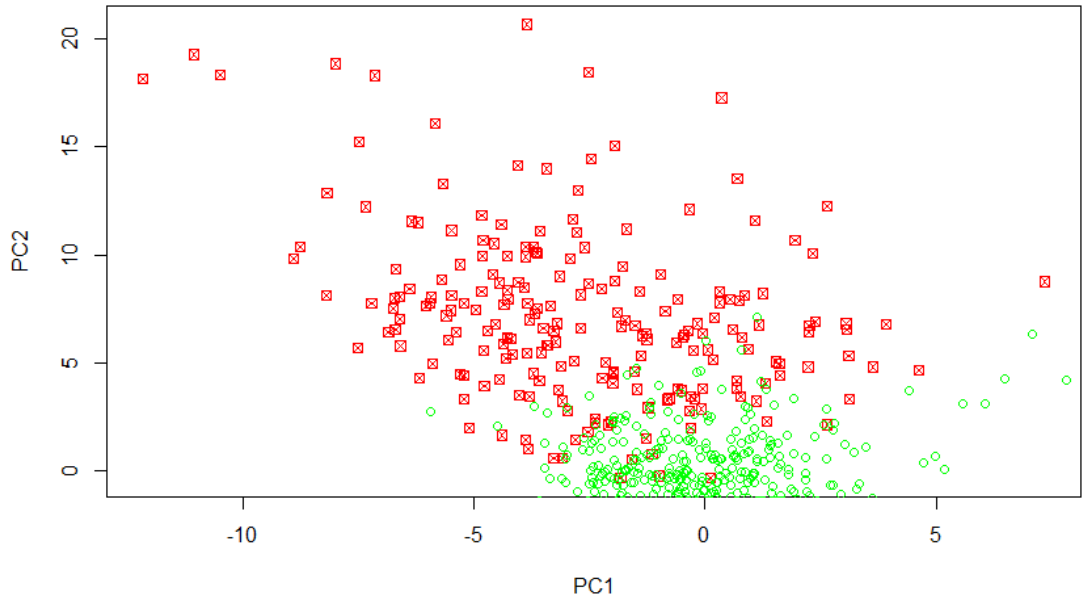

Figura 2.2: Gráfico de la aplicación de la transformación de las componentes principales de células benignas a todos los datos

3

Correlaciones canónicas

Las correlaciones canónicas son un método estadístico multivariante que nos permite estudiar las relaciones que pueden existir entre dos grupos de variables. Son dos combinaciones lineales de estos grupos de variables con máxima correlación lineal entre ellas, mostrando cuales de las variables individuales de cada grupo tiene una mayor importancia en la correlación total.

3.1. Tipos de correlación

Dependiendo de si queremos describir el grado de relación lineal entre variables aleatorias o vectores aleatorios debemos utilizar diferentes tipos de correlación. Las correlaciones canónicas son una generalización de las correlaciones simple y múltiple.

Tenemos tres posibilidades de relacionar dos variables:

- La correlación lineal de Pearson si X, Y son dos variables.
- La correlación múltiple si Y es una variable y $\mathbf{X}$ es un vector de variables. Puede considerarse una extensión del coeficiente de correlación lineal de Pearson entre Y y las componentes de $\mathbf{X}$.
- La correlación canónica si $\mathbf{X}$ e $\mathbf{Y}$ son dos vectores de variables.

Nos centraremos en la última, dado que nuestro objetivo es conseguir relacionar dos grupos de variables.

3.2. Correlación canónica

Sean $\mathbf{X} = (X_1, \ldots, X_p)'$ e $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ dos vectores de variables que observamos en una muestra de n individuos. Buscamos dos vectores $\mathbf{a} = (a_1, \ldots, a_p)'$ y $\mathbf{b} = (b_1, \ldots, b_q)'$ que definan las combinaciones lineales

$$U = \mathbf{a}'\mathbf{X} = a_1X_1 + ... + a_pX_p \quad \text{y} \quad V = \mathbf{b}'\mathbf{Y} = b_1Y_1 + ... + b_qY_q$$

de tal manera que cor(U, V), la correlación muestral entre U y V sea máxima.

Denotamos $\mathbf{S}_{11}$ y $\mathbf{S}_{22}$ las matrices de covarianzas muestrales de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente. Y las matrices de covarianzas cruzadas $\mathbf{S}_{12} = \mathbf{S}_{21}'$:

$$\mathbf{S}_{11} = \text{var}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' \quad \text{y} \quad \mathbf{S}_{22} = \text{var}(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})$$

$$\mathbf{S}_{12} = \mathbf{S}_{21}' = \text{cov}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{y}_i - \overline{\mathbf{y}})'.$$

Para que el problema de maximizar cor(U, V) tenga solución, imponemos var$(U) = \mathbf{a}'\mathbf{S}_{11}\mathbf{a} = 1$ y var$(V) = \mathbf{b}'\mathbf{S}_{22}\mathbf{b} = 1$. De esta manera, el problema se transforma en maximizar cor$(U, V) =$ cov$(U, V) = \mathbf{a}'\mathbf{S}_{12}\mathbf{b}$. Los vectores $\mathbf{a}, \mathbf{b}$ que cumplen esto son los primeros vectores canónicos. La máxima correlación entre U, V es la primera correlación canónica. Veremos con los teoremas que siguen como relacionar éstas con la descomposición en valores singulares.

Teorema 3.1 *Los primeros vectores canónicos satisfacen las ecuaciones:*

$$\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a} = \lambda\mathbf{S}_{11}\mathbf{a}$$
$$\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b} = \lambda\mathbf{S}_{22}\mathbf{b}$$

Demostración: Queremos maximizar la función $f(\mathbf{a}, \mathbf{b}) = \mathbf{a}'\mathbf{S_{12}}\mathbf{b}$ sujeta a las restricciones:

$$g_1(\mathbf{a}, \mathbf{b}) = \mathbf{a}'\mathbf{S}_{11}\mathbf{a} - 1 = 0$$
$$g_2(\mathbf{a}, \mathbf{b}) = \mathbf{b}'\mathbf{S}_{22}\mathbf{b} - 1 = 0$$

Utilizando multiplicadores de Lagrange [6], tenemos lo siguiente:

$$\nabla f(\mathbf{a}, \mathbf{b}) = \mu_1\nabla g_1(\mathbf{a}, \mathbf{b}) + \mu_2\nabla g_2(\mathbf{a}, \mathbf{b}) \implies \begin{bmatrix} \frac{\partial \mathbf{a}'\mathbf{S}_{12}\mathbf{b}}{\partial \mathbf{a}} \\ \frac{\partial \mathbf{a}'\mathbf{S}_{12}\mathbf{b}}{\partial \mathbf{b}} \end{bmatrix} = \mu_1 \begin{bmatrix} \frac{\partial \mathbf{a}'\mathbf{S}_{11}\mathbf{a}}{\partial \mathbf{a}} \\ 0 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 \\ \frac{\partial \mathbf{b}'\mathbf{S}_{22}\mathbf{b}}{\partial \mathbf{b}} \end{bmatrix}$$

del que obtenemos el sistema de ecuaciones:

$$\mathbf{S}_{12}\mathbf{b} - 2\mu_1\mathbf{S}_{11}\mathbf{a} = 0 \tag{3.1}$$
$$\mathbf{S}_{21}\mathbf{a} - 2\mu_2\mathbf{S}_{22}\mathbf{b} = 0 \tag{3.2}$$
$$\mathbf{a}'\mathbf{S}_{11}\mathbf{a} - 1 = 0 \tag{3.3}$$
$$\mathbf{b}'\mathbf{S}_{22}\mathbf{b} - 1 = 0 \tag{3.4}$$

Multiplicando (3.1) y (3.2) por $\mathbf{a}'$ y $\mathbf{b}'$, respectivamente, obtenemos:

$$\begin{cases} \mathbf{a}'\mathbf{S}_{12}\mathbf{b} - 2\mu_1\mathbf{a}'\mathbf{S}_{11}\mathbf{a} = 0 \\ \mathbf{b}'\mathbf{S}_{21}\mathbf{a} - 2\mu_2\mathbf{b}'\mathbf{S}_{22}\mathbf{b} = 0 \end{cases} \Longrightarrow \begin{cases} \mathbf{a}'\mathbf{S}_{12}\mathbf{b} - 2\mu_1 = 0 \\ \mathbf{b}'\mathbf{S}_{21}\mathbf{a} - 2\mu_2 = 0 \end{cases} \Longrightarrow \quad 2\mu_1 = 2\mu_2 = \mu \tag{3.5}$$

ya que queremos maximizar $\mathbf{a}'\mathbf{S}_{12}\mathbf{b} = \mathbf{a}'\mathbf{S}_{21}\mathbf{a}$.

Ahora utilizamos las ecuaciones (3.2) y (3.5) para obtener $\mathbf{b} = \mu^{-1}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}$ y sustituyendo en la equación (3.1) tenemos $\mu^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a} - \mu\mathbf{S}_{11}\mathbf{a} = 0$. De aquí obtenemos $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a} = \lambda\mathbf{S}_{11}\mathbf{a}$ utilizando la notación $\mu^2 = \lambda$. Análogamente obtenemos $\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b} = \lambda\mathbf{S}_{22}\mathbf{b}$.

Observemos, por tanto, que $\mathbf{a}$ y $\mathbf{b}$ son autovectores de $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ y $\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$, respectivamente. □

Teorema 3.2 *Los vectores canónicos que satisfacen las ecuaciones (3.3) y (3.4), están relacionados por:*

$$\mathbf{a} = \lambda^{-1/2}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}$$
$$\mathbf{b} = \lambda^{-1/2}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}$$

y la primera correlación canónica es $r_1 = \sqrt{\lambda_1}$, donde λ_1 es el primer valor propio de $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$

Demostración: Tenemos del Teorema 3.1 que $\mathbf{b} = \mu^{-1}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}$ y que $\mathbf{a} = \mu^{-1}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}$. Esto prueba la primera afirmación. De la ecuación (3.3) se sigue que:

$$1 = \mathbf{a}'\mathbf{S}_{11}\mathbf{a} = \mu^{-1}\mathbf{a}'\mathbf{S}_{11}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b} = \mu^{-1}\mathbf{a}'\mathbf{S}_{12}\mathbf{b} \tag{3.6}$$

La correlación es $r_1 = \mathbf{a}'\mathbf{S_{12}}\mathbf{b}$ y usando (3.6) y $\mu^2 = \lambda$, siendo λ el valor primer valor propio de $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ (ver demostración del Teorema 3.1) se cumple $r_1 = \sqrt{\lambda}$. □

Además, las ecuaciones del Teorema 3.1 tienen otras soluciones. En concreto, si $\mathbf{X}$ tiene p variables e $\mathbf{Y}$ tiene q, existen $m = \min(p, q)$ pares de vectores canónicos solución $\mathbf{a}_1, \mathbf{b}_1, \ldots, \mathbf{a}_m, \mathbf{b}_m$ que proporcionan las variables y correlaciones canónicas:

$$\begin{array}{ccc} U_1 = \mathbf{X}\mathbf{a}_1, & V_1 = \mathbf{Y}\mathbf{b}_1, & r_1 = \text{cor}(U_1, V_1) \\ \vdots & \vdots & \vdots \\ U_m = \mathbf{X}\mathbf{a}_m, & V_m = \mathbf{Y}\mathbf{b}_m, & r_m = \text{cor}(U_m, V_m) \end{array}$$

Teorema 3.3 *Supongamos $r_1 > r_2 > \; > r_m$ las correlaciones canónicas de dos vectores de variables, $\mathbf{X}$ e $\mathbf{Y}$. Entonces:*

1.– Tanto las variables canónicas $U_1, ..., U_m$ como las variables canónicas $V_1, ..., V_m$ son incorreladas.

2.– La primera correlación canónica $r_1 = \text{cor}(U_1, V_1)$ es la máxima correlación entre una combinación lineal de $\mathbf{X}$ y una combinación lineal de $\mathbf{Y}$.

3.– La segunda correlación canónica $r_2 = \text{cor}(U_2, V_2)$ es la máxima correlación entre las

combinaciones lineales de $\mathbf{X}$ incorreladas con U_1 y las combinaciones lineales de $\mathbf{Y}$ incorreladas con V_1.

4.– cor$(U_i, V_j) = 0$ si $i \neq j$.

Demostración: Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ los autovalores de $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ y a_i el autovector correspondiente a λ_i. Por tanto, para cada i, tenemos $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_i = \lambda S_{11}\mathbf{a}_i$. Para cualquier a_j con $\lambda_i \neq \lambda_j$ se cumple entonces que

$$\mathbf{a}_j'\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_i = \lambda_i\mathbf{a}_j'\mathbf{S}_{11}\mathbf{a}_i \quad \text{y} \quad \mathbf{a}_i'\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_j = \lambda_j\mathbf{a}_i'\mathbf{S}_{11}\mathbf{a}_j.$$

Si restamos ambas ecuaciones tenemos que $(\lambda_i - \lambda_j)\mathbf{a}_i'\mathbf{S}_{11}\mathbf{a}_j = 0$, es decir, $\mathbf{a}_i'\mathbf{S}_{11}\mathbf{a}_j = 0$ y , por tanto, cov$(U_i, U_j) = 0$. Análogamente, tenemos que cor$(V_i, V_j) = 0$ siendo todas las variables canónicas incorreladas entre sí.

Para ver que las variables U_i son incorreladas con las V_i observemos que

$$\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_i = \lambda_i\mathbf{a}_i \quad \text{y} \quad \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}_j = \lambda_j\mathbf{b}_j.$$

Si multiplicamos por $\mathbf{b}_j'\mathbf{S}_{21}$ y por $\mathbf{a}_i'\mathbf{S}_{12}$, respectivamente, llegamos a que

$$\mathbf{b}_j'\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_i = \lambda_i\mathbf{b}_j'\mathbf{S}_{21}\mathbf{a}_i \quad \text{y} \quad \mathbf{a}_i'\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}_j = \lambda_j\mathbf{a}_i'\mathbf{S}_{12}\mathbf{b}_j.$$

Restando nuevamente, tenemos que $(\lambda_i - \lambda_j)\mathbf{a}_i'\mathbf{S}_{11}\mathbf{b}_j = 0$, es decir, $\mathbf{a}_i'\mathbf{S}_{11}\mathbf{b}_j = 0$ y, por tanto, cor$(U_i, V_j) = 0$ con lo que queda demostrado que las U_i y las V_i están incorreladas. □

De esta manera, hemos concluido que mediante las matrices anteriormente definidas ($\mathbf{S}_{11}, \mathbf{S}_{22}$ y $\mathbf{S}_{12}$) podemos encontrar las correlaciones canónicas de dos muestras multivariantes.

3.3. Descomposición en valores singulares y correlación canónica

Utilizando la descomposición en valores singulares de una matriz podemos llegar a una fórmula que nos genera los vectores canónicos y correlaciones canónicas.

Consideramos la matriz $p \times q$:

$$\mathbf{Q} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2},$$

Si hallamos su descomposición en valores singulares siguiendo la notación de la Sección 1.3 se tiene lo siguiente.

Teorema 3.4 *Los vectores canónicos y correlaciones canónicas son*

$$\mathbf{a}_i = \mathbf{S}_{11}^{-1/2}\mathbf{u}_i, \;\; \mathbf{b}_i = \mathbf{S}_{22}^{-1/2}\mathbf{u}_i, \;\; r_i = \lambda_i,$$

donde $\mathbf{u}_i$ *son las columnas de* $\mathbf{U}$ *y* $\mathbf{D} = diag(\lambda_1, \lambda_2, \ldots)$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots$.

Demostración:
Utilizando la descomposición en valores singulares, tenemos la siguiente igualdad:

$$\mathbf{QQ}' = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}\mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2} = \mathbf{UDV}'.$$

De esta manera, podemos identificar los autovectores y autovalores de $\mathbf{QQ}'$ y llegar al resultado:

$$\mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}\mathbf{u}_i = \lambda_i^2\mathbf{u}_i \implies \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}(\mathbf{S}_{11}^{-1/2}\mathbf{u}_i) = \lambda_i(\mathbf{S}_{11}^{-1/2}\mathbf{u}_i).$$

□

3.4. Aplicación de correlación canónica a datos reales

Una vez desarrollada la parte teórica con respecto a las correlaciones canónicas, se prosigue con la aplicación de esta técnica de análisis estadístico a datos reales. Primero se realizará una descripción de los datos indicando como se seleccionaron y, posteriormente, se expondrán los resultados y conclusiones obtenidas.

3.4.1. Descripción de los datos

Los datos utilizados se han sacado de la base de datos del Banco Mundial correspondiente a *World Development Indicators* (`databank.worldbank.org/databases`). En este caso se necesitaban dos muestras multivariantes de las que pudiéramos tener datos de los mismos individuos. Se seleccionaron datos acerca del nivel educativo de las mujeres y de la libertad y poder de decisión que tienen en diferentes países. En todo caso, se han utilizado los datos más actualizados en la medida de lo posible desde el año 2015, siendo los países los individuos sobre los que haremos el análisis.

Los datos a nivel educativo encontrados en un inicio fueron un total de 17 variables, de las que finalmente nos quedamos con 7 debido a que las muestras no eran completas en todos los países. Son las siguientes:

- Ratio de alfabetismo en mujeres jóvenes (% de mujeres entre 15 y 24 años). (ALF-JOVEN)
- Ratio de alfabetismo en mujeres adultas (% de mujeres mayores de 15 años). (ALF-ADULTA)
- Ratio estudiante-profesor en educación primaria. (EST-PROF-PRIM)

- Ratio estudiante-profesor en educación secundaria. (EST-PROF-SEC)
- Porcentaje de inscripción femenino en educación primaria. (INSC-PRIM)
- Porcentaje de inscripción femenino en educación secundaria. (INST-SEC)
- Porcentaje de profesoras en la educación secundaria. (PROFESORAS)

Su distribución se puede observar en la Figura 3.1.

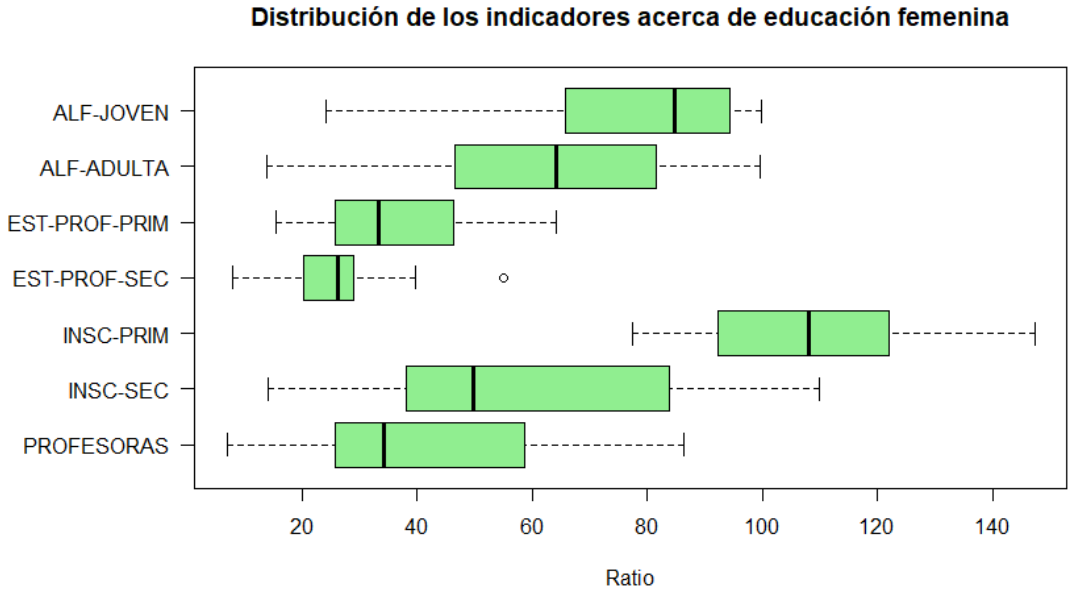

Figura 3.1: Distribución de los indicadores de educación femeninos por países.

En cuanto a los datos relacionados con los derechos y libertades de las mujeres, estos provienen de una encuesta realizada por Unicef. Se corresponden a 5 preguntas con respuesta afirmativa o negativa, de las que se toma el porcentaje de respuestas afirmativas. Las variables son las mostradas a continuación:

- Mujeres que creen que está justificado que su marido la golpee si ella discute con él. (DISCUTIR)
- Mujeres que creen que está justificado que su marido la golpee si a ella se le quema la comida.(COMIDA)
- Mujeres que creen que está justificado que su marido la golpee si ella desatiende a los hijos. (HIJOS)
- Mujeres que creen que está justificado que su marido la golpee si ella sale de casa sin consultárselo. (SALIR)
- Mujeres que creen que está justificado que su marido la golpee si ella se niega a mantener relaciones sexuales con él. (RELACIONES)

Su distribución se puede observar en la Figura 3.2.

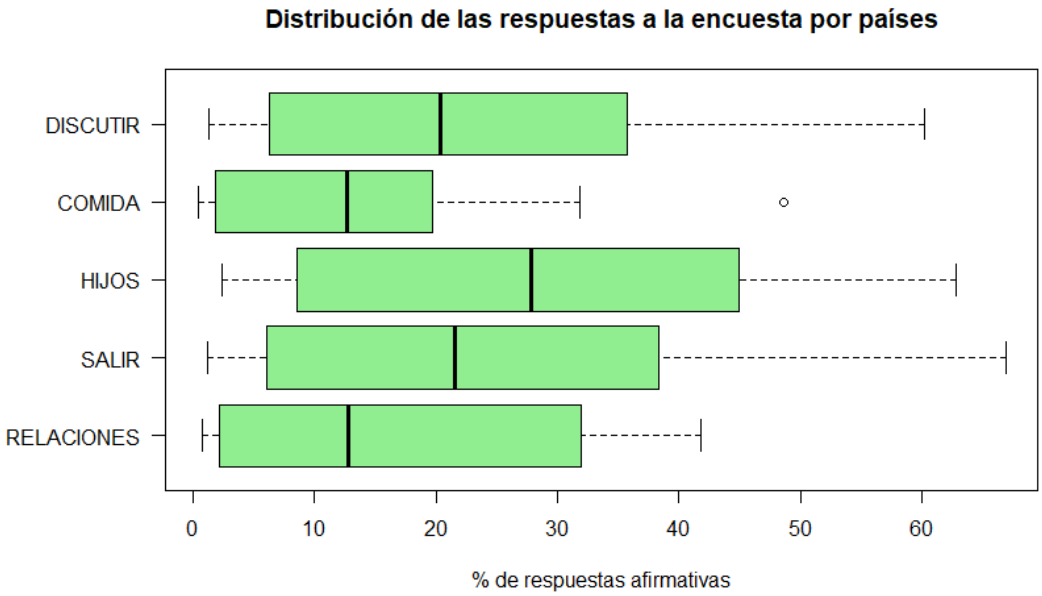

Figura 3.2: Distribución de las respuestas a la encuesta por países

Aunque la cantidad de países de los que se tenía alguno de estos datos era numeroso (264 países), se ha tenido que reducir el número utilizado para las pruebas dado que no todos los países tenían información de todas las variables. Se utilizan únicamente 19 países.

3.4.2. Resultados y conclusiones

Con este planteamiento y estos datos pasamos a realizar el análisis mediante el uso de las correlaciones canónicas. Los resultados obtenidos se muestran en la Tabla 3.1. Para cada variable, se indica su peso en la variable canónica correspondiente. En la última fila se encuentran los valores de las correlaciones canónicas.

Como se puede observar, tenemos dos pares de variables canónicas con una correlación alta $r_1 = 0.928$ y $r_2 = 0.856$, lo que muestra que las variables seleccionadas guardan una relación considerable.

En ambas variables canónicas hay elementos de distinto signo. Esto puede resultar llamativo dado que todas las variables seleccionadas relacionadas con educación son *positivas* para el desarrollo de los derechos de las mujeres y las preguntas relacionadas con la libertad de las mujeres son todas *negativas* para el desarrollo de sus derechos. Es precisamente por esto, porque todas las variables aportan en la misma dirección, respectivamente, por lo que vemos pesos muy pequeños en magnitud y que pueden tener distinto signo.

Fijándonos en detalle en las dos primeros pares de variables canónicas, vemos que ambos son

Variable	CC1	CC2	CC3	CC4	CC5
ALF-JOVEN	-0.023	-0.030	0.002	0.020	0.005
ALF-ADULTA	0.009	0.019	-0.008	-0.016	-0.002
EST-PROF-PRIM	0.006	-0.010	-0.027	0.007	0.007
EST-PROF-SEC	0.004	0.007	0.012	0.009	0.014
INSC-PRIM	-0.002	0.011	0.002	-0.012	0.000
INSC-SEC	0.008	0.000	-0.009	0.012	-0.004
PROFESORAS	-0.001	0.006	0.003	-0.002	0.008
DISCUTIR	-0.027	-0.028	-0.014	0.020	-0.054
COMIDA	0.032	0.010	-0.007	-0.016	-0.021
HIJOS	-0.031	0.001	0.013	-0.012	-0.015
SALIR	0.040	0.024	0.026	0.008	0.059
RELACIONES	-0.002	-0.019	-0.020	-0.014	0.022
Correlación canónica	0.928	0.856	0.603	0.474	0.367

Tabla 3.1: Resultados de la aplicación del método de correlaciones canónicas.

relativamente parecidas. Esto se observa especialmente en los factores relacionados con la educación, que mantienen relativamente estables. En estas variables, vemos pesos similares y no existe ninguna que destaque por su magnitud, por lo que se consideran igualmente importantes (salvo los casos de RELACIONES en la primera variable e HIJOS en la segunda).

Con respecto a los factores relacionados con la educación, destaca la alfabetización de mujeres jóvenes (entre 14 y 25 años) sobre todo en la primera variable canónica donde el resto tienen pesos menores. Se observa que son también significativas, aunque en menor medida la alfabetización de mujeres adultas, el ratio de profesores por alumno y la inscripción de mujeres a educación primaria. Esto nos hace ver que la alfabetización es especialmente importante dado que esta depende, a su vez, de la inscripción en educación primaria de mujeres.

Podemos concluir que la relación entre una educación de calidad de las mujeres y las libertades y derechos que tienen es alta. Aunque los datos usados no son especialmente amplios y las variables utilizadas podrían ser más numerosas, podemos usar los resultados para llegar a estas conclusiones dada la alta correlación que muestra el análisis.

4

Aproximación y completación de matrices

La aproximación de una matriz consiste en encontrar otra de rango menor, pero que conserve la mayoría de características de la matriz original en la medida de lo posible. De esta manera, reducimos la dimensión del problema, lo que es especialmente útil para matrices muy grandes.

Podemos usar este método para completar datos faltantes en una matriz, partiendo de una suposición inicial. Así, basándonos en los datos que ya tenemos, podremos estimar aquellos que están incompletos.

Un caso especialmente interesante es el de Netflix, que realiza un procedimiento similar con su sistema de recomendación de contenido. En este capítulo realizaremos un sistema similar a menor escala.

4.1. Teorema de la aproximación

El problema de la aproximación consiste en, dada una matriz, aproximarla por otra que tenga un rango inferior y que sea similar a ella de alguna manera. Este problema fue resuelto por Schmidt, ver [4]. A continuación, mostraremos una demostración del resultado.

Teorema 4.1 *Dada una matriz $\mathbf{A}$ de dimensiones $m \times n$, la matriz de rango menor o igual que k y expresada en la forma*

$$\sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i'$$

que mejor aproxima a $\mathbf{A}$, en el sentido de que minimiza el siguiente error de aproximación:

$$\left|\left| \mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i' \right|\right|$$

donde $||\mathbf{A}|| = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 = tr(\mathbf{A}\mathbf{A}')$ es la norma de Frobenius, viene dada por los k primeros términos de la descomposición en valores singulares de la matriz $\mathbf{A}$.

Si denotamos

$$\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i'$$

(utilizando la notación de la Sección 1.3), tenemos que

$$||\mathbf{A} - \mathbf{A_k}|| = \min \left|\left| \mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i' \right|\right|.$$

Demostración:

Se cumple entonces la siguiente igualdad:

$$\begin{aligned} ||\mathbf{A} - \mathbf{A}_k||^2 &= \left|\left| \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i' - \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i' \right|\right|^2 = \left|\left| \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i' \right|\right|^2 \\ &= \sum_{i=k+1}^{r} \sigma_i^2 = \sum_{i=1}^{r} \sigma_i^2 - \sum_{i=1}^{k} \sigma_i^2 = ||\mathbf{A}||^2 - \sum_{i=1}^{k} \sigma_i^2. \end{aligned}$$

Por lo tanto, si podemos mostrar que para cualquier $\mathbf{x}_i$ y $\mathbf{y}_i$ arbitrarios se cumple que

$$\left|\left| \mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i' \right|\right|^2 \geq ||\mathbf{A}||^2 - \sum_{i=1}^{k} \sigma_i^2$$

entonces $\mathbf{A}_k$ será la aproximación que minimizaría la diferencia.

Podemos asumir que los vectores $\mathbf{x}_1, \ldots, \mathbf{x}_k$ son ortonormales. Si no lo fueran, bastaría con aplicar el algoritmo de Gram-Schmidt. Por tanto, sin perdida de generalidad tenemos que:

$$\begin{aligned} \left|\left| \mathbf{A} - \sum_{i=i}^{k} \mathbf{x}_i \mathbf{y}_i' \right|\right|^2 &= \text{tr}\left[(\mathbf{A} - \sum_{i=i}^{k} \mathbf{x}_i \mathbf{y}_i')'(\mathbf{A} - \sum_{i=i}^{k} \mathbf{x}_i \mathbf{y}_i') \right] \\ &= \text{tr}\left[\mathbf{A}'\mathbf{A} + \sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{A}'\mathbf{x}_i)(\mathbf{y}_i - \mathbf{A}'\mathbf{x}_i)' - \sum_{i=1}^{k} \mathbf{A}'\mathbf{x}_i \mathbf{x}_i' \mathbf{A} \right] \\ &= \text{tr}(\mathbf{A}'\mathbf{A}) + \sum_{i=1}^{k} \text{tr}[(\mathbf{y}_i - \mathbf{A}'\mathbf{x}_i)(\mathbf{y}_i - \mathbf{A}'\mathbf{x}_i)'] - \sum_{i=1}^{k} \text{tr}(\mathbf{A}'\mathbf{x}_i \mathbf{x}_i' \mathbf{A}) \\ &\geq \sum_{i=1}^{k} \sigma_i^2 - \left|\left| \mathbf{A}'\mathbf{x}_i \right|\right|^2 = \|\mathbf{A}\|^2 - \sum_{i=1}^{k} \left|\left| \mathbf{A}^{\mathbf{T}}\mathbf{x}_i \right|\right|^2. \end{aligned}$$

Por lo tanto, queda probar que

$$\sum_{i=1}^{k} \left|\left| \mathbf{A}'\mathbf{x}_i \right|\right|^2 \leq \sum_{i=1}^{k} \sigma_i^2.$$

Utilizando la descomposición en valores singulares, definimos $\mathbf{U} = (\mathbf{U}_1 \mathbf{U}_2)$ donde $\mathbf{U}_1$ tiene k

columnas y $\mathbf{U}_2$ tiene $n-k$. De manera análoga definimos $\mathbf{D} = (\mathbf{D}_1\mathbf{D}_2)$ como una partición de la matriz diagonal $\mathbf{D}$.

Como se cumple que

$$\begin{aligned} 0 &= \sigma_k^2 - \sigma_k^2 = \sigma_k^2 - \sigma_k^2\|\mathbf{x}_i\|^2 \\ &= \sigma_k^2 - \sigma_k^2\|\mathbf{U}'\mathbf{x}\sigma_i\|^2 = \sigma_k^2(1 - \|\mathbf{U}'\mathbf{x}_i\|^2) \\ &= \sigma_k^2(1 - \|\mathbf{U}_1'\mathbf{x}_i\|^2 - \|\mathbf{U}_2'\mathbf{x}_i\|^2), \end{aligned}$$

podemos expresar $\|\mathbf{A}'\mathbf{x}_i\|^2$ de la siguiente manera:

$$\begin{aligned} \|\mathbf{A}'\mathbf{x}_i\|^2 &= \|\mathbf{D}\mathbf{U}'\mathbf{x}_i\|^2 \\ &= \|\mathbf{D}_1\mathbf{U}_1'\mathbf{x}_i\|^2 + \|\mathbf{D}_2\mathbf{U}_2'\mathbf{x}_i\|^2 \\ &= \|\mathbf{D}_1\mathbf{U}_1'\mathbf{x}_i\|^2 + \|\mathbf{D}_2\mathbf{U}_2'\mathbf{x}_i\|^2 + \sigma_k^2(1 - \|\mathbf{U}_1'\mathbf{x}_i\|^2 - \|\mathbf{U}_2'\mathbf{x}_i\|^2) \\ &= \sigma_k^2 + (\|\mathbf{D}_1\mathbf{U}_1'\mathbf{x}_i\|^2 - \sigma_k^2\|\mathbf{U}_1'\mathbf{x}_i\|^2) - (\sigma_k^2\|\mathbf{U}_2'\mathbf{x}_i\|^2 - \|\mathbf{D}_2\mathbf{U}_2'\mathbf{x}_i\|^2) \end{aligned}$$

donde el último término es no negativo. De esta manera calculamos

$$\begin{aligned} \sum_{i=1}^{k}\|\mathbf{A}'\mathbf{x}_i\|^2 &\leq k\sigma_k^2 + \sum_{i=1}^{k}(\|\mathbf{D}_1\mathbf{U}_1'\mathbf{x}_i\|^2 - \sigma_k^2\|\mathbf{U}_1'\mathbf{x}_i\|^2) \\ &= k\sigma_k^2 + \sum_{i=1}^{k}\sum_{j=1}^{k}(\sigma_j^2 - \sigma_k^2)\|\mathbf{u}'\mathbf{x}_i\|^2 \\ &= \sum_{j=1}^{k}(\sigma_k^2 + (\sigma_j^2 - \sigma_k^2)\sum_{i=1}^{k}\|\mathbf{u}'\mathbf{x}_i\|^2) \\ &\leq \sum_{j=1}^{k}(\sigma_k^2 + (\sigma_j^2 - \sigma_k^2)) \\ &= \sum_{j=1}^{k}\sigma_j^2 \end{aligned}$$

con lo que queda demostrado el teorema. □

Con esto queda demostrado que, tomando la descomposición en valores singulares de una matriz $\mathbf{A}$ y quedándonos solo con los k primeros valores, hallamos una aproximación de rango k que es la mejor posible en términos de la norma de Frobenius.

4.2. Aproximación y compleción de matrices

En esta sección nos centraremos en, dada una matriz $\mathbf{A}$, encontrar una matriz $\hat{\mathbf{A}}$, de estructura más *simple* (en algún sentido) que la aproxime de manera óptima (respecto a alguna medida de error). El objetivo de esto es entender la matriz $\mathbf{A}$ a través de la aproximación $\hat{\mathbf{A}}$ que tiene una forma más sencilla y el de rellenar valores faltantes que pudiera tener $\mathbf{A}$.

El problema de optimización se presenta de la siguiente manera:

$$\mathbf{A} = \underset{\mathbf{M}\in\mathbb{R}^{m\times n}}{\arg\min} \|\mathbf{A}-\mathbf{M}\|^2 \text{ sujeto a } \Phi(\mathbf{M}) \leq c, \tag{4.1}$$

donde la función Φ sirvce para establecer una restricción que hace que la matriz $\hat{\mathbf{A}}$ tenga muchos ceros (*sparse*). Como se menciona en [5], la manera en la que imponemos que la matriz $\hat{\mathbf{A}}$ tenga muchos ceros nos lleva a una amplia variedad de procedimientos y variaciones. En nuestro caso, utilizaremos la descomposición en valores singulares y el teorema de la aproximación.

Hemos visto que, utilizando los k primeros términos de la descomposición en valores singulares, se halla la aproximación de rango k que minimiza la diferencia entre la matriz $\mathbf{A}$ y su aproximación respecto de la norma de Frobenius. Así, si tomamos

$$\mathbf{M} = \mathbf{A}_k,\ \Phi(\mathbf{M}) = \text{rank}(\mathbf{M}) \text{ y } c = k \leq \text{rank}(\mathbf{A})$$

resolvemos el problema (4.1) para reducir el rango utilizando la descomposición en valores singulares. La función Φ asegura la presencia de muchos ceros en la matriz $\hat{\mathbf{A}} = \mathbf{A}_k$, ya que se igualan a cero todos los valores y vectores singulares que no son los k primeros. Evidentemente, el cálculo de esta $\hat{\mathbf{A}}$ sólo lo podemos llevar a cabo si conocemos todas las componentes de $\mathbf{A}$.

Abordamos ahora el problema de completar valores faltantes en la matriz $\mathbf{A}$. Partiendo del mismo escenario que en el problema de reducción de rango, lo que queremos encontrar es una matriz que mantenga los valores ya observados en $\mathbf{A}$ y que complete los faltantes de una manera adecuada. Para ello, como se trata de un problema mal planteado, debemos imponer alguna restricción adicional, como una restricción en el rango de la aproximación.

Más formalmente, supongamos que tenemos la matriz $\mathbf{A}$ de dimensiones $m \times n$ de la que tenemos observaciones en un subconjunto $\Omega \subset \{1,\dots,m\} \times \{1,\dots,n\}$. Esto es, los datos faltantes de $\mathbf{A}$ que queremos completar estan en $\Omega^{\perp}$. Lo que buscamos es la matriz $\hat{\mathbf{A}}$ de menor rango que interpole los valores observados de $\mathbf{A}$. Esto es, minimizar el rango de $\mathbf{M}$ sujeto a $m_{ij} = z_{ij}\ \forall (i,j) \in \Omega$.

Como se explica en [5], en contraposición a la versión con la matriz completa, el problema de minimización de rango para matrices incompletas es más complejo y alberga más dificultades computacionales. Se trata de un problema NP-complejo, que no puede ser resuelto incluso para matrices moderadamente grandes. Además, dependiendo de la cantidad de datos faltantes, forzar la interpola-

ción de todos ellos puede provocar sobreajuste, por lo que puede ser mejor permitir a la matriz cierto error en los datos ya observados. Por esta razón, sustituimos el problema original en el que se interpolan las componentes observadas de $\mathbf{A}$ por una versión menos restrictiva.

Consideramos entonces los siguientes problemas de optimización:

$$\text{mín } \text{rank}(\mathbf{M}) \text{ sujeto a } \sum_{(i,j)\in\Omega} (a_{ij} - m_{ij})^2 \leq \delta$$

o su equivalente

$$\min_{\text{rank}(\mathbf{M})\leq r} \sum_{(i,j)\in\Omega} (a_{ij} - m_{ij})^2.$$

La familia de soluciones generada al variar δ es la misma que al variar r. Sin embargo, los problemas de optimización propuestos son no convexos y las soluciones globales no son posibles de manera general. Aún así, existen diferentes heurísticas que permiten encontrar mínimos locales de manera efectiva. Por ejemplo, utilizar unos valores esperados para los datos faltantes como punto de partida e iterar con los que van resultando de cada aproximación. Esto es lo que hace el paquete de R, `SoftImpute`.

Con las condiciones ya mencionadas, supongamos $P_\Omega(\mathbf{A})$ la matriz con los valores de Ω mantenidos como en $\mathbf{A}$ y el resto igualadas a 0. Partiendo de $\mathbf{M}$, una estimación inicial de los valores faltantes de $\mathbf{A}$ el proceso iterativo que realiza `SoftImpute` es:

1.– $\mathbf{Z} = P_\Omega(\mathbf{A}) + P_{\Omega^\perp}(\mathbf{M})$

2.– Hallamos la descomposición en valores singulares de $\mathbf{Z}$, esto es, $\mathbf{Z} = \mathbf{UDV}'$

3.– Obtenemos la matriz $\mathbf{D}_k$. Esta es igual a la matriz $\mathbf{D}$, pero tiene igualadas a cero todos los valores singulares diferentes de los k primeros (reduciendo así el rango)

4.– Construimos $\mathbf{Z}_k = \mathbf{UD}_k\mathbf{V}'$

5.– Actualizamos la matriz $\mathbf{M} = \mathbf{Z}_k$

De esta manera, tenemos un proceso iterativo que se repetirá hasta la convergencia. En [5] ilustra como utilizando esta técnica junto con más restricciones y otros procedimientos alternativos se llega a resultados más satisfactorios. Nosotros utilizaremos este método, aplicándolo a datos reales en la siguiente sección.

4.3. El premio Netflix: Aplicación de la compleción de matrices

En [5], se expone como se utilizó la compleción de matrices en la resolución de un problema planteado por Netflix para crear un sistema de recomendación de contenido. El sistema ganador utilizaba, una versión mejorada del método de compleción de matrices expuesto. Lo que realizamos en esta sección es una adaptación a menor escala de este tipo de sistemas de recomendación.

Pulp fiction	Los juegos del hambre	Interestellar	El corredor del laberinto	El lobo de Wall Street	El viaje de Chihiro	Seven
4	3	5	2	4	5	5
4	2	5	2	5	4	4
5	3	5	3	4	4	4
3	3	5	2	4	5	4
4	3	1	1	3	4	N/A
4	3	N/A	N/A	3	4	4
5	2	5	2	4	5	5
4	3	5	3	4	3	3
5	3	4	1	3	4	4
4	2	5	4	N/A	5	N/A
4	4	4	4	4	4	4
5	4	5	N/A	5	5	5
5	1	4	N/A	4	N/A	4

Tabla 4.1: Resultados de la encuesta acerca de diferentes películas.

Para ello se realizó una encuesta a 13 personas para que dieran una valoración entre 1 y 5 de un total de 7 películas como se puede ver en la Tabla 4.1. En un comienzo, ya se disponían de 8 datos faltantes pues no todos los individuos habían visualizado todas las películas. Sin embargo, y para poder tener una perspectiva de como de preciso es el sistema, se eliminan más datos para tener un total de 16 datos faltantes. Para medir la efectividad del sistema se compararán los valores aproximados con los reales.

Dado que nuestra matriz $\mathbf{A}$ tiene dimensiones 13×7, la reducción de rango se puede hacer para $r = 1, \ldots, 6$. Aplicando `SoftImpute` a la matriz de datos faltantes con todos estos rangos se obtienen los resultados mostrados en la Tabla 4.2. Se considera que el algoritmo converge cuando la diferencia en la norma de Frobenius de dos aproximaciones consecutivas es menor que un umbral elegido, en este caso 10^{-5}

Podemos observar que la reducción de rango que mejor ha funcionado es la que reduce a rango 1

Rango máximo	Iteraciones hasta convergencia	Diferencia total	Diferencia media	Diferencia mínima	Diferencia máxima
1	6	8.505	1.063	0.288	1.891
2	12	17.157	2.145	0.254	6.112
3	26	19.073	2.384	0.052	7.626
4	24	29.593	3.699	0.248	6.662
5	50	27.593	3.449	1.234	6.491
6	15	30.165	3.771	0.661	6.542

Tabla 4.2: Resultados de la aplicación de `SoftImpute` a valoraciones de películas.

por varios motivos. En primer lugar, es la que converge de una manera más rápida, únicamente en 6 iteraciones del algoritmo. Además, es la que presenta menores diferencias con la matriz original.

Estas diferencias se han calculado utilizando los 8 valores que hemos establecido como faltantes, pero que habíamos observado. De esta manera, por cada uno de estos datos faltantes obtenemos una diferencia entre lo aproximado y el valor real. Cuando imponemos que el máximo rango sea 1, llegamos a la mínima diferencia con respecto a los datos originales. Además, observando las aproximaciones, vemos que en aquellas con rangos máximos mayores que 1 se incluyen valores negativos, lo que nos hace reforzar la idea de quedarnos con la de rango máximo 1. Así, este será el criterio que utilizaremos para completar nuestra matriz con únicamente 8 datos faltantes.

Cabe resaltar que el error cometido incluso en la mejor aproximación es significativo. Recordemos que las películas están valoradas con puntuaciones entre 1 y 5, y lo mejor que hemos conseguido un error medio de 1 punto en nuestras aproximaciones. Sin embargo, lo que se quería mostrar con este sistema a pequeña escala, en el que hay que tener en cuenta que los datos utilizados son pocos, es un procedimiento para hacer una compleción de matrices. Este mismo procedimiento se puede exportar a mayores escalas con matrices de tamaños más grandes, donde tiene más sentido hacer una reducción de rango.

5

Conclusiones

La conclusión principal a la que se ha llegado tras realizar este trabajo de fin de Grado es que la descomposición en valores singulares es una herramienta robusta con una fuerte base matemática en la que apoyarse. Esto provoca que se pueda utilizar en una gran variedad de ámbitos y situaciones en el marco de la Estadística Multivariante. Todo esto unido a la facilidad computacional que requiere su calculo y las grandes cantidades de datos que se recaban en la actualidad, hacen de la descomposición en valores singulares una técnica estadística muy potente y con multitud de aplicaciones.

Se ha visto como puede utilizarse para obtener las componentes principales de una matriz y, utilizando esto en la matriz de covarianzas muestral de un conjunto de datos, analizar como se relacionan las variables observadas. Se consiguió realizar una clasificación de diferentes tipos de aves atendiendo a su tamaño y forma, pudiendo aplicar además el teorema de Perron debido al carácter alométrico de los datos. Además, se pudo obtener una clasificación de células relacionadas con el cáncer de mama. Diferenciando en este caso entre células malignas e inocuas atendiendo también a su tamaño y forma.

También se pudo aplicar la descomposición en valores singulares a la obtención de las correlaciones canónicas de dos muestras multivariantes. En este caso la descomposición nos da las correlaciones canónicas de mayor a menor. Se pudo argumentar gracias a esto como el desarrollo y la calidad de la educación en un país esta íntimamente relacionado con los derechos y libertades que disfrutan las mujeres en estos países.

Por último, vimos como la descomposición en valores singulares nos puede ayudar en la aproximación de matrices reduciendo su rango o nos genera un algoritmo iterativo con el que podemos reconstruir datos faltantes en matrices de datos de una manera satisfactoria. Esto lo comprobamos aplicándolo a un sistema de recomendación y predicción de gustos a pequeña escala, pero que se puede aumentar manteniendo el mismo procedimiento.

Con estos ejemplos concretos, queda demostrada la variedad de aplicaciones que tiene la descomposición en valores singulares. Pudiendo ser aplicada en campos tan diversos como la Biología, la Medicina, la Sociología o el tratamiento masivo de datos. De esta manera, podemos posicionar a la descomposición en valores singulares como una técnica útil, potente y con fundamento a la par que sencilla de aplicar.

Bibliografía

[1] Cuadras, C.M.: *Nuevos métodos de análisis multivariante.* CMC Editions. (2010).

[2] Izenman, A.J.: *Modern Multivariate Statistical Techniques. Regression, Classification and Manifold Learning* Springer. (2008).

[3] Bapat & Raghavan: *Nonnegative Matrices and Applications.* Cambridge University Press (1997).

[4] Stewart, G.W.: On the early history of the singular value decomposition *SIAM Review 35, 4,551-566.* (1993).

[5] Hastie, T.; Tibshirani, R.; Wainwright, M.: *Statistical Learning with Sparsity. The Lasso and Generalizations.* Chapman and Hall/CRC (2016).

[6] Marsden, J.E.: *Elementary Classical Analysis* W. H. Freeman and Company. (1974).


UAM
UNIVERSIDAD AUTONOMA
DE MADRID