USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Análisis Estadístico Exploratorio de Datos Complejos

Ariadna Quiroga Doamo

Curso 2023/2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Traballo Fin de Grado

Análisis Estadístico Exploratorio de Datos Complejos

Ariadna Quiroga Doamo

Septiembre, 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística
Título: Análisis Estadístico Exploratorio de Datos Complejos
Breve descripción del contenido
Se pretende describir y desarrollar algunos procedimientos estadísticos exploratorios, diseñados para objetos estadísticos. Abordamos el estudio de los métodos más clásicos diseñados para datos vectoriales y su correspondiente extensión al caso de objetos estadísticos diversos.
Recomendaciones
Guión aproximado: 1) Análisis descriptivo exploratorio de datos vectoriales y datos con estructura compleja. 2) Análisis de Componentes Principales y sus aplicaciones en la visualización de datos complejos. 3) Técnicas de clasificación supervisada y no supervisada y sus extensiones recientes a estructuras de datos más complejos. 4) Aplicación e ilustración con datos reales.
Otras observaciones
Se pretende que el alumno dedique aproximadamente cuatro meses al estudio de las técnicas correspondientes a los apartados 1), 2) y 3) del guión. Un mes para el desarrollo de la aplicación de los datos reales.

Índice

Resumen

El principal objetivo de este trabajo es desarrollar técnicas estadísticas exploratorias, centrándose en el análisis de objetos estadísticos diversos. En el primer capítulo se presentan los conceptos básicos de ADOO y su terminología. En el segundo, se explican diversas técnicas de análisis descriptivo exploratorio, aplicados a datos complejos. El Análisis de Componentes Principales y su aplicación a la visualización se presenta en el capítulo tercero. En los capítulos cuarto y quinto se tratan técnicas de clasificación supervisada y no supervisada, respectivamente. En el último capítulo se introduce un problema de datos reales en el que se aplican algunas de las técnicas vistas anteriormente.

Abstract

This project aims to develop exploratory statistical techniques, focusing on the analysis of various statistical objects. The first chapter the basic concepts of OODA and its terminology are presented. In the second, various exploratory descriptive analysis techniques are explained, applied to complex data. Principal Component Analysis and its application to visualization is presented in the third chapter. The fourth and fifth chapters discuss supervised and unsupervised classification techniques, respectively. In the last chapter, a real data problem is introduced in which some of the techniques seen previously are applied.

Introducción

Actualmente las áreas de estadística, ciencias y análisis de datos han experimentado un crecimiento masivo, tanto de las capacidades computacionales como de conocimientos y comprensión, aplicables a campos tan diferentes como la inteligencia artificial o la predicción de desastres naturales. Para comprender la evolución de estas áreas es importante la noción de matriz de datos, ya que, en particular, el contexto de Big Data tiene distintos enfoques, desde áreas con baja dimensión-gran tamaño muestral (que es la base del pensamiento estadístico clásico, como son las encuestas por muestreo), pasando por alta dimensión-gran tamaño muestral (como los conjuntos de datos a escala de internet) hasta contextos de gran dimensión y bajo tamaño muestral (empleados en áreas de la genética y otros tipos de medidas ricas pero computacionalmente caras). Así, esta apremiante necesidad de analizar datos en una amplia gama de situaciones ha generado ideas nuevas y enfoques interesantes.

Sin embargo, un vistazo a estos desarrollos sugiere que la organización de los datos en una matriz puede estar imponiendo limitaciones. En particular, existe una creciente conciencia de que los desafíos que presenta Big Data están siendo eclipsados por los desafíos, quizá mucho mayores, de los *datos complejos*, que normalmente no se representan fácilmente como una matriz sin restricciones. El *Análisis de Datos Orientado a Objetos Complejos* (ADOO) proporciona un marco general útil para la consideración de muchos tipos de datos, como pueden ser formas (como por ejemplo, empleadas para la segmentación de imágenes médicas para encontrar un tratamiento adecuado), curvas, sonidos (por ejemplo, para el análisis del habla humana) o imágenes (por ejemplo, distinguir entre hombres y mujeres a partir de los rasgos faciales).

Frente a un conjunto de datos complejos, la manera de proceder será elegir el objeto de datos, ya que una elección u otra pueden resultar en conclusiones muy diferentes. Para encontrar alguna estructura poblacional en los datos conviene proceder con análisis exploratorio, se recurre a matrices de diagramas de dispersión, gráficos de fluctuación y mapas de calor entre otros. Para confirmar lo visto gráficamente, se aplican técnicas ya conocidas como son Análisis de Componentes Principales o ciertas técnicas de Clasificación Supervisada, como los métodos Kernel o Discriminación Ponderada por Distancia, y No Supervisada, como agrupaciones en K-medias o

jeráquicas, pero para el caso de datos complejos.

Capítulo 1

Introducción a ADOO

El *Análisis de Datos Orientado a Objetos* (ADOO) consiste en el análisis de datos complejos, que son aquellos que normalmente no se representan fácilmente como una matriz sin restricciones. Para comprender su aplicación y extensión, consideremos los *objetos de datos* como átomos de un análisis estadístico, donde átomo se entiende como partícula elemental. Así, en un curso básico de estadística, los átomos son números, y el objetivo es desarrollar métodos para comprender la variación en poblaciones de números. Un curso más avanzado, denominado *análisis multivariante*, generaliza los átomos, es decir, los objetos de datos, de números a vectores, y se estudian una serie de métodos para gestionar la incertidumbre en este contexto. A continuación podríamos pensar en funciones como objetos de datos, lo que se denomina *Análisis de Datos Funcionales* (ADF). Aquí, el objetivo consistiría en analizar la variación en una población de curvas. Finalmente, el ADOO proporciona el siguiente paso en términos de complejidad de los átomos de un análisis estadístico, ya que ahora consideramos una amplia gama de objetos más complicados, como caras, formas, sonidos, imágenes, etc. Cabe destacar que cada una de las áreas mencionadas contiene a la anterior como caso particular. Por ejemplo, el análisis multivariante es un caso particular de ADF, que a su vez es un caso particular de ADOO. Generalmente, podemos dividir en tres fases principales este tipo de análisis de datos:

- *Definición de Objetos:* Esta el la fase en la que se aborda la cuestión fundamental de cuál debe ser el objeto de datos.

- *Análisis Exploratorio:* Aquí el objetivo es encontrar una estructura poblacional en los datos, a menudo utilizando algún tipo de método de visualización.

- *Análisis Confirmatorio:* En esta fase nos centramos en validar los descubrimientos hechos en el apartado anterior.

1.1. Terminología ADOO

Una complicación que surge al trabajar con datos complejos, es que con frecuencia no es obvio cómo manejar los datos. Nuestra problema radica en la siguiente pregunta: *¿Cuál debe ser nuestro objeto de datos?* Esta elección depende de dos componentes importantes, la *determinación* de los objetos de datos y su *representación numérica*. La determinación depende del enfoque del análisis. Para ello es útil considerar la noción de *espacio de objetos*, que es el espacio conceptual que contiene a todos los potenciales objetos de datos. En este tipo de análisis debemos considerar simultáneamente la noción paralela de *espacio de características*, que contiene las representaciones numéricas prácticas, tales como *vectores característicos*.

Estos vectores característicos luego se agregan en una matriz de datos, que es una herramienta útil para organizar ideas del análisis de datos. Una de las dimensiones de esta matriz, normalmente representa los *casos*, es decir, los *elementos* de una muestra estadística, también denominados *observaciones* o *individuos*. La otra dimensión se utiliza para clasificar las *características* o descriptores numéricos de cada objeto de datos, también denominadas *variables*. En este trabajo, las variables van asociadas a las filas y los objetos de datos a las columnas. También emplearemos la letra n para denotar el tamaño muestral y d para la *dimensión* de los vectores de objetos de datos, obteniendo una matriz de datos de $d \times n$. A continuación, definiremos un concepto que será especialmente útil a la hora de definir el objeto de datos de nuestra muestra.

Definición 1.1. Un **modo de variación** de una muestra de objetos de datos es un conjunto de miembros potenciales del espacio de objetos que proporciona un resumen simple de un componente de la variación. Este resumen es unidimensional, es decir, razonablemente representable por un único parámetro real.

Ahora planteamos un ejemplo para entender el proceso de elección del objeto de datos y su aplicación en el análisis, así como la utilidad de los modos de variación.

Ejemplo 1.2. *Datos Mortalidad Española*: Este conjunto de datos, disponible en el *Human Mortality Database* de Wilmoth y Shkolnikov (2008), se centra en la mortalidad de los hombres en España desde 1908 hasta 2002. Las filas y columnas de la matriz de datos se encuentran indexados por años y edades. Así, cada entrada de dicha matriz contiene la probabilidad de que una persona de esa edad muera en ese año. Esto se calcula dividiendo el número de muertes entre el número de personas totales para ese par *año-edad*.

Ahora se presenta el problema de la elección del objeto de datos. En primer lugar, debemos tener en cuenta el rango de los curvas, que fluctúan en varios órdenes, por lo que una transformación logarítmica puede ser útil, en especial si aplicamos la transformación $\log_{10}$, como observamos en la Figura II.1. En el panel izquierdo se muestran las curvas originales y en el derecho después

de aplicarle la transformación logarítmica. En este último vemos que se revelan estructuras ocultas, especialmente en edades jóvenes. Por otro lado, existen dos formas de plantear la matriz de datos. Una de ellas considerando los objetos de datos como curvas de mortalidad en función de las edades e indexadas por años, o al revés. En este caso, nos quedaremos con la primera. Hecha ya esta elección, obtenemos un conjunto de $n = 95$ curvas cada una asociada a un año, desde 1908-2002, considerando la franja de edad de 0-98.

Fijándonos en las gráficas, vemos que nacer constituye una actividad de riesgo, disminuyendo a su vez hasta la adolescencia, donde vuelve a aumentar. Continúa aumentando paulatinamente hasta la vejez, donde la tasa de mortalidad es considerablemente alta. De todas formas no podemos apreciar una estructura temporal al tener las curvas los colores por defecto del programa. Así, en la Figura II.2, clasificamos las curvas temporalmente, donde los colores indican los años, ordenados como se indican en la gráfica. Ahora sí que apreciamos una mejora de la tasa de mortalidad, disminuyendo sustancialmente entre los jóvenes.

Normalmente, en los conjuntos de datos ADOO, el análisis prosigue descomponiendo la variación presente en las curvas, a través del *Análisis de Componentes Principales* (Capítulos 2 y 3). Será especialmente útil a la hora de ver cómo se relacionan los objetos de datos entre sí. Podemos interpretarlo como si los datos estuviesen en una nube puntos de un espacio $d = 99$-dimensional, dónde las proyecciones de baja dimensión muestran estas relaciones (agrupaciones de los datos). A menudo se comienza con un *centrado medio*, que consiste en trasladar la nube de puntos de forma que estén centrados en el origen, como se ilustra en la Figura II.3.

En el panel izquierdo vemos la curva media, es decir, la media de las curvas de la Figura II.2. El derecho muestra las medias residuales, que se calculan restando la media a cada curva, conservando el color de referencia del año. La curva media contiene muchas de las características de los datos originales, en particular, los relacionados con la edad. Así, una mortalidad baja para los jóvenes, junto con una cada vez mayor para los mayores, son propiedades de la media. Como se trata de propiedades poblacionales, éstas no se ven reflejadas en la gráfica de los residuos. En la curva media vemos unas desviaciones que parecen aleatorias, sin embargo, se trata de un redondeo de edad causado por el deficiente registro de nacimientos, lo que repercute en un desconocimiento de la edad de muerte. También vemos una mejoría clara de la mortalidad a lo largo del tiempo, en la que los jóvenes son los más beneficiados.

El ACP consiste en la descomposición de los datos centrados medios en *modos de variación* útiles. El primero de esos modos será la variación que revela la primera componente principal (Figura 1.1), en el que cada curva (objeto de datos) es un punto. Dichos modos se obtienen a partir de las direcciones ortogonales de máxima variación de la nube de puntos. Así, la primera *dirección* CP es el vector unitario, basado en la media muestral, que maximiza la varianza de los datos proyectados en ese vector. Se define como el primer vector propio de la matriz de

covarianzas muestrales (definida en (I.5)). Las entradas del vector indican cómo se relaciona con las variables (es decir, las características) y se denominan *pesos*. Su representación (en *gráfica de modos variación*) se muestra en el panel izquierdo de la Figura 1.1 donde el eje horizontal representa a las variables (edad), y las curvas son múltiplos del vector propio. Son las columnas de la matriz de rango 1, producto del vector columna de los pesos por el vector fila de las puntuaciones (*scores*), que, a su vez, son los coeficientes de proyección de cada objeto de datos sobre el vector propio. Esta representación del CP1 destaca el modo de variación dominante, en el que se refleja la importante mejora general de la mortalidad. A lo largo del estudio, mejora el registro de nacimientos y defunciones, lo que se refleja en una disminución del redondeo. Esto se muestra parcialmente en la media de la Figura II.3.

La gráfica de distribución de las puntuaciones (der. Figura 1.1), se emplea frecuentemente para mostrar información sobre cómo los objetos de datos se relacionan entre sí. Cada círculo representa una puntuación, usando la misma asociación año-color (magenta 1908, rojo 2002). La coordenada horizontal representa las puntuaciones y la vertical el orden del conjunto de datos. La tendencia hacia la izquierda implica nuevamente la mejora de la mortalidad a lo largo del tiempo. La curva negra es una *estimación de la densidad del núcleo*, donde el eje vertical muestra la altura de la curva. Hay una concentración mayor de puntuaciones en las regiones altas y bajas, lo que implica que el cambio de una mortalidad más alta a una más baja fue muy rápido. Tanto en la pandemia de gripe de 1918 (círculo violeta de la derecha) como en la Guerra Civil Española (derecha, azul claro) vemos un aumento de mortalidad.

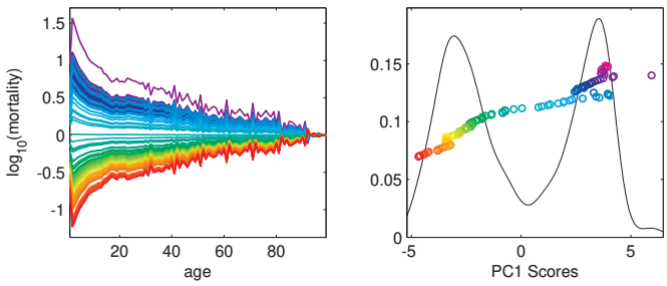

Figura 1.1: *Gráfica del primer modo de variación CP1 (izq.) y la gráfica de la distribución de las puntuaciones (der.). Los colores de las curvas se han organizado en función de los años (magenta*-1908*, rojo*-2002*). Se aprecia una mejora de la tasa de mortalidad, además bastante rápida, salvo el año de la Gripe Española en 1918, que se corresponde con el atípico morado en ambos paneles.*

Ahora planteamos el segundo modo de variación como el segundo CP (Figura 1.2), que es la dirección de la segunda variación más fuerte (ortogonal a la primera). Esta se calcula como el segundo vector propio de la matriz de covarianzas muestrales. La gráfica del modo de variación

CP2 (izq.) destaca la diferencia entre el rango de 20-45, con la unión de los jóvenes y los viejos. El patrón de color es mucho más difícil de interpretar, pero se aprecia bien en la gráfica de distribución de puntuaciones (der.). Otra interpretación que obtenemos, es que los hombres de 20-45 fueron los que más sufrieron los efectos de la guerra y la pandemia, ya que la mortalidad es bastante mayor que el del resto. También lo es entre los 1960 y 1980 debido al creciente uso de automóviles, y por tanto, de accidentes. Esta tendencia disminuye en los años siguientes debido a la mejora de la seguridad automovilística y de las carreteras.

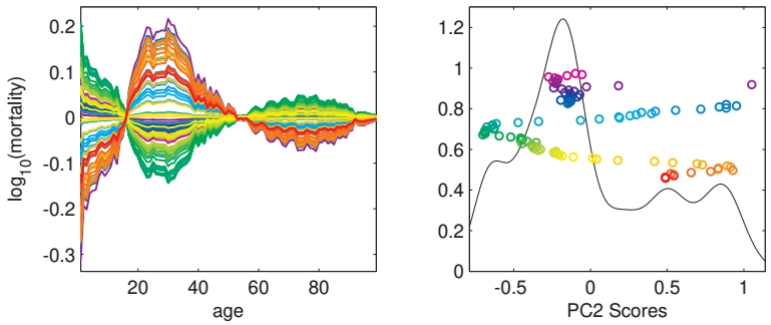

Figura 1.2: *Gráfica del segundo modo de variación CP2 (izq.) y la gráfica de la distribución de las puntuaciones (der.), usando el mismo formato que la Figura 1.1. Se muestra un contraste entre los hombres de* 20-45 *con el resto. En el panel derecho se observan los efectos de la pandemia de gripe.*

La Figura II.4 muestra un diagrama de dispersión de la distribución bivariante de las puntuaciones de CP1 y CP2, que proporciona un resumen de la mayor parte de la estructura de este conjunto de datos. Así, la distribución unidimensional de las puntuaciones CP1 (der. Figura 1.1) se encuentra en el eje horizontal y la de CP2 (der. Figura 1.2) en el eje vertical. Esta es la proyección bidimensional de los datos en el plano con máxima variación. Además, los objetos de datos están unidos por una recta en orden, lo que facilita la interpretación del paso del tiempo. Al igual que en las otras gráficas, se aprecia una mejora clara de la mortalidad, a excepción de los años de la pandemia de gripe y la guerra. En el eje vertical, también se aprecia el contraste entre las personas de 20 a 45 años y el resto, destacando los efectos mencionados anteriormente (gripe, accidentes, guerra, etc.). □

Para este ejemplo, los dos primeros modos de variación son los más interesantes. Dependiendo del conjunto de datos, se obtendrán más modos de variación o no. Si se tienen más de dos CP, resulta útil obtener una matriz de estos diagramas de dispersión, en el que la diagonal muestre una distribución unidimensional, como la combinación de *gráficos de fluctuación*, que se estudian en la Sección 2.1.

Capítulo 2

Análisis descriptivo exploratorio de datos con estructura compleja

El análisis descriptivo exploratorio de datos es un método que consiste en analizar e investigar conjuntos de datos y resumir sus principales características mediante métodos de visualización de datos. Su principal objetivo es descubrir patrones, detectar anomalías, probar una hipótesis o comprobar un supuesto. Además, permite conocer mejor las variables del conjunto de datos y las relaciones entre ellas, así como determinar si las técnicas estadísticas que se están considerando para el análisis de datos son apropiadas. Así, en este Capítulo nos centraremos en la fase dos (*Análisis Exploratorio*) que mencionamos anteriormente, combinando lo ya estudiado (*gráfica de distribución de puntuaciones*) junto con técnicas gráficas (*gráficas de fluctuación, mapas de calor y gráficas de distribuciones marginales*).

2.1. Análisis exploratorio: Descubrimiento de estructura en los datos

La utilidad de la visualización de datos no solo se centra en el análisis exploratorio de datos y en entender cómo están relacionados entre sí los objetos de datos, sino que también se emplea frecuentemente para la elección efectiva de estos, además de su comprobación en el análisis.

Como veremos en el Capítulo 3, el ACP (*Análisis de Componentes Principales*) es una herramienta efectiva para estudiar los modos de variación. Estos brindan información sobre cómo se relacionan los objetos de datos entre sí, como la exploración de potenciales *clusters*, o grupos. Ahora, estudiaremos un ejemplo para entender el concepto de descomposición en modos de variación.

Ejemplo 2.1. *Parábolas Inclinadas.* Tendremos en cuenta un conjunto de datos de tamaño $n = 50$ en forma de curva (estamos en el marco de ADF), que se muestran en el panel de arriba a la izquierda (Figura II.5). Dichas curvas han sido simuladas de forma aproximada a una parábola, pero se han incluido variaciones de distintos tipos. En realidad, cada parábola es una gráfica de coordenadas paralelas de una colección de vectores de dimensión 10.

En este caso, no podemos representar de forma explícita el espacio característico, ya que está contenido en $\mathbb{R}^{10}$. Aún así, es importante no olvidar este detalle cuando observemos el espacio de objetos correspondiente (es decir, las curvas). El panel de arriba en el centro muestra la media muestral de las curvas (es decir, la media de las curvas del panel izquierdo), mientras que el panel de la derecha muestra los residuos de la media, que son una visualización de las curvas que corresponden al desplazamiento de la nube de puntos en $\mathbb{R}^{10}$ de forma que tenga una media centrada en el origen. Gracias a estos residuos vemos que la forma parabólica de las curvas está determinada totalmente por la media, y no por la variabilidad respecto de la media.

Las siguientes tres filas de la figura muestran los resultados de la descomposición en modos de variación utilizando ACP. El primer modo de variación se muestra en el panel de la izquierda de la segunda fila de la figura. Dicho modo se calcula obteniendo la dirección de mayor variación proyectada en el espacio de características (que recordemos es $\mathbb{R}^{10}$), proyectando cada curva de los residuos de la media en dicha dirección, para luego mostrar el conjunto de curvas resultante. Cabe destacar que este conjunto de curvas son las columnas de una matriz de rango 1, por lo que todas son múltiplos de la misma curva (que es la curva representación del vector director en el espacio de características). Por tanto, el primer modo de variación es un desplazamiento vertical, como observamos en la nube de curvas de los datos originales.

El panel de la derecha de la segunda fila muestra la distribución de los coeficientes de proyección, es decir, las puntuaciones. Cada puntuación se representa con un círculo, que tendría el mismo color que el de la curva de la que proviene, y la curva negra sería un histograma suavizado. Como las parábolas no están ordenadas, la altura de las puntuaciones en esta gráfica se podría considerar aleatoria (lo que se conocería como *jitter plot*, gráfico de fluctuación). El panel del centro de la segunda fila, muestra las correspondientes curvas residuales CP1 (1ª componente principal), cada una de las cuales es el residuo centrado menos su proyección CP1. También se puede interpretar como las proyecciones de los residuos de la media sobre el hiperplano ortogonal a la dirección CP1.

La tercera fila muestra el segundo modo de variación. El panel de la izquierda es la representación del espacio de objetos de los residuos de CP1 en la segunda dirección de componentes principales en $\mathbb{R}^{10}$. Este modo de variación es mucho más difícil de ver tanto en el conjunto de datos original como en los residuos de la media, demostrando la capacidad del Análisis de Componentes Principales de encontrar modos de variación que no son visibles a primera vista

en los datos originales. Las puntuaciones CP2, en el panel de la derecha, muestran mucha menos variación que en el primer modo de variación. Vemos que esta variación disminuye con el tercer modo de variación, como observamos en la distribución de los coeficientes de proyección.

Otro punto de vista útil proviene de la suma de los cuadrados de las proyecciones. Las proyecciones CP1, que explican la mayor parte de la varianza de los datos, se cuantifican como la suma de los cuadrados de las proyecciones CP1. Estas representan el 86 % de la suma de los residuos de la media al cuadrado. Análogamente, las proyecciones CP2 representan el 10,4 %, mientras que la suma de las proyecciones CP restantes representarían solo el 3,6 %, confirmando que la variación restante es muy pequeña. En particular, los datos originales pueden obtenerse como una suma de la media muestral de los datos, los tres modos de variación y los residuos. □

Como veremos en el Capítulo 3, los vectores CP usados para la descomposición en este conjunto de datos, se pueden calcular a partir de los autovalores y autovectores de la matriz de varianzas-covarianzas, o equivalentemente usando la *Descomposición en Valores Singulares* de la matriz de residuos.

A pesar de que los ejemplos de juguete son importantes para entender conceptos, también lo es considerar conjuntos de datos de la vida real, como el siguiente:

Ejemplo 2.2. *Caso de estudio: Datos Cáncer de Pulmón.* Para este caso, empleamos el conjunto de curvas, que reciben el nombre de datos *Lung Cancer RNAseq*, procedentes de un estudio del cáncer de pulmón *The Cancer Genome Atlas (TCGA)*, Weinstein et al. (2013). Este estudio se centra en el gen CDKN2A, que está involucrado en la aparición de muchos tipos de cáncer.

En la Figura II.6, el eje horizontal representa la región del cromosoma que se emplea para producir el ARN medido (utilizando la tecnología descrita en Wang et al. (2009)). Para cada una de las $d = 1709$ localizaciones, el eje vertical muestra los recuentos (en escala $\log_{10}(\cdot + 1)$) de moléculas de ARN amplificadas que coinciden con el cromosoma en esa localización. Hay $n = 180$ curvas. Aquí hemos empleado la transformación de $\log_{10}$ porque los recuentos oscilan en magnitudes de 3 órdenes. En este caso, un recuento pequeño, por ejemplo, 1 o 2 ocupan la mayor parte inferior de la gráfica, ya que $\log_{10}(1+1) \approx 0{,}301$ y $\log_{10}(2+1) \approx 0{,}477$. Las curvas parecen algo más bajas en algunos intervalos, debido a que estas regiones de código no están en una región continua del cromosoma, sino que están separadas en intervalos, que reciben el nombre de *exones*. En el eje horizontal se utilizó la unión de estas regiones exónicas.

En la gráfica apreciamos mucha variación en las curvas, lo que complica encontrar algún patrón en los datos. Por tanto, construimos un diagrama de dispersión de puntuaciones ACP para tratar de entender mejor las relaciones entre los objetos de datos (Figura II.7). Como observamos en la figura, las distribuciones unidimensionales de puntuaciones, que serían los paneles correspondientes a la diagonal, muestran una estructura multimodal, como podemos

comprobar gracias a los histogramas suavizados. Pero, los diagramas de dispersión fuera de la diagonal, descubren otra agrupación de datos más, que quedaba oculta ya que los dos grupos superiores se combinan en las puntuaciones de CP1, mientras que los dos de la derecha en las de CP2. Solo se muestran las dos primeras componentes porque son las que más variación explican (82 % la primera, y 8 % la segunda, las siguientes ya explicarían un porcentaje menor del 5 %). También conviene mencionar otro patrón interesante, ya que parece que los puntos que del segundo y cuarto cuadrante forman una recta.

Para llegar a comprender las conductas de estos grupos emplearemos una técnica llamada *brushing* (Becker and Cleveland (1987)), que consiste en usar colores para realizar un seguimiento de los grupos de datos en las gráficas, como observamos en la Figura II.8, donde cada grupo se ha resaltado con un color. Además, observamos que los paneles correspondientes a las distribuciones unidimensionales se han reforzado con *subdensidades*, donde el área bajo cada una de ellas será proporcional al tamaño del grupo, y la suma de dichas subdensidades será 1.

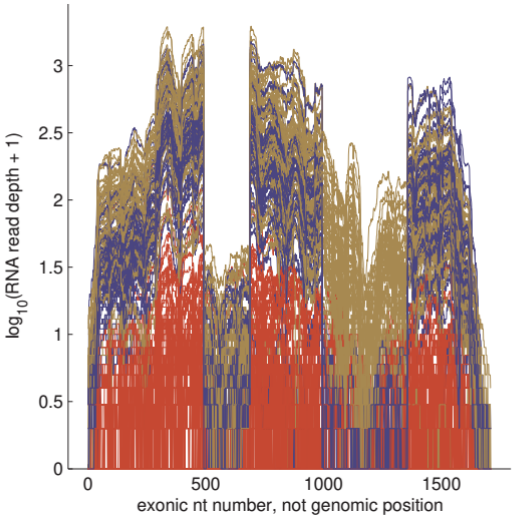

Figura 2.1: Mismas curvas de datos que en la Figura II.6, aplicando la técnica de *brushing* y usando los mismos colores que en la Figura II.8.

Aplicamos esta técnica a las curvas originales, que resulta en la Figura 2.1. Como observamos, las curvas rojas están en la parte inferior, lo que denota un nivel de expresión mucho más bajo de moléculas de ARN (debido a la transformación $\log_{10}(\cdot+1)$). Para la mayor parte de los exones las curvas azules y marrones tienen valores de expresión altos. Pero ocurre un hecho muy interesante con un exón de la derecha, donde los casos marrones tienen valores altos mientras que los azules no tienen expresión. Este fenómeno recibe el nombre de empalme alternativo (*alternate splicing*), donde individuos distintos producen diferentes versiones de ARNm a partir de la misma región cromosómica. Este descubrimiento fue muy importante para el desarrollo de nuevos tratamientos

contra el cáncer, porque este fenómeno se puede tratar con medicamentos adaptados. □

Una dificultad que surge es que la mayor parte de los métodos automáticos de agrupación encuentran siempre muchos grupos, pero estos no tienen por qué ser importantes. Otro aspecto importante a tener en cuenta en la representación de los datos es la *escala* y los problemas de *normalización*, que forman parte de lo que se denomina *preprocesamiento de ADOO*. Como vimos en estos ejemplos, ACP es una herramienta importante con la que descubrimos estructuras de los datos, pero esta técnica pierde efectividad en situaciones donde las variables tienen distintas escalas, ignorando las de menor escala, y el problema se acentúa cuando las variables tienen unidades distintas.

2.2. Preprocesamiento ADOO

Esta sección se centra en la descripción de algunas formas útiles de comprender problemas de datos ocultos y algunas soluciones que se extienden de manera razonable a conjuntos de datos más grandes, incluso aquellas con muchas variables. Un término general que engloba todos estos problemas es la *procedencia de los datos*, que incluyen información sobre las fuentes y los procesos que llevaron a la creación y la representación de los datos.

2.2.1. Visualización de distribuciones marginales

Un importante paso de preprocesamiento de ADOO, que ayuda a evitar problemas, es la visualización de distribuciones marginales. Como, por lo general, tratamos con un gran número de variables, se recomienda emplear *gráficas de distribuciones marginales*, seleccionando un subconjunto *representativo* de las variables para analizarlas, recurriendo a un resumen estadístico unidimensional, como por ejemplo la media de las variables.

Ejemplo 2.3. *Caso de estudio: Datos Mortalidad Española.* En la Figura II.9 mostramos una gráfica de distribuciones marginales para el conjunto de datos de *Mortalidad Española*, que estudiamos en la Sección 1.1. Recordemos que consiste en la matriz de datos donde las columnas (objetos de datos) se indexan por años $(1908 - 2002)$ y las filas (variables) se corresponden a las edades, donde cada entrada de la matriz es la probabilidad de que un hombre muera con esa edad y en el año dado.

El panel superior izquierdo de la Figura II.9 muestra la mortalidad media por edades, ordenadas de forma creciente. La primera mitad de estas medias parecen muy pequeñas, con valores mucho mayores para la segunda. Esto es consistente con la impresión visual de la Figura II.1, de que alrededor de la mitad de las edades tienen mortalidades varios órdenes menores que el resto.

Este conjunto de variables medias ordenadas es la clave para encontrar un conjunto *representativo* de variables (edades, en este caso).

Una noción de representativo, es fijarse en los subconjuntos igualmente espaciados (entre las edades medias ordenadas), indicado por las rectas verticales discontinuas. Los paneles restantes muestran las 8 distribuciones marginales de las edades correspondientes a esas 8 rectas. En particular, los círculos se corresponden con los años (es decir, los objetos de datos), con los colores representando los años (magenta-1908, rojo-2002). Como coordenada horizontal empleamos la mortalidad, mientras que la vertical indica el orden en el conjunto de datos, donde la curva negra es una estimación de densidad kernel de los datos.

Si observamos las primeras edades, 11 y 19, tienen mortalidades muy pequeñas, del orden de 10^{-3}. Las de la fila del medio, 32, 40 y 59, del orden de 10^{-2}. En la fila de abajo, todas las edades tienen mortalidades grandes. En general, los conjuntos de datos que tienen variables con escalas que difieren varios órdenes pueden resultar problemáticos para muchas formas de análisis estadístico. Además, estos gráficos de distribución marginal muestran problemas de asimetría en la mayoría de los gráficos además de algún dato atípico, correspondiente al año 1918 (pandemia de Gripe Española).

Para abordar el problema de los órdenes de magnitud de las variables, aplicamos una transformación logarítmica para ajustar los datos. Los resultados se aprecian en la Figura II.10, donde vemos que, aunque existe todavía una variación natural de la media, esta ya no ocupa varios órdenes de magnitud. Esta es la razón por la que la impresión visual de variación en el panel derecho de la Figura II.1 es mucho más reveladora que en el izquierdo. Otro resultado de la transformación es que las distribuciones que antes presentaban asimetría se han transformado en distribuciones principalmente bimodales, además de que el impacto del dato atípico del año 1918 ha disminuido sustancialmente. □

Encontrar un conjunto representativo de variables mediante la clasificación según sus medias es una forma muy eficaz de comprender aspectos críticos sobre este conjunto de datos. Otros resúmenes estadísticos también pueden resaltar nociones diferentes y muy reveladoras de nuestras variables representativas.

2.2.2. Estandarización: Escalado Lineal

Como ya mencionamos, es necesario escalar cuando las distintas variables no sean *conmensurables*, es decir, cuando tengan distintas unidades, y también cuando tengan distintos órdenes de magnitud. Una solución es estandarizar cada variable, proceso conocido como *pre-whitening*, que consiste en restar su media y dividir entre su desviación típica. En ACP, esta operación implica

sustituir la matriz de covarianzas muestral por la matriz de correlaciones muestral.

Ejemplo 2.4. *Curvas de Dos Escalas*: El conjunto de datos *Curvas de Dos Escalas* es un ejemplo de juguete diseñado para explicar la cuestión de escalado, que se muestra en Figura II.11. Las $n = 200$ curvas de $d = 100$ dimensiones aparecen en el panel superior izquierdo. Observamos que las primeras 20 variables presentan mayor variación que las 80 restantes. Como la media es prácticamente 0, los residuos de la media serán iguales a los datos originales. El primer modo de variación CP, que sería la segunda fila, claramente engloba esas 20 variables, reflejando como fluctúan arriba y abajo entre ellas. De forma similar, el segundo modo de variación, es un contraste entre las 10 primeras variables y las 10 primeras variables de escala menor. Obviamente este será ortogonal al primer modo de variación y refleja a su vez mucha menos variación, que irá disminuyendo con los modos de variación siguientes.

Ahora, escalamos los datos, es decir, restamos la media y dividimos entre la desviación típica, y repetimos el análisis, que representamos en la Figura II.12. Observando la gráfica con los datos escalados, vemos que las últimas 80 variables tienen la misma importancia que las otras antes de estandarizarlas. Los modos de variación también serán distintos, ya que se enfocarán únicamente en estas últimas variables. Por el contrario, las 20 primeras variables solo aparecerán en el tercer modo de variación. Como ahora las magnitudes de los dos tipos de variables son comparables entre sí, las últimas 80 variables predominan en el análisis, además de que es un conjunto mayor, por lo que contribuirá más a la variación total (fueron simuladas para ser independientes por tanto la variación esencialmente se mueve en direcciones ortogonales).

En la Tabla II.20 se exponen los porcentajes de los cuadrados de las sumas de cada modo de variación respecto de los residuos de la media. Como las curvas de los datos sin estandarizar tenían toda la variación hacia la izquierda, no es sorprendente que parte del rango impulse dos componentes CP de gran tamaño, explicando casi toda la variación de los datos, como vemos en la primera fila. La fila inferior de la tabla muestra una variación más extendida, consistente con la visión de las gráficas que se representaron al estandarizar los datos. En particular, la tabla nos indica que la variación de la derecha es la dominante ya que contiene el 80 % (suma de los dos primeros modos de variación) de la variación total. □

2.2.3. Transformación: Escalado No Lineal

Aunque la magnitud de las variables es una consideración importante en ADOO, de manera similar, la forma de las distribuciones marginales también puede serlo, como se ve en los datos de *Mortalidad Española* en las Figuras II.9 y II.10. Como vimos también para dicho conjuntos de datos, una transformación logarítmica de cada variable es muy útil, ya que tiende a reducir la influencia de los datos que son varios órdenes de magnitud, mayores o menores, que las otras.

Una variación de la transformación logarítmica es la *transformación de logaritmo desplazado* de la forma $\log(\cdot - c)$, donde los datos se desplazan una cantidad c antes de la aplicación del logaritmo. Esta transformación es muy útil sobre todo para datos que toman valores negativos o 0, y en el caso de $c < 0$ es también útil para datos con una relativa asimetría.

Otro tipo de transformaciones ampliamente utilizada es la familia Box-Cox

$$f(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log(x) & \text{for } \lambda = 0 \end{cases}$$

propuesto por Box y Cox (1964). Esencialmente es una familia de potencias unida a una transformación lineal. Debemos tener en cuenta que un cálculo del límite $\lambda \to 0$ muestra que esta es una función continua del parámetro de λ.

2.2.4. Registro: Alineamiento

En algunos tipos de ADOO pueden surgir *problemas de alineación* o registro. Para demostrar este punto, tomemos un ejemplo de juguete de ADF, que se muestra en la Figura II.13. Los datos originales se muestran en el panel izquierdo, donde vemos que cada curva tiene dos picos, pero existe una variación importante tanto en la ubicación como en la altura de estos. Con el objetivo de resaltar la variación de la amplitud, las curvas están codificadas con un esquema de colores utilizando la altura del pico de la izquierda siendo el magenta, el más alto, y el rojo, el más bajo. Las diferencias en las ubicaciones de los picos crea problemas importantes, debido a que se ignora la fuerte variación de fase, como podemos observar en la curva media (curva discontinua negra), que no es para nada representativa de la población de curvas. En particular, sus picos son mucho más bajos que cualquiera del conjunto de datos y además el pico izquierdo aparece como dos modos.

El panel derecho, muestra los resultados de aplicar un método de registro de curvas (en concreto el de Fisher-Rao). Las curvas tienen las mismas alturas que en el panel izquierdo solo que, ahora, el eje horizontal del panel derecho se ha deformado para que queden perfectamente alineadas. La curva media ahora sí que se se ajusta a la población de curvas. En particular, el ACP de las curvas alineadas proporciona una representación mucho más intuitiva y compacta (empleando solo un modo de variación) del conjuntos de datos.

2.3. Visualización de Datos

En esta sección estudiaremos la visualización ADOO más a fondo. Aquí nos centraremos en objetos de datos que tienen una representación del espacio de características Euclidiana, donde

el conjunto de datos se resume, convenientemente, en una matriz.

2.3.1. Mapas de Calor para Matrices de Datos

El objetivo de un mapa de calor es representar la estructura de una matriz de datos construyendo una imagen a partir de una "mancha" de color, asociada a cada entrada de la matriz, en una cuadrícula rectangular. Así, los patrones que se observan en los colores nos pueden dar perspectivas útiles sobre el conjunto de datos representado por la matriz.

El orden de los elementos de la matriz de datos tiene una influencia muy importante a la hora de visualizar el mapa de calor. Lo ilustraremos con el siguiente ejemplo, basado en el conjunto de datos *Dos Grupos*. En la Figura II.14, vemos tres paneles, basados en dicho conjunto de datos. En la gráfica de la izquierda no apreciamos ningún patrón ni en las filas ni en las columnas. En la del centro, hemos aplicado un algoritmo de agrupamiento a las columnas, por lo que ahora podemos apreciar una clara prominencia de colores grises hacia la derecha. Y, finalmente, en la de la derecha, hemos aplicado el algoritmo anterior también a las filas, lo que revela un patrón que relaciona la mitad superior izquierda e inferior derecha de la gráfica. El contraste entre los lados izquierdo y derecho revela dos grupos en el conjunto de datos (por columnas) con diferente estructura media de grupos en las variables reordenadas.

Otra cuestión importante para la visualización de mapas de calor es la escala, como planteamos en las siguientes gráficas, ambas representadas a partir de la misma matriz de datos. También veremos el impacto de la elección de la escala de colores. La Figura II.15 muestra cómo la elección típica de una escala de colores equiespaciada puede oscurecer estructuras importantes de los datos, como observamos en este caso, dónde lo único visible en el mapa de calor es un único punto blanco. Estudiamos las causas de este comportamiento a partir de la distribución de los valores de la matriz, situada en el panel derecho. En particular, las entradas de la matriz, que se encuentran en el intervalo $[0, 20]$, se muestran como $50 \times 50 = 2500$ círculos, donde la coordenada horizontal es la entrada de la matriz y la vertical es el orden en la versión vectorizada de los datos donde las columnas de la matriz se han concatenado en un única columna. Tanto el mapa de calor como la distribución de los círculos muestran un conjunto de datos asimétrico. Las líneas de puntos verticales del panel derecho representan los límites de las 20 regiones de escala de grises empleadas en dicho mapa de calor. Así, tres cuartas partes de la gráfica están vacías, lo que se refleja en la ausencia de píxeles grises o blancos en el mapa de calor. Pero, la forma de distribución de los círculos sugiere que puede haber una estructura oculta en el mapa de calor, esencialmente porque todos los puntos se encuentran en la región negra.

Una forma de hacer un mejor uso de la escala de grises es transformar la distribución de las entradas de la matriz para que los datos sean menos sesgados. Por tanto, realizamos una

transformación logarítmica, que es la que mejor se ajusta a los datos y observamos su efecto en la Figura II.16. Ahora, en el mapa de calor se aprecian otros tres círculos grises más, aunque el impacto de la transformación se observa en la distribución de las entradas de la matriz. El círculo gris más brillante del mapa de calor se corresponde con el pico más alto y disperso de los valores de la matriz. Los otros círculos menos brillantes también serían picos, pero con la mitad de altura. Aún así, podemos apreciar otra estructura adicional en los datos, como una cuadrícula, lo que en el panel derecho se observa como una banda gruesa de círculos en el lado izquierdo.

Podemos seguir mejorando la apariencia del mapa de calor, aplicando alguna técnica de escalado, ya que la mitad del rango está dedicada al pico más alto de la matriz, y por tanto a la mitad más brillante de la escala de grises. Entonces, aplicaremos lo que se denomina como *escalado cuantil* o *igualación del histograma*. Esta técnica se asegura de que en cada nivel de gris se utilice aproximadamente el mismo número de píxeles, colocando las líneas verticales en el panel de la derecha en cuantiles equiespaciados de la distribución de valores de la matriz. Aunque ahora se aprecie el patrón del fondo en el mapa de calor (Figura II.17), revelar estos patrones tiene un precio, ya que existe mucho menos contraste entre las alturas de los distintos picos.

Debemos tener en cuenta, que al ser el logaritmo una función monótona, el escalado cuantil dará los mismos resultados tanto para los datos originales como para los transformados, solo que la distribución de los valores de la matriz será más sencilla de interpretar para los datos transformados.

Consideremos ahora otro ejemplo basado en el conjunto de datos *Dos Clases Gaussianas*, cuyo objetivo es demostrar que el mapa de calor puede pasar por alto importantes estructuras en los datos, a pesar de proporcionar perspectivas muy útiles. Dicho conjunto consiste en una matriz de $n = 200$ vectores de objetos de datos de longitud $d = 20000$. Para mejorar la estructura del mapa de calor de la Figura II.18, aplicamos un método de agrupamiento jerárquico usando enlace promedio y distancia Euclidiana en filas y columnas de la matriz. El problema es que al tener un tamaño tan grande y dado que la capacidad de visualización de los píxeles es de muy pocos de miles, el mapa de calor será difícilmente visible. Por eso, en la Figura II.18, solo se representa un subconjunto de los datos consistente en las 200 primeras filas y columnas de la matriz.

En este conjunto habrá una estructura presente, ya que los $n_1 = 100$ primeras columnas están generadas por una $\mathcal{N}(0{,}04, 1)$, mientras que las restantes por una $\mathcal{N}(-0{,}04, 1)$, ambas independientemente distribuidas e independientes entre sí. Como el mapa de calor está dominado por el ruido, planteamos otro tipo de visualización, como una matriz de diagramas de dispersión ACP (Figura II.19). Aquí se distinguen perfectamente los dos grupos. Sin embargo, debemos ser escépticos en cuanto a si estos aspectos, visualmente aparentes, de un diagrama de dispersión representan una verdadera estructura subyacente o son simplemente artefactos de muestreo

irreproducibles, comprobándolo con otras técnicas analíticas.

En resumen, los mapas de calor son técnicas muy poderosas que tienen un sólido historial a la hora de encontrar grupos en una única visualización, pero también pueden pasar por alto patrones en situaciones donde hay mucho ruido. Además debemos tener en cuenta que los grupos aparentes en un mapa de calor pueden no representar descubrimientos reproducibles.

2.3.2. Matrices de Gráficas de Curvas y Modos de Variación

Las gráficas de curvas revelan aspectos bastante diversos de la variación en una muestra de objetos de datos. El valor de estas gráficas recae sobre todo en la visualización de modos de variación. Algunos ejemplos que ya hemos visto son:

- *Figura II.1 Datos Mortalidad Española:* Estas visualizaciones demuestran el valor de la transformación logarítmica, en términos de hacer que gran parte de la variación de los datos esté disponible para el análisis.

- *Figura II.3 Datos Mortalidad Española:* Esta gráfica se centra en la variación de la media, para la cual muchos de los impactos de la edad sobre la mortalidad permanecieron invariantes a lo largo de tiempo.

- *Figura 1.1 Datos Mortalidad Española:* El primer modo de variación es una mejora general de la mortalidad, que es más espectacular entre los jóvenes. Se aprecia la mejora en los registros de nacimientos y defunciones.

- *Figura 1.2 Datos Mortalidad Española:* El segundo modo de variación es un contraste entre los hombre de entre 20-45 años y el resto de la población. Este modo refleja la pandemia de Gripe Española junto con la Guerra Civil, entre otros efectos.

- *Figura II.5 Datos Parábolas Inclinadas:* Hay una serie de gráficas en este análisis que dan una indicación clara de la variación de los datos, los residuos medios y varios componentes principales junto con sus respectivos residuos. El gráfico izquierdo de la segunda fila muestra que el desplazamiento vertical es el modo de variación dominante (un aspecto común de muchos conjuntos de datos funcionales que se explora más profundamente en el Capítulo 3). En la tercera fila se revela el modo de variación menos obvio. La última fila sugiere que la variación restante es aleatoria.

- *Figuras II.6 y 2.1 Datos Cáncer de Pulmón:* La primera muestra una gran cantidad de variación que es difícil de analizar visualmente. La técnica de *brushing* (es decir, la coloración) de estas curvas que se muestra en la segunda revela importantes modos de variación.

- *Figuras II.11 y II.12 Curvas de Dos Escalas:* Estas gráficas demuestran el impacto potencial de la normalización de los datos (a través del escalado de la desviación estándar). En particular, esta elección puede poner de relieve modos de variación completamente diferentes.

Estos ejemplos demuestran los beneficios de dichas gráficas, especialmente cuando se usan a la par que ACP, u otro enfoque que ayude a revelar modos de variación. Sin embargo, hay situaciones en las que empleo de las curvas gráficas no es muy efectivo, como en los conjuntos de datos con mucho ruido presente.

2.3.3. Centrado de Datos y Vistas Combinadas

El *mean centering* (o centrado medio) de los datos es una cuestión fundamental para el análisis ADOO, cuyo impacto parece bastante obvio y rutinario. Sin embargo, al contemplar mapas de calor, los efectos pueden ser sorprendentemente importantes e incluso difíciles de comprender intuitivamente. Esto hace que se considere el centrado medio para las filas y columnas de la matriz de datos.

La importancia de realizar el centrado antes de buscar modos de variación, se aprecia en la Figura 2.2. Esto contrasta el ACP con un análisis totalmente descentrado basado en la Descomposición de Valores Singulares (DVS) directa de la matriz de datos. Muchos aspectos de DVS y su relación con ACP se estudian en el Capítulo 3.

La Figura 2.2 proporciona una comparación entre DVS no centrado y ACP en el contexto de un conjunto de datos de un ejemplo de juguete bidimensional, donde los datos se muestran en el espacio de características como los círculos negros en los dos paneles superiores. El panel izquierdo muestra la aproximación DVS de este conjunto de datos. La recta roja es el subespacio unidimensional (recta que pasa por el origen) que mejor se ajusta a los datos. Las cruces de color magenta son las proyecciones de los datos sobre dicho subespacio, y son la mejor aproximación de rango uno de los datos. En particular, esta dirección minimiza la suma de los cuadrados de los residuos proyectados, que serán las rectas azules. Las primeras puntuaciones DVS son los coeficientes de estas proyecciones, que aparecen en el eje horizontal en el panel inferior izquierdo. Las longitudes de las rectas azules son los coeficientes de las proyecciones sobre el segundo vector singular, es decir, las segundas puntuaciones DVS, que se muestran en el eje vertical del panel inferior izquierdo. Como DVS ignora el centro de datos, no logra resumir ni mostrar de manera eficiente el modo de variación dominante en este conjunto de datos, donde, en cambio, se divide entre ambos modos.

Los paneles de la derecha muestran los mismos datos solo que aplicándole ACP. Debemos

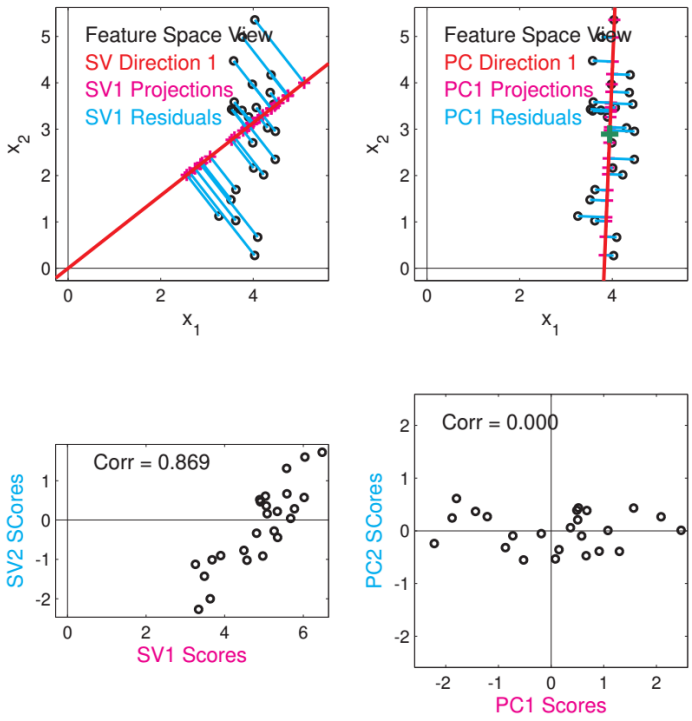

Figura 2.2: *Comparación de DVS y ACP para el ejemplo de juguete bidimensional. Las filas superiores muestran el espacio de características de los objetos de datos (círculos negros) con aproximaciones de DVS (izq.) y ACP (der.). En la fila inferior se muestran las correspondientes representaciones de las puntuaciones. Muestran que el centrado medio del objeto columna de ACP permite una buena aproximación de los datos de rango bajo.*

tener en cuenta que la diferencia con DVS es el centrado del objeto de datos medio (se muestra con una cruz verde). Esto da como resultado que la recta de mejor aproximación, que se muestra en rojo, se elija ahora entre los vectores dirección basados en ese punto. Las aproximaciones ACP de rango 1 se muestran como signos + magenta, que claramente proporcionan un resumen de los datos mucho mejor que el proporcionado por DVS, porque esta dirección maximiza la varianza de las proyecciones. Esta dirección ahora refleja apropiadamente el modo de variación vertical dominante. En particular, estas aproximaciones de rango 1 ahora se encuentran en el medio de la nube de puntos. Los coeficientes de estas proyecciones son las puntuaciones CP1, trazadas en el eje horizontal del diagrama de dispersión de las puntuaciones en el panel inferior derecho. Las rectas azules en el panel superior derecho muestran los residuos de esta aproximación (que por supuesto tienen una suma mínima de cuadrados). Las longitudes de éstas son las puntuaciones CP2 utilizadas en el eje vertical en el panel inferior derecho.

La Figura 2.2 también destaca otro aspecto del ACP (en relación con DVS de datos no centrados) que vale la pena señalar: las puntuaciones no están correlacionadas. Básicamente,

esto se debe a que cuando hay correlación, como se muestra en el panel inferior izquierdo, la varianza de la proyección más grande puede aumentar aún más mediante la rotación adecuada. De manera similar, la suma de los cuadrados de los residuos (que se muestran en azul en la fila superior) se puede reducir mediante esta rotación. Aunque parezca sorprendente, la falta de correlación en las puntuaciones de CP es una consecuencia del centrado medio del objeto de datos de la columna. Esto lo estudiamos profundamente en el Capítulo 3.

2.3.4. Matrices de Diagramas de Dispersión de Puntuaciones

Las matrices de diagramas de dispersión proporcionan otro tipo de perspectivas útiles, esta vez destacando las relaciones entre objetos de datos. Esto también lo hemos observado en los ejemplos planteados hasta ahora:

- *Figura II.7 Datos Cáncer de Pulmón:* Esta matriz de diagrama de dispersión ACP reveló la importante estructura de los grupos de los datos, lo que motivó el desarrollo de un método novedoso para encontrar empalmes alternativos utilizando datos de expresión genética.
- *Figura II.8 Datos Cáncer de Pulmón:* Esta versión mejorada de la Figura II.7 definió un esquema de color cuya aplicación condujo a la Figura 2.1, que reveló la naturaleza de los grupos de este conjunto de datos, en términos de empalme alternativo.
- *Figura II.19 Ejemplo Dos Clases Gaussianas:* En este caso, las matrices de diagramas de dispersión revelan estructuras de los datos que no eran fácilmente visibles en sus correspondientes mapas de calor.

En la mayoría de las matrices de diagramas de dispersión anteriores, los gráficos son sencillos porque las direcciones que determinan los ejes son ortogonales. Pero este no es siempre el caso. Una forma de manejar esto es simplemente dibujar las puntuaciones en los ejes vertical y horizontal. Un inconveniente importante es que el gráfico de puntuaciones ya no mantiene las posiciones relativas de los datos en el espacio Euclidiano subyacente y, por lo tanto, puede ofrecer una visión bastante distorsionada, especialmente cuando los vectores de dirección están lejos de ser ortogonales.

Capítulo 3

Análisis de Componentes Principales y su aplicación en la visualización

Uno de los objetivos del análisis multivariante de datos es reducir, o resumir los datos en un número de dimensiones menor que el original sin perder información esencial. Fue Pearson (1901) quien consideró este problema, pero Hotelling (1933) propuso una solución: En vez de tratar cada variable por separado, consideraremos combinaciones de las variables, como por ejemplo, la media de todas las variables. Pero surge una pregunta natural: *¿Cómo debemos escoger estas combinaciones?* La propuesta de Hotelling consistía en *encontrar combinaciones lineales de las variables que expliquen mejor la variabilidad de los datos*, ya que éstas son fáciles de interpretar. También podemos preguntarnos, *¿cuántas combinaciones de este tipo debemos escoger?*. La respuesta a esta pregunta no es tan fácil de responder, en parte porque depende de la pregunta anterior. Así, este número de combinaciones debe representar un compromiso entre *precisión* y *eficiencia*, es decir, separar la *señal* del *ruido*. Así, en este Capítulo estudiamos más a fondo el Análisis de Componentes Principales, y cómo combina las variables originales en un número reducido de ellas para facilitar la interpretación de los datos.

3.1. Puntos de Vista de ACP

Hay muchas maneras de considerar ACP, muchas de las cuales proporcionan distintas perspectivas. En esta sección, tendremos en cuenta la siguiente notación.

Recordemos de (I.4), una matriz $d \times n$, $\widetilde{\mathbf{X}}$ (el uso de tildes indica cantidades aleatorias) donde los vectores columnas de objetos de datos (también llamados *casos* o *muestras*) son $\widetilde{\mathbf{x}}_j = [\widetilde{x}_{1,j}, \cdots, \widetilde{x}_{j,n}]^t \in \mathbb{R}^d$ en el que cada $\widetilde{x}_{i,j}$ es la variable (o *característica*) para $j = 1, \ldots, n$ y

$i = 1, \ldots, d$. Todo esto ocurre en el espacio vectorial $\mathbb{R}^{d\times n}$. También, la matriz identidad $d \times d$ se denota por $\mathbf{I}_d$, $\mathbf{1}_{d,n}$ la matriz $d \times n$ de unos y $\mathbf{0}_{d,n}$ la matriz $d \times n$ de ceros. La operación *proyección* (I.8) también es muy importante en ACP. En la Sección 3.1.1 , estudiaremos las operaciones de centrado, usando las proyecciones en $\mathbb{R}^{d\times n}$ respecto a la norma de Frobenius $||\cdot||_F$ (I.10).

Dos subespacios de $\mathbb{R}^{d\times n}$ se llamarán *vectores planos*, si sus entradas son todas iguales. Esto permite interpretar las medias univariantes como proyecciones. Primero tengamos en cuenta que un vector de base uno del subespacio de vectores planos en $\mathbb{R}^n$ es $\frac{1}{\sqrt{n}}\mathbf{1}_{n,1}$. Por eso, dado $\mathbf{x} = (x_1, \cdots, x_n)^t \in \mathbb{R}^n$, el coeficiente de proyección $\mathbb{R}^n$ de $\mathbf{x}$ sobre el subespacio plano es el producto interno $\frac{1}{\sqrt{n}}\mathbf{1}_{n,1}\mathbf{x} = \sqrt{n}\bar{x}$. Multiplicando ese coeficiente por el vector base resulta en la proyeción de $\mathbb{R}^n$ de $\mathbf{x}$ sobre el subespacio de vectores planos definida como

$$\frac{1}{\sqrt{n}}\mathbf{1}_{n,1}\left(\frac{1}{\sqrt{n}}\mathbf{1}_{n,1}\mathbf{x}\right) = \frac{1}{n}\mathbf{1}_{n,n}\mathbf{x} = \mathbf{1}_{n,1}\overline{x}, \tag{3.1}$$

que es el vector plano cuya entrada es $\overline{x}$.

Para entender el centrado de datos como proyecciones en el espacio de matrices $\mathbb{R}^{d\times n}$, una notación útil es $\mathcal{S}_{FT}$ que denota el subespacio de $\mathbb{R}^{d\times n}$ en el que todos los vectores fila son planos, donde FT viene de *Flat Traits* (característica plana), es decir,

$$\mathcal{S}_{FT} = \left\{\mathbf{u}\mathbf{1}_{1,n} \ : \ \mathbf{u} \in \mathbb{R}^d\right\}. \tag{3.2}$$

De forma similar, definimos el subespacio (de $\mathbb{R}^{d\times n}$) consistente en matrices compuestas por vectores de objetos columna planas (*Flat Objects*) como

$$\mathcal{S}_{FO} = \left\{\mathbf{1}_{d,1}\mathbf{v}^t \ : \ \mathbf{v} \in \mathbb{R}^n\right\}. \tag{3.3}$$

Aplicando la traspuesta de (3.1) a cada fila de la matriz de datos $\widetilde{\mathbf{X}}$ obtenemos la proyección

$$P_{\mathcal{S}_{FT}}\left(\widetilde{\mathbf{X}}\right) = \widetilde{\mathbf{X}}\left(\frac{1}{\sqrt{n}}\mathbf{1}_{n,1}\right)\left(\frac{1}{\sqrt{n}}\mathbf{1}_{1,n}\mathbf{x}\right) = \widetilde{\mathbf{X}}\left(\frac{1}{n}\mathbf{1}_{n,n}\right) = \overline{\mathbf{x}}_{CO}\mathbf{1}_{1,n} \tag{3.4}$$

que esencialmente extiende la media columna $d\times 1$ $\overline{\mathbf{x}}_{CO}$, definida como $\overline{\mathbf{x}}_{CO} = (\overline{x}_{1,A}, \cdots, \overline{x}_{d,A})^t \in \mathbb{R}^d$ con $\overline{x}_{i,A}$ definida en (I.7), a un elemento $d \times n$ de $\mathcal{S}_{FT}$. La versión centrada del objeto columna de la matriz de datos también se puede escribir como una proyección sobre el subespacio complementario ortogonal (respecto al producto interno de Frobenius (I.10)) $\mathcal{S}_{FT}^{\perp} = \left\{\mathbf{M} \in \mathbb{R}^{d\times n} : \mathbf{M} \perp \mathcal{S}_{FT}\right\}$ como

$$P_{\mathcal{S}_{FT}^{\perp}}\left(\widetilde{\mathbf{X}}\right) = \widetilde{\mathbf{X}} - \overline{\mathbf{x}}_{CO}\mathbf{1}_{1,n} = \widetilde{\mathbf{X}}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n,n}\right). \tag{3.5}$$

De igual manera, el vector medio de características fila, definido como $\overline{\mathbf{x}}_{RT} = (\overline{x}_{A,1}, \cdots, \overline{x}_{A,n})$, $\overline{\mathbf{x}}_{RT} \in \mathbb{R}^n$, cuyas entradas son $\overline{x}_{A,j} = d^{-1}\sum_{i=1}^{d} x_{i,j}$ para $j = 1, \cdots, n$, se puede expresar en términos de proyecciones sobre el subespacio de objetos planos $\mathcal{S}_{FO}$

$$P_{\mathcal{S}_{FO}}\left(\widetilde{\mathbf{X}}\right) = \mathbf{1}_{d,1}\overline{\mathbf{x}}_{RT}^t = \left(\frac{1}{d}\mathbf{1}_{d,d}\right)\widetilde{\mathbf{X}}, \tag{3.6}$$

y su complementario ortogonal es $\mathcal{S}_{FO}^{\perp}$

$$P_{S_{FO}^{\perp}}\left(\widetilde{\mathbf{X}}\right) = \widetilde{\mathbf{X}} - \mathbf{1}_{d,1}\overline{\mathbf{x}}_{RT}^{t} = \left(\mathbf{I}_d - \frac{1}{d}\mathbf{1}_{d,d}\right)\widetilde{\mathbf{X}} \tag{3.7}$$

3.1.1. Centrado de Datos

Como ya dijimos, el centrado parece una cuestión simple, pero puede ser resbaladiza. Aquí, nos centramos en una visión más profunda basada en las ideas anteriores de las proyecciones sobre $\mathbb{R}^{d\times n}$ con el ejemplo de juguete basado en el conjunto de datos *Onda Sinusoidal* que se muestra en la Figura II.21. Este conjunto de datos fue generado (para $i = 1, \cdots, 20$ y $j = 1, \cdots, 10$) como

$$x_{i,j} = T_1 + T_2 + T_3 - 3 + \mathcal{N}(0, 10^{-6}) \tag{3.8}$$

donde

$$\begin{aligned} T_1 &= sin(5 \cdot \pi \cdot (i-1)/19), \\ T_2 &= 0{,}3 \cdot (j - 5{,}8)^2, \\ T_3 &= 0{,}005 \cdot (i - 10{,}5) \cdot (j - 5{,}5) \end{aligned} \tag{3.9}$$

Sean $\mathbf{T}_1, \mathbf{T}_2$ y $\mathbf{T}_3$ las versiones matriciales de los tres primeros términos de la ecuación (3.9), donde cada uno de ellos es un modo de variación. En particular, $\mathbf{T}_1 = \mathbf{u}\mathbf{1}_{1,n}$ ($\mathbf{u} \in \mathbb{R}^d$ genera la onda sinusoidal) tiene filas planas, por lo que $\mathbf{T}_1 \in \mathcal{S}_{FT}$ (es decir, es un modo característico plano) y la variación entre filas sigue una onda sinusoidal. Análogamente, $\mathbf{T}_2 = \mathbf{1}_{d,1}\mathbf{v}^t$ ($\mathbf{v} \in \mathbb{R}^n$ genera una parábola) tiene columnas planas, por lo que $\mathbf{T}_2 \in \mathcal{S}_{FO}$ (es un objeto plano) y la variación entre columnas sigue una parábola. Ambos modos son visibles en la Figura II.21. El tercer modo $\mathbf{T}_3$ es un producto de dos vectores lineales (ambos con media cero), tiene un coeficiente mucho más pequeño para que su contribución a la variación total sea mucho menor. Tenemos que $\mathbf{T}_3 \in \mathcal{S}_{FT}^{\perp} \cap \mathcal{S}_{FO}^{\perp}$, lo que dará como resultado $\mathbf{T}_3$ después de aplicar ambas operaciones de centrado (por filas y por columnas). Finalmente, se añade también un nivel muy bajo de variación normal independiente.

En el panel izquierdo tenemos un mapa de calor en el que se aprecian perfectamente ambos modos, la onda sinusoidal vertical y la forma parabólica horizontal. Para obtener otro punto de vista, podemos fijarnos en el panel central, que muestra los objetos de datos columna como $n = 10$ curvas (ondas sinusoidales). Esta gráfica sugiere que este sería el modo de variación dominante. El panel derecho se centra en los vectores característica fila $n = 20$ en $\mathbb{R}^n$. Aquí, sin embargo, la gráfica sugiere que el modo de variación dominante es el de la curva parabólica. Llegamos a la conclusión de que la variación de la altura de las ondas sinusoidales sigue una parábola, y la de la parábola una onda sinusoidal. Otro punto importante es que ninguno de los dos es un modo único de variación, según la definición dada en el Capítulo 1, ya que no son

matrices de rango 1. De hecho, ambos son la suma de dos modos que se muestran como $\mathbf{T}_1 + \mathbf{T}_2$ en (3.11). El primer modo $\mathbf{T}_1 = \mathbf{u}\mathbf{1}_{1,n}$ extiende la onda sinusoidal $\mathbf{u}$ de una forma horizontal plana o bien proporciona una variación sinusoidal a la recta plana $\mathbf{1}_{1,n}$. De manera análoga, el modo $\mathbf{T}_2 = \mathbf{1}_{d,1}\mathbf{v}^t$ proporciona variación parabólica a la recta vertical $\mathbf{1}_{d,1}$ y también una extensión plana vertical de la parábola $\mathbf{v}^t$.

El efecto del centrado medio del objeto columna, es decir, $P_{\mathcal{S}_{FT}^{\perp}}\left(\widetilde{\mathbf{X}}\right)$ es visible en la Figura II.22. Como $\mathbf{T}_1 \in \mathcal{S}_{FT}$, esa resta de la media del objeto de datos de la columna elimina la onda sinusoidal vertical del mapa de calor, lo que hace más evidente la estructura parabólica horizontal del modo $\mathbf{T}_2$. El panel central, muestra que la estructura de onda sinusoidal se ha eliminado, dejando en su lugar un rectas horizontales, con alturas determinadas por la parábola. Además, el modo $\mathbf{T}_3$ comienza a ser visible en estas curvas. En el panel derecho, se elimina la mayor parte de la variación impulsada por $\mathbf{T}_1$, dejando un conjunto de parábolas casi iguales.

De manera análoga, la Figura II.23 muestra el efecto del centrado medio de la característica fila, es decir, $P_{\mathcal{S}_{FO}^{\perp}}\left(\widetilde{\mathbf{X}}\right)$. En el panel izquierdo observamos que el efecto de la parábola horizontal $\mathbf{T}_2$ se eliminó ya que $\mathbf{T}_2 \in \mathcal{S}_{FO}$ dejando franjas horizontales que siguen el patrón de las ondas sinusoidales verticales. Los ejes verticales en los paneles central y derecho de las Figuras II.21, II.22 y II.23 utilizan la misma escala para mostrar que la variación parabólica $\mathbf{T}_2$ tiene mayor magnitud en relación con la componente de onda sinusoidal $\mathbf{T}_1$. Finalmente, el panel derecho muestra que esta operación elimina la estructura parabólica $\mathbf{T}_2$. Ahora, en ambos gráficos, el componente $\mathbf{T}_3$ es un poco más visible.

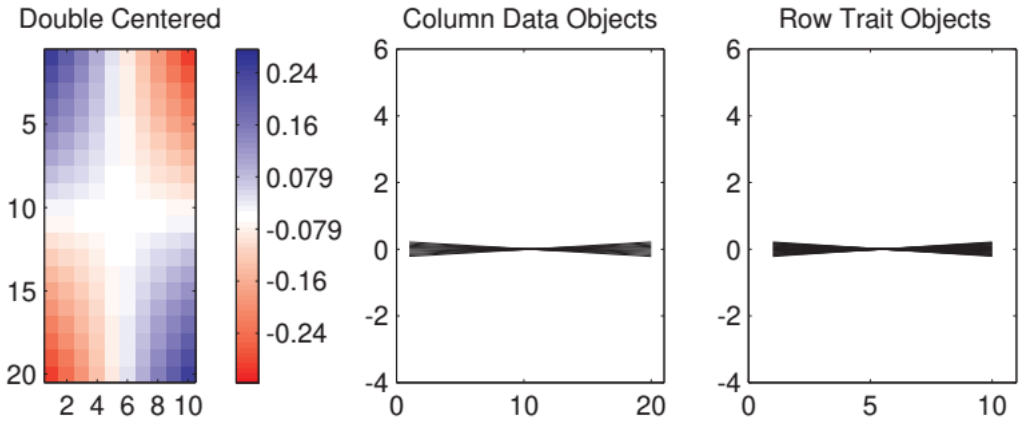

Figura 3.1: *Resultado de aplicar el centrado doble al conjunto de datos Onda Sinusoidal. Vemos que desaparece la mayor parte de la estructura de los datos.*

Otra opción de centrado de datos es el *centrado doble* (Figura 3.1). La forma más simple de obtenerla es mediante la eliminación simultánea de ambos tipos de media a partir de la

composición de proyecciones $P_{\mathcal{S}_{FT}^{\perp}}\left(P_{\mathcal{S}_{FO}^{\perp}}\left(\widetilde{\mathbf{X}}\right)\right)$. Otras formas de expresar el doble centrado son

$$\begin{aligned} P_{\mathcal{S}_{FT}^{\perp}}\left(P_{\mathcal{S}_{FO}^{\perp}}\left(\widetilde{\mathbf{X}}\right)\right) &= \left(\widetilde{\mathbf{X}} - P_{\mathcal{S}_{FO}}\left(\widetilde{\mathbf{X}}\right)\right) - P_{\mathcal{S}_{FT}}\left(\widetilde{\mathbf{X}} - P_{\mathcal{S}_{FO}}\left(\widetilde{\mathbf{X}}\right)\right) = \\ &= \widetilde{\mathbf{X}} - P_{\mathcal{S}_{FO}}\left(\widetilde{\mathbf{X}}\right) - P_{\mathcal{S}_{FT}}\left(\widetilde{\mathbf{X}}\right) + P_{\mathcal{S}_{FT}\cap\mathcal{S}_{FO}}\left(\widetilde{\mathbf{X}}\right) \end{aligned} \tag{3.10}$$

Escribiendo esto en forma matricial, obtenemos la versión doblemente centrada de los datos,

$$\begin{aligned} \overline{\overline{\mathbf{X}}} = \widetilde{\mathbf{X}} - \overline{\mathbf{x}}_{CO}\cdot\mathbf{1}_{1,n} - \mathbf{1}_{d,1}\cdot\overline{\mathbf{x}}_{RT}^{t} + \overline{x}_{AA}\cdot\mathbf{1}_{d,n} = \\ = \left(\mathbf{I}_d - \frac{\mathbf{1}_{d,d}}{d}\right)\widetilde{\mathbf{X}}\left(\mathbf{I}_n - \frac{\mathbf{1}_{n,n}}{n}\right) \end{aligned} \tag{3.11}$$

donde la *media total* (es decir, la media de todas las entradas de la matriz de datos) es el escalar

$$\overline{x}_{AA} = d^{-1}\sum_{i=1}^{d}\overline{x}_{i,A} = n^{-1}\sum_{j=1}^{n}\overline{x}_{A,j} = (nd)^{-1}\sum_{i=1}^{d}\sum_{j=1}^{n}\widetilde{x}_{i,j} = (nd)^{-1}\mathbf{1}_{1,d}\mathbf{X}\mathbf{1}_{n,1} \tag{3.12}$$

Cabe destacar que el término $\overline{x}_{AA}$ se vuelve a agregar en (3.11) porque en un desplazamiento vertical de la matriz de datos, dicha media se restará dos veces. Otra cosa que debemos tener en cuenta en la Figura 3.1 es que los efectos horizontales y verticales que dominaban la Figura II.21 se han eliminado, lo que deja el producto de los componentes lineales generados por $\mathbf{T}_3 \in \mathcal{S}_{FT}^{\perp}\cap\mathcal{S}_{FO}^{\perp}$ mostrándose como un conjunto de rectas cuyas pendientes cambian linealmente de positiva a negativa.

3.1.2. Descomposición en Valores Singulares

Una concepción típica de ACP es que es secuencial por naturaleza a través de los componentes (modos de variación), donde en cada paso se busca una dirección de variación máxima en el subespacio ortogonal a las direcciones anteriores, para luego utilizar las proyecciones de los objetos de datos sobre estas direcciones para definir los modos. Una propiedad interesante de ACP es que encontrar el conjunto completo de modos puede realizarse en una única operación matricial. Para ello, empleamos la *Descomposición en Valores Singulares* (DVS) ya que proporciona un enfoque directo y simple para calcular ACP y comprender sus propiedades. La versión de la matriz completa DVS de una matriz de datos $d \times n$, $\widetilde{\mathbf{X}}$, es

$$\widetilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^{t}, \tag{3.13}$$

donde $\mathbf{U} \in \mathcal{O}(d)$ (conjunto de matrices ortogonales) es una matriz básica ortonormal $d \times d$ de $\mathbb{R}^d$, $\mathbf{D}$ es una matriz diagonal $d \times n$ de *valores singulares* no negativos ordenados en orden decreciente y $\mathbf{V} \in \mathcal{O}(n)$ es una matriz básica ortonormal $n \times n$. Podemos escribir esta versión matricial de DVS en términos de submatrices. Si $d \neq n$, la matriz diagonal $\mathbf{D}$ tiene como máximo $d \wedge n$ ($\wedge$ denota el mínimo) elementos distintos de cero, por lo que habrá filas o columnas que

no se emplearían para obtener $\widetilde{\mathbf{X}}$. Entonces, DVS se puede formular con las submatrices $\mathbf{U}$ $((n \times (d \wedge n)))$, $\mathbf{D}$ $((d \wedge n) \times (d \wedge n))$ y $\mathbf{V}$ $((d \wedge n) \times n)$. Aún podemos simplificarlo más si tenemos en cuenta que $\widetilde{\mathbf{X}}$ puede tener rango $r < (d \wedge n)$, porque entonces algunos valores singulares son 0, por lo que no tendrán efecto en el producto. En este caso, las dimensiones de las submatrices son de la forma, $\mathbf{U}$ de $n \times r$, $\mathbf{D}$ de $r \times r$ y $\mathbf{V}$ de $r \times n$, que, a su vez, son la versión más eficiente de DVS, obteniéndose una representación exacta de $\widetilde{\mathbf{X}}$. Se llamarán *vectores singulares izquierdo* y *derecho*, respectivamente, a las columnas de

$$\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_r], \ \mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_r] \tag{3.14}$$

Tenemos que DVS proporciona soluciones para distintos problemas de optimización, entre ellos existe uno que es clave para comprender los modos de variación generados por ACP, y se puede ver enunciándolo como una suma de matrices de rango 1:

$$\widetilde{\mathbf{X}} = \sum_{l=1}^{r} s_l \mathbf{u}_l \mathbf{v}_l^t \tag{3.15}$$

donde s_l es el l-ésimo *valor singular* (que son los elementos diagonales de $\mathbf{D}$), $\mathbf{u}_l$ y $\mathbf{v}_l$ definidos en (3.14). En particular, para $k \leq r$ (rango de $\widetilde{\mathbf{X}}$), partiendo del orden de los valores singulares $s_1 \geq \cdots \geq s_r > 0$, se deduce que $\widetilde{\mathbf{X}}_k = \sum_{l=1}^{k} s_l \mathbf{u}_l \mathbf{v}_l^t$ es la mejor aproximación de rango k de $\widetilde{\mathbf{X}}$ en el sentido de que

$$\widetilde{\mathbf{X}}_k = arg \min_{\mathbf{M} \in \mathcal{R}_k} ||\widetilde{\mathbf{X}} - \mathbf{M}||_F, \tag{3.16}$$

donde $\mathcal{R}_k$ es el subconjunto de matrices de rango $\leq k$. Debemos tener en cuenta que $\widetilde{\mathbf{X}}_k$ es de la forma $\widetilde{\mathbf{X}}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^t$, donde $\mathbf{U}_k$, $\mathbf{D}_k$ y $\mathbf{V}_k^t$ son las k primeras columnas de $\mathbf{U}$, la subdiagonal superior $k \times k$ de $\mathbf{D}$, y las primeras k filas de $\mathbf{V}^t$, respectivamente. Esto también muestra cómo para cada rango k, DVS puede verse como una proyección en $\mathbb{R}^{d \times n}$ sobre $\mathcal{R}_k$. Además, cada matriz de rango 1 $s_l \mathbf{u}_l \mathbf{v}_l^t$ es la proyección de la matriz de datos $\widetilde{\mathbf{X}}$ sobre el subespacio unidimensional generado por $\mathbf{u}_l \mathbf{v}_l^t$, y la aproximación de rango k es la proyección sobre el subespacio de k dimensiones generado por $\mathbf{u}_1 \mathbf{v}_1^t, \cdots, \mathbf{u}_k \mathbf{v}_k^t$. Finalmente, debemos tener en cuenta que cada matriz de rango 1 $s_l \mathbf{u}_l \mathbf{v}_l^t$ (para $l = 1, \cdots, k$) es un modo de variación que contiene tanto a los pesos, $\mathbf{u}_l$, como a las puntuaciones, $s_l \mathbf{v}_l^t$.

La representación (3.15) proporciona una conexión muy útil entre los pesos y las puntuaciones de DVS. En particular, las puntuaciones son proyecciones de los datos sobre el subespacio generado por los respectivos vectores de pesos en $\mathbb{R}^d$ en el sentido de que para $l = 1, \cdots, r$

$$\mathbf{u}_l^t \widetilde{\mathbf{X}} = \mathbf{u}_l^t \sum_{l'=1}^{r} s_{l'} \mathbf{u}_{l'} \mathbf{v}_{l'}^t = s_l \mathbf{v}_l^t. \tag{3.17}$$

De forma similar, los pesos tienen una representación en términos de puntuaciones como

$$\widetilde{\mathbf{X}} \mathbf{v}_l = \sum_{l'=1}^{r} s_{l'} \mathbf{u}_{l'} \mathbf{v}_{l'}^t \mathbf{v}_l = s_l \mathbf{u}_l \tag{3.18}$$

La versión matricial de (3.17) es la proyección de cada objeto de datos sobre cada uno de los vectores singulares, mediante el cálculo del producto interno de la matriz

$$\mathbf{U}^t\widetilde{\mathbf{X}} = \mathbf{U}^t\mathbf{U}\mathbf{D}\mathbf{V}^t = \mathbf{D}\mathbf{V}^t. \tag{3.19}$$

y con estos productos internos obtenemos los coeficientes de las proyecciones, o puntuaciones DVS, base de muchas visualizaciones, como, por ejemplo, las matrices de diagramas de dispersión del Capítulo 2. Podríamos entender ACP como DVS de los datos centrados en el objeto columna,

$$\breve{\mathbf{X}} = n^{-1/2}\left(\widetilde{\mathbf{X}} - \overline{\mathbf{x}}_{CO}\mathbf{1}_{1,n}\right) = n^{-1/2}\widetilde{\mathbf{X}}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n,n}\right), \tag{3.20}$$

escrito como

$$\breve{\mathbf{X}} = \breve{\mathbf{U}}\breve{\mathbf{D}}\breve{\mathbf{V}}^t. \tag{3.21}$$

Para $l = 1, \cdots, r$, el l-ésimo conjunto de puntuaciones ACP, es decir, los *l-ésimos componentes principales*, están en l-ésima fila de la matriz $r \times n$, $\breve{\mathbf{D}}\breve{\mathbf{V}}^t$. Además, la l-ésima columna de $\breve{\mathbf{U}}$ es el l-ésimo vector de pesos, que es la dirección en $\mathbb{R}^d$ de la l-ésima variación más grande de los datos. Para $i = 1, \cdots, d$, la i-ésima entrada del vector de pesos refleja la dirección y magnitud de la influencia de la i-ésima variable en la l-ésima dirección. Los pesos y puntuaciones CP están relacionadas entre sí a través de proyecciones. En particular, como

$$\breve{\mathbf{U}}^t\breve{\mathbf{X}} = \breve{\mathbf{U}}^t\breve{\mathbf{U}}\breve{\mathbf{D}}\breve{\mathbf{V}}^t = \breve{\mathbf{D}}\breve{\mathbf{V}}^t \tag{3.22}$$

la l-ésima fila de la matriz de puntuaciones $\breve{\mathbf{D}}\breve{\mathbf{V}}^t$ es el vector $n \times 1$ de productos internos (esencialmente coeficientes de proyección) del l-ésimo vector de pesos con los objetos de datos centrados (columnas de $\widetilde{\mathbf{X}}$). De forma similar

$$\breve{\mathbf{X}}\left(\breve{\mathbf{D}}\breve{\mathbf{V}}^t\right)^t\breve{\mathbf{D}}^{-2} = \breve{\mathbf{U}}\breve{\mathbf{D}}\breve{\mathbf{V}}^t\breve{\mathbf{V}}\breve{\mathbf{D}}\breve{\mathbf{D}}^{-2} = \breve{\mathbf{U}} \tag{3.23}$$

muestra que los pesos se pueden representar como una normalización (reescalado por las varianzas inversas) de los productos internos de la matriz de datos centrada y las puntuaciones. En resumen, (3.22) y (3.23) muestran que tanto los pesos como las puntuaciones se pueden obtener entre sí mediante un producto con la matriz de datos centrada y escalada.

La operación de centrado de columnas en (3.20) proporciona una mejor interpretabilidad en varias situaciones. Una de ellas, resultante del centrado de objetos columna se ilustró en los paneles inferiores de la Figura 2.2, comparando DVS no centrado con ACP en el ejemplo de juguete planteado. En particular, para ACP hace que obtengamos diagramas de dispersión de puntuaciones no correlacionados más fácilmente interpretables. Como vimos en la Sección 2.3.3, los diagramas de dispersión de puntuaciones ACP tienen correlación 0, fenómeno que proviene del cálculo de la matriz $r \times r$ de productos internos de los vectores de puntuaciones, y por la ortonormalidad de las columnas de $\breve{\mathbf{V}}$ (filas de $\breve{\mathbf{V}}^t$) y el hecho de que $\breve{\mathbf{D}}$ es diagonal,

$$\breve{\mathbf{D}}\breve{\mathbf{V}}^t\left(\breve{\mathbf{D}}\breve{\mathbf{V}}^t\right)^t = \breve{\mathbf{D}}\breve{\mathbf{V}}^t\breve{\mathbf{V}}\breve{\mathbf{D}}^t = \breve{\mathbf{D}}^2. \tag{3.24}$$

Entonces, la ortogonalidad de los vectores de puntuaciones (filas de $\breve{\mathbf{D}}\breve{\mathbf{V}}^t$), se obtiene del hecho de que $\breve{\mathbf{D}}^2$ también es una matriz diagonal.

Sin el centrado de la media del objeto columna, generalmente existe alguna correlación de muestra (es decir, distinta de cero) en los gráficos de puntuaciones DVS. Sin embargo, en muchas situaciones, esto no nos perjudica, porque frecuentemente uno de los vectores singulares (generalmente el primero) apunta, de forma aproximada, en la dirección de la media del objeto columna de la muestra. En particular, muchos conjuntos de datos tienden a tener un modo de variación plano importante, por ejemplo $\mathbf{T}_2$, como se ilustra en (3.9). Ese modo plano suele estar en la primera dirección DVS, por lo que los modos de variación DVS suelen ser similares a los de un ACP estándar.

Una vez más, este centrado da como resultado puntuaciones de CP no correlacionadas. La simetría de la DVS en términos de filas y columnas sugiere una relación paralela: el centrado medio de la característica fila implica que dichos diagramas de dispersión de los pesos de ACP tampoco están correlacionados. Mientras que el centrado medio de objetos de columna es muy común para ACP, el centrado medio de característica fila (o el centrado medio doble) no lo es. Como se muestra en la Figura 2.2, un centrado apropiado garantizará que las puntuaciones no estén correlacionadas, lo que no parece muy obvio, ya que típicamente en los análisis, el centrado medio del objeto columna elimina la gran media $\overline{x}_{AA}$, que a menudo es un factor importante en la falta de correlación de las puntuaciones.

También es útil formular estas cantidades en términos de la matriz de covarianza muestral $\widehat{\mathbf{\Sigma}}$ definida en (I.5). Dada una matriz de datos $\widetilde{\mathbf{X}}$, las entradas de $\widehat{\mathbf{\Sigma}}$ son las varianzas y covarianzas muestrales, $\widehat{var}_i$ y $\widehat{cov}_{i,i'}$ definidas en (I.6). Así, la matriz de covarianzas muestrales está relacionada con ACP representándola como el producto externo

$$\widehat{\mathbf{\Sigma}} = \breve{\mathbf{X}}\breve{\mathbf{X}}^t. \tag{3.25}$$

Usando la representación en valores singulares (3.21) obtenemos

$$\widehat{\mathbf{\Sigma}} = \breve{\mathbf{U}}\breve{\mathbf{D}}\breve{\mathbf{V}}^t\breve{\mathbf{V}}\breve{\mathbf{D}}^t\breve{\mathbf{U}}^t = \breve{\mathbf{U}}\breve{\mathbf{\Lambda}}\breve{\mathbf{U}}^t \tag{3.26}$$

denominada *descomposición de valores propios* o *análisis propio*, donde $\breve{\mathbf{\Lambda}} = \breve{\mathbf{D}}\breve{\mathbf{D}}^t$ es la matriz diagonal cuyas entradas son los cuadrados de las entradas de $\breve{\mathbf{D}}$, es decir, los cuadrados de los valores singulares $\breve{s}_l$. Esto revela que ACP también se puede considerar como un análisis de valores propios (autovalores) de la matriz de covarianzas de la muestra.

3.1.3. Visualización de Componentes Principales

Como ya vimos en el Capítulo 2, la representación de ACP facilita la interpretación de los datos, además de ayudarnos a encontrar estructuras y patrones de los conjuntos de datos. A

lo largo de este trabajo estudiamos distintas técnicas gráficas que se apoyaban en ACP. En los casos que enunciamos a continuación utilizaremos lo visto en la Sección anterior, es decir, los autovalores y autovectores correspondientes de la matriz de covarianzas muestrales.

Gráficos de Sedimentación, Autovalores y Varianza

Comenzamos introduciendo los gráficos de sedimentación, que nos muestran la distribución de los autovalores y el decrecimeineto de la varianza de las puntuaciones. El tamaño real de los autovalores puede que no sea muy importante, así que la proporción de varianza total nos da una estandarización conveniente de los autovalores. Estas pueden mostrar un *kink*, es decir, un codo o pliegue. Se dice que el índice k en el que aparece dicho codo sería el número de componentes que representan adecuadamente los datos, es decir la dimensión de los datos reducidos o componentes principales. Pero la existencia de estos codos no está garantizada, e incluso no existe una justificación real par usar ese índice como dimensión de los datos de componentes principales.

Ejemplo 3.1. *Datos Cáncer de Mama y rendimientos de Dow Jones:* Los datos de *cáncer de mama* de Blake y Merz (1998) consisten en 569 registros y 30 variables. Los rendimientos de Dow Jones consiste en 30 acciones en 2,528 días durante el periodo comprendido entre Enero de 1991 y Enero de 2001. En los datos del cáncer de mama surgen dos grupos: 212 casos malignos y 357 casos benignos. En todos los registros se conoce este estado, no estamos interesados en ellos, sino en la matriz de covarianzas muestrales. Las observaciones de Dow Jones son los *rendimientosdiarios*, es decir, las diferencias de precios de registro tomadas en días consecutivos.

En la Figura II.24 vemos las contribuciones a la varianza total en el panel de arriba y las contribuciones acumulativas en el de abajo, respecto al índice de los CPs en el eje x. Los autovalores de los datos del cáncer decrecen más rápido que los de Dow Jones. Para el cáncer de mama, el primer CP contribuye al 44 %, y el segundo el 19 %. Para $k = 17$ supera el 99 %. Este rápido crecimiento sugiere que las componentes principales superiores a 18 son despreciables. Para Dow Jones, el primer CP explica el 25,5 % de la varianza total. Para llegar al 95 %, se necesitarán los 26 primeros CP. Las gráficas de estos conjuntos de datos no tienen codos. La ausencia de codos es más común que su presencia. □

Gráficas de Puntuaciones

Nos referimos a ellas como gráficas de puntuaciones de componentes principales o gráficas de puntuaciones CP. Éstas ayudan a resumir los datos y muestran patrones en los datos, como grupos que no se aprecien a simple vista. Ya vimos muchas gráficas de este tipo a lo largo del Capítulo anterior, formando parte de matrices de gráficas de puntuaciones.

Aunque podemos hacer gráficas de puntuaciones CP de cualquier CP, nos centraremos en aquellos valores CP que tengan autovalores grandes, ya que como tienen más peso, se espera que la estructura presente en los datos sea más visible para estas gráficas.

Ejemplo 3.2. *Datos reconocimiento de vino:* El conjunto de datos *reconocimiento de vino* de Corina et al. se obtienen como resultado de un análisis químico de 3 tipos de vino que crecieron en la misma región de Italia, pero derivados de cultivos tres cultivos distintos. Este análisis ha resultado en 13 variables, llamadas *constituyentes*. De las 178 observaciones, 59 pertenecen al primer cultivo (negro), 71 al segundo (rojo) y las restantes 48 al tercero (azul). Examinamos las componentes principales de estos datos en la Figura II.25. Empleamos una gráfica en 3D para representar las puntuaciones CP. En concreto representamos las puntuaciones CP1, CP3 y CP4. Se obtiene una separación razonable, aunque no perfecta, de los cultivos. □

Gráficas de Proyección y Estimación de la Densidad de las Puntuaciones

Las proyecciones de componentes principales son matrices $d \times n$. Para cada índice $k \leq d$, la i-ésima columna de dicha matriz representa la contribución de la i-ésima observación en la dirección del k-ésimo autovector. Las proyecciones de las componentes principales "inventan los datos" en el sentido de que podemos reconstruir (arbitrariamente de forma muy aproximada) los datos a partir de esas proyecciones.

La proyección k-ésima de componentes principales muestra como el autovector k-ésimo ha sido modificado por la puntuación k-ésima, por lo que resulta natural ver la distribución de dichas puntuaciones, en forma de estimación de densidad. Ésta nos da información muy importante sobre la distribución de las puntuaciones.

Ejemplo 3.3. *Datos Cáncer de Mama y rendimientos de Dow Jones:* Continuamos con el análisis del conjunto de datos del Ejemplo 3.1. Dow Jones tiene 5 veces más observaciones que las de cáncer de mama. Ahora estudiaremos diagramas coordinados paralelos y estimaciones de densidad de las proyecciones del primer y segundo componente principal para ambos conjuntos de datos, como observamos en la Figura 3.2.

En ambos conjuntos de datos, todas las entradas del primer autovector tienen el mismo signo. Verificamos eso en las gráficas de proyecciones en la 1ª y 3ª fila, donde cada observación o es positiva para cada variable o se mantiene negativa para todas las variables. Este comportamiento es inusual, y nos permite dividir los datos en dos grupos, los positivos y los negativos. Además, no hay ninguna variable que destaque, en ambos conjuntos el mayor peso es de 0,26.

Las gráficas de proyección de los segundos autovectores muestran un patrón más común, con entradas positivas y negativas, que muestran los efectos contrarios de las variables. Para Dow Jones, las variables 3, 13, 16, 17 y 23 tienen pesos negativos. Éstas se corresponden con

Information Technology Companies (IT). Salvo una, estas 4 compañías tienen los 4 pesos más grandes (en valor absoluto). Por tanto, CP2 separa claramente estas compañías del resto. Para los datos de cáncer de mama, las puntuaciones están mucho más espaciadas, tanto para CP1 como para CP2, además del rango de los valores de y en las gráficas de proyección y los valores de x en las gráficas de densidad. Recordemos que estas dos componentes principales explican, para Dow Jones, el 33 % de la varianza total y para cáncer de mama, el 60 %.

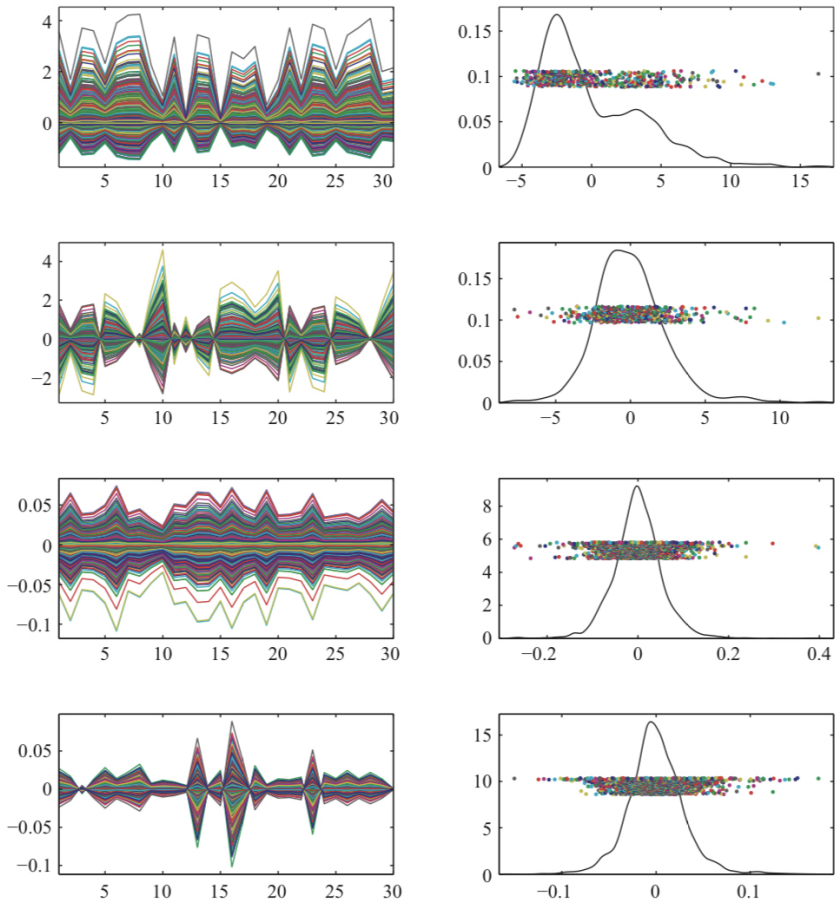

Figura 3.2: *Gráficas de proyección (izq.) con sus respectivas estimaciones de densidad (der.) de las puntuaciones CP1 (primera y tercera fila) y CP2 (segunda y cuarta) para los datos de cáncer de mama (dos primeras filas) y los de Dow Jones (dos últimas filas).*

Las gráficas de la columna de la derecha muestran las puntuaciones y sus estimaciones de densidad no paramétricas. La puntuación de cada observación viene dada por su valor en el eje x. Para entenderlo mejor visualmente, los valores reales de las puntuaciones son mostradas a alturas aleatorias, y, como puntos de colores, y cada observación es representada por el mismo color en ambas gráficas de la misma fila. El valor atípico en el extremo derecho en el gráfico de densidad de cáncer de mama de CP1 corresponde con a la curva más positiva en la gráfica de

proyección correspondiente. Observando la estimación densidad muestra que las puntuaciones CP1 del cáncer de mama se desvían sustancialmente de la distribución normal. Esto puede ser debido a que en el conjunto existen observaciones malignas y benignas (forma binomial). Estas puntuaciones no siguen una distribución Gaussiana, aunque las otras 3 densidades parecen simétricas y normales. □

En resumen, estas gráficas nos ayudan a desenmascarar subgrupos, encontrar valores atípicos o deducir que los datos no son Gaussianos. Todas estas propiedades nos ayudan a comprender y redactar análisis posteriores.

3.1.4. Punto de Vista de Probabilidad Normal

Hasta ahora nos hemos centrado en el desarrollo *no paramétrico* o centrado de datos de métodos analíticos para ADOO, por lo que ahora nos centraremos en los enfoques de probabilidad resultantes. Para ellos emplearemos modelos de *factores latentes* de rango bajo ya que proporcionan un enfoque tradicional de este tipo para ACP. En particular, supongamos que la matriz de datos $d \times n$ se puede escribir en la forma

$$\widetilde{\mathbf{X}} = \boldsymbol{\mu}\mathbf{1}_{1,n} + \mathbf{LS} + \widetilde{\mathbf{E}} \tag{3.27}$$

donde para algunos r de rango bajo, $\mathbf{L}$ es una matriz de pesos ortonormales $d \times r$, $\mathbf{S}$ es una matriz de puntuaciones $r \times n$ cuyas filas son ortogonales y suman 0 y $\widetilde{\mathbf{E}}$ es una matriz $d \times n$ de errores cuyas entradas se suponen $\mathcal{N}(0, \sigma^2)$, independientes. A menudo se piensa que las filas de $\mathbf{S}$ son factores latentes no observados y su estimación es el objetivo principal del *análisis factorial*. Los cálculos muestran que para un r dado, usando la definición de $\overline{x}_{CO}$ y (3.21), las estimaciones de máxima verosimilitud de $\boldsymbol{\mu}$, $\mathbf{L}$, $\mathbf{S}$, son $\overline{\mathbf{x}}_{CO}$, $\breve{\mathbf{U}}$ y $\breve{\mathbf{D}}\breve{\mathbf{V}}^t$ respectivamente.

Aunque este enfoque es matemáticamente elegante, tiene el grave inconveniente de que ha alimentado una idea errónea, que es que ACP sólo es útil cuando los datos siguen, aproximadamente, una distribución normal multivariante, aunque, como se ve a lo largo de este trabajo, ACP proporciona visualizaciones de datos muy útiles en muchas situaciones no normales. Además, aunque no son lo mismo, el análisis factorial está estrechamente relacionado con ACP. La diferencia es que el análisis factorial incluye la estimación de la varianza residual σ^2 como parte del cálculo de probabilidad, lo que da como resultado los mismos pesos (vectores columna de $\breve{\mathbf{U}}$), pero las puntuaciones se ven algo afectadas por dicha estimación de σ^2.

Capítulo 4

Técnicas de Clasificación Supervisada

El proceso de *clasificación* estadística o *discriminación* está basado en un *conjunto de prueba* consistente en individuos con etiquetas conocidas, que además incluye un vector de características para cada individuo. El objetivo es dividir a dichos individuos, de acuerdo a una regla basada en los datos, en K clases. Aún así, para explicar la notación tomaremos $K = 2$ clases (caso de *clasificación binario*) por lo que, suponemos los siguientes *datos de prueba*,

$$(\widetilde{\mathbf{x}}_1, \widetilde{y}_1), \cdots, (\widetilde{\mathbf{x}}_n, \widetilde{y}_n), \tag{4.1}$$

donde $\widetilde{\mathbf{x}}_i \in \mathbb{R}^d$ y $\widetilde{y}_i \in \{-1, 1\}$ para $i = 1, \cdots, n$ (la tilde sobre los vectores y escalares indica cantidades aleatorias). Tomemos $(\widetilde{\mathbf{x}}_0, \widetilde{y}_0)$ y el objetivo es encontrar una *regla de clasificación*, usando $\widetilde{\mathbf{x}}_0$, para predecir la etiqueta de la clase correspondiente $\widetilde{y}_0$, es decir, una función $c(\mathbf{x}) : \mathbb{R}^d \to \{-1, 1\}$. Suponemos que los datos de prueba se extraen, de forma independiente, de una distribución de probabilidad conjunta $f(\mathbf{x}, y)$, desconocida. Es útil pensar desde un punto de vista Bayesiano, con una función de pérdida que asigna 0 a una clasificación correcta y 1, en otro caso. Las distribuciones marginales $p_+ = P\{\widetilde{y} = +1\}$ y $1 - p_+ = P\{\widetilde{y} = -1\}$ se denominan *probabilidades a priori*. El rendimiento de un clasificador, c, se evalúa empleando el riesgo de Bayes (pérdida esperada), $P\{c(\widetilde{\mathbf{x}}) = \widetilde{y}\}$. La regla de razón de verosimilitud,

$$c_{LR}(\mathbf{x}) = \begin{cases} +1 & \text{si } \frac{f(\mathbf{x},+1)}{f(\mathbf{x},-1)} \geq 1 \\ -1 & \text{si } \frac{f(\mathbf{x},+1)}{f(\mathbf{x},-1)} < 1 \end{cases} \tag{4.2}$$

es el riesgo de Bayes óptimo, que minimiza dicho riesgo sobre todas las opciones de c. Otro método importante de clasificación general son los *clasificadores lineales*. Dado un vector dirección $\mathbf{w} \in \mathbb{R}^d$ ($||\mathbf{w}|| = 1$) y un intercepto $\beta \in \mathbb{R}$, definimos

$$c_{\mathbf{w},\beta}(\mathbf{x}) = 1 - 2 \cdot 1_{\{\mathbf{w}^t\mathbf{x} < \beta\}}, \tag{4.3}$$

donde $1_{\{\cdot\}}$ denota la *función indicadora*. Debemos tener en cuenta que $c_{\mathbf{w},\beta}$ asigna $\mathbf{x}$ a $+1$ cuando está en el lado positivo (respecto a la dirección de $\mathbf{w}$) del hiperplano cuyo vector normal es $\mathbf{w}$ y el intercepto es β, por lo que hay una *superficie de separación* entre las dos regiones de clases.

La clasificación es un área en la que se pueden aplicar una gran cantidad de métodos disponibles, cada uno con sus respectivas fortalezas y debilidades.

4.1. Métodos Clásicos

Un clasificador muy simple es la *Diferencia de Medias* (DM) (Sección 2.3.2), cuyo objetivo es tomar la clase más cercana a su media, aunque también se puede considerar como un clasificador lineal (4.3). En particular, el hiperplano de separación tiene la recta entre las medias de clase como dirección normal y el intercepto como punto medio de ellas. Un beneficio notacional de las etiquetas de clase ± 1 es que $\widehat{\mathbf{w}} = \frac{\sum_{i=1}^{n} \widetilde{y}_i \widetilde{\mathbf{x}}_i}{\| \sum_{i=1}^{n} \widetilde{y}_i \widetilde{\mathbf{x}}_i \|}$ y $\widehat{\beta} = n^{-1} \sum_{i=1}^{n} \widetilde{\mathbf{x}}_i$, donde el símbolo del sombrero indica una cantidad estimada. Cuando ambas clases son normales con covarianza identidad, la DM es la razón de verosimilitud, por lo que el riesgo de Bayes es óptimo.

En situaciones donde las variables pueden tener escalas diferentes, es sensato reescalar primero cada variable en el conjunto de datos de prueba completo (variante de DM denominada *Naive Bayes* por Domingos y Pazzani (1997)). Este es el riesgo de Bayes óptimo en entornos Gaussianos equilibrados donde la matriz de covarianza es diagonal.

Los métodos DM y Naive Bayes funcionan bien cuando no hay correlación entre las variables, pero se pueden mejorar sustancialmente, como se demuestra en el ejemplo de *Gaussianas Correlacionadas Desplazadas* en la Figura II.26. El panel izquierdo muestra un conjunto de datos de prueba $d = 2$-dimensional con $n_+ = 20$ puntos de clase $+1$, en rojo y $n_- = 20$ de -1 en azul, para un total de $n = n_+ + n_- = 40$ puntos. Los signos $\times$ en un círculo verde indican las dos medias muestrales y el signo más en un círculo es la media general. La recta verde muestra la dirección DM. El hiperplano de separación DM (normal a la dirección DM) se muestra como la recta verde discontinua. En el panel derecho se muestran las proyecciones de los datos de prueba sobre el subespacio (recta) generado por la dirección DM. La superposición sustancial de los datos proyectados indica que esta no es una dirección buena para la clasificación lineal. Una dirección más ortogonal a las nubes de puntos podría ofrecer un rendimiento mucho mejor.

Las ideas anteriores se pueden utilizar para dar una variante gráfica del *Análisis Discriminante Lineal* (ADL). La Figura II.26 sugiere que DM fracasó porque sólo utiliza las medias de clase e ignora la información de la covarianza. ADL fue propuesto por Fisher (1936), quien hizo la observación clave de que cuando se supone que ambas clases son normales con matrices de covarianza comunes, el clasificador de razón de verosimilitud es lineal. Por esta razón, casi todas

las variantes (de 2 clases) de ADL se realizan desde un punto de vista de probabilidad Gaussiana. Esto puede llevar a la impresión errónea de que ADL requiere normalidad para ser efectivo. Para contrarrestar esto, en la Figura II.27 se ofrece una introducción no paramétrica a ADL.

En la Figura II.27 se utiliza el mismo conjunto de datos que en la Figura II.26. Los componentes críticos de ADL son las medias de clase $\overline{\mathbf{x}}_+$ y $\overline{\mathbf{x}}_-$, que se muestran como signos $\times$ encerrados en un círculo en la parte inferior izquierda de la Figura II.27. También son importantes las dos matrices de covarianza muestrales de las clases, $\widehat{\mathbf{\Sigma}}^+$ y $\widehat{\mathbf{\Sigma}}^-$. Bajo el supuesto de que son iguales, tiene sentido estimar una matriz común agrupando las matrices de covarianza anteriores para obtener $\widehat{\mathbf{\Sigma}}^w$, tomando un promedio ponderado (según el tamaño de la clase) de $\widehat{\mathbf{\Sigma}}^+$ y $\widehat{\mathbf{\Sigma}}^-$. $\widehat{\mathbf{\Sigma}}^w$ se denomina *covarianza muestral dentro de la clase*. Esta matriz de covarianzas soluciona los problemas de DM en la Figura II.26, a partir de una operación *esférica*, es decir, transformando cada punto en

$$\widetilde{\mathbf{z}}_i = \left(\widehat{\mathbf{\Sigma}}^w\right)^{-1/2} \widetilde{\mathbf{x}}_i. \tag{4.4}$$

El éxito de esto se ve en el panel superior derecho de la Figura II.27, donde cada clase, de hecho, parece distribuida esféricamente. Esa es la situación ideal para el clasificador DM, que se aplica en el panel inferior derecho. Las medias de clase transformadas, $\overline{\mathbf{z}}^+$ y $\overline{\mathbf{z}}^-$, se muestran como signos $\times$ en un círculo magenta, y la media general se muestra como $+$ en un círculo. Pero, es más revelador realizar la clasificación de los datos de este espacio volviendo al espacio original. Debemos tener en cuenta que el punto medio $+$ en un círculo se transforma nuevamente en $\left(\widehat{\mathbf{\Sigma}}^w\right)^{-1/2}\left(\frac{1}{2}\overline{\mathbf{z}}^+ + \frac{1}{2}\overline{\mathbf{z}}^-\right)$, mientras que el vector normal transformado es proporcional a $\left(\widehat{\mathbf{\Sigma}}^w\right)^{-1/2}(\overline{\mathbf{z}}^+ - \overline{\mathbf{z}}^-)$. Ahora, invirtiendo la transformación esférica (4.4) se obtiene el punto central ADL, $\widehat{\boldsymbol{\mu}}_{ADL} = \left(\frac{1}{2}\overline{\mathbf{x}}^+ + \frac{1}{2}\overline{\mathbf{x}}^-\right)$ y el vector de dirección

$$\widehat{\mathbf{w}}_{ADL} = \frac{\left(\widehat{\mathbf{\Sigma}}^w\right)^{-1}(\overline{\mathbf{x}}^+ - \overline{\mathbf{x}}^-)}{\left\|\left(\widehat{\mathbf{\Sigma}}^w\right)^{-1}(\overline{\mathbf{x}}^+ - \overline{\mathbf{x}}^-)\right\|} \tag{4.5}$$

Por lo tanto, este es un clasificador lineal (4.3) donde el intercepto es el producto interno $\widehat{\beta}_{ADL} = \widehat{\mathbf{w}}^t_{ADL}\widehat{\boldsymbol{\mu}}_{ADL}$. El hiperplano de separación correspondiente se muestra como la recta discontinua magenta en el panel inferior izquierdo.

Por lo general, ADL funciona mejor que DM, salvo en el caso, cada vez más importante, de alta dimensión donde $d > n$. Esto se convierte en un desafío ya que la matriz de covarianzas $\widehat{\mathbf{\Sigma}}^w$ no es invertible. La inversa generalizada de Moore-Penrose (también llamada pseudoinversa) es una versión bien definida de la inversa que consiste en invertir valores propios distintos de cero dejando los iguales a 0 como están. Si bien, proporciona una versión bien definida de ADL, el rendimiento en dimensiones altas es bastante pobre.

Otro método, no muy común, es la dirección de Acumulación Máxima de Datos (AMD, Ahn

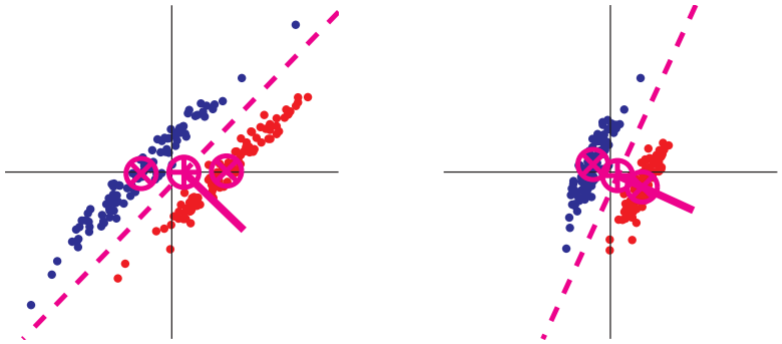

Figura 4.1: *Introducción visual a AMD, usando el conjunto de datos Gaussianas Correlacionadas Desplazadas. Muestra cómo para $d < n$ los resultados generales de la transformación esférica son los mismos que para ADL.*

y Marron (2010)). AMD tiene propiedades bastante diferentes de los dos métodos anteriores. En dimensiones altas la generalización de AMD es bastante comparable a la de la regla de razón de verosimilitud óptima, DM. Además, es sorprendentemente similar a ADL, donde la única diferencia es que la matriz de covarianza dentro de la clase $\widehat{\mathbf{\Sigma}}^w$ se reemplaza por la matriz de covarianza general $\widehat{\mathbf{\Sigma}}$, de modo que

$$\widehat{\mathbf{w}}_{AMD} = \frac{\left(\widehat{\mathbf{\Sigma}}\right)^{-1}(\overline{\mathbf{x}}^+ - \overline{\mathbf{x}}^-)}{\left\|\left(\widehat{\mathbf{\Sigma}}\right)^{-1}(\overline{\mathbf{x}}^+ - \overline{\mathbf{x}}^-)\right\|} \tag{4.6}$$

Además, para $d \leq (n-2)$ los vectores $\widehat{\mathbf{w}}_{ADL}$ y $\widehat{\mathbf{w}}_{AMD}$ son iguales.

En la Sección 2.1 vimos que usar las alturas de los datos transmite más información que emplear unas aleatorias. En la Figura 4.1 se ilustra esto utilizando el mismo conjunto de datos datos *Gaussianas Correlacionadas Desplazadas* que en las Figuras II.26 y II.27, mostrando solo la fila inferior. La *formación de esferas* que se produce en el panel derecho se realiza mediante $\left(\widehat{\mathbf{\Sigma}}\right)^{-1/2}$. En lugar de dividir individualmente cada clase por separado, ahora se divide el conjunto de datos completo. Las dos nubes de puntos todavía son evidentes, pero se han ajustado para que la matriz de covarianza sea la identidad. Además, esta transformación también aborda el desplazamiento de la clase, que era el problema original con DM en la Figura II.26. En particular, las medias muestrales que siguen a esta transformación se muestran como signos $\times$ en un círculo magenta, y el hiperplano de separación se indica con la recta magenta discontinua. El resultado, siguiendo el formato de la Figura II.27, da como resultado la regla de clasificación lineal del panel inferior izquierdo, que es la misma que en el caso ADL.

Como conclusión, observamos que ADL debería sustituir a AMD como clasificador de elección, debido a que son iguales si $n < d$, y en otro caso, AMD es superior y más sencillo de calcular.

Un punto que será relevante en la Sección 4.3 es que se puede considerar que todos los métodos desarrollados en esta sección se basan en distribuciones de probabilidad.

Vale la pena mencionar una extensión de los métodos de clase $K > 2$ de ADL. Ahora el enfoque de verosimilitud ya no es útil, pero está disponible un enfoque basado en la dirección. De las Figuras II.27 y (4.5) queda claro que el vector de dirección ADL apunta en la dirección de la diferencia de medias de clase después de realizar el ajuste apropiado mediante la estructura de covarianza intraclase agrupada. Para medias múltiples, la dirección única entre medias puede ser reemplazada de manera útil por un ACP del conjunto de vectores medios. Esto se calcularía como un análisis de autovalores de la matriz de covarianza del conjunto de medias de clase, que a veces se denomina matriz de covarianza entre clases $\widehat{\mathbf{\Sigma}}^b$. Esto da como resultado un conjunto de vectores de dirección llamado *Análisis de Variantes Canónicos* (ACV) en Mardia et al. (1979) y *Análisis Discriminante Múltiple* en Duda et al. (2001).

4.2. Métodos Kernel

Los métodos Kernel se consideran útiles como un tipo de transformación, cuyo objetivo es hacer que los datos sean más accesibles al análisis lineal. Uno de esos métodos es la *incrustación del núcleo polinomial*, donde los datos se transforman, de forma no lineal, en un espacio de dimensiones superior en el que luego se aplica ADL. La incorporación de polinomios también tiene algunos inconvenientes que incluyen una posible falta de flexibilidad, así como una mala interpretabilidad y una elección no obvia del grado del polinomio. Esto se ilustra en la Figura II.28, que comienza con un conjunto de datos en $\mathbb{R}^2$. El rendimiento del clasificador, basado en cada incrustación polinomial, se muestra tratando cada píxel de fondo como un nuevo punto y clasificándolo como amarillo, que indica una clasificación + (rojo) o el cian, − (azul). Ambos colores se mueven hacia el blanco para los píxeles cercanos al límite. Los datos siguen un patrón de tablero de ajedrez. Cada clase consta de 8 normales estándar de 25 puntos cada uno, espaciados 6 unidades en ambas direcciones. La incrustación cúbica es $(x_{1,i}, x_{2,i}, x_{1,i}^2, x_{2,i}^2, x_{1,i}^3, x_{2,i}^3)$ que se muestra en el panel izquierdo de la Figura II.28, y proporciona resultados de muy pobre desempeño en la clasificación.

En el panel derecho de la Figura II.28 se aprecian mejores resultados. Se utiliza una incrustación muy diferente, utilizando ideas de estimación de la densidad kernel. En particular, los datos están integrados en $\mathbb{R}^{49}$ asignando $(x_{1,i}, x_{2,i})^t$ a

$$\left(\phi\left(\left(x_{1,i} - g_{1,1}\right)/h\right) \cdot \phi\left(\left(x_{2,i} - g_{2,1}\right)/h\right), \cdots, \phi\left(\left(x_{1,i} - g_{1,49}\right)/h\right) \cdot \phi\left(\left(x_{2,i} - g_{2,49}\right)/h\right)\right)^t \quad (4.7)$$

donde ϕ es la densidad normal estándar, donde $(g_{1,1}, g_{2,1}), \cdots, (g_{1,49}, g_{2,49})$ es una cuadrícula de 7×7 de puntos centrales kernel, y donde la dispersión de cada kernel normal está controlado

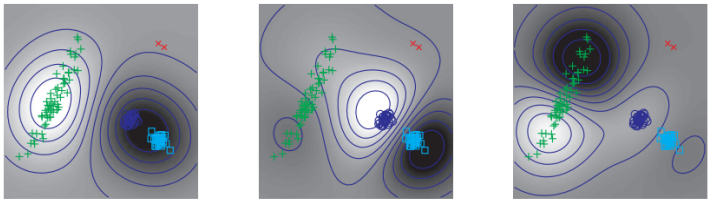

Figura 4.2: *Ejemplo de Cuatro Subgrupos para ACP Kernel. Se muestran los* 3 *primeros modos de variación. El primero (izq.) contrasta el grupo verde con el azul y cian. El segundo (centro) contrasta los grupos azul y cian, y el tercero (der.) separa el grupo verde.*

por el ancho de banda $h = 1$. Esta incorporación da como resultado una clasificación excelente donde cada punto de prueba se clasifica correctamente.

Una variación de la idea de incorporación del núcleo es ACP Kernel, propuesta por Schölkopf et al. (1997). Esto se ilustra en la Figura 4.2, basada en el conjunto de datos *Cuatro Subgrupos*, que es la base de varios ejemplos en el Capítulo 5. Este conjunto de datos tiene dos grupos esféricos mostrados como círculos azules y cuadrados cian, un grupo alargado de signos + verdes, y un grupo de dos puntos que se muestra como signos × rojos. ACP Kernel comienza con ACP convencional realizado en el espacio del kernel, con la visualización realizada en el espacio objeto original, que proviene de proyectar cada punto en él sobre cada vector de dirección y colorear la imagen con la puntuación, usando gris para 0 y negro (blanco) intenso para la magnitud de positivo (negativo, respectivamente).

El primer componente (es decir, el modo de variación, que se muestra en el panel izquierdo) resalta un agrupamiento que proporciona un contraste entre el grupo verde y los otros puntos, al tiempo que trata la unión del azul y el cian como un único componente, mientras que el segundo separa los grupos azul y cian. El tercer modo de variación CP divide el grupo verde en dos grupos. Solo las componentes superiores (no mostradas aquí) conseguirán separar el grupo rojo. El ACP Kernel normal y también otros métodos kernel dependen de la elección del ancho de la ventana. En la Figura 4.2 esto se eligió mediante prueba y error para dar una buena interpretación de los modos de variación.

Aún así, esta visión del ACP Kernel no tiene mucho impacto en el análisis de datos reales, ya que esta visualización requiere un espacio de objetos de baja dimensión. Una visualización mucho más exitosa, basada en ACP Kernel, es el enfoque t-SNE (*t-distribution Stochastic Neighbor Embedding*) de Maaten y Hinton (2008), que consiste en un tipo de inversión del ACP Kernel normal, utilizando un núcleo de Cauchy (es decir, colas muy pesadas). La idea clave es encontrar un conjunto de representantes (de cada objeto de datos) cuya posición se ajuste mejor al conjunto de distancias entre los objetos de datos. En t-SNE, las incorporaciones de los representantes del

núcleo de Cauchy se ajustan a las posiciones de los objetos de datos en el espacio kernel. Las colas del núcleo de Cauchy, dan una representación que mantiene los puntos cerca entre ellos, mientras que aleja los que están menos cerca, lo que acentúa visualmente los grupos, mejorando considerablemente las gráficas, aunque depende de la elección de un parámetro de ajuste. Sin embargo, no resuelve problemas como qué grupos representan la verdadera estructura subyacente.

La Figura II.29 muestra tres aplicaciones de t-SNE al mismo conjunto de datos, *Cuatro Subgrupos*. Los paneles muestran diferentes valores de *perplejidad*, que es un parámetro de sintonización. La perplejidad predeterminada de 30 se muestra en el panel central y brinda una vista razonable de los datos, excepto por la ubicación de los dos puntos rojos en el pequeño grupo aislado. El panel derecho, con perplejidad 60, quizá algo mejor, aunque los grupos azul y cian parecen estar demasiado cerca. La visualización de perplejidad 12 en el panel izquierdo tiene los puntos rojos aún más separados y dos de los puntos azules separados del resto. Un aspecto importante es que la estructura global es bastante arbitraria, pero los grupos tienden a distinguirse de forma razonable, que es el objetivo de t-SNE.

Una cuestión fundamental es cómo se relaciona la incorporación del núcleo con los productos internos de muchos métodos de análisis de datos, como ACP y ADL. En la Figura II.28 se utiliza una *incrustación explícita del kernel* donde primero se realiza la incrustación en el espacio del kernel y luego se aplica explícitamente ADL en ese espacio. Pero existen importantes ventajas computacionales al calcular primero la matriz de productos internos de los objetos de datos y luego asignarlos al espacio del núcleo, lo que se denomina útilmente *incrustación implícita* del núcleo. Sin embargo, la gran flexibilidad de los métodos kernel conlleva un gran peligro de sobreajuste. Además, existe una limitación interesante, que es que en el caso de dimensiones muy altas, los métodos kernel ofrecen el mismo rendimiento de clasificación que los métodos lineales convencionales.

4.3. Máquinas de Vectores de Soporte

Hasta ahora, los métodos vistos en la Sección 4.1, basados en distribuciones de probabilidad, están muy bien desarrollados y se entienden mejor en espacios Euclidianos planos de $\mathbb{R}^d$ y en la Sección 4.2, se enunciaron mejoras en el mapeo de los datos sobre una variedad curva. Aún así, como las distribuciones de probabilidad en variedades están mucho menos desarrolladas, tiene sentido considerar enfoques no probabilísticos para la invención de métodos de clasificación.

Tales consideraciones motivaron la invención de la *Máquina de Vectores de Soporte* (MVS) por Vapnik (1982, 1995). Aunque están dirigidos a variedades curvas, sus características se comprenden mejor en un espacio Euclidiano simple, como el del ejemplo bidimensional de la Figura 4.3. Estos datos son similares a los de *Gaussianas Correlacionadas Desplazadas*, solo que ahora

hay $n_+ = 15$ círculos rojos en la clase + y $n_- = 15$ círculos en la −. Estos datos son *separables*, es decir, existen varios hiperplanos que dividen a las clases, y MVS busca el mejor de esos hiperplanos de separación. La noción de mejor de MVS se ilustra en el panel derecho, donde dado un candidato a hiperplano de separación, los residuos $\widetilde{r}_i$ de la proyección de cada punto $\mathbf{x}_i$ en el plano se muestran como segmentos magentas y MVS escoge el que está más lejos de cada punto, en el sentido de maximizar la más pequeña de estas distancias proyectadas (panel derecho). La distancia minimizadora se llama *margen*, y los planos se muestran como rectas discontinua negras paralelas al hiperplano de separación. Como observamos, hay 3 puntos sobre estas rectas, que serán aquellos cuya distancia es mínima. Se denominan *vectores de soporte* y se muestran con cajas negras.

Para formular MVS como un problema de optimización, consideremos la matriz $d \times n$, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$, así como las etiquetas de clase como un vector $n \times 1$, $\mathbf{y} = [y_1, \cdots, y_n]^t$. Sea $\mathbf{Y}$ la matriz diagonal $n \times n$ con $y_1, \cdots, y_n$ en su diagonal. Dado un vector director unitario $\mathbf{w} \in \mathbb{R}^d$ y un intercepto $\beta \in \mathbb{R}$, el n-vector magenta de longitudes de los residuos se escribe como $\check{\mathbf{r}} = \mathbf{Y}\mathbf{X}^t\mathbf{w} + \beta\mathbf{y}$. En el caso separable, el problema de optimización MVS, se formula como $\underset{\mathbf{w},\beta}{max}\underset{i}{min}$ $\check{r}_i$, que se puede resolver mediante programación cuadrática, introduciendo una nueva variable τ, y maximizarla sujeta a $\check{r}_i \geq \tau$ para $i = 1, \cdots, n$. Dado que $\check{\mathbf{r}}$ es proporcional a $\mathbf{w}$ y β, maximizar τ sujeto a $||\mathbf{w}|| = 1$ es equivalente a minimizar $\mathbf{w}$ sujeto a $\tau = 1$. Cuando los datos no se suponen separables, los residuos se mantienen no negativos mediante el empleo de variables de holgura $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_n]^t$, que se incorporan a la matriz de residuos modificados $\mathbf{r} = \mathbf{Y}\mathbf{X}^t\mathbf{w} + \beta\mathbf{y} + \boldsymbol{\xi}$. Dado un parámetro de ajuste λ, estos residuos no negativos se limitan a ser no negativos en la versión más general de la optimización MVS:

$$\underset{\mathbf{w},\beta,\boldsymbol{\xi}}{min} \left(||\mathbf{w}||^2 - \lambda \mathbf{1}_{1,n}\boldsymbol{\xi} \right) \tag{4.8}$$

sujeto a las restricciones $r_i \geq 0$, $\xi_i \geq 0$ para $i = 1, \cdots, n$. Valores pequeños de λ dan una solución llamada *margen duro*, que intenta poner tantos datos como sea posible fuera de los márgenes. Por el contrario, valores grandes de λ relajan ese objetivo permitiendo más *transgresores* (puntos dentro de los márgenes).

Otro tema importante es la extensión de MVS al caso $K > 2$. Existen dos enfoques comunes. El primero plantea escoger un clasificador de cada clase, compararlos con el resto y elegir la clase con mejor resultado (en términos de proyección sobre el vector dirección MVS). El segundo ejecuta todos los pares de clasificación posibles y escoge el mejor. Por lo general este último es mejor. Aún así, MVS tiene un defecto para altas dimensiones que es la pérdida de efectividad en la clasificación y visualización. Un método de clasificación lineal parecido, motivado por este problema de MVS es la *Discriminación Ponderada por Distancia* (DPD), que se estudia a continuación. Este método soluciona los problemas de clasificación de MVS, además de que las clases

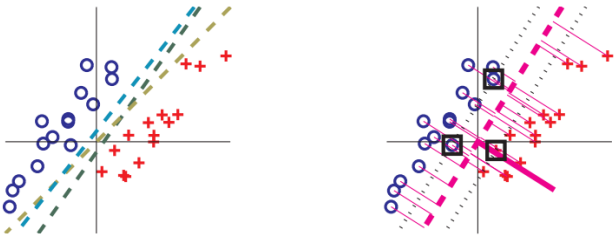

Figura 4.3: *Introducción gráfica a MVS. Se muestran los datos de juguete junto con varios hiperplanos de separación (izq.). A la derecha, el vector de dirección MVS (segmento magenta) y el hiperplano de separación ortogonal (recta discontinua magenta). Las rectas finas magenta son los residuos y los margenes son las líneas de puntos negras. Los vectores de soporte se señalan en cajas negras.*

proyectadas tienen una distribución aproximadamente normal.

4.4. Discriminación Ponderada por Distancia

El problema de efectividad de MVS puede solucionarlo DPD. Además su uso es recomendable para tareas de visualización donde la atención se centra en las diferencias potenciales entre pares de grupos de datos. DPD consigue esta mejora mediante una modificación del problema de optimización MVS en (4.8).

En el caso separable, la idea es modificar la forma en que los residuos, $\check{r}_i$, impactan en el resultado, permitiendo que todos ellos influyan, no solo los más pequeños. Éstos jugarán un papel importante, a partir del estudio de la inversa de los residuos, $\frac{1}{\check{r}_i}$. Así, a medida que se aproxime a 0, la inversa tiene un valor grande, por lo que su suma empuja el hiperplano de separación hacia el centro de la región entre los conjuntos de datos. De este modo, la versión separable de DPD usa el problema de optimización $\underset{\mathbf{w},\beta}{min} \sum_{i=1}^{n} \frac{1}{\check{r}_i}$. Utilizando el mismo enfoque de variable de holgura que MVS, el problema de optimización DPD es, dado un parámetro de ajuste λ

$$\underset{\mathbf{w},\beta,\boldsymbol{\xi}}{min} \left(\sum_{i=1}^{n} \frac{1}{\check{r}_i} + \lambda \mathbf{J}_{1,n} \boldsymbol{\xi} \right) \tag{4.9}$$

sujeto a $\check{r}_i \geq 0$, $\xi_i \geq 0$ para $i = 1, \cdots, n$. Esto es una versión lineal de DPD, aunque también existen versiones apropiadas kernel.

La solución numérica de (4.9) motivó el desarrollo de un algoritmo llamado FastDWD por Lam et al. (2018), que utiliza un enfoque basado en *método de multiplicadores de dirección alterna* y es más efectivo que otras implementaciones de MVS. Otro problema es la selección del parámetro de ajuste λ. Marron et al (2007) recomendó la elección de $\lambda = \frac{100}{d_i^2}$, donde 100 se

considera como número grande y d_t es una noción útil de escala de los datos calculados como $d_t = median\{||\mathbf{x}_i - \mathbf{x}_{i'}|| : y_i = +1, y_{i'} = -1\}$.

En la Figura II.30 se muestra un ejemplo de juguete que plantea una situación del área de ajuste por lotes. Los + representan los $n_+ = 200$ datos de un laboratorio y los círculos los $n_- = 200$ del otro. Los colores distinguen dos grupos, cuya diferencia es el foco del experimento. Otra diferencia es el tamaño de los subtipos con un desequilibrio de 4 : 1. Para compensarlo, podemos restar la media de los datos de cada laboratorio, pero en realidad esto causa más problemas. Restar la media de cada laboratorio se corresponde a unirlos desplazándose uno hacia el otro a lo largo de la dirección DM (recta verde). Esto se muestra en el panel inferior derecho. Vemos que este ajuste es claramente insatisfactorio debido a que la diferencia entre los colores se reduce hasta el punto de superponerse.

Se obtiene un mejor rendimiento mediante el uso de DPD. La dirección DPD se muestra en el panel izquierdo de la Figura II.30 como la recta rosa discontinua, que está mucho más próxima a la dirección vertical ideal. Los resultados de deslizar los laboratorios a lo largo de la dirección DPD se muestran en el panel superior derecho. Se aprecian mejores resultados ya que ahora existe una separación sustancial entre los colores. La dirección DPD funciona mejor para el ajuste de lotes porque está impulsada por el hiperplano de separación DPD (recta rosa). Los $\frac{1}{\tilde{r}_i}$ de cada punto empujan este hiperplano fuera, resultando en una dirección mucho mejor para el ajuste de lotes.

Wang y Zou (2018) desarrollaron un *DPD generalizado* como una variante de DPD con importantes ramificaciones para ajuste de lotes y robustez contra la heterogeneidad, donde el impacto de $\frac{1}{\tilde{r}_i}$ se controla usando una potencia q, $\frac{1}{\tilde{r}_i^q}$. Así, valores grandes de q magnifican el efecto ilustrado en la Figura II.30, en el sentido de empujar la recta rosa más hacia la recta horizontal. Esto resulta en que la dirección de separación se aproxima más a la dirección vertical óptima, es decir, mejora de la robustez. También proporciona una conexión entre DPD y MVS que es esencialmente el límite $q \to \infty$.

Capítulo 5

Técnicas de Clasificación No Supervisada

La *agrupación* (o técnicas de clasificación no supervisada) es una operación muy útil para muchos propósitos, cuya idea principal es resaltar subconjuntos de datos que, en algún sentido, tienen algo en común. Comparte la idea de clasificar los datos como en el Capítulo 4, pero con la diferencia de que aquí el objetivo es determinar las etiquetas de clases. Existen muchos métodos y variantes de estas técnicas, aunque en este Capítulo solo veremos algunos.

5.1. Agrupación en K-medias

Un enfoque muy intuitivo de agrupación es el llamado *k-medias* propuesto por Steinhaus (1956). Dado un conjunto de datos aleatorio $\widetilde{\mathbf{x}}_1, \cdots, \widetilde{\mathbf{x}}_n$ en $\mathbb{R}^d$, la idea principal consiste en elegir conjuntos de índices de agrupación $C_1, \cdots, C_k$ que *particionen* al total de índices $\{1, \cdots, n\}$ (es decir, cada índice está contenido en exactamente uno de los C_j), de forma que minimice la *Suma de Cuadrados dentro del Grupo* (SCDG),

$$SCDG = \sum_{j=1}^{k} \sum_{i \in C_j} ||\widetilde{\mathbf{x}}_i - \overline{\mathbf{x}}_j||^2, \tag{5.1}$$

donde cada media del grupo se denota por $\overline{\mathbf{x}}_j = \frac{1}{\#(C_j)} \sum_{i \in C_j} \widetilde{\mathbf{x}}_i$. Además, se puede reescalar dividiendo por el total de los residuos generales al cuadrado de la media (STC),

$$STC = \sum_{i=1}^{n} ||\widetilde{\mathbf{x}}_i - \overline{\mathbf{x}}||^2, \tag{5.2}$$

donde $\overline{\mathbf{x}}$ es la media total del conjunto de datos. Este reescalado resulta en el *Índice de Conglomerado*

$$IC = \frac{SCDG}{STC} \tag{5.3}$$

Del panel derecho de la Figura II.31, vemos que si $IC = 0$ entonces cada punto está en la media de su grupo, y si $IC = 1$ todas las medias de grupos son iguales a la media total. Valores pequeños de IC indican grupos "próximos" o parecidos. Algunas propiedades de IC se ilustran en la Figura II.32, usando un conjunto de datos de una dimensión, consistente en 4 grupos normales, dos de los cuales tienen 5 puntos centrados en ± 20, respectivamente, y los otros 2, 1000 puntos centrados en ± 2 (asteriscos negros). El comportamiento de IC se muestra en la curva roja, cuyo valor es la altura de IC de los grupos determinado por la coordenada horizontal. Por ejemplo, la recta azul discontinua divide al primer grupo (conjunto de puntos a la izquierda) del segundo (los de la derecha). El IC de este grupo es 0,84 determinado por la altura del círculo azul. A medida que movemos la recta azul horizontalmente, el círculo se desplaza a lo largo de la recta roja. Tenemos que $IC = 1$ en los extremos porque $SCDG = STC$ (uno de los grupos es el vacío). Hay dos picos en ± 2, ya que los puntos del grupo contribuyen a ambos términos de $SCDG$, lo que impacta en ambas medias del grupo. En la región central, ambas medias de grupos están esencialmente en el medio de los respectivos grupos, lo que hace que ambos términos del $SCDG$ sean mucho más pequeños. Esto muestra cómo IC puede tener muchos mínimos locales, y la posibilidad de que los métodos se bloqueen buscando a éste. Si esto ocurre con el caso más simple, es esperable que se complique para dimensiones mayores.

El algoritmo estándar de k-medias comienza con un conjunto de candidatos a medias de grupos, para luego iterar a través de la asignación de cada punto a la media más próxima, seguido de un nuevo cálculo de las medias de grupos. Este proceso iterativo converge, normalmente, a un óptimo local. Planteamos un ejemplo para ilustrar las propiedades de la agrupación por k-medias, recordando el conjunto de *Cuatro Subgrupos* del Capítulo 4 que se ilustra en las Figuras II.33 y II.31. El panel superior izquierdo muestra el caso $k = 2$. La partición óptima en 2 medias pone los + verdes en un grupo y los restantes en el otro. Para $k = 3$ (superior derecho) muestra las 3 etiquetas para este conjunto de datos, que los separa en 3 grupos, que son los + verdes, cuadrados cian y círculos azules, solo que ahora este último es magenta para reflejar la unión de los círculos azules y las cruces rojas.

Cabría esperar que para $k = 4$ se muestren los colores originales de la Figura II.31, pero como observamos se produce una división entre el grupo verde (el IC es muy pequeño) mientras que los otros grupos se mantienen igual. Tenemos que llegar a $k = 5$ para que las cruces rojas sean un grupo, con la desventaja de que los círculos y los cuadrados se consideran un único grupo. Esto revela un aspecto importante de las agrupaciones de k-medias, que es que los resultados pueden ser impredecibles. Por tanto su uso se debe complementar con confirmaciones visuales, como ACP. Otra desventaja es que los grupos no necesitan estar relacionados para un k y $k + 1$

dado. Este problema lo solventa el siguiente método que planteamos.

5.2. Agrupaciones Jerárquicas

La agrupación jerárquica proporciona un dendrograma completo de agrupaciones, como vemos en la Figura II.34 usando otra vez el conjunto de datos de *Cuatro Subgrupos*. Un dendrograma se puede crear de dos formas.

La primera se crea de forma *divisiva* o *arriba-abajo*. Comenzamos con el conjunto total de datos en un grupo, seguido de una serie de divisiones binarias, hasta llegar a grupos de un solo elemento. Tales divisiones se representan con rectas horizontales, mientras que las verticales representan los subgrupos. La longitud de éstas indican la fuerza de un grupo. Los colores se corresponden con los grupos de la Figura II.31, y vemos que en la primera división, el primer grupo se corresponde con las cruces rojas. La siguiente es entre el grupo verde y la unión de cian y azul, que son divididos a continuación. Las rectas verticales que separan estos grupos son grandes porque hay pocas similitudes entre ellos (cian y azul son los más similares por lo que son las más cortas).

Otra manera de crear un dendrograma es de forma *aglomerativa* o *abajo-arriba*, en la que comenzamos únicamente con los datos, y secuencialmente los vamos combinado en pares de grupos hasta que solo nos queda un grupo, el total. Ambas maneras de razonar resultan en el mismo dendrograma, pero existe una diferencia importante que es el coste computacional.

Hay muchas formas de construir dendrogramas jerárquicos, la mayor parte están indexados por una *distancia* (entre objetos de datos) y por *enlaces*. Para ver la relación que existe entre distintos enlaces, estudiaremos el siguiente ejemplo usando los datos *Gaussianas de Alta-Dimensión*, en la Figura II.35. El conjunto de datos consiste en $n = 30$ vectores de dimensión $d = 500$ simulados de la distribución normal estándar. La agrupación en ambos paneles está basada en la Distancia Euclidiana, con enlaces únicos en el panel izquierdo, que esencialmente toma la distancia entre grupos como la distancia mínima por pares entre puntos. En el panel derecho se aplica el enlace de Ward (Ward (1963)), que aplica divisiones iguales (minimizando SCDG como en k-medias).

El análisis de enlace único comienza dividiendo un elemento, con los restantes formando el otro subgrupo. Repetimos este proceso a lo largo de todo el dendrograma. Como el objetivo no es conseguir muchos grupos de un elemento, este método no es muy efectivo en altas dimensiones. Por otra parte, el enlace de Ward muestra un comportamiento diferente. Este análisis favorece divisiones relativamente equilibradas, que pueden ser más útiles, dependiendo de la aplicación. De todas formas, es muy recomendable aplicar varios tipos de enlaces y quedarse con el que

obtengamos mejores resultados visuales.

Más información sobre este comportamiento de agrupación proviene de la vista de matriz de diagramas de dispersión del conjunto de datos en la Figura II.36, donde las direcciones son escogidas para destacar los dos primeros grupos de un elemento (líneas verticales negras de la Figura II.35). Para ello elegimos los ejes, que son esencialmente vectores unitarios apuntando en la dirección de esos dos puntos (se muestran como círculos negros). Sorprendentemente, estos puntos parecen más alejados del resto para ser datos normales. En la tercera fila, para contrastar, encontramos las puntuaciones CP1 (es decir, las proyecciones sobre el autovector CP1). Obviamente la desviación estándar de estas puntuaciones es claramente más grande en la dirección de estos subgrupos. Ahora planteamos otra matriz de diagramas de dispersión de los mismos datos basada en los enlaces de Ward (Figura II.37). Los colores son los mismos que los del dendrograma de la Figura II.35. Utilizamos la dirección MD como eje para visualizar estos grupos. Otra vez visualizamos en la tercera fila las puntuaciones CP. Como antes, la desviación típica estándar en la dirección CP1 es grande para cualquiera de esas dos diferencias de grupos.

En resumen, la agrupación jerárquica es un conjunto muy flexible de métodos con un gran potencial para descubrir grupos interesantes. Sin embargo, como no hay un mejor método de agrupación como tal, se necesitan conocimientos previos para usarlos de forma efectiva. Como hay tantas elecciones y opciones disponibles, es necesario el análisis confirmatorio.

5.3. Métodos Basados en Visualización

Para comprender este tipo de métodos plantearemos el siguiente ejemplo basado en el conjunto de datos *Fujo Máximo*, proporcionado por Enrica Bellone del Centro Nacional de Investigación Atmosférica, en la Figura II.38. Cada curva representa una nube. La fila superior muestra la operación de centrado de objetos inicial, donde las curvas se muestran a la izquierda, su media en el centro y sus residuos a la derecha. El primer modo de variación se muestra en el panel inferior izquierdo, mostrando predominancia hacia la izquierda. El panel del centro muestra la curva media, además de la curva máxima (discontinua) y mínima (puntos) de las curvas del panel inferior izquierdo. Debido a la naturaleza inclinada de la media, se observa en la gráfica que este modo de variación tiene en cuenta la altura de las curvas y la localización de sus picos. Las puntuaciones se muestran en el panel inferior derecho, del que observamos 3 picos, lo que puede sugerir la existencia de grupos en este conjunto de datos.

A pesar de todo, los métodos visual para encontrar grupos tiene sus límites, por lo que en situaciones complicadas, es útil profundizar, es decir, tomar cada grupo que se encontró hasta el momento y analizarlos cada uno por separado.

Capítulo 6

Ilustración sobre datos simulados

En este último capítulo aplicaremos los métodos explicados a un conjunto de datos real de R, llamado *aemet*, que lo encontramos en el paquete *fda.usc*. Este conjunto de datos consiste en el estudio del clima español desde 1980 hasta 2009, en 73 estaciones meteorológicas a lo largo del país. Así, cada estación proporciona información geográfica y la media de temperatura, precipitación y velocidad del viento diaria. Para los años bisiestos se calcula la media entre el 28 y 29 de Febrero para que coincida el número de días del año. Para este ejemplo, trataremos de clasificar las estaciones meteorológicas en función de la temperatura diaria.

Nuestro objeto de datos serán las curvas medias de temperatura diaria media desde 1980 hasta 2009, donde cada una de ellas representa una estación meteorológica. Así, el espacio de objetos correspondiente sería el conjunto de las curvas, 73 en total. El espacio característico será de dimensión 365, una por cada día del año. Si representamos los datos originales, como se muestra en Figura II.39, no observamos ningún patrón de color, ya que cada estación tiene un color aleatorio asociado. Podemos ver si existe una relación entre la temperatura y la provincia en la que se encuentra la estación. Para ello, aplicamos la técnica de *brushing*, es decir, organizamos las estaciones por colores, donde cada color representa una provincia, es decir, un subgrupo. Al representar ahora los datos, como vemos en la Figura 6.1, observamos cierta estructura. Las curvas azules y rosa fuerte (*deep pink* en R), correspondientes a las estaciones de Santa Cruz de Tenerife y las Palmas de Gran Canaria, tienen una temperatura con menos variación que el resto, ya que tienen un clima tropical. Son las estaciones más alejadas del resto de la Península, por lo que, aparentemente, tendríamos dos subgrupos. Sin embargo, una de estas estaciones está considerablemente por debajo de las restantes. Esto podría explicarse por la altura de ésta. Para ello representamos las alturas de las estaciones en la Figura II.40. Así, el punto más alto es el correspondiente a Izaña, el Instituto de Astrofísica de Canarias, que se encuentra a 2390 metros de altura, poniendo de relieve el efecto obvio de la altura en la temperatura. El resto de curvas

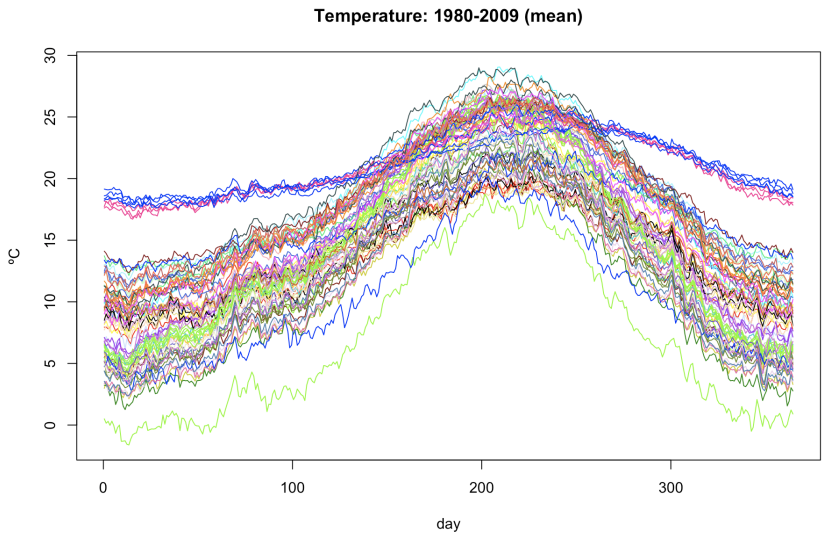

Figura 6.1: *Se representan los mismos datos que en la Figura II.39, aplicando una clasificación por colores según la provincia de la estación meteorológica*

tienen una estructura similar, de acuerdo al clima, mayoritariamente, mediterráneo de España, estando la mínima en Madrid, en concreto en el Puerto de Navacerrada (altura de 1898 metros). A continuación, calculamos la media de estas curvas (Figura II.41, izq.) y sus residuos respecto a la media (Figura II.41, der.). Vemos que las curvas asociadas Tenerife y Las Palmas son las que presentan residuos mayores en los extremos debido a su clima elevado pero con poca variación. También vemos que la forma de las curvas está determinada totalmente por la media.

Aplicamos ahora ACP a nuestro conjunto de datos y representamos los resultados en la Figura II.42. Probamos con las dos primeras componentes principales, que explican el 98,78 % de la variabilidad total (CP1, 85,57 %, CP2, 13,21 %). La tercera componente principal explica menos del 1 % (0,46 %), por lo que nos quedamos con las dos primeras componentes. La gráfica de la derecha representa los pesos de CP1 (negro) y CP2 (rojo). La de la izquierda a las puntuaciones CP1 respecto a los de CP2. Apreciamos una separación razonable en varios subgrupos. Vemos uno hacia la derecha, compuesto de puntos azules y rosas, que se corresponden con las estaciones de las Islas Canarias. Serán un subgrupo diferenciado al tener, como ya dijimos, un clima diferente del resto del país. En la parte de arriba de la gráfica observamos otra nube de puntos, en la que estarían las estaciones correspondientes a A Coruña, Guipúzcoa, Vizcaya. Todas estas estaciones están en el norte, que tenían un clima más cercano al atlántico que el mediterráneo. El resto de puntos parece que se agrupan en torno a una recta que va desde el segundo cuadrante al cuarto, y otra agrupación de puntos situada a la derecha, aunque esta separación no es tan clara como las anteriores.

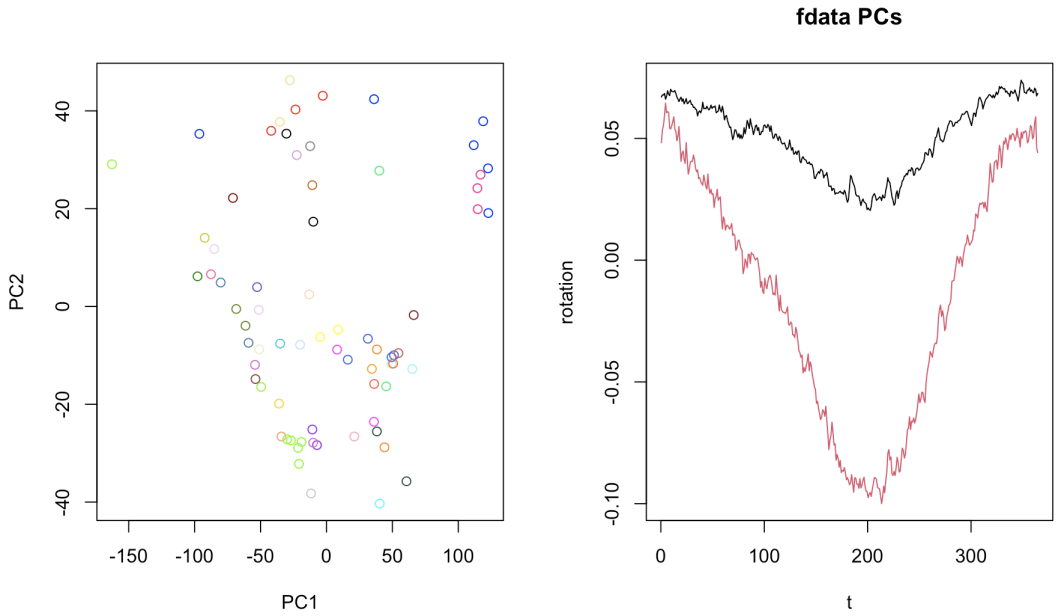

Figura 6.2: *En la izquierda, gráfica de CP1 respecto a CP2 de nuestro conjunto de datos. Se aprecia una separación razonable, pero no perfecta. A la derecha se representan los pesos de CP1 y CP2.*

Finalmente, aplicamos a nuestro conjunto de datos el algoritmo de *k-medias*. Probamos distintas particiones de medias, comenzando por $k = 2$ (Figura II.42). La partición óptima en 2 medias divide el conjunto de curvas en las consistentes en los centros de Tenerife y Las Palmas (a excepción de la de Izaña) y las restantes, en la que su curva media está situada en la curva mínima, que recordemos era la correspondiente al Puerto de Navacerrada, Madrid. Esto es congruente con lo que analizamos anteriormente, es decir, que esos centros presentan un clima tropical, distinto al de los otros centros meteorológicos. Continuamos con $k = 3$, como vemos en la Figura II.43. En este caso se produce un desdoble en el grupo de los centros de clima tropical, dividiéndolas en dos subgrupos. Además, la curva verde cambia ligeramente respecto al caso anterior. La curva roja no varía. Para $k = 4$ (Figura II.44), se vuelve a unificar a los centros de clima tropical en una única curva media, mientras que el resto se desdoblan en tres subgrupos. Ahora se produce una partición más clara entre el resto del territorio, con la curva azul dividiendo provincias como Castellón, Almería, Valencia, con temperaturas más elevadas, frente a la curva roja, con Guipúzcoa, A Coruña, con un clima más frío, muy similar al atlántico. Vemos que en este caso, el centro de Izaña, se presenta como curva media (en cian). Por último, para $k = 5$, que se muestra en la Figura 6.3, se produce una separación entre las curvas que tiene sus extremos más despegados del resto, es decir, que tienen temperaturas más elevadas en los meses más fríos, como Mallorca, Cádiz. Se sigue manteniendo a las curvas asociadas a Tenerife y Las Palmas como un subgrupo, así como las que estaban a mayor altitud, mencionadas anterior-

mente. Las curvas roja y magenta separan a las curvas que presentan un clima similar, pero con diferentes temperaturas. En la roja, estarían contenidas provincias como Castellón o Barcelona, con clima más mediterráneo, mientras que en la magenta, estarían Asturias o Pontevedra, con clima atlántico.

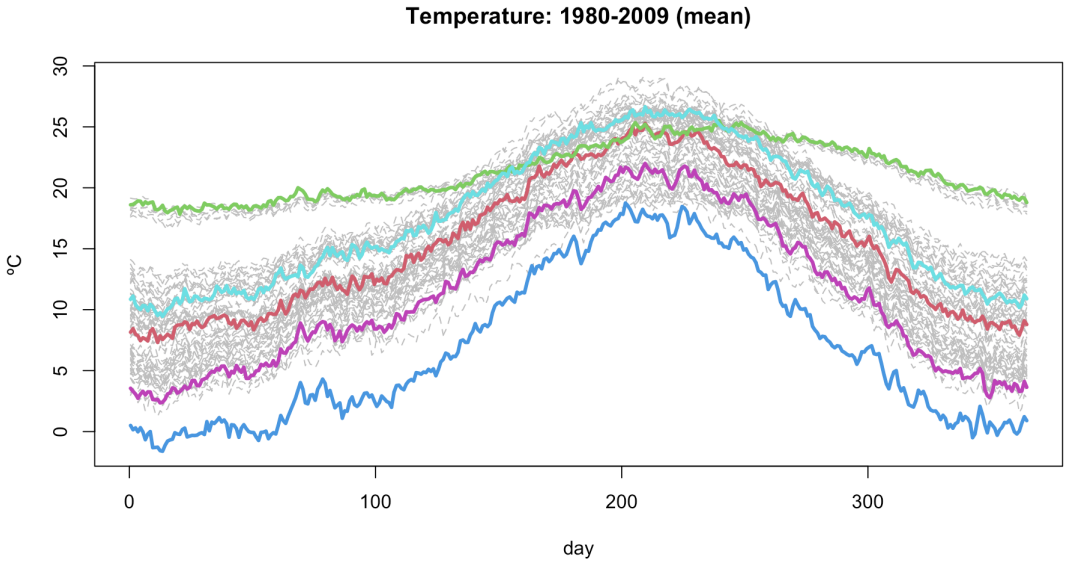

Figura 6.3: Agrupación en k-medias para $k = 5$.

Anexo I

Notación Matemática

Introducimos la notación que será utilizada a lo largo de este trabajo.

- El conjunto de los números reales, $\mathbb{R}$
- El espacio Euclidiano d-dimensional estándar de vectores columna,

$$\mathbb{R}^d = \left\{ \mathbf{x} : \mathbf{x} = (x_1 \ \dots \ x_d)^t, \ x_1, \dots, x_d \in \mathbb{R} \right\} \tag{I.1}$$

- Indicaremos la matriz traspuesta con el superíndice t.
- La esfera cuya superficie dimensional es d (esencialmente los puntos de la superficie de la esfera unidad sólida en $\mathbb{R}^{d+1}$),

$$\mathbb{S}^d = \left\{ \mathbf{u} \in \mathbb{R}^{d+1} \ : \ \|\mathbf{u}\|_2 = 1 \right\} \tag{I.2}$$

- La norma L^p en $\mathbb{R}^d$,

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^{d} |x_j|^p \right)^{\frac{1}{p}} \tag{I.3}$$

- El producto cartesiano, $\times$.
- El conjunto de matrices $d \times n$,

$$\mathbb{R}^{d \times n} = \left\{ \mathbf{X} \ : \ \mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{d,1} & \cdots & x_{d,n} \end{bmatrix}, \ x_{1,1}, \cdots, x_{d,n} \in \mathbb{R} \right\} \tag{I.4}$$

- Dada una matrix de datos $d \times n$, $\mathbf{X}$, la matriz de covarianzas muestral,

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \widehat{var}_1 & \widehat{cov}_{1,2} & \cdots & \widehat{cov}_{1,d} \\ \widehat{cov}_{2,1} & \widehat{var}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \widehat{cov}_{d-1,d} \\ \widehat{cov}_{d,1} & \cdots & \widehat{cov}_{d,d-1} & \widehat{var}_d \end{bmatrix} \tag{I.5}$$

  donde $\widehat{var}_i$ denota la varianza muestral y $\widehat{cov}_{i,i'}$ la covarianza muestral de las filas i e i' de la matriz $\mathbf{X}$, definida como

$$\begin{aligned} \widehat{var}_i &= \frac{1}{n}\sum_{j=1}^{n}(x_{i,j} - \overline{x}_{i,A})^2 \\ \widehat{cov}_{i,i'} &= \frac{1}{n}\sum_{j=1}^{n}(x_{i,j} - \overline{x}_{i,A})(x_{i',j} - \overline{x}_{i',A}) \end{aligned} \tag{I.6}$$

  con

$$\overline{x}_{i,A} = n^{-1}\sum_{j=1}^{n} x_{i,j} \tag{I.7}$$

  para $i = 1, \cdots, d$.

- Dado cualquier espacio métrico con distancia δ, un subconjunto $\mathcal{S}$, y un elemento $\mathbf{x}$, la proyección de $\mathbf{x}$ sobre $\mathcal{S}$ es el punto más próximo en $\mathcal{S}$ a $\mathbf{x}$, es decir,

$$P_{\mathcal{S}} = argmin_{s\in\mathcal{S}}\delta(\mathbf{s}, \mathbf{x}). \tag{I.8}$$

- Definimos la norma de Frobenius $||\cdot||_F$, que es la raíz cuadrada de la suma de las entradas de las matrices al cuadrado:

$$||\mathbf{X}||_F = \sqrt{traza(\mathbf{X}^t\mathbf{X})} = \left(\sum_i \sum_j x_{i,j}^2\right)^{1/2} \tag{I.9}$$

  Definimos también el producto interno de Frobenius en $\mathbb{R}^{d\times n}$ como sigue:

$$\langle \mathbf{M}, \mathbf{N}\rangle_F = \sum_{i=1}^{d}\sum_{j=1}^{n} M_{i,j}N_{i,j} \tag{I.10}$$

Anexo II

Gráficas

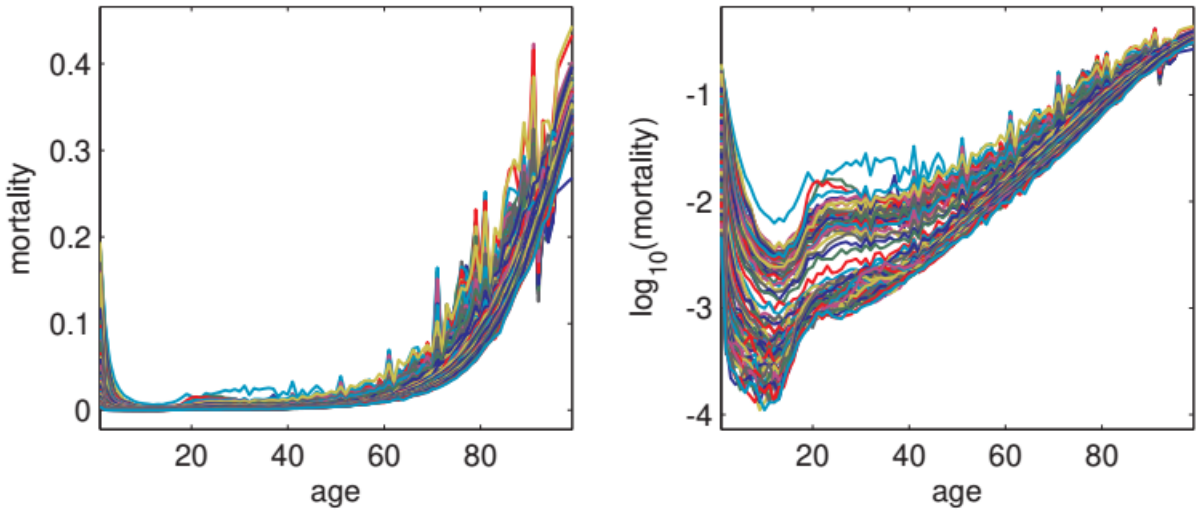

Figura II.1: Curvas del conjunto de datos *Mortalidad Española*. Los datos originales se muestran en el panel izquierdo, aplicando una transformación logarítmica en el derecho.

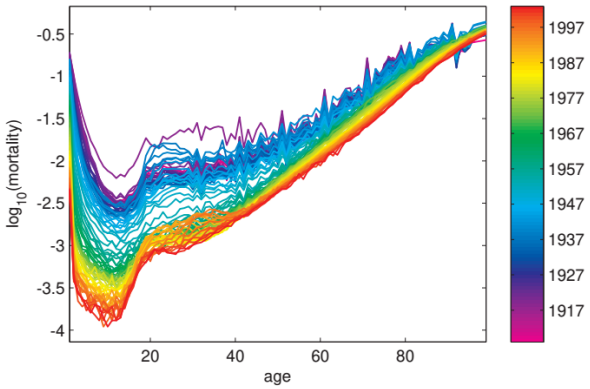

Figura II.2: Curvas del conjunto de datos *Mortalidad Española* ahora utilizando un esquema de color que indicando el paso de los años (de 1908 a 2002). Se aprecia una mejora de la mortalidad a lo largo del tiempo.

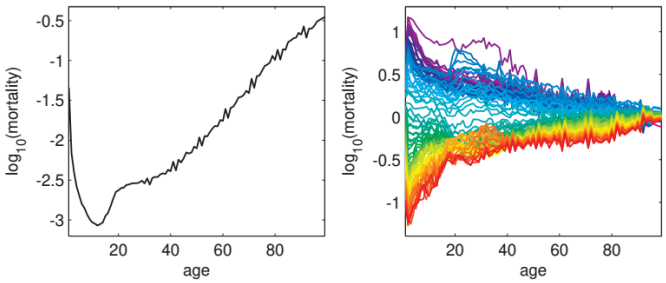

Figura II.3: En la izquierda tenemos la curva de mortalidad media. A la derecha tenemos los residuos de la media. Muestra que los efectos de la edad son los mismos a lo largo del tiempo. Se aprecia una mejora de la mortalidad, especialmente en los jóvenes.

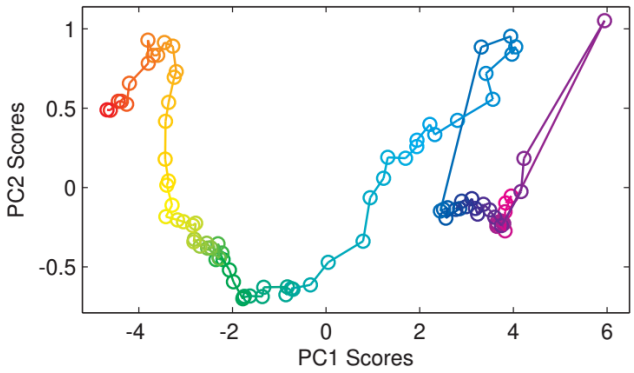

Figura II.4: Diagrama de dispersión de las puntuaciones CP1 y CP2 para el conjunto de datos *Mortalidad Española*.

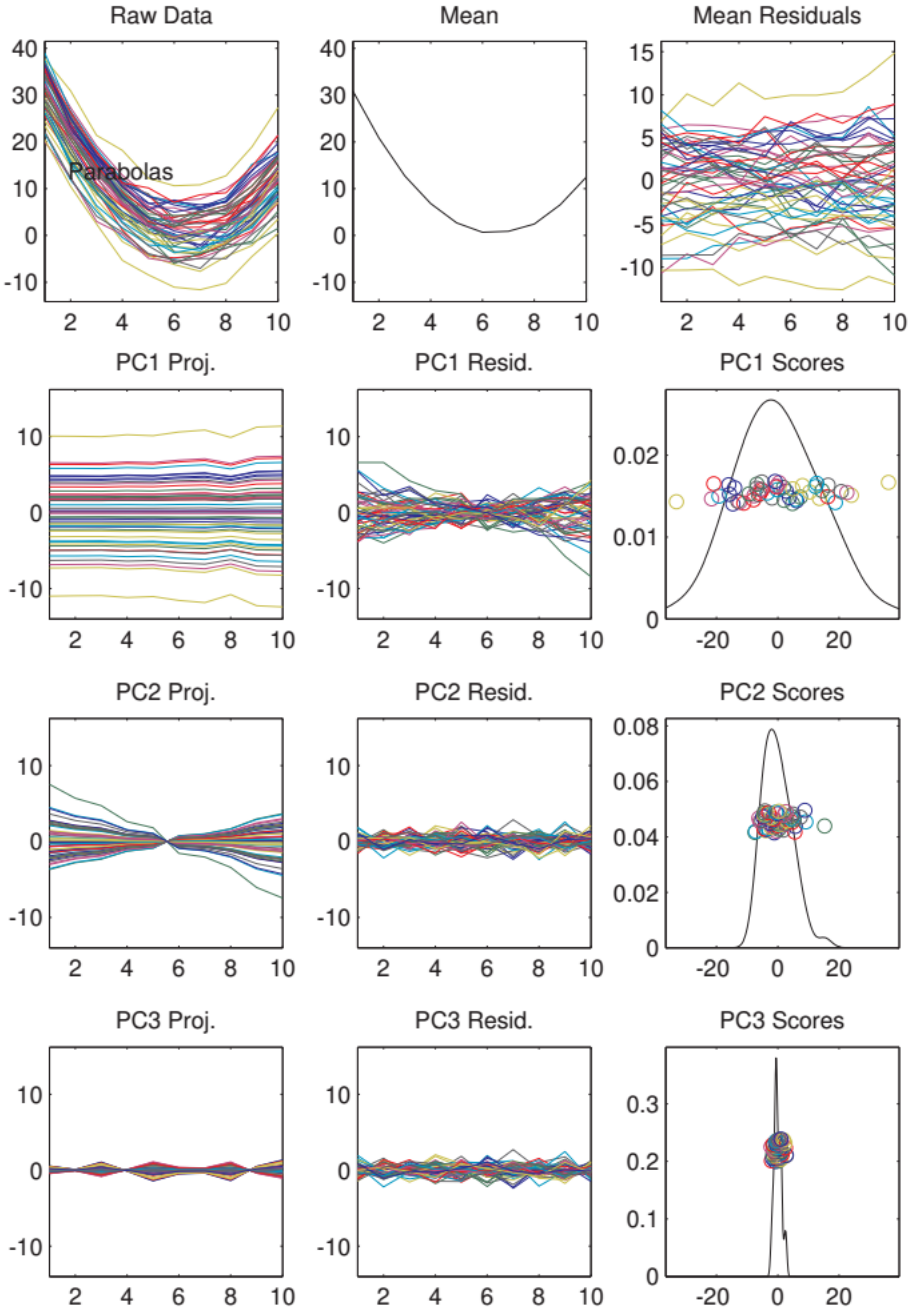

Figura II.5: Ejemplo de juguete *Parábolas Inclinadas* en 10-d que ilustra el concepto de modos de variación. La primera fila muestra las curvas de datos a la izquierda, su media en el centro y los residuos de la media a la derecha. Las filas restantes muestran las componentes CP, con las gráficas de modos de variación (proyecciones) a la izquierda, sus residuos en el centro, y las distribuciones de las puntuaciones (coeficientes de proyección) a la derecha.

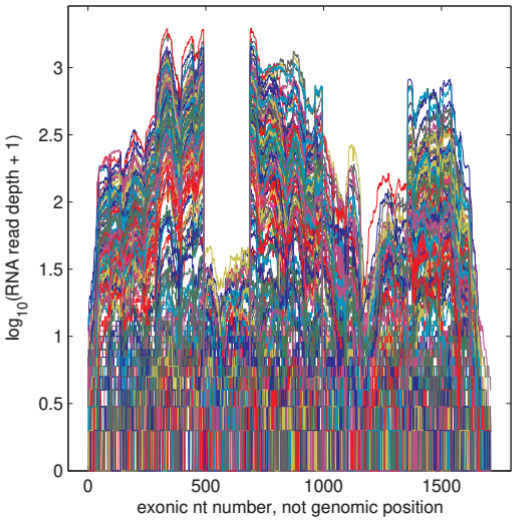

Figura II.6: Curvas $\log_{10}$ de recuento del conjunto de datos *Lung Cancer RNAseq*. Los colores son los estándar del programa informático por lo que no se distingue ninguna estructura poblacional.

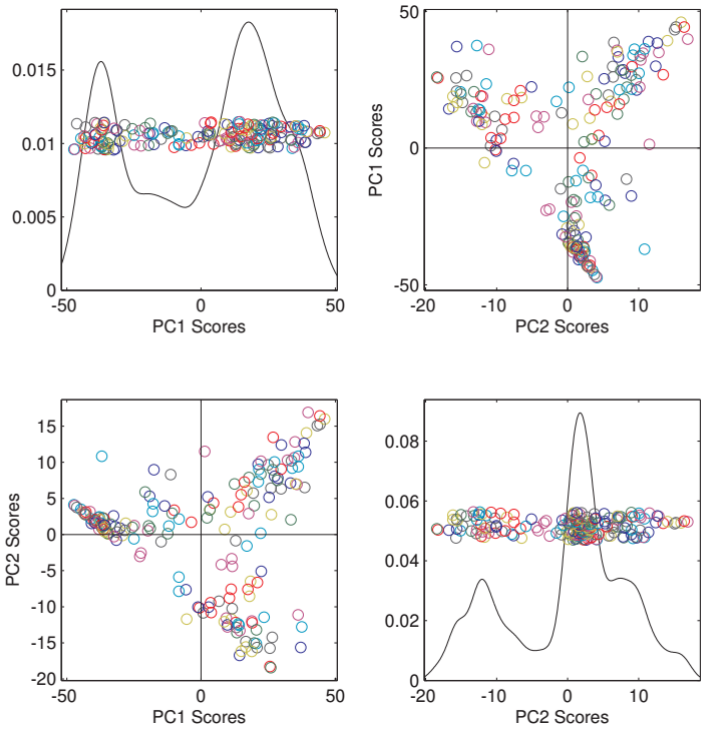

Figura II.7: Matriz de diagramas de dispersión de los datos *Lung Cancer RNAseq*. Se representa el espacio de características de los datos en el que cada círculo se corresponde con una curva de la Figura II.6. Se pueden apreciar tres grupos.

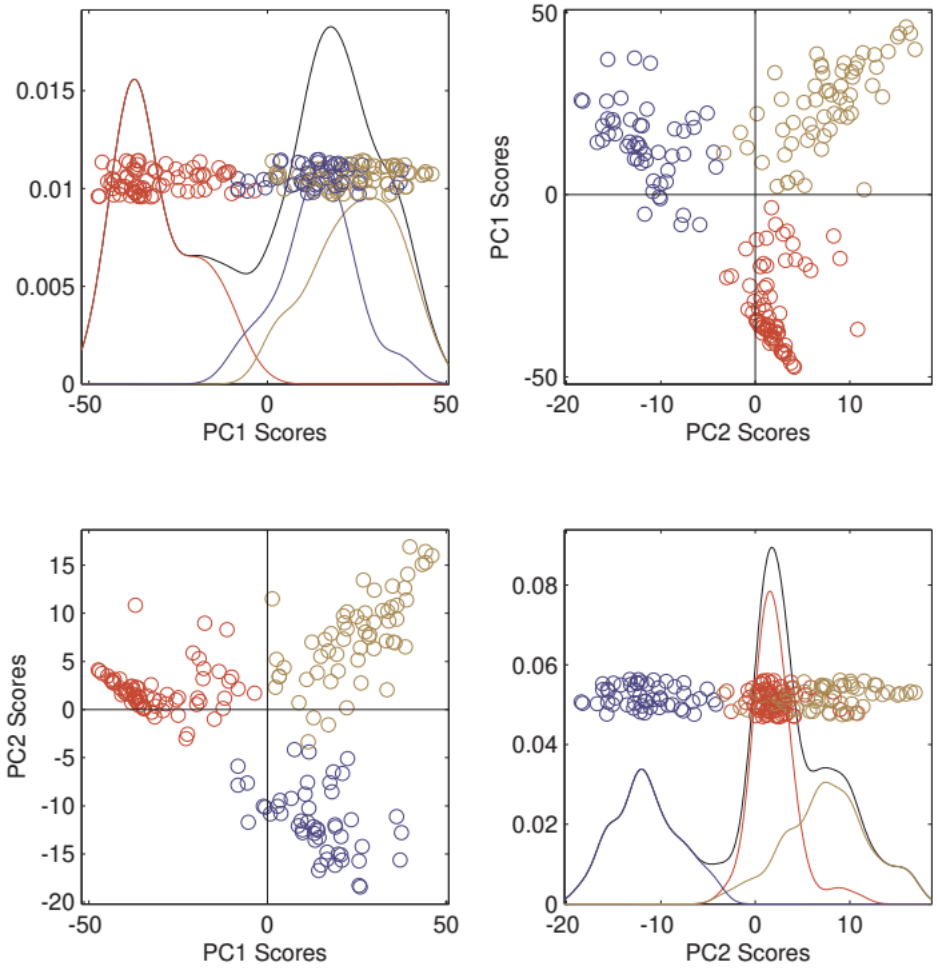

Figura II.8: Matriz de diagramas de dispersión de los datos *Lung Cancer RNAseq* aplicando la técnica de *brushing*. Se han empleado tres colores distintos para clasificar los grupos, además en las distribuciones de la diagonal también se muestran estimaciones de las subdensidades.

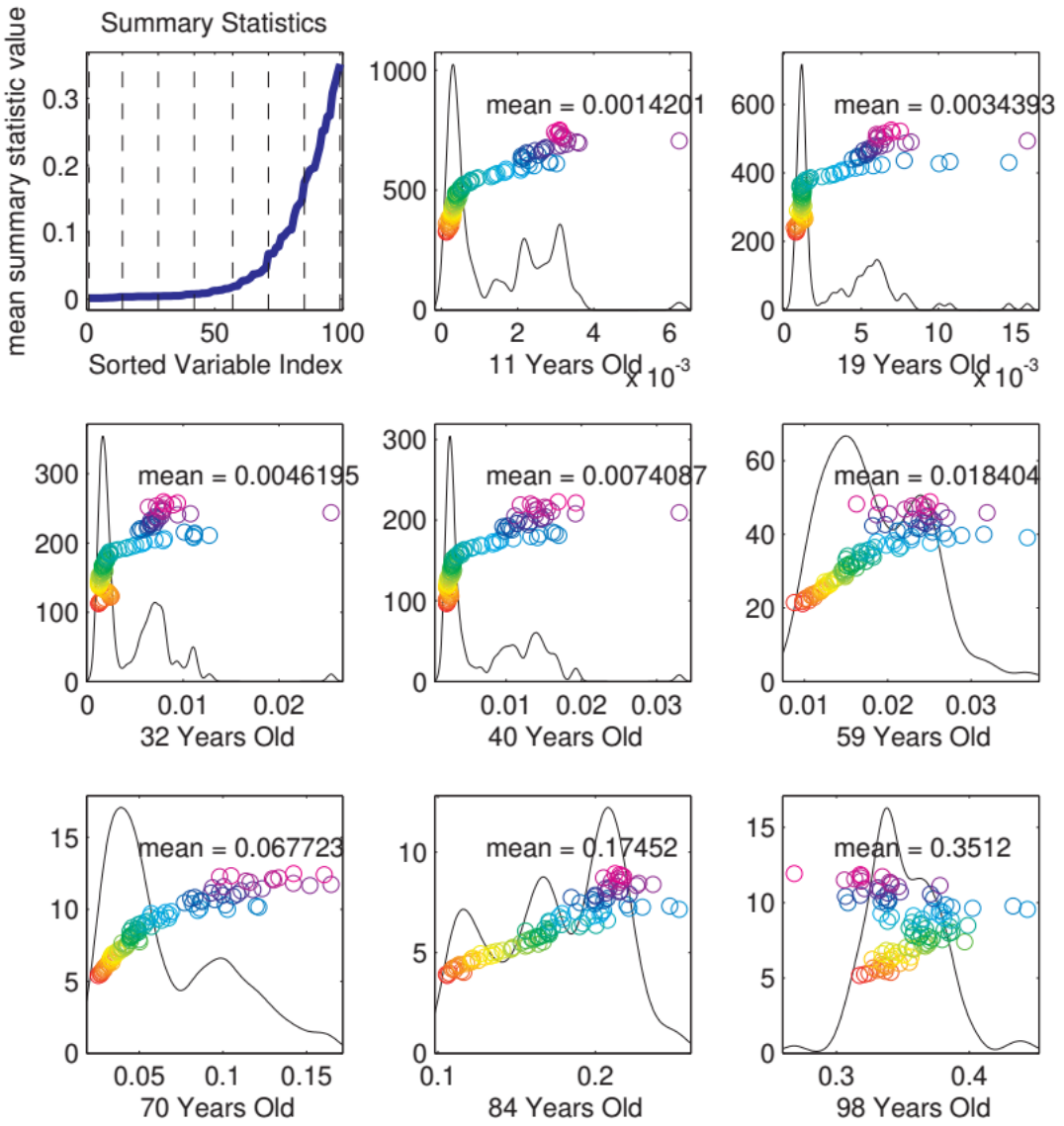

Figura II.9: Gráfica de distribuciones marginales del conjunto de datos *Mortalidad Española*. En el panel superior izquierdo se muestra las medias de las variables (curva azul). En los paneles restantes se muestran las distribuciones marginales de un conjunto de variables representativo igualmente espaciado (rectas discontinuas, primer panel). Se aprecian valores atípicos así como una fuerte asimetría.

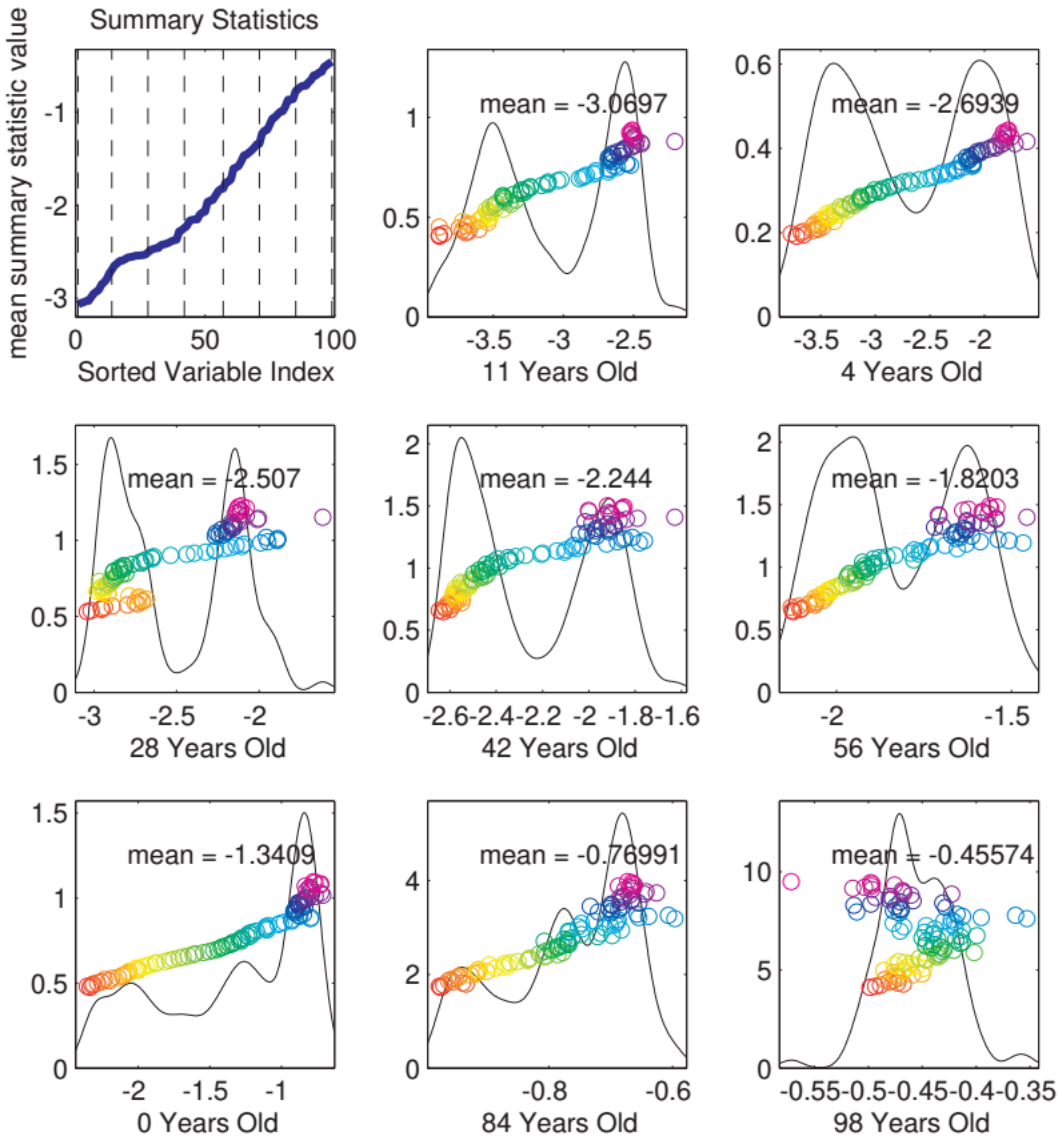

Figura II.10: Muestra las distribuciones marginales del conjunto de datos Mortalidad Española en el mismo formato que para la Figura II.9 pero aplicando $\log_{10}$.

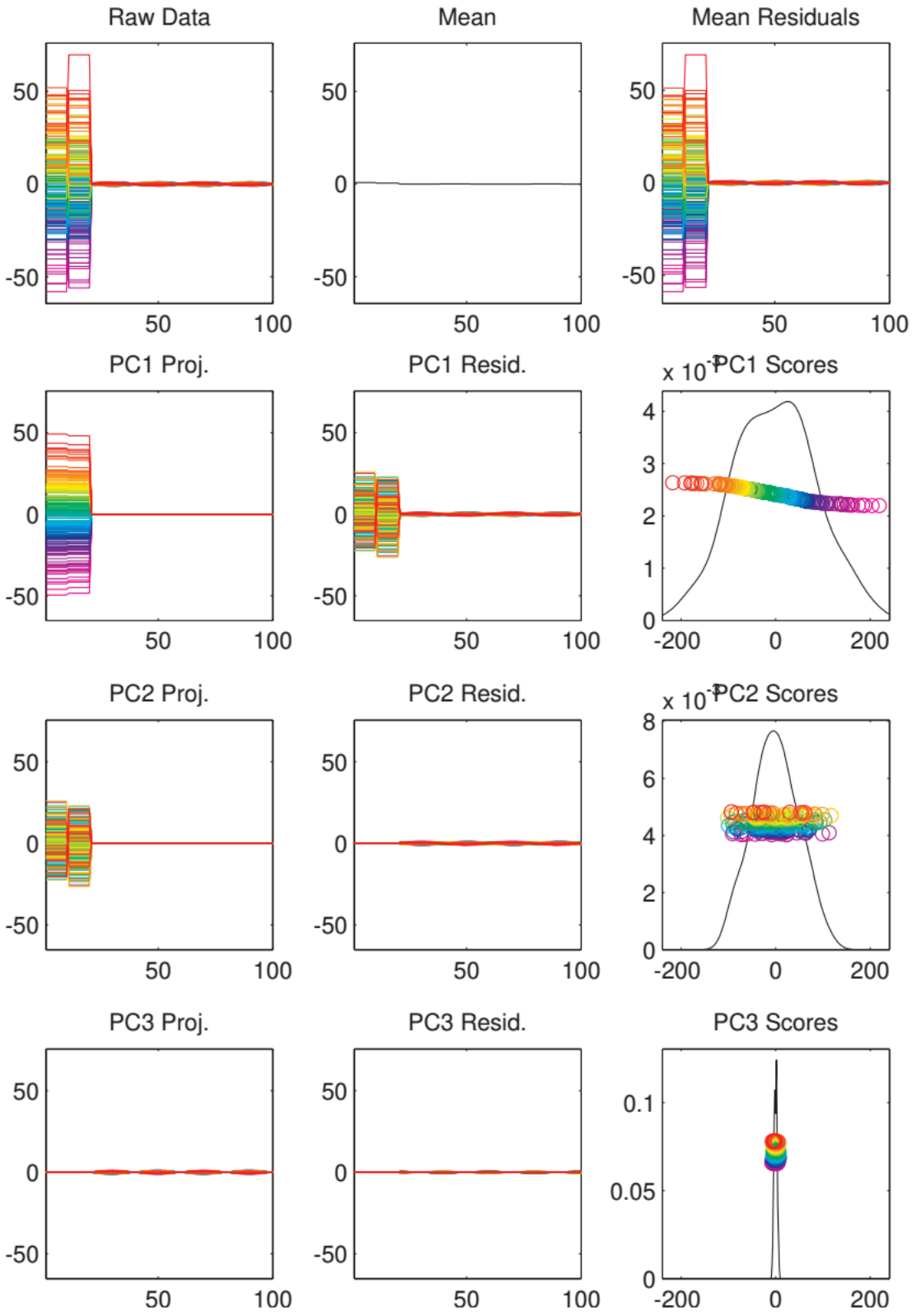

Figura II.11: Ejemplo de juguete de *Curvas de Dos Escalas*, en el que se ilustra el problema de diferentes escalas en las variables. Las 20 primeras variables tiene mayor variación por lo que dominan las dos primeras CP.

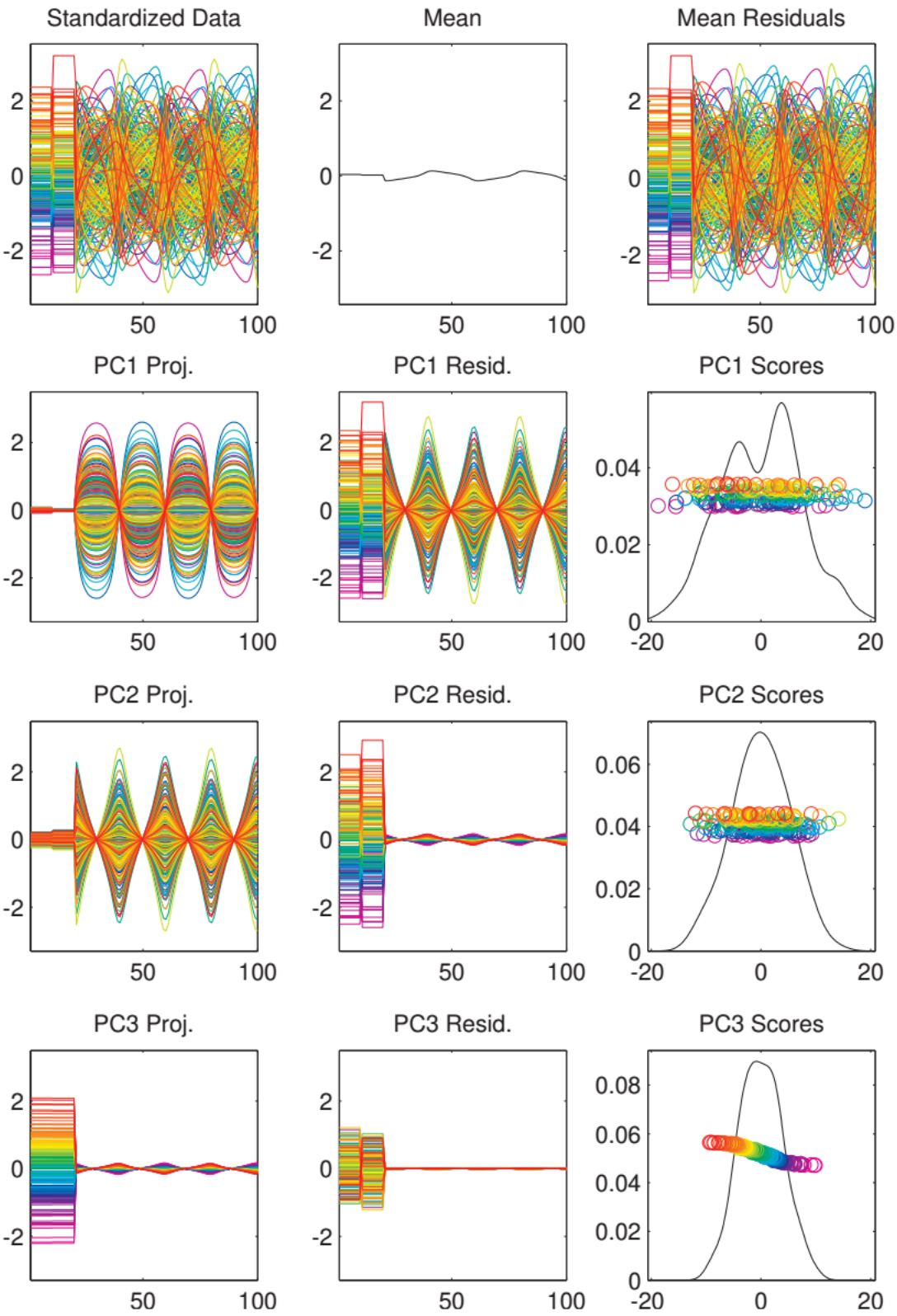

Figura II.12: Resultado de escalar el conjunto de datos *Curvas de Dos Escalas*. Ahora las 80 últimas variables dominan la variación, resultando en conclusiones muy distintas.

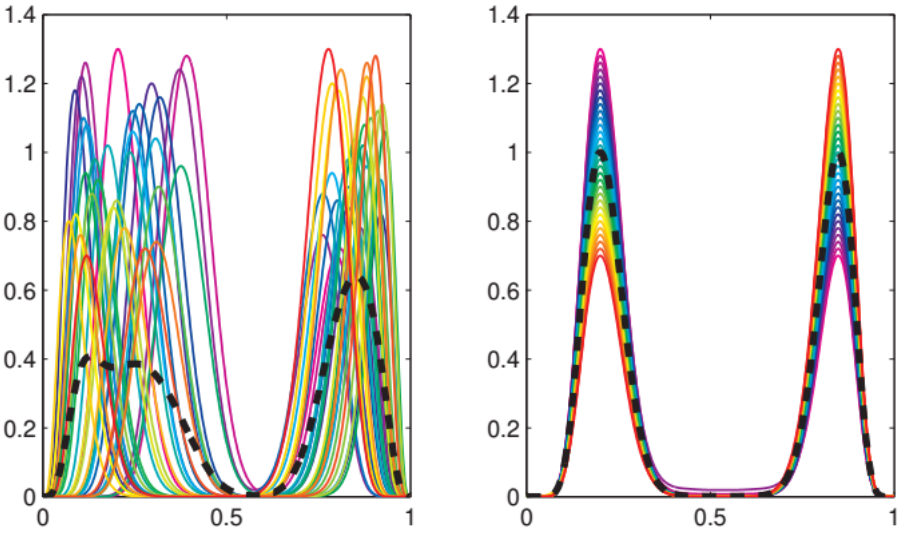

Figura II.13: Ejemplo de juguete, en el que cada curva tiene dos picos para ilustrar la importancia del registro de curvas. El panel izquierdo muestra las curvas originales, en el que color de las curvas va en función de la altura de su pico izquierdo. El panel de la derecha muestra el resultado de alinear las curvas. La curva media (curva negra discontinua) es mucho más representativa para el conjunto derecho.

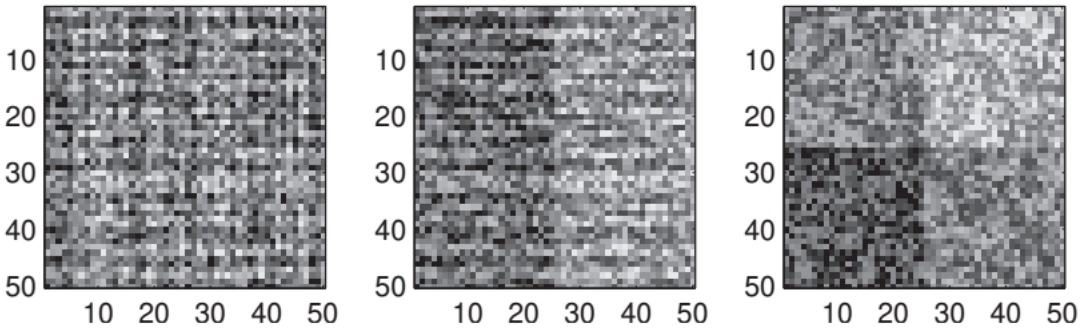

Figura II.14: Ejemplo de juguete que muestra la importancia de la organización jerárquica de filas y columnas en mapas de calor. A la izquierda vemos un orden totalmente aleatorio, en el centro, aplicamos agrupación por columnas y a la derecha, agrupación por filas y columnas. En este último ya distinguimos cierta estructura.

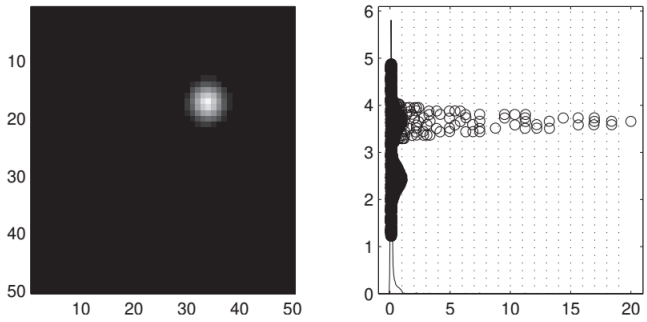

Figura II.15: Ejemplo de juguete ilustrando los problemas de la escala de grises. El mapa de calor se muestra a la izquierda y las distribuciones de las entradas de la matriz a la derecha, con las rectas verticales de puntos indicando las regiones de grises. Vemos que, debido a la asimetría, la escala de grises empleada no es efectiva.

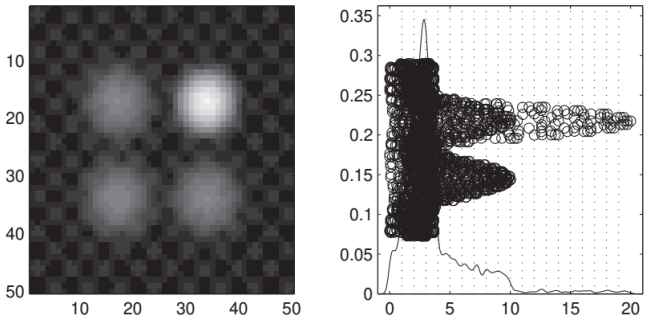

Figura II.16: Aplicamos otra escala de grises para las mismas entradas de la matriz que en la Figura II.15 (las rectas de puntos indican, como antes, los límites en la escala de grises) aplicando una transformación logarítmica, cuya distribución se muestra en el panel derecho. Se aprecia una mayor estructura.

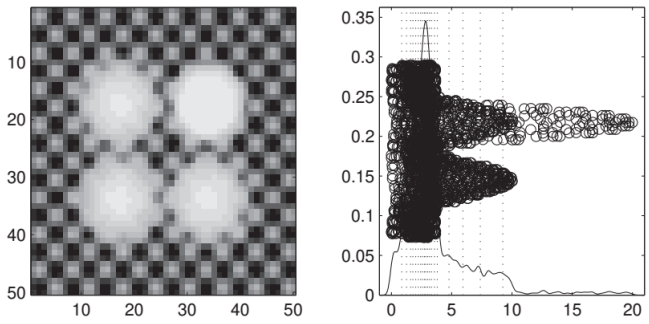

Figura II.17: Otro mapa de calor de la misma matriz de datos que en las Figuras II.15 y II.16. Se muestra otra estructura de datos empleando una escala de grises equilibrada gracias al escalado cuantil.

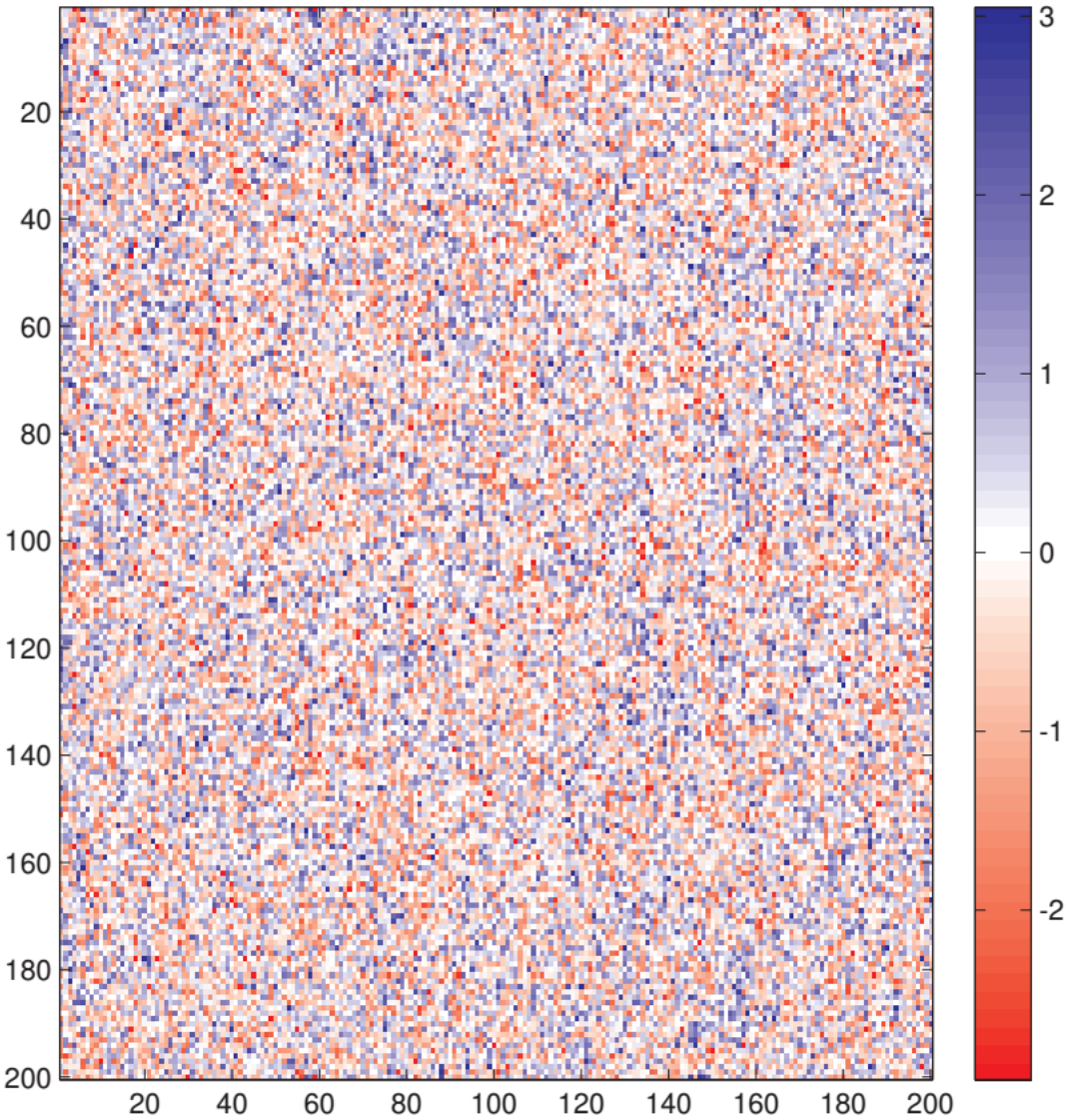

Figura II.18: Mapa de calor del conjunto de datos *Dos Clases Gaussianas*. El blanco indica las entradas 0 con azul para valores positivos y rojo para negativos. No se aprecia ninguna estructura aparente.

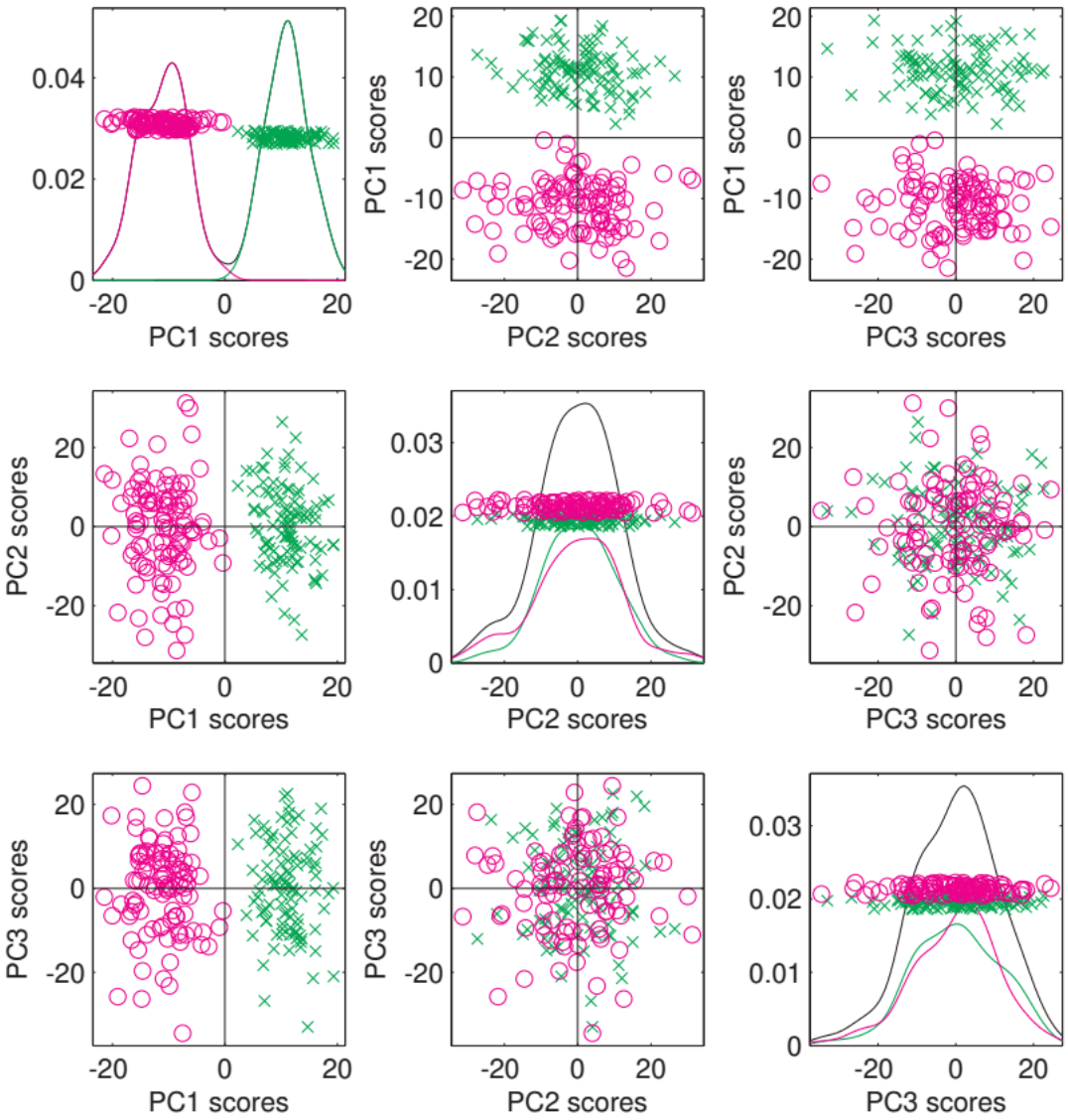

Figura II.19: Matriz de diagramas de dispersión ACP del conjunto de datos *Dos Clases Gaussianas*. Vemos que se aprecian dos grupos, indicados con distintos colores y símbolos, a pesar de la cantidad de ruido presente en el conjunto de datos.

	PC 1	PC 2	PC 3	PC 4
Raw PCA	76%	24%	0.1%	0.03%
Standardized PCA	53%	27%	15%	5%

Figura II.20: Porcentaje de la suma de cuadrados explicada por cada componente CP para el conjunto *Curvas de Dos Escalas*. Se muestra como las componente de los datos originales se centran en la estructura de la izquierda, mientras que los estandarizados se centran en los derecha.

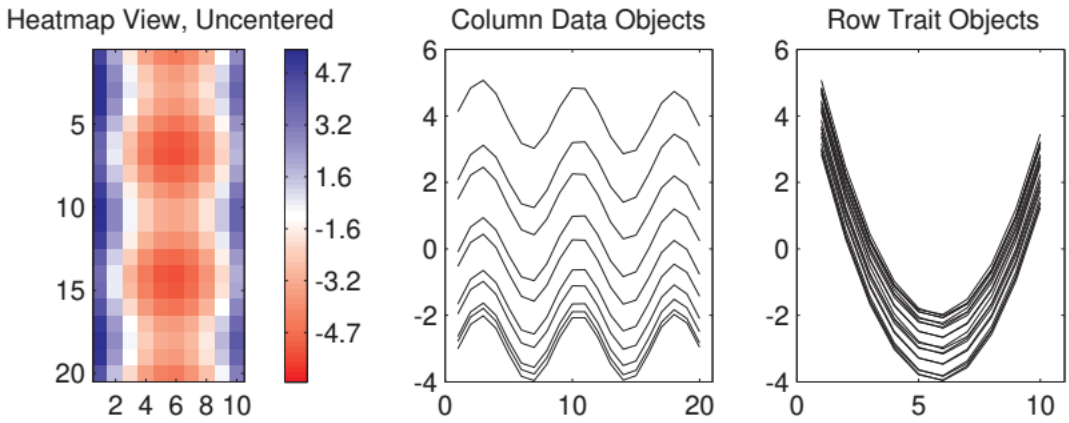

Figura II.21: Matriz de los datos originales del conjunto *Onda Sinusoidal*. Se muestra un mapa de calor (izq.), las gráficas de las curvas de dicho conjunto (centro), con las columnas como objetos de datos y los vectores de característica fila (der.). Es complicado obtener una interpretación común a los tres paneles.

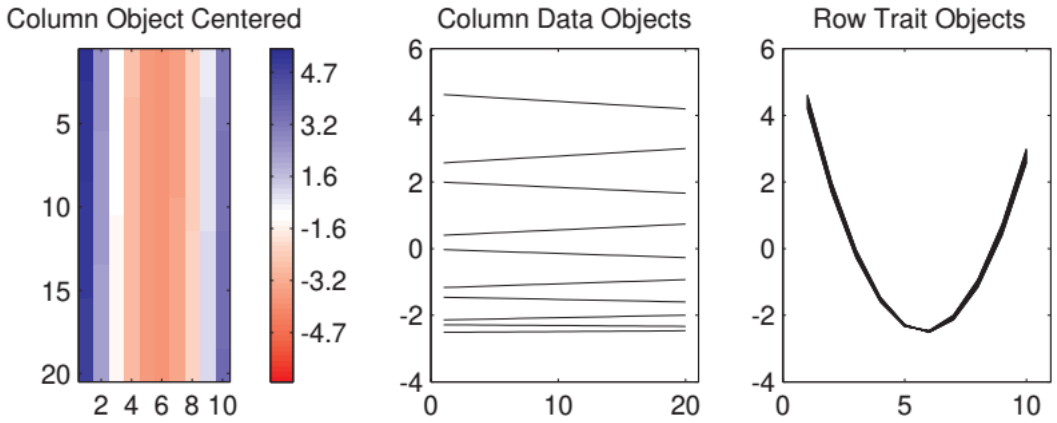

Figura II.22: Resultado de aplicar el centrado por columnas para el conjunto de datos *Onda Sinusoidal*, usando el mismo formato que la Figura II.21. Se muestra la eliminación de la componente de onda sinusoidal vertical, dejando columnas constantes cuyas alturas están determinadas por la parábola.

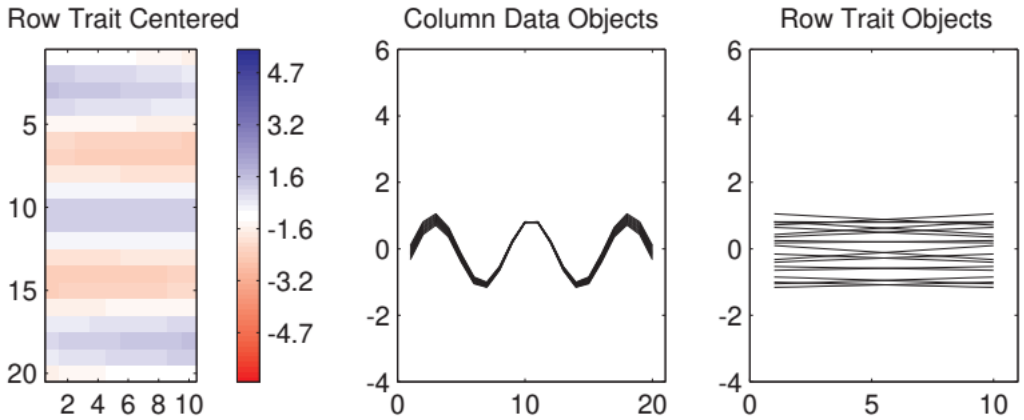

Figura II.23: Aplicación del centrado por filas del conjunto de datos *Onda Sinusoidal*, usando el mismo formato que la Figura II.21. Análogamente, se muestra la eliminación de la componente parabólica horizontal, dejando rectas horizontales determinadas por la onda sinusoidal.

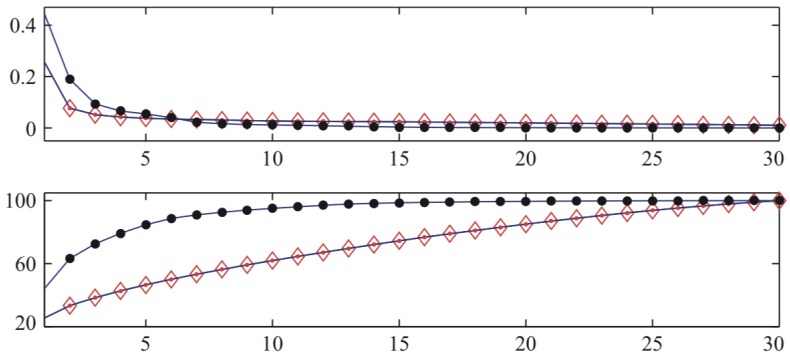

Figura II.24: Gráfica de sedimentación (arriba) y las contribuciones totales a la varianza (abajo) para el conjunto de datos de *cáncer de mama* (puntos negros) y los rendimientos de Dow Jones (triángulos rojos). En ninguno de los dos conjuntos se aprecian codos.

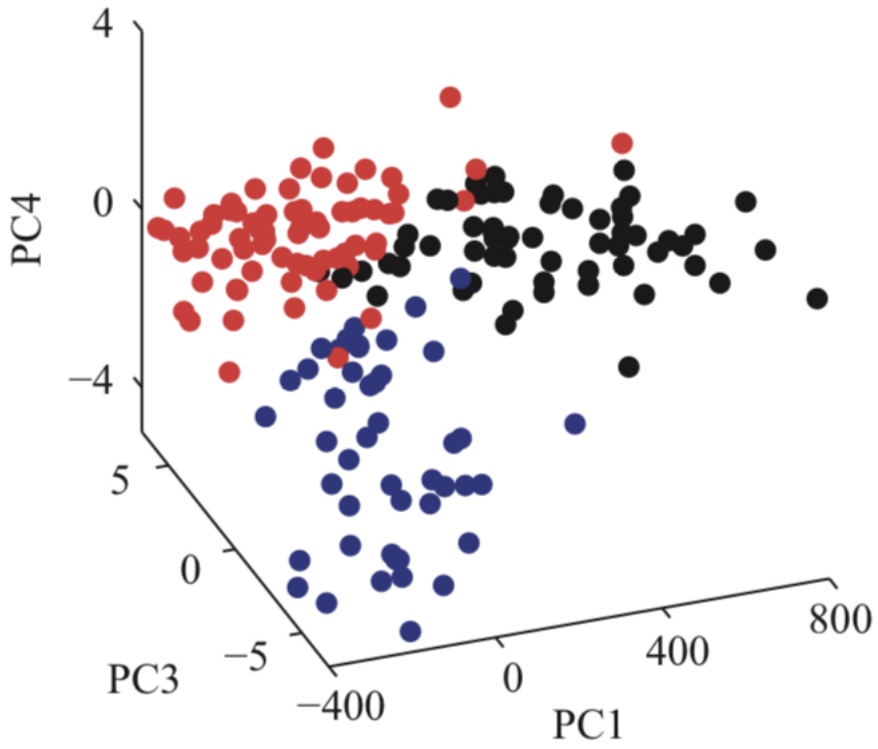

Figura II.25: Gráfica de puntuaciones en 3D para el conjunto de datos *reconocimiento del vino*. Se obtiene una separación aceptable.

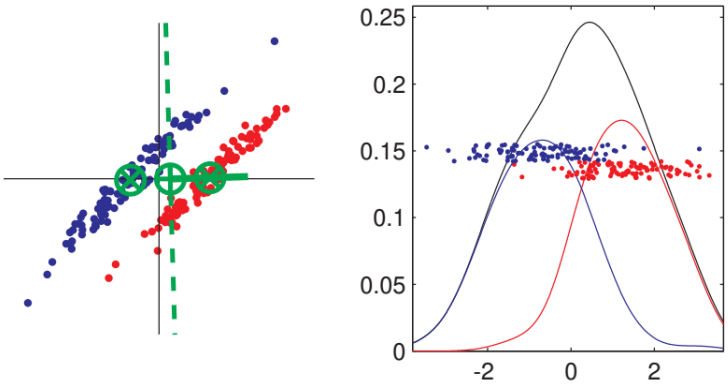

Figura II.26: En el panel izquierdo se muestra el conjunto de datos *Gaussianas Correlacionadas Desplazadas* donde los colores indican las clases. El vector de dirección DM se muestra en la recta verde y el hiperplano de separación es la recta verde discontinua. El panel derecho muestra las proyecciones en la dirección DM. En este caso, la separación de clases es muy pobre debido a que las variables de prueba están correlacionadas.

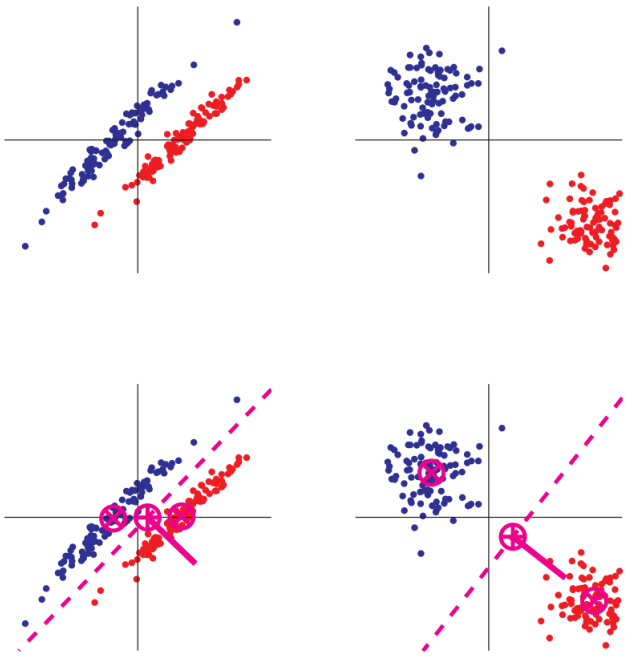

Figura II.27: En el panel superior izquierdo se vuelven a mostrar los mismos datos que para la Figura II.26. El panel superior derecho muestra los resultados de la transformación *esférica* entre clases. El panel inferior derecho muestra el resultado de aplicar DM a los datos transformados. El panel inferior izquierdo muestra la efectividad de DM, que no es más que la inversa de la transformación *esférica*.

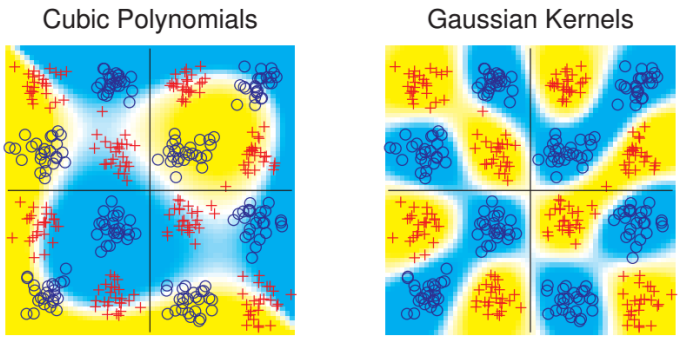

Figura II.28: Ejemplo *Patrón de Ajedrez* 2-*d*. Para incrustación polinómica (izq.) vemos una clasificación deficiente, mientras que obtenemos un rendimiento mucho mejor con ADL (der.).

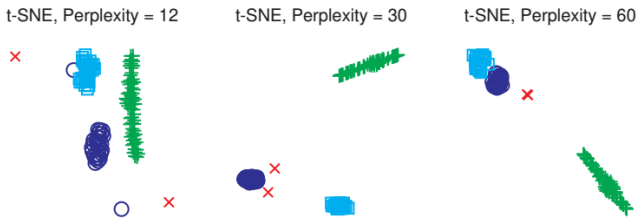

Figura II.29: Ejemplo de visualización t-SNE del conjunto de datos *Cuatro Subgrupos*, para diferentes valores del parámetro de perplejidad. La elección razonable parece 60 que se muestra en el panel derecho.

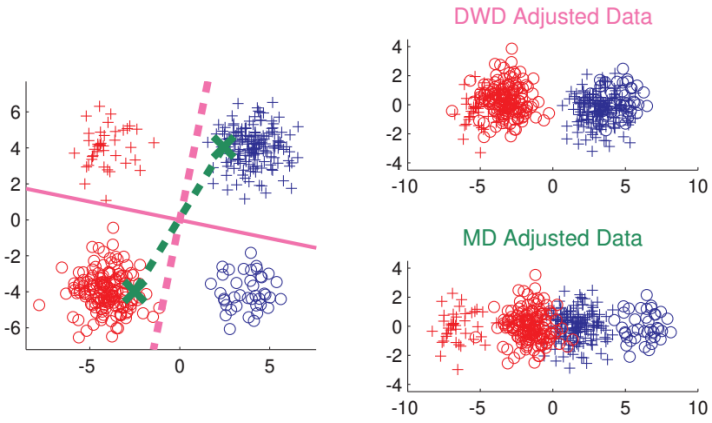

Figura II.30: Conjunto de datos de juguete (izq.) mostrando el valor DPD del ajuste por lotes relativo al enfoque DM. Los colores representan distintos focos de experimento, mientras que los símbolos reflejan efectos de lotes insignificantes. DM se muestra en verde, del que obtenemos resultados deficientes. La dirección DPD (recta magenta discontinua) con resultados mucho mejores, ya que consigue separar los lotes.

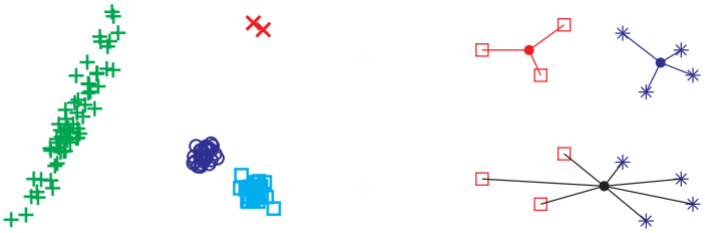

Figura II.31: Izquierda: Visualización del conjunto de datos *Cuatro Subgrupos*. Derecha: Ejemplo de juguete que proporciona información sobre IC. Arriba, la suma de longitudes de los segmentos de recta al cuadrado es la *Suma de Cuadrados dentro del Grupo*, SCDG. Abajo, se demuestra de manera análoga que es la *Suma Total de Cuadrados*, STC.

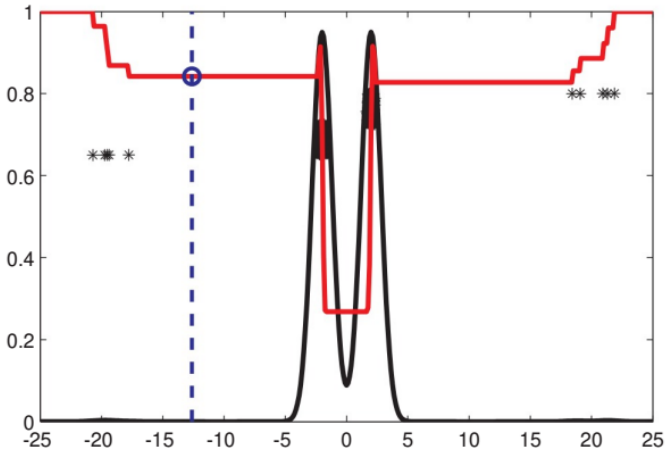

Figura II.32: Ejemplo unidimensional, que resalta los problemas del mínimo local de IC. Muchos métodos se confunden escogiendo un mínimo muy distante del mínimo global IC.

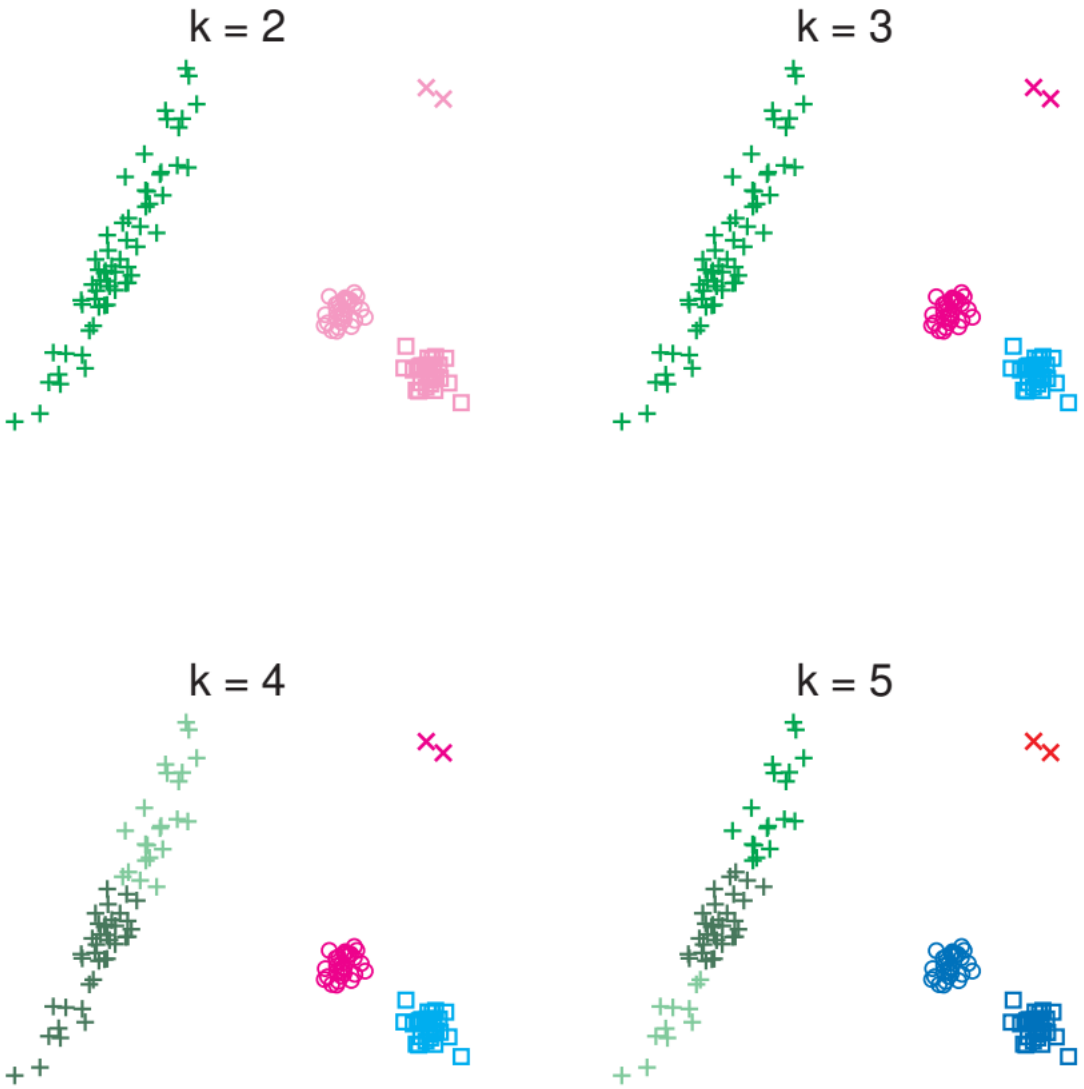

Figura II.33: Demostración de la agrupación de k-medias, para el conjunto *Cuatro Subgrupos*, con $k = 2, 3, 4, 5$. Muestra que se debe tener cuidado con la interpretación.

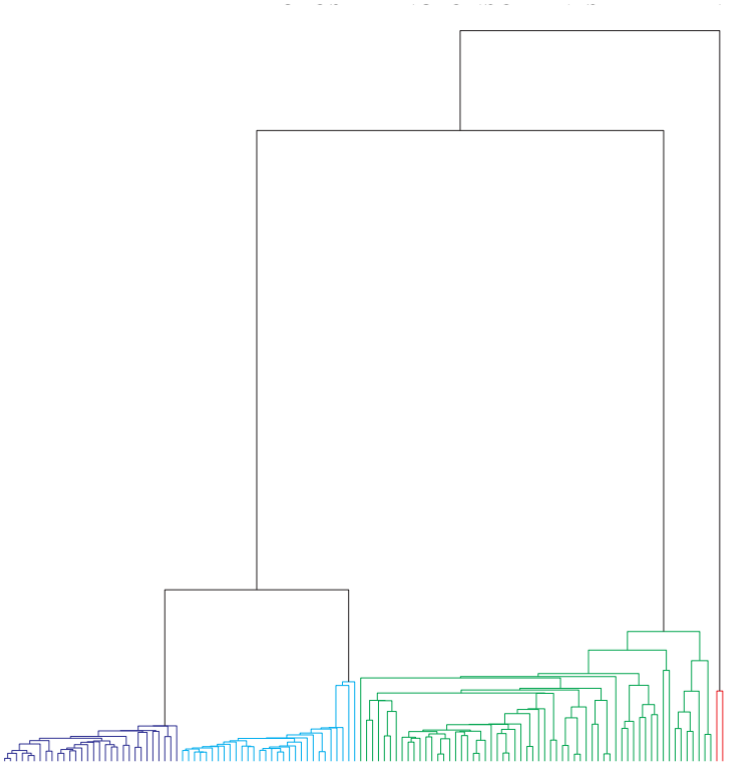

Figura II.34: Dendrograma del conjunto de datos *Cuatro Subgrupos*, basado en distancia Euclidiana y enlace único. Muestra como partimos del grupo total, en la parte superior, a grupos de un único elemento en la inferior.

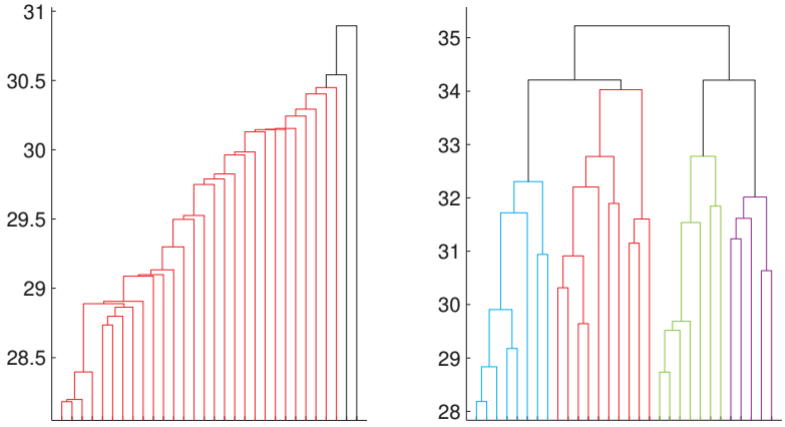

Figura II.35: Dendrograma del conjunto de datos *Gaussianas de Alta-Dimensión*, basado en distancia Euclidiana. Se realiza un contraste entre enlace único (izq.), que separa individuos secuencialmente, con enlace de Ward (der.), que hace particiones más equilibradas.

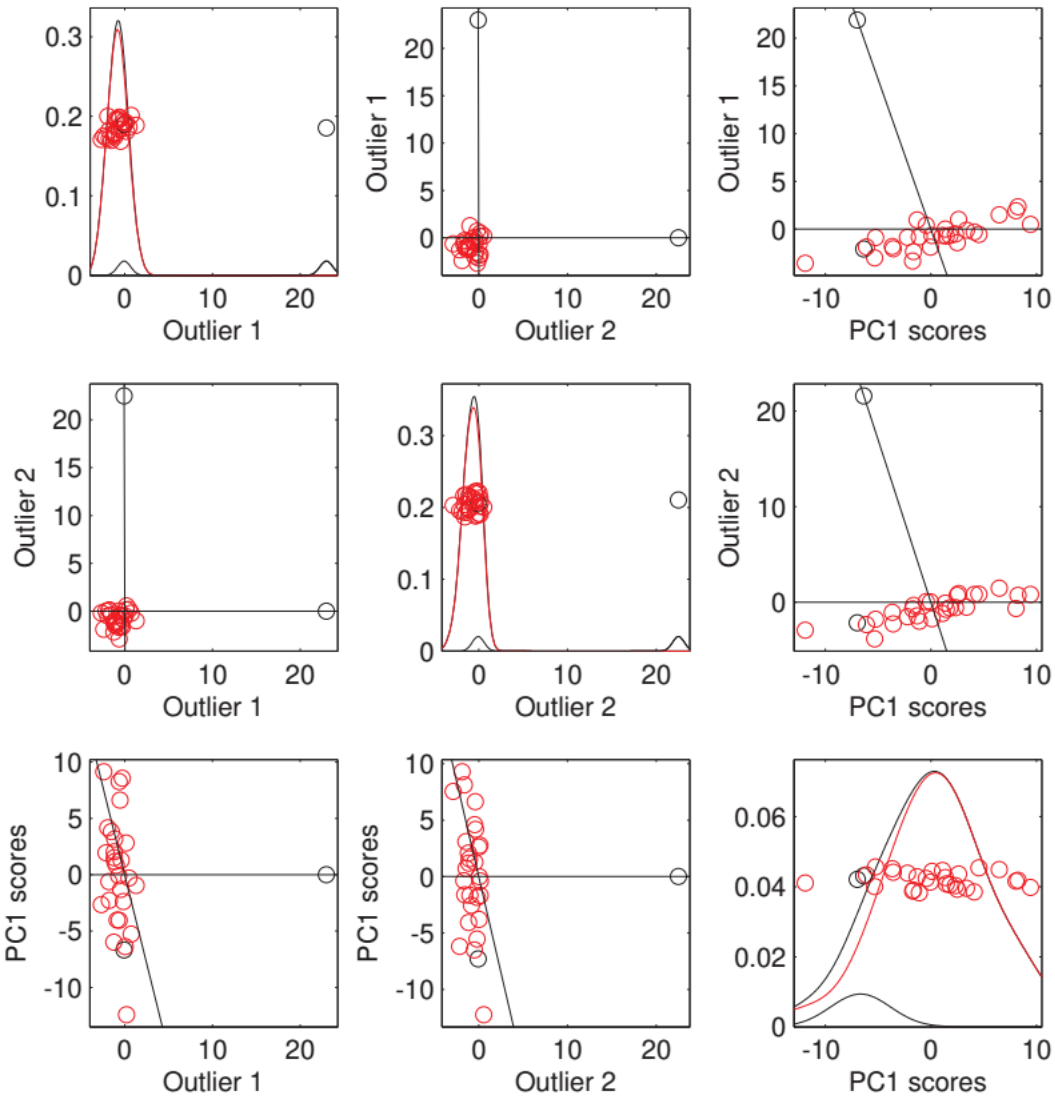

Figura II.36: Matriz de diagramas de dispersión de conjunto de datos *Gaussianas de Alta-Dimensión*, usando los colores del panel izquierdo de la Figura II.35 (enlace único). Los dos primeros paneles muestran qué tan bien están separados los dos primeros grupos de un elemento del resto de los datos. El tercer eje es la dirección CP1, para contrastar.

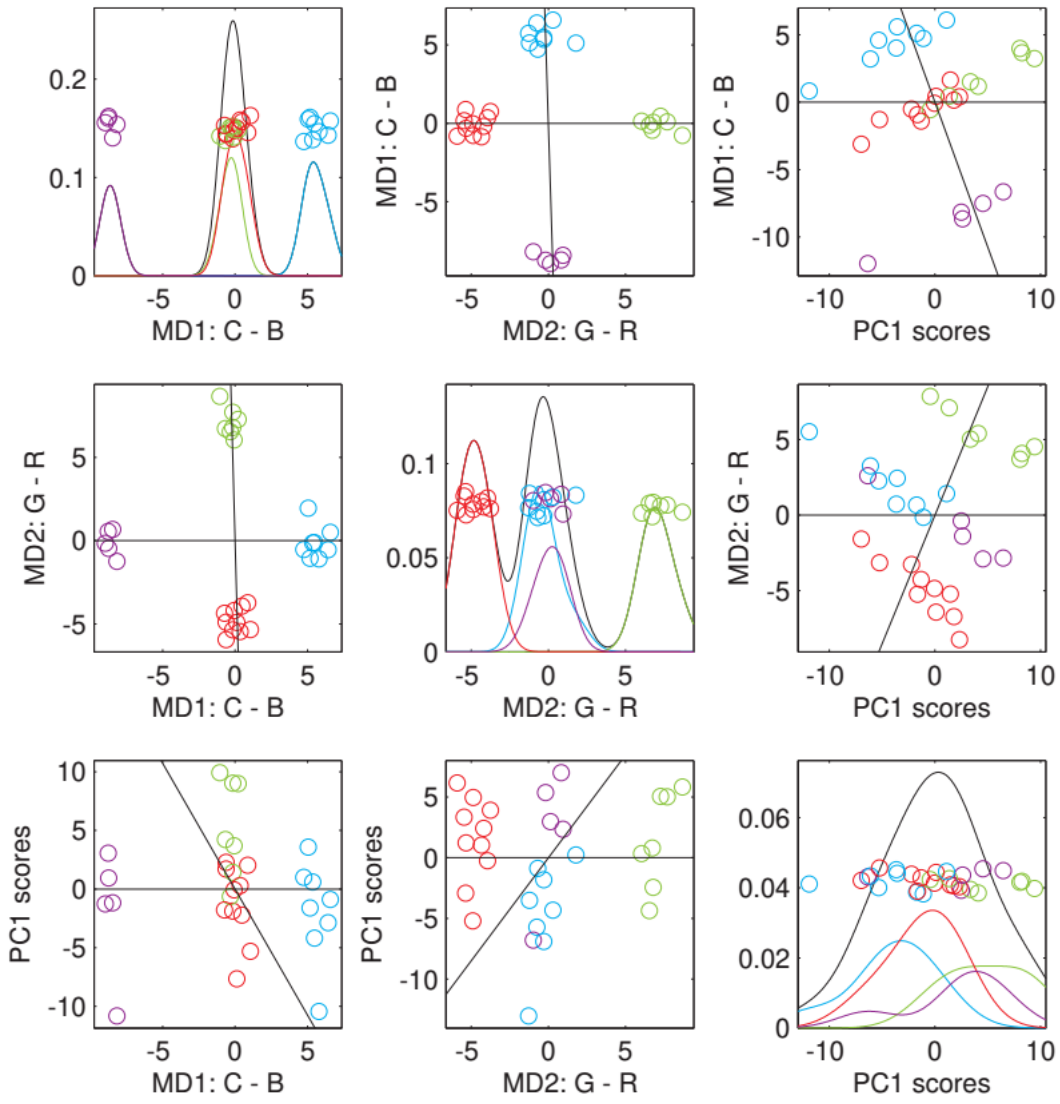

Figura II.37: Otra matriz de diagramas de dispersión para *Gaussianas de Alta-Dimensión*, esta vez empleando el enlace de Ward, con colores del panel derecho de la Figura II.35. Los dos primeros ejes están basados en DM, Cian-Azul en la primera dirección, Verde-Rojo, en la segunda. Otra vez el tercero muestra las puntuaciones CP1.

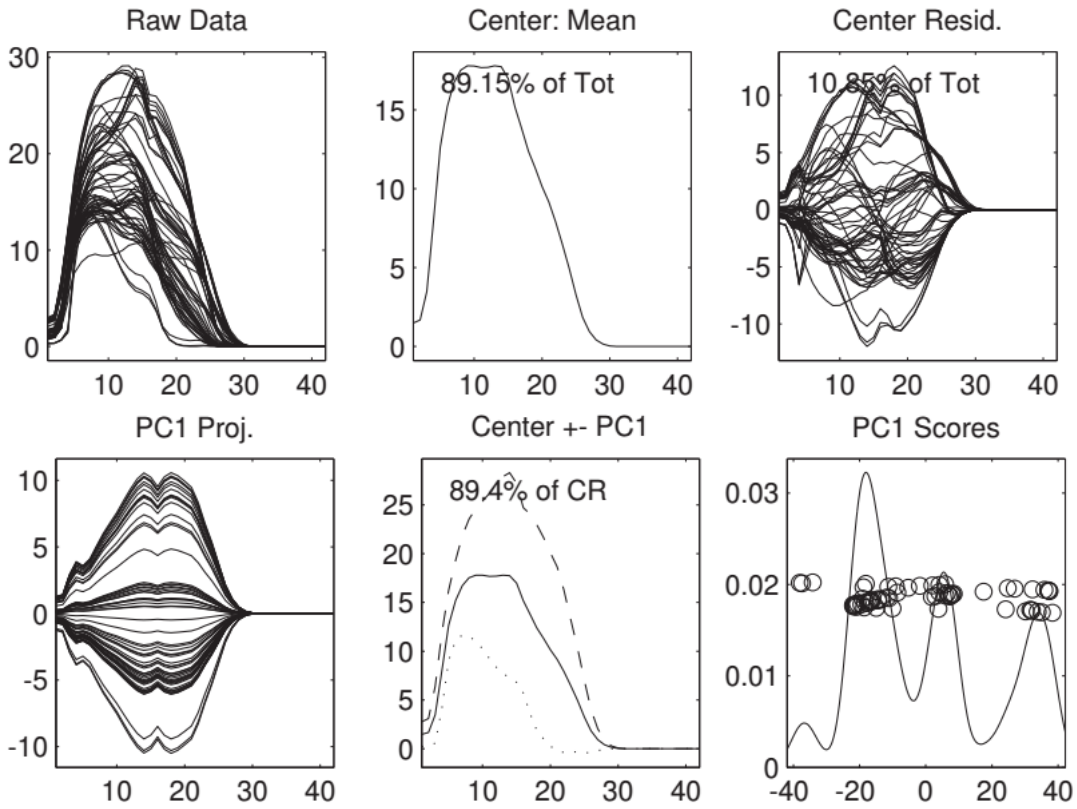

Figura II.38: ACP del conjunto de datos *Flujo Máximo*. La fila superior muestra el efecto del centrado de medias (datos originales, media y residuos). El primer modo de variación se muestra en la fila inferior izquierda, el centro muestra la relación con la media, y la derecha muestra las puntuaciones, en la que distinguimos 3 grupos.

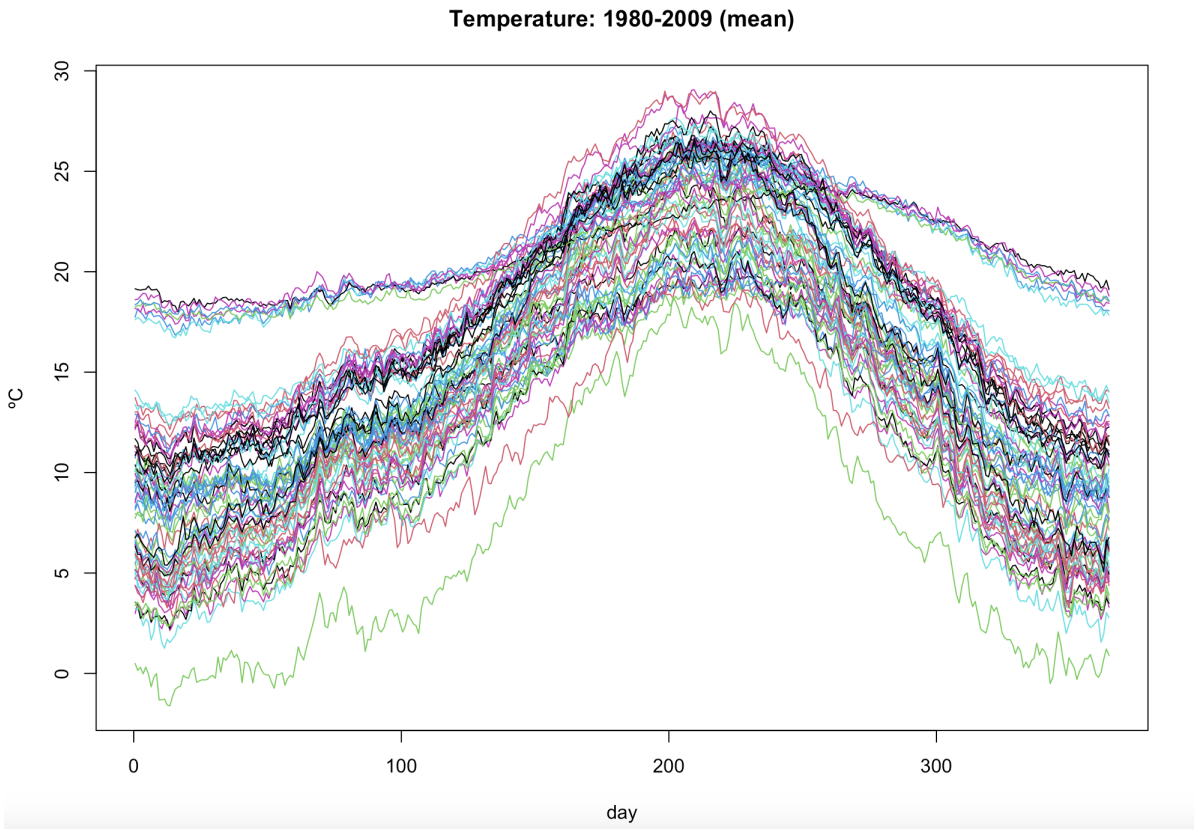

Figura II.39: Curvas de la temperatura media diaria de España desde 1980 hasta 2009 obtenidas de las 73 estaciones meteorológicas. No se observa ninguna estructura de los datos.

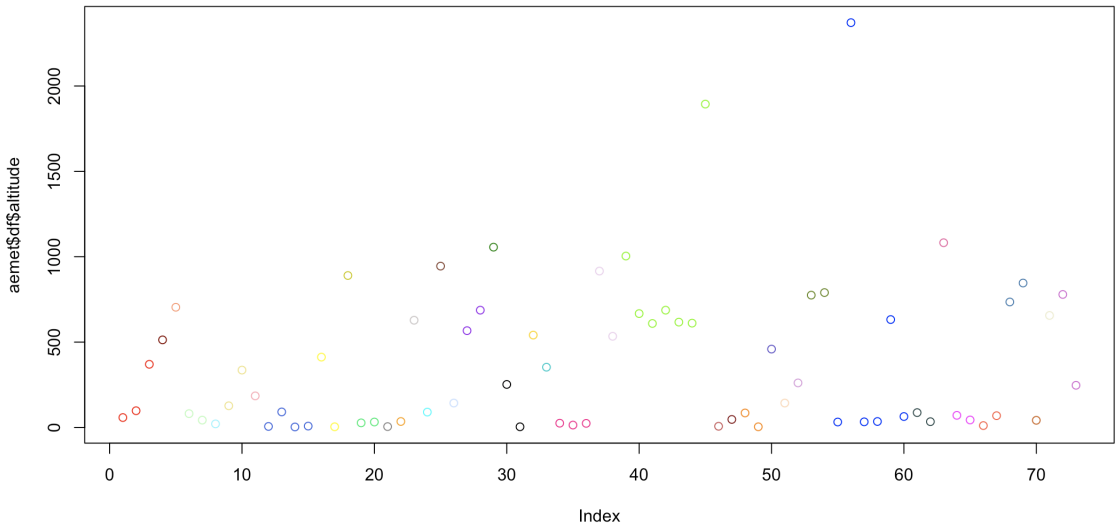

Figura II.40: Representación de las alturas de las estaciones meteorológicas respecto a su índice. Los colores representan las provincias en las que se encuentra cada estación. El punto más alto es la estación de Izaña, en Santa Cruz de Tenerife.

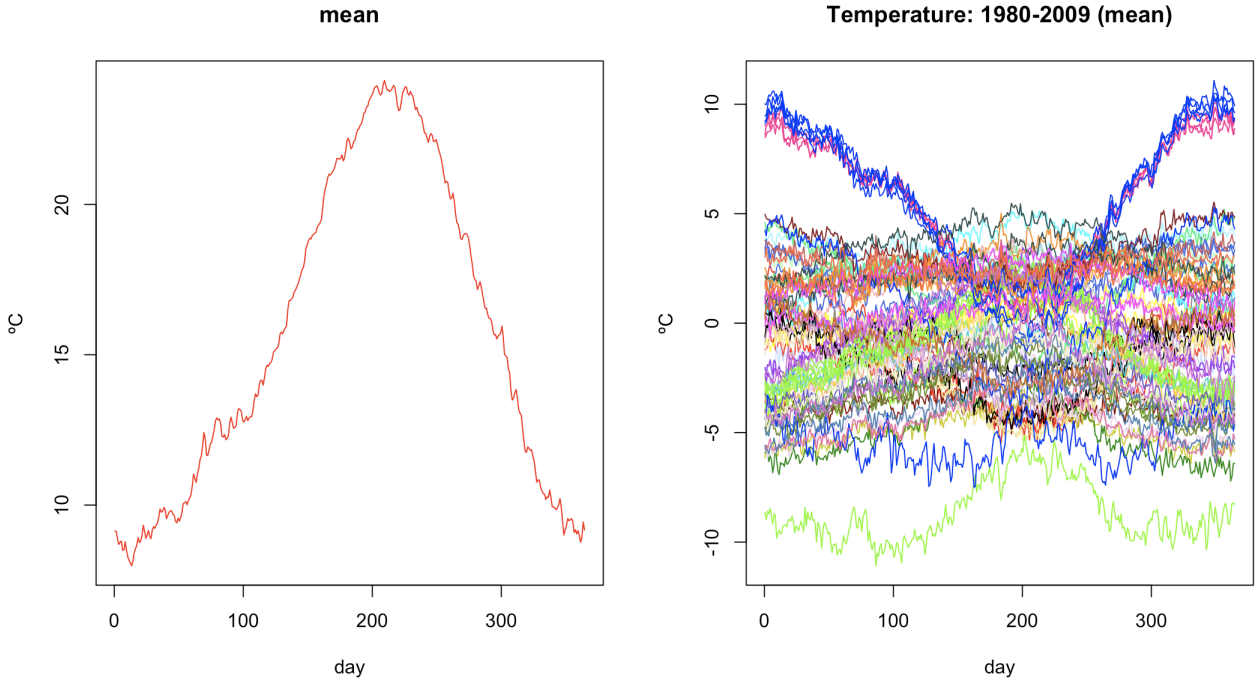

Figura II.41: En el panel izquierdo se representa la curva media de estas funciones de la temperatura diaria. En el derecho se representan los residuos de la media, empleando la misma clasificación de colores establecida en la Figura 6.1.

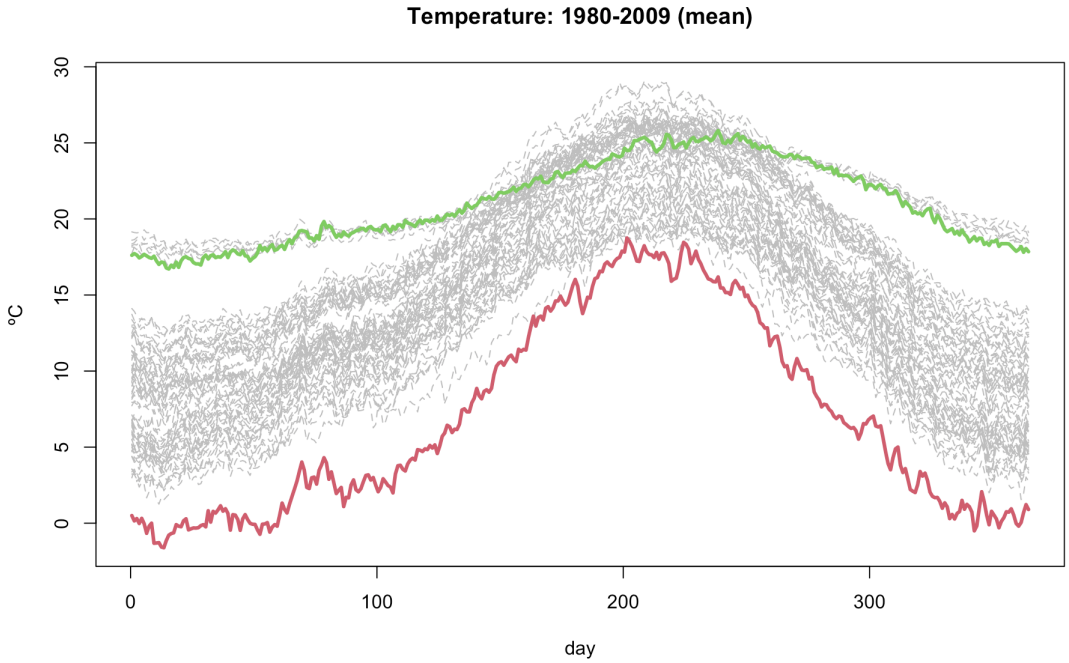

Figura II.42: Agrupación en k-medias para $k = 2$.

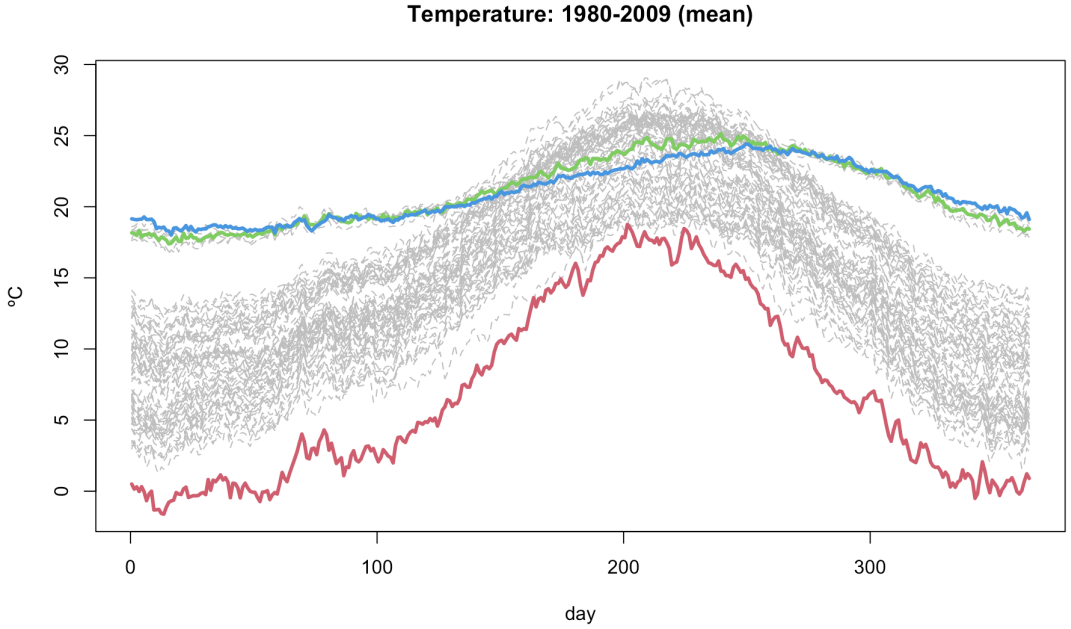

Figura II.43: Agrupación en k-medias para $k = 3$.

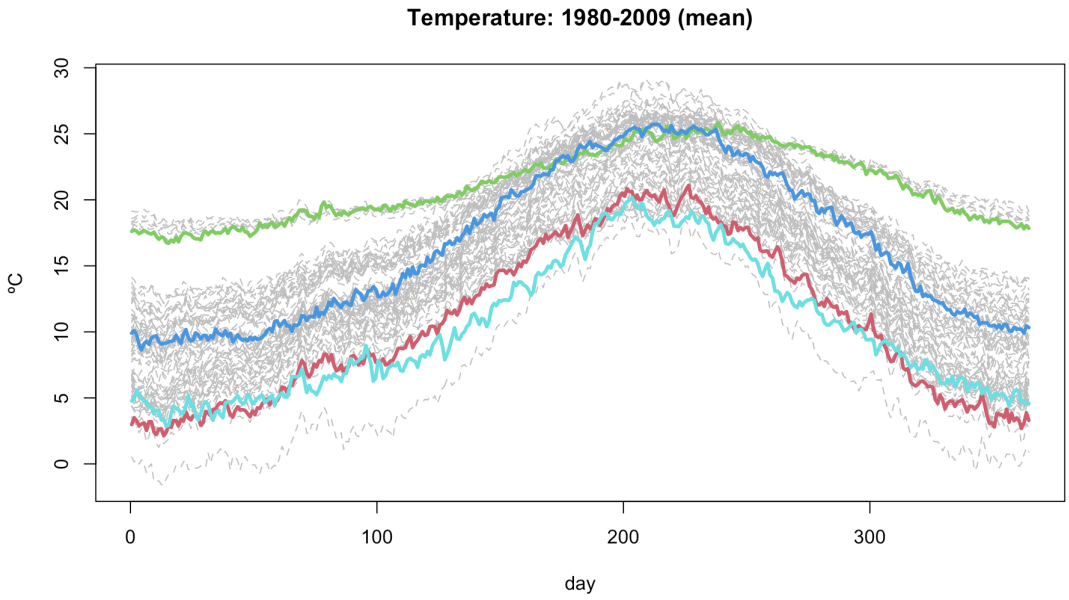

Figura II.44: Agrupación en k-medias para $k = 4$.

Anexo III

Script de R

En este Apéndice se recoge el código de R necesario para generar los resultados y gráficas de nuestro conjunto de datos reales del Capítulo 6.

```
# Paquetes Necesarios
install.packages('rainbow',dep=TRUE)
library(rainbow)
library(MASS)
library(pcaPP)
library(RCurl)
library(fds)
library(fda.usc)


# Lectura de datos
data(aemet)
summary(aemet)
datos<-aemet$temp
dim(datos)
plot.fdata(datos)


aemet$df$province
# Creamos un vector con los colores para agruparlos por provincias (brushing)
colores<-c('red','red','red','red4','lightsalmon','darkseagreen1','darkseagreen1'
,'cadetblue1','khaki','khaki','lightpink1','royalblue','royalblue',
'royalblue','royalblue','yellow','yellow','yellow3','springgreen2',
'springgreen2','snow4','orange','snow3','cyan','salmon4',
```

```
'lightsteelblue1','purple','purple','green4','black','black','gold'
,'cyan3','deeppink','deeppink','deeppink','thistle2','thistle2'
,'chartreuse','chartreuse','chartreuse','chartreuse','chartreuse'
,'chartreuse','chartreuse','indianred','firebrick4','darkorange'
,'darkorange','slateblue','peachpuff','plum','olivedrab','olivedrab'
,'blue','blue','blue','blue','blue','blue','darkslategray'
,'darkslategray','hotpink2','magenta','magenta','tomato','tomato'
,'steelblue','steelblue','chocolate','lightyellow2','orchid',
'orchid')


plot.fdata(datos,col=colores)
# Ahora las curvas están organizadas por provincia.


plot(aemet$df$altitude,col=colores)
# También por la altura de la estación meteorológica
# Más altas: datos[56,] <- Izaña (Tenerife)
# datos[45,] <- Puerto de Navacerrada (Madrid)


par(mfrow=c(1,2))
# Calculamos la media y la representamos
media<-func.mean(datos)
plot.fdata(media,col=colores)


# Calculamos los residuos y los representamos
datoscen<-fdata.cen(datos,meanX=media)$Xcen
plot(datoscen,col=colores)


# Calculamos los modos de variación (Componentes Principales)
# Probamos con 2
PC<-fdata2pc(datos,ncomp=2,norm=TRUE)
summary(PC,biplot=TRUE)
# Explican el 98.78% de la variabilidad


score<-PC$x # scores de CP1 y CP2
loadings<-PC$rotation # loadings
```

```
# Los representamos
par(mfrow=c(1,2))
plot(score,col=colores)
plot(loadings)


# Agrupación en k-medias
kmeans.center.ini(datos, draw=TRUE) # K-medias con k=2
kmeans.center.ini(datos, ncl=3, draw=TRUE) # K-medias con k=3
kmeans.center.ini(datos, ncl=4, draw=TRUE) # K-medias con k=4
kmeans.center.ini(datos, ncl=5, draw=TRUE) # K-medias con k=5
```

Bibliografía

[1] Marron, J. S. y Dryden, Ian. L. (2021). *Object Oriented Data Analysis*, 1st ed., Monographs on Statistics and Applied Probability, 169, Chapman and Hall/CRC, New York.

[2] Härdle, W. K. y Simar L. (2019). *Applied Multivariate Statistical Analysis*, 5th ed., Springer Science & Business Media.

[3] Koch I. (2013). *Analysis of Multivariate and High-Dimensional Data*, 1st ed., Cambridge University Press, New York.