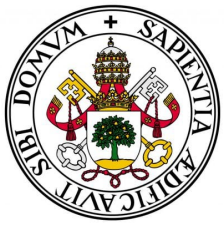


Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Matemáticas

Procedimientos de clasificación con recortes

Autor:
Pablo Andrés Salgado

Tutor:
Carlos Gabriel Matrán Bea

Año: 2025

Agradecimientos

A mis padres, Iván y Marisa, por su incansable apoyo y paciencia, y por ser los pilares de mi vida. Gracias por vuestro sacrificio y esfuerzo, y por brindarme una educación inmejorable.

A mi hermano, Iván, por ser mi mayor referente a seguir. Gracias por enseñarme a disfrutar del proceso y a valorarme más a mi mismo.

A mi tutor, Carlos Matrán, por su dedicación y atención durante la realización de este trabajo. Gracias por despertar mi interés y motivarme a profundizar en una materia tan apasionante.

A mis amigos del Colegio Mayor de Santa Cruz, por converitirse en mi segunda familia. Gracias por todos los momentos vividos, por las enseñanzas aprendidas y por hacer de estos años los mejores de mi vida.

Valladolid, 11 de julio de 2025

Índice general

Resumen

La clasificación y/o la agrupación de individuos a partir de los datos obtenidos de acuerdo con la medición de algunas de sus características es un objetivo fundamental en casi cualquier tipo de estudio. Entre los procedimientos de clasificación, el análisis "cluster" basado en k-medias es sin duda el más popular por su sencillez conceptual, sin embargo es bien conocido que su comportamiento ante la presencia de datos atípicos puede ser catastrófico, lo que limita su uso en la práctica. Este trabajo incluye y discute algunos de los procedimientos más relevantes del clustering robusto, diseñados con el objetivo de evitar esta falta de robustez, especialmente los basados en recortes de los datos.

Abstract:

The classiffication and/or grouping of individuals from the meassurement of some of their characteristics is one of the main goals in almost any type of study. Among classiffication procedures, the cluster analysis based on k-means is undoubtedly the most popular one due to its conceptual simplicity. However, it is well known its catastrophic behavior in presence of atypical data, which limits it practical utility. This paper includes and discusses some of the most relevant robust clustering procedures, designed to avoid this lack of robustness, especially those based on data trimming.

Introducción

Desde que somos pequeños, los seres humanos tendemos a establecer relaciones entre aquello que nos rodea: clasificamos los objetos por formas y colores, identificamos a los animales por especies y catalogamos las situaciones de nuestro día a día para tomar decisiones consecuentemente. Esta necesidad de ordenación y clasificación constituye una característica principal del pensamiento humano, y nos ayuda a conocer mejor el mundo en el que vivimos.

En la antigüedad, los primeros humanos debían identificar qué alimentos eran comestibles y cuales venenosos, o qué animales resultaban ser inofensivos y cuáles peligrosos. La clasificación, responde así a una necesidad evolutiva de comprender nuestro entorno y adaptarnos al mismo, la cuál se mantiene vigente aún en nuestros días.

Esta necesidad de clasificación no se restringe únicamente a la vida contidiana, sino que se encuentra también presente en el ámbito científico, consituyendo una base esencial de campos como la biología o la sociología. A menudo, la cantidad de información recogida en este tipo de estudios es inabarcable para un ser humano, y se debe recurrir a técnicas que nos permitan sintetizar esta información mediante la formación de grupos de observaciones.

En este contexto, surge el análisis cluster como una herramienta estadística fundamental en el tratamiento de estas grandes cantidades de datos. El objetivo del análisis cluster consiste en identificar agrupaciones naturales de individuos de un conjunto de datos, de forma que los del mismo grupo sean lo más homogéneos posibles, y distintos de los del resto de grupos.

Entre los métodos más utilizados en el análisis cluster, el algoritmo de k-medias destaca por su sencillez conceptual y eficiencia computacional. De forma similar a como la media es la mejor representación unipuntual de un grupo de datos, las k-medias extienden esta idea, y se busca la mejor representación por k elementos. De esta forma, una vez encontrada esta representación, se genera una partición natural de los individuos, asignando cada uno de estos al grupo del representante más similar a él.

No obstante, una de las principales limitaciones del método de k-medias es su sensibilidad a la presencia de valores atípicos, lo que puede afectar significativamente la calidad de la agrupación. Esta falta de consistencia justifica la necesidad de aplicación de técnicas de estadística robusta al clustering.

El objetivo de este trabajo es elaborar un resumen crítico de la evolución de los métodos de clustering robusto basados en el recorte, abarcando desde sus motivaciones iniciales hasta algunos de los desarrollos y extensiones más recientes. Para ello, se ha realizado una revisión bibliográfica de los distintos procedimientos del análisis cluster con recorte apoyada en el trabajo recopilatorio [17], del que se ha seguido una organización y estructura similares.

En cada uno de los métodos expuestos durante el desarrollo de este trabajo se han incluído ejemplos de elaboración propia, empleando conjuntos de datos existentes (los cuáles se especifican tras su utilización) o elaborados artificialmente. Algunos de estos ejemplos están inspirados en los aplicados en estos trabajos, y otros constituyen una aportación personal, representando situaciones de aplicabilidad real de los métodos presentados.

Igualmente, durante la exposición de los correspondientes procedimientos se han discutido las motivaciones de su implementación y su aplicación en la práctica, así como la comparación de estos modelos entre sí, analizando su adecuación a determinados contextos.

En cuanto a las demostraciones presentes este trabajo, se tratan de resultados tomados de la bibliografía mencionada, unificando la notación y simplificándola siempre que sea posible. Además, se han añadido comentarios y razonamientos que facilitan el seguimiento de estas pruebas, completando pasos intermedios que en la versión original se omitían o presuponían. Por último, se han generalizado una serie de resultados clave en la demostración de la existencia de las k-medias recortadas, los cuales estaban expuestos para el caso $k = 1$ en la bibliografía original.

El presente trabajo consta así de cuatro capítulos. En el primero de ellos se expone la falta de robustez del método de k-medias, y se introduce una primera extensión de este algoritmo, incluyendo un recorte de los datos. Este modelo, conocido como k-medias recortadas, supone una solución conceptualmente sencilla del problema de contaminación, que será la base de métodos posteriores más flexibles y sofisticados. Por último, se analizarán algunas de sus propiedades como la existencia de soluciones o la consistencia del método.

En el segundo capítulo se introduce una segunda generación de métodos de clustering robusto, que generalizan a las k-medias recortadas. Estos modelos se sitúan dentro de la metodología TCLUST, que introduce un marco de trabajo más flexible, permitiendo tratar con conjuntos de datos de estructuras más variadas. Además, se describe un modelo de mezcla adaptado al clustering, que permite establecer una clasificación suave de las observaciones.

El tercer capítulo recoge una serie de herramientas y estrategias clave en la elección de los parámetros de los algoritmos presentados previamente. Asimismo, se analiza un conjunto de indicadores de la calidad de la clasificación obtenida por estos algoritmos, permitiendo así identificar un posible error en la determinación de los parámetros.

En el cuarto y último capítulo se recopila una colección de extensiones de los métodos presentados en los capítulos anteriores, adaptados a ramas específicas del análisis

cluster como el clustering sobre espacios lineales o el co-clustering.

Las imágenes y gráficas que aparecen en este trabajo han sido realizadas con el lenguaje de programación R [24], ampliamente conocido y utilizado en el ámbito de la estadística. Para la implementación de algunos de estos ejemplos se ha recurrido al empleo de los paquetes 'tclust'[20], 'blockcluster'[23] y 'palmerpenguins'[1], además de los incluídos ya de base en R.

Capítulo 1

k-medias con recorte

El término "k-medias" aparece por primera vez en 1967, de la mano de James MacQueen, sin embargo, la idea de agrupar puntos en función de su distancia a un "centroide" se le atribuye a Hugo Steinhaus, quien propuso este concepto en 1956. El algoritmo estándar fue desarrollado por Stuart Lloyd en 1957 como una técnica para la modulación por impulsos codificados, aunque su publicación fuera de los laboratorios Bell no ocurrió hasta 1982. Desde entonces, se ha escrito abundante literatura acerca de este método, cubriendo aspectos del mismo como su consistencia o la inicialización óptima del algoritmo, lo que lo han convertido en uno de los procedimientos más importantes del análisis cluster. Es de hecho la base de una gran cantidad de métodos de clustering posteriores, y aún a día de hoy existen líneas de investigación abiertas basadas en este problema.

En este capítulo comenzamos describiendo la idea detrás del método a través de un ejemplo sencillo con el fin de asentar los conceptos fundamentales, y presentamos una de sus mayores deficiencias: la falta de robustez ante la presencia de datos átipicos, la cual constituye el eje principal de este trabajo.

A continuación, presentaremos una de las primeras soluciones planteadas a dicho problema, basada en el recorte de los datos, discutiendo la motivación detrás de esta herramienta. Por último, generalizaremos este problema al caso poblacional, lo que permitirá estudiar las características matemáticas del mismo, como la existencia de soluciones o la consistencia del método.

1.1. El problema de agrupación: método de las k-medias

Como mencionamos en la introducción, el objetivo del análisis clúster es encontrar agrupaciones naturales dentro de un conjunto de datos, de forma que se puedan identificar relaciones entre los individuos, agrupando aquellos que comparten características similares. Esto nos permite, además, elegir un individuo que actúe como representante de cada grupo, resumiendo de la mejor forma posible sus propiedades comunes. Así, conseguimos reducir significativamente la cantidad de datos con la que trabajamos sin perder demasiada información relevante.

Para ello, resulta lógico pensar en alguna medida de centralización que sintetice bien la información de cada grupo. Entre las posibles opciones, la media se presenta como la más adecuada, especialmente si usamos el criterio de mínimos cuadrados.

A partir de estas consideraciones, surge de forma natural la idea de encontrar la mejor representación de un conjunto de datos por k elementos, tomando la distancia euclídea entre dos puntos como medida de su similitud. Es decir, supongamos que disponemos de un conjunto de n individuos $\{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ de los que hemos medido p propiedades. Nuestro objetivo es encontrar k centros óptimos $m_1^*, ..., m_k^* \in \mathbb{R}^p$ tales que:

$$\{m_1^*, ..., m_k^*\} = \underset{m_1,...,m_k}{\arg\, min} \sum_{i=1}^{n} \underset{j=1,...,k}{min} \|x_i - m_j\|^2 \tag{1.1}$$

Ahora disponemos de una representación óptima del conjunto de datos por k elementos, que induce de forma natural una partición, conocida como partición de Voronoi, asignando cada individuo al cluster del elemento m_i^* más próximo. Se definen por lo tanto k clusters de la forma:

$$Cluster\ J = \left\{x_i : \|x_i - m_J^*\|^2 \leq \left\|x_i - m_j^*\right\|^2 \quad para \quad j \neq J\right\} \tag{1.2}$$

Nótese que pueden existir datos que se encuentren en la frontera entre dos o más clusters. Estos pueden ser asignados indistintamente a cualquiera de estos clusters, tomando un criterio común, como por ejemplo, escoger siempre el de menor índice.

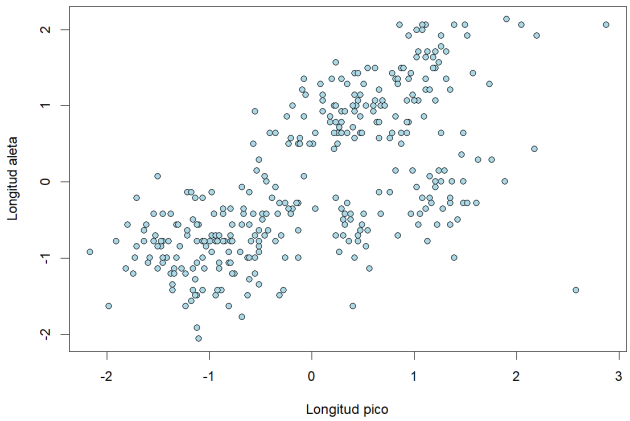
Encontrar los centros que minimizan la expresión (1.1) es un problema computacionalmente muy complejo, luego se requiere de algoritmos que aproximen la solución óptima. El algoritmo estándar de k-medias, presentado por Lloyd (1957), consta de los siguientes sencillos pasos:

Algoritmo clásico de k-medias

1. Se escogen aleatoriamente k centroides iniciales $H = \{h_1, .., h_k\}$.
2. Para cada punto x_i, $i = 1, \ldots, n$:
   (a) Para cada $j = 1, \ldots, k$, se calcula la distancia $d_{i,j}$ de x_i al centro h_j.
   (b) Se asigna x_i al cluster C_j cuyo centro h_j es más cercano.
3. Para cada centro h_j, $j = 1, \ldots, k$, se recalcula el centro de masa de cada cluster mediante la media de las observaciones y se asigna a h_j dicho valor. Esto es
$$h_j \coloneqq \frac{1}{|C_j|} \sum_{x \in C_j} x$$
4. Se repiten los pasos 2) y 3) hasta que H se estabilice.

Con el objetivo de ilustrar la efectividad de este método, se expone a continuación un ejemplo sencillo. Supongamos que disponemos de las mediciones de dos variables (longitud del pico y longitud de la aleta) de 342 pingüinos, y se quiere realizar una clasificación agrupando los pingüinos con características similares.

En la figura 1.1 se han representado las medidas de estas variables, tomadas del famoso conjunto de datos "penguins" de R, de la librería "palmerpenguins", los cuales



	Longitud pico	Longitud aleta
1	-0.88320467	-1.41627153
2	-0.80993901	-1.06069609
3	-0.66340769	-0.42066030
4	-1.32279862	-0.56289047
5	-0.84657184	-0.77623574
6	-0.91983750	-1.41627153
7	-0.86488825	-0.42066030
8	-1.79902541	-0.56289047
9	-0.35202864	-0.77623574
10	-1.12131806	-1.06069609
11	-1.12131806	-1.48738661
12	-0.51687637	-1.34515644

Figura 1.1: Conjunto de pinguinos con dos variables: longitud de pico y de aleta.

han sido escalados y redondeados para facilitar su posterior representación. Este conjunto de datos contiene también la información de la especie a la que pertenece cada pingüino, no obstante, en un problema real de análisis no supervisado no sabríamos a priori las clases (en este caso especies) existentes. De hecho, no sabríamos ni si existen distintas clases. La elección de este conjunto de datos se ha realizado con el fin de comprobar posteriormente la efectividad del método de k-medias, en el que obviamente no tomamos la información de las especies. En caso de que los pingüinos de distintas especies presenten medidas dispares en estas variables, podríamos esperar que el método de k-medias realizase una clasificación que no se alejara demasiado de la clasificación real.

Para la aplicación del algoritmo de k-medias, se debe prefijar el numero de clusters k a determinar, desconociendo el número real de agrupaciones existentes en el conjunto. En nuestro caso particular, donde solo contamos con dos variables, la representación de los datos permite realizar una elección inteligente de k, sin embargo, en dimensiones mayores, la determinación de este parámetro se convierte en uno de los problemas más estudiados acerca de este método. En el capítulo 3 describiremos alguna de las técnicas más utilizadas para la elección sensata de k.

Observando la gráfica de la figura 1.1 podemos identificar un grupo de pinguinos con el tamaño de las aletas mayor que el resto. A su vez, dentro de los que tienen un tamaño más reducido de aletas, parece razonable dividirlos en dos grupos distintos en función de la longitud de su pico, quedando una partición en 3 grupos distintos.

La figura 1.2 muestra el resultado de aplicar el método de k-medias al conjunto de datos descrito previamente para k=2,3,4. El caso k=3 es el que mejor parece clasificar estos individuos, como bien habíamos deducido previamente. Para k=2 el método agrupa conjuntos claramente diferenciados en el mismo cluster, mientras que para k=4 separa en dos clusters distintos los pingüinos con tamaño de aletas grandes, cuando a simple vista parecen conformar un solo grupo y no hay razones para efectuar dicha escisión.

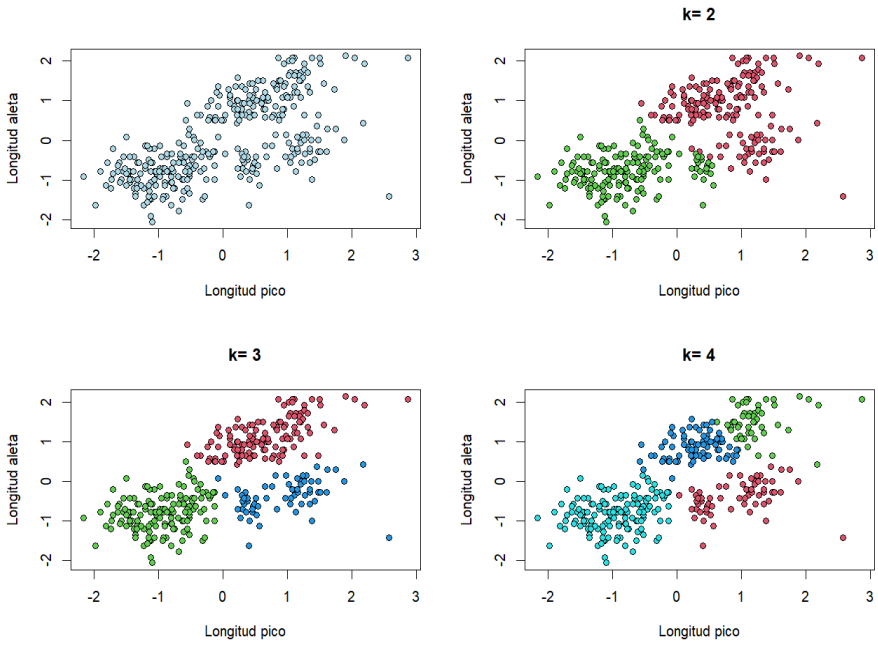
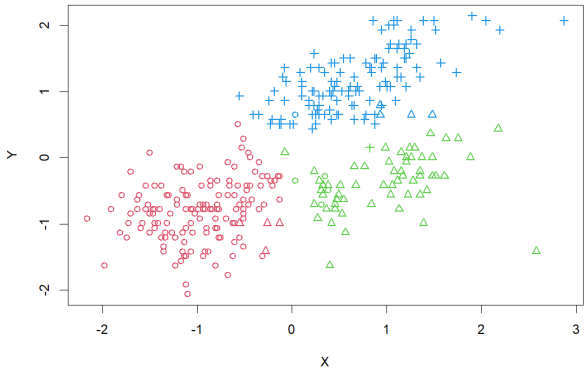


Figura 1.2: Método de k-medias para k=2,3,4 sobre el conjunto de pingüinos.

Haciendo uso ahora de la información de las especies de estos pingüinos, se puede comprobar la similitud entre nuestra partición por k-medias y la clasificación real. Los pingüinos de este conjunto de datos pertenecen a 3 especies distintas (Adelie, Gentoo y Chinstrap), lo que confirma nuestra sospecha de la existencia de 3 grupos de pingüinos diferenciados.

En la figura 1.3 se puede observar la efectividad del método para k=3, donde se han representado con distintos colores los clusters obtenidos por el método de k-medias, y con distintos símbolos los grupos originales de especies de pingüinos. También se ha generado una tabla de contingencia que muestra la coincidencia entre las especies y los clusters resultantes del método. Se puede comprobar como la inmensa mayoría de pingüinos, en torno al 96 por ciento, han sido correctamente agrupados con los de su misma especie, luego el método de k-medias ha sido capaz de identificar la partición subyacente de los datos.

No obstante, este ejemplo ha sido elegido con el fin de ilustrar una aplicación del método, es decir, se ha tomado un conjunto de datos que se adapta a las suposiciones realizadas en el método de k-medias con el objetivo de mostrar su efectividad. Basta observar la definición de los clústeres en (1.2) para deducir que este procedimiento no resulta adecuado cuando los grupos a identificar no están linealmente separados. Ahora bien, podemos confiar en que ofrecerá una clasificación precisa siempre que las clases o especies sean separables mediante hiperplanos.



	Adelie	Gentoo	Chinstrap
Cluster 1	146	0	5
Cluster 2	1	122	4
Cluster 3	4	1	59

Figura 1.3: Comparación entre la clasificación real y la clasificación obtenida por el método de k-medias sobre el conjunto de pinguinos.

1.1.1. Problemas e inconvenientes del método: falta de robustez

El método de k-medias es un procedimiento ampliamente utilizado por su sencillez conceptual y eficiencia. No obstante, presenta algunas deficiencias que limitan su uso, especialmente cuando el conjunto de datos no es apropiado. Las restricciones más evidentes de este modelo son consecuencia de las asunciones de convexidad, tamaño similar, y separabilidad por hiperplanos de los grupos. Estas suposiciones convierten al método de k-medias en un procedimiento inadecuado para la clasificación de grupos con formas arbitrarias o clústeres solapados. Otro de los inconvenientes más conocidos, que es el que nos ocupa en este trabajo, es su falta de robustez, es decir, su sensibilidad a la presencia de datos atípicos o ruido.

En muchos estudios es habitual la aparición de observaciones desviadas de la norma debido a mediciones erróneas o valores atípicos. Por ello, es fundamental que el método de clustering utilizado sea robusto y capaz de mantener un buen rendimiento incluso en presencia de este tipo de irregularidades. Sin embargo, el algoritmo clásico de k-medias, calcula los centroides como la media aritmética de los datos pertenecientes a un cluster, y una sola observación suficientemente lejana puede desviar el centroide significativamente, alterando por completo la clasificación. De hecho, este tipo de datos tienen más relevancia en el resultado, en el sentido de determinar los clusters, que cualquiera de los datos regulares, lo cual es lo contrario a nuestro propósito.

Consideramos de nuevo el conjunto de datos presentado anteriormente, al que se ha añadido contaminación artificial con el fin de analizar su efecto sobre el método de k-medias. La figura 1.4 muestra la aplicación del algoritmo con k=3 sobre el conjunto mencionado con dos tipos distintos de contaminación. En la gráfica de la izquierda se han añadido 30 observaciones generadas por una distribución normal de gran dispersión centrada en el origen. En la derecha, en cambio, se han incorporado solamente 4 observaciones adicionales, con valores atípicamente bajos para la variable correspondiente a la longitud del pico. En ambos casos puede apreciarse como la presencia de estos datos anómalos provoca la agrupación artificial de clusters distintos en uno solo, evidenciando así la falta de robustez del algoritmo. Este efecto es especialmente llamativo en el

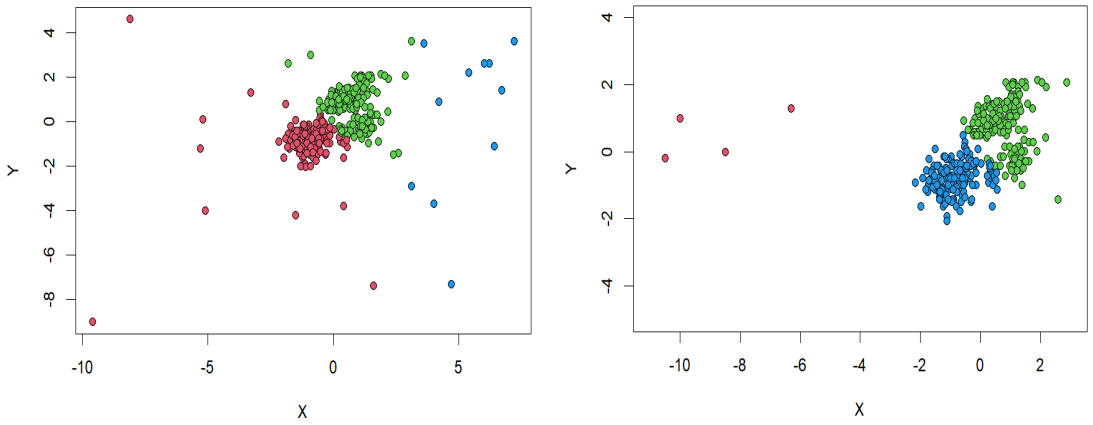


Figura 1.4: Efecto del ruido sobre la clasificación por k-medias del conjunto de pingüinos.

segundo caso, donde apenas un número muy reducido de datos atípicos bastan para alterar significativamente el resultado.

Una primera idea para solucionar este inconveniente podría consistir en aumentar el número de clusters k. Esta solución es interesante y funciona en numerosos contextos, especialmente cuando los datos atípicos aparecen agrupados, como en el segundo caso de nuestro ejemplo. Sin embargo, este método no resulta válido en el caso general, y, aún cuando funciona correctamente, puede dar lugar a la aparición de pequeños clusters aislados con difícil interpretación en la práctica. A esto se le suma el hecho de que, a menudo, se define el número de grupos a buscar con un determinado propósito, en función de las características de nuestro problema, lo que impide considerar clusters adicionales.

Otra posible solución consistiría en modelar explícitamente el ruido, por ejemplo, considerando componentes distribuídas uniformemente o asociadas a distribuciones de colas pesadas. Si bien esta técnica también puede resultar efectiva en numerosos escenarios, implicaría realizar suposiciones sobre la contaminación de los datos, lo cual puede ser problemático si la contaminación real dista significativamente de nuestra hipótesis. El objetivo es desarrollar un método flexible y adaptable a numerosos contextos, independientememte de la naturaleza impredecible de los datos irregulares.

Existen en la literatura numerosas propuestas de clustering robusto resistentes a la presencia de datos atípicos. En este trabajo nos centraremos en aquellas basadas en el recorte de los datos, una de las técnicas más antiguas de la estadística robusta, cuyo uso sigue vigente hoy en día debido a su efectividad y sencillez. La idea detrás de este procedimiento se basa en aplicar herramientas estadísticas usuales una vez se han descartado los datos más susceptibles de ser atípicos o contaminantes.

Un ejemplo típico del recorte de datos es la denominada media recortada univariante, que efectúa la media de un conjunto de datos una vez descartadas la proporción $\frac{\alpha}{2}$ de las observaciones más grandes y la proporción $\frac{\alpha}{2}$ de las observaciones más pequeñas, evitando que el resultado obtenido se vea muy afectado por valores atípicamente extremos.

No obstante, no existe una generalización directa del recorte para el caso multivariante debido a la ausencia de un orden natural de los datos, es decir, no podemos hablar de observaciones "grandes" o "pequeñas". Además, la "anomalía" de una observación es un concepto extrínseco a la misma, dicho de otra manera, es el contexto el que determina si una observación es irregular, y no sus propias características. Por lo tanto, debemos considerar un recorte "imparcial", esto es, debe ser el propio conjunto de datos el que determine cuales son las observaciones más atípicas, que serán recortadas, y no la elección de una dirección o norma predefinida.

En la siguiente sección veremos un primer método de clustering basado en el recorte imparcial de los datos, extendiendo así el método de k-medias. Este enfoque busca solucionar la falta de robustez del algoritmo clásico, manteniendo una buena detección de agrupaciones en presencia de ruido.

1.2. Introducción del recorte

Un primer método robusto de agrupación de datos basado en el recorte aparece de la mano de Cuesta-Albertos, J. A., Gordaliza, A., Matrán, C.[4], donde se introducen las denominadas k-medias recortadas como una extensión de las k-medias clásicas, de una manera análoga a como la media recortada extiende a la media clásica. Este trabajo se apoya en la idea de Gordaliza (1991a)[18] sobre la aproximación a variables aleatorias mediante el recorte, adaptándolo al contexto del clustering, y encontrando así la mejor aproximación a un vector aleatorio por k elementos.

La idea detrás del las k-medias recortadas se basa en aplicar el método de agrupación clásico tras descartar una proporcion α de las observaciones, resaltando la proporcion $1-\alpha$ restante, conformada por las observaciones consideradas más "creíbles". No esperamos que todas las observaciones recortadas sean datos atípicos, pero sí que la gran mayoría de datos atípicos hayan sido descartados. De esta manera, los datos más anómalos no participarían en la determinación de los clusters, evitando que los centroides se vean desviados por estas observaciones extremas. Con esta idea, se define a continuación el problema de k-medias para el caso muestral.

Problema de k-medias recortadas (caso muestral)

Dada una muestra $x_1, x_2, ..., x_n$ en $\mathbb{R}^p$, queremos encontrar k centros $m_1, m_2, ..., m_k$ (denominados centros de k-medias recortadas) y una partición $\{R_0, R_1, ..., R_k\}$ de los índices $\{1, 2, ..., n\}$ con R_0 conteniendo una proporción $[n\alpha]$ de los índices y minimizando:

$$\sum_{j=1}^{k} \sum_{i \in R_j} \|x_i - m_j\|^2 \tag{1.3}$$

Nótese que el sumatorio va desde 1 hasta k, luego la proporción α de observaciones contenidas en R_0 han sido omitidas.

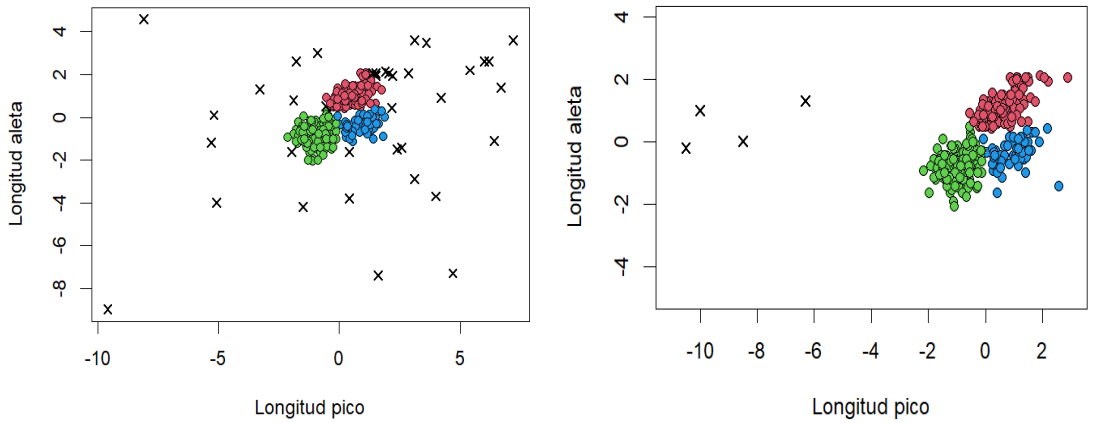


Figura 1.5: Resultado de aplicar 3-medias recortadas al conjunto de datos de la figura 1.4 con $\alpha = 0.1$ (izquierda) y $\alpha = 0.012$ (derecha) para distintas contaminaciones.

Para comprobar la potencia del método, se ha aplicado el procedimiento de 3-medias recortadas a los conjuntos de datos de la figura 1.4, tomando $\alpha = 0.1$ y $\alpha = 0.012$ respectivamente. El resultado obtenido se muestra en la figura 1.5, donde los datos recortados han sido representados mediante cruces. Se puede observar como ahora el método sí que detecta correctamente los clusters buscados, pues la mayoría de los datos contaminantes que introdujimos artificialmente han sido recortados.

Observando la función objetivo (1.3) a minimizar, se deduce que el conjunto de observaciones recortadas R_0 está formado por la fracción α de las observaciones más distantes a su centro más cercano. Esta regla descarta los datos potencialmente atípicos, priorizando aquellos que se ajustan a la estructura de los datos. Por otro lado, los centros óptimos se corresponden a los obtenidos tras aplicar el método de k-medias clásico sobre el conjunto de datos no recortados. En consecuencia, el método de k-medias recortadas se reduce al método de k-medias clásico para $\alpha = 0$.

1.3. Estudio teórico de las k-medias recortadas

Hasta ahora hemos abordado el problema de agrupación desde una perspectiva muestral, es decir, partiendo de un conjunto finito de datos y buscando los k representantes que mejor lo describen. En adelante buscaremos generalizar el problema a un marco teórico que nos permita no solo tratar conjuntos de datos, sino distribuciones de probabilidad.

En este nuevo enfoque, el objetivo ya no es solo encontrar los k puntos que mejor representan a una muestra, sino identificar los k representantes óptimos de una distribución de probabilidad dada. Consideramos por lo tanto un vector aleatorio X definido en un espacio probabilístico (Ω, σ, P), siendo P_X la distribución de probabilidad asociada a X. Esta generalización permite analizar la influencia de los datos atípicos no solo en una muestra particular, sino en el comportamiento global del procedimiento de agrupación en un contexto más abstracto.

Nótese además que, este planteamiento, incluye el caso en que disponemos de un conjunto de n individuos, $\{x_1, \ldots, x_n\} \subset \mathbb{R}^p$, que es el que hemos tratado hasta ahora. Basta considerar la probabilidad muestral P_n, que asigna una probabilidad $\frac{1}{n}$ a cada punto x_i.

El paso del caso muestral al caso poblacional requiere además sustituir la función objetivo en (1.3) por un análogo teórico definido sobre una distribución de probabilidad. La media, como mejor representante de un conjunto de datos, se generaliza a través del concepto de esperanza matemática, que se define como

$$E(X) = \int x \; dP_X(x) \tag{1.4}$$

La elección de la esperanza matemática como mejor representante de una distribución está fundamentada por el criterio de mínimos cuadrados:

$$E(X) = \underset{a \in \mathbb{R}^d}{arg\,\text{mín}} \int_{\mathbb{R}^d} \|x - a\|^2 \; dP_X(x) \tag{1.5}$$

No obstante, podemos generalizar aún más el modelo, considerando otros representantes asociados a otras medidas de disimilaridad. Sea una función $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$, creciente, contínua, con $\phi(0) = 0$ y tal que $\phi(x) < \phi(\infty)$ para todo x, mediremos entonces la discrepancia entre dos puntos x e y como $\phi(\|x-y\|)$, permitiéndonos extender el concepto de media al de ϕ-media:

Definición 1. *Sea $X : (\Omega, \sigma, P) \longrightarrow \mathbb{R}^d$ un vector aleatorio y sea P_X la probabilidad inducida por X en $\mathbb{R}^d$. Sea una función de penalización $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$, creciente, contínua, con $\phi(0) = 0$ y tal que $\phi(x) < \phi(\infty)$ para todo x. Suponemos que $\exists\ a \in \mathbb{R}^d$ tal que $\int \phi(\|x - a\|) \; dP_X(x) < \infty$. Entonces se denomina ϕ-media de X a un vector $m \in \mathbb{R}^d$ que verifique:*

$$m = \underset{a \in \mathbb{R}^d}{\arg\,\text{mín}} \int \phi(\|x - a\|) \; dP_X(x)$$

Con estos ingredientes, podemos generalizar la función objetivo a minimizar para el caso poblacional. Una forma natural de formalizar esta idea es definir la variación de un conjunto de k elementos, M, respecto un conjunto de Borel, A.

Definición 2. *Sean $\alpha \in (0, 1)$, $k \in \mathbb{N}$ y ϕ una función de penalización. Para todo conjunto A tal que $P(A) \geq 1 - \alpha$ y todo conjunto de cardinal k, $M = \{m_1, m_2, ..., m_k\}$, en $\mathbb{R}^p$, se define la variación respecto de M dado A como:*

$$V_\phi^A(M) \coloneqq \frac{1}{P(A)} \int_A \phi \left(\underset{i=1,\ldots,k}{inf} \|X - m_i\| \right) \; dP_X$$

Esta variación mide qué tan bien representa el conjunto M la distribución de la probabilidad P condicionada a A. Nuestro objetivo por lo tanto es encontrar la mejor de las representaciones de la distribución en el conjunto más "adecuado" que contenga

la masa de probabilidad dada. Así pues, se puede definir el problema de k-ϕ-medias recortadas como sigue.

Definición 3. *Sea $X : (\Omega, \sigma, P) \longrightarrow \mathbb{R}^p$ un vector aleatorio y sea P la probabilidad inducida por X en $\mathbb{R}^d$. Sean $\alpha \in (0,1)$, k un número natural y $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función creciente, continua y tal que $\phi(0) = 0$. Si denotamos:*

$$V_{\phi,\,k}^{A} \coloneqq \inf_{\substack{M \subset \mathbb{R}^p \\ \#M = k}} V_{\phi}^{A}(M)$$

$$V_{k,\,\phi,\,\alpha} \coloneqq \inf_{\substack{A \in \beta^p \\ P(A) \geq 1-\alpha}} V_{\phi,\,k}^{A}$$

El problema de k-ϕ-medias recortadas consiste en encontrar un conjunto A_0 con $P(A_0) \geq 1 - \alpha$ y un conjunto de cardinal k, $M_0 = \{m_1, m_k, ..., m_k\}$, tales que:

$$V_{\phi}^{A_0}\,(M_0) = V_{k,\,\phi,\,\alpha} \tag{1.6}$$

En caso de existir, M_0 recibe el nombre de k-ϕ-media recortada. No obstante, para facilitar la notación lo denominaremos k-media recortada, y hablaremos del problema de k-medias recortadas.

Una vez presentado el problema de k-medias recortadas, nuestro objetivo es estudiar alguna de sus propiedades básicas como la existencia de soluciones y la consistencia del método. Para ello, seguiremos el enfoque utilizado en Gordaliza(1991a)[18], basado en el uso de funciones de recorte, que generalizan a los conjuntos de recorte, permitiendo una mayor flexibilidad en su tratamiento. No obstante, veremos que, una vez estudiada la mejor solución posible por funciones de recorte, esta es en esencia la función indicatriz de una unión de bolas con el mismo radio.

Para el manejo de este tipo de funciones debemos introducir una serie de nociones y resultados que faciliten los desarrollos posteriores.

1.3.1. Funciones de recorte: notación y resultados clave

En adelante trabajaremos en un espacio probabilístico (Ω, σ, P), donde $X : \Omega \longrightarrow \mathbb{R}^p$ es un vector aleatorio con ley de probabilidad P_X en la σ-álgebra de Borel, $\mathscr{B}^p$. La función de penalización $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ se supone contínua, no decreciente y tal que $\phi(0) = 0$ y $\phi(x) < \phi(\infty)$ para todo x.

Denotaremos por $d(x, y)$ la distancia euclídea en $\mathbb{R}^p$. Para $x \in \mathbb{R}^p$ y $C,\ D \subset \mathbb{R}^p$, denotaremos

$$d(x, C) \;=\; \inf_{y\,\in\,C} d(x, y)$$

y también

$$d(C, D) \;=\; \sup \left\{ \sup_{x\,\in\,C} d(x, D),\ \sup_{y\,\in D} d(C, y) \right\}$$

Nótese que esta última distancia, que usaremos como medida de disparidad de dos conjuntos de k elementos, coincide con la distancia de Haussdorff para conjuntos acotados.

Introducimos ahora la noción de funciones de recorte, que constituyen la herramienta clave para nuestros propósitos. Sea $\alpha \in (0,1)$, τ_α denota el conjunto de funciones de recorte para X de nivel α, es decir:

$$\tau_\alpha = \left\{ \tau : \mathbb{R}^p \longrightarrow [0,1], \text{ medible y tal que } \int \tau(X)\ dP_X = 1-\alpha \right\}$$

y, τ_{α^-} denota el conjunto de funciones de recorte de nivel $\beta \leq \alpha$, esto es,

$$\tau_{\alpha^-} = \left\{ \tau : \mathbb{R}^p \longrightarrow [0,1], \text{ medible y tal que } \int \tau(X)\ dP_X \geq 1-\alpha \right\} = \bigcup_{\beta \leq \alpha} \tau_\beta$$

Estas familias de funciones son una generalización de las funciones indicatrices de conjuntos con probabilidad $1-\alpha$ (resp. al menos $1-\alpha$), incluyendo la posibilidad de que algunos puntos participen parcialmente en el recorte.

Esta generalización permite extender de forma natural el problema de k-medias recortadas a través de la definición de variación respecto M dada una función de recorte.

Definición 4. *Sean $\alpha \in (0,1)$, $k \in \mathbb{N}$ y ϕ una función de penalización. Para todo función $\tau \in \tau_{\alpha^-}$ y todo conjunto de cardinal k, $M = \{m_1, m_2, ..., m_k\}$ en $\mathbb{R}^p$, se define la variación respecto M dado τ como:*

$$V_\phi^\tau(M) \coloneqq \frac{1}{\int \tau(X)\ dP_X} \int \tau(X)\ \phi\left(d(X,M)\right)\ dP_X$$

Esta noción permite definir el problema de k-medias recortadas en el marco de las funciones de recorte de manera análoga al problema para conjuntos de recorte. Si denotamos por

$$V_{\phi,\ k}^\tau \coloneqq \inf_{\substack{M \subset \mathbb{R}^p \\ \#M = k}} V_\phi^\tau(M)$$

$$V_{k,\ \phi,\ \alpha} \coloneqq \inf_{\tau \in \tau_{\alpha^-}} V_{\phi,\ k}^\tau$$

el objetivo es encontrar una función de recorte τ_0 y un conjunto de k elementos M_0, si existen, tales que

$$V_\phi^{\tau_0}(M_0) = V_{k,\ \phi,\ \alpha} \tag{1.7}$$

Es claro que la solución obtenida por funciones de recorte, será mejor que la obtenida por conjuntos de recorte, pues si A_0 es el conjunto para el que se cumple (1.6) entonces tomando $I_{A_0} \in \tau_{\alpha^-}$, se tendrá $V_\phi^{A_0}(M_0) = V_\phi^{I_{A_0}}(M_0) \geq V_{k,\ \phi,\ \alpha}$.

A continuación, se exponen una serie de resultados esenciales en la demostración de la existencia de las k-medias recortadas, adaptados de [18] y generalizados para k elementos.

Lema 1.1. *Sean $M = \{m_1, ..., m_k\} \subset \mathbb{R}^p$ y $\beta \in (0,1)$. Denotaremos la bola generalizada centrada en M y de radio $r \geq 0$ por*

$$B(M,r) = \bigcup_{i=1}^{k} B(m_i, r),$$

y sean

$$r_\beta(M) = inf\left\{r \geq 0 : P_X(B(M,r)) \leq 1-\beta \leq P_X(\overline{B}(M,r))\right\}$$

y

$$\tau_{M,\ \beta} = \left\{\tau \in \tau_\beta : I_{B(M,r_\beta(M))} \leq \tau \leq I_{\overline{B}(M,r_\beta(M))}\right\}$$

entonces, para todo $\tau \in \tau_{M,\ \beta}$ se tiene:

(a) $\int \tau(X)\phi(d(X,M))\ dP_X \leq \int \tau'(X)\phi(d(X,M))\ dP_X$ *para todo* $\tau' \in \tau_\beta$

(b) Si ϕ es estrictamente creciente, entonces la desigualdad en (a) es estricta si y solo si $\tau' \in \tau_\beta - \tau_{M,\ \beta}$

Demostración. Para facilitar la notación durante la demostración denotaremos $B = B(M, r_\beta(M))$. Sean $\tau' \in \tau_\beta$ y $\tau \in \tau_{M,\ \beta}$. Teniendo en cuenta que $I_B \leq \tau \leq I_{\overline{B}}$, es fácil ver que:

$$\tau(x)(1-\tau'(x)) = 0 \qquad para\ todo \quad x \notin \overline{B} \tag{1.8}$$

$$\int \tau(X)(1-\tau'(X))\ dP = \int \tau'(X)(1-\tau(X))\ dP \tag{1.9}$$

y

$$\tau'(x)(1-\tau(x)) = 0 \qquad para\ todo \quad x \in B \tag{1.10}$$

Nótese que en (1.9) se ha utilizado que $\tau, \tau' \in \tau_\beta$, y en consecuencia, $\int \tau(X)\ dP = \int \tau'(X)\ dP = 1-\beta$. Ahora, aplicando (1.8), (1.9) y (1.10) sucesivamente se tiene que:

$$\begin{aligned}
&\int \tau(X)(1-\tau'(X))\ \Phi(d(X,M))\ dP_X \\
&\leq \Phi(r_\beta(M)) \int \tau(X)(1-\tau'(X))\ \ dP_X \qquad (1.11)\\
&= \Phi(r_\beta(M)) \int \tau'(X)(1-\tau(X))\ \ dP_X \\
&\leq \int \tau'(X)(1-\tau(X))\Phi(d(X,M))\ \ dP_X \qquad (1.12)
\end{aligned}$$

Y, por la tanto, deducimos que

$$\begin{aligned}\int \tau(X)\,\Phi(d(X,M))\,dP_X &= \int \tau(X)\,\tau'(X)\,\Phi(d(X,M)\,dP_X \\ &\quad + \int \tau(X)(1-\tau'(X))\,\Phi(d(X,M))\,dP_X \\ &\leq \int \tau(X)\,\tau'(X)\,\Phi(d(X,M))\,dP_X \\ &\quad + \int \tau'(X)(1-\tau(X))\,\Phi(d(X,M))\,dP_X \\ &= \int \tau'(X)\,\Phi(d(X,M))\,dP_X;\end{aligned}$$

y hemos probado la primera parte del lema. Obsérvese que la igualdad se cumple si y solo si se da la igualdad en (1.11) y (1.12). Ahora bien, la igualdad en (1.11) se cumple si y solo si

$$\int_B \tau(x)(1-\tau'(x))\ dP = 0,$$

equivalentemente,

$$\int_B (1-\tau'(x))\ dP = 0,$$

es decir,

$$I_B \leq \tau' \qquad P-c.t.p$$

De forma análoga, la igualdad en (1.12) se da si y solo si

$$\int_{\overline{B}^C} \tau'(x)(1-\tau(x))\ dP = 0,$$

equivalentemente,

$$\int_{\overline{B}^C} \tau'(x)\ dP = 0,$$

esto es,

$$\tau' \leq I_{\overline{B}} \qquad P-c.t.p$$

Por lo tanto, la igualdad de (a) se cumple si y solo si $I_B \leq \tau' \leq I_{\overline{B}}$ para casi todo x con respecto a la medida P_X, es decir, si $\tau' \in \tau_{M,\,\beta}$. □

A partir de este lema se deduce que la variación β-recortada sobre M, definida por:

$$V_{\Phi,\,\beta}(M) \coloneqq \frac{1}{1-\beta}\int \tau(X)\Phi(d(X,M))\ dP$$

coincide para cualquier función en $\tau_{M,\,\beta}$, por tanto, usaremos esta misma notación

para referirnos a cualquier función de $\tau_{M,\ \beta}$. Nótese que el nombre de β-variación sobre M es consistente, pues no depende de la función de recorte τ tomada.

El siguiente lema nos ayudará a comprobar un resultado sencillo en concepto, pero no tanto en su prueba: aumentar la proporción de recorte siempre conduce a disminuir el valor de la función objetivo.

Lema 1.2. *Manteniendo la notación del Lema 1.1, si $\beta \leq \alpha$, se tiene que:*

(a) $V_{\Phi,\alpha}(M) \leq V_{\Phi,\beta}(M)$

(b) si Φ es estrictamente creciente, entonces la igualdad en (a) se da si y solo si $r_\alpha(M) = r_\beta(M)$ *y* $P_X[B(M, r_\alpha(M))] = 0$.

Demostración. Primero nótese que, para todo $\alpha \in (0,1)$, sean $\tau, \tau' \in \tau_{M,\ \alpha}$ se tiene

$$\int \tau(X)\, \Phi(d(X,M))\, dP = \int \tau'(X)\, \Phi(d(X,M))\, dP.$$

Podemos asumir por lo tanto sin pérdida de generalidad que $\tau_{M,\beta}(X) \geq \tau_{M,\alpha}(X)$ para casi todo x.

De hecho, siempre es posible elegir $\tau_{M,\beta}$ y $\tau_{M,\alpha}$ tales que $\tau_{M,\beta} \geq \tau_{M,\alpha}$ en todos los puntos. En consecuencia,

$$\int (\tau_{M,\beta}(X) - \tau_{M,\alpha}(X))\ \ \Phi(d(X,M))\, dP \geq \Phi(r_\alpha(M)) \int (\tau_{M,\beta}(X) - \tau_{M,\alpha}(X))\, dP. \tag{1.13}$$

Además, como $\tau_{M,\ \alpha} \leq I_{\overline{B}(M, r_\alpha(M))}$, también se deduce que

$$\Phi(r_\alpha(M)) \int \tau_{M,\alpha}(X)\, dP \geq \int \tau_{M,\alpha}(X)\, \Phi(d(X,M))\, dP. \tag{1.14}$$

Ahora, aplicando (1.13) y (1.14) sucesivamente:

$$\begin{aligned}
&\int \tau_{M,\alpha}(X)\, dP \int (\tau_{M,\beta}(X) - \tau_{M,\alpha}(X))\, \Phi(d(X,M))\, dP \\
&\geq \int \tau_{M,\alpha}(X)\, dP\ \ \Phi(r_\alpha(M)) \int (\tau_{M,\beta}(X) - \tau_{M,\alpha}(X))\, dP \\
&\geq \int \tau_{M,\alpha}(X)\, \Phi(d(X,M))\, dP \int (\tau_{M,\beta}(X) - \tau_{M,\alpha}(X))\, dP
\end{aligned}$$

Por lo tanto

$$\begin{aligned}
&\int \tau_{M,\alpha}(X)\,dP \int \tau_{M,\beta}(X)\,\Phi(d(X,M))\,dP \\
&= \int \tau_{M,\alpha}(X)\,dP \int \tau_{M,\alpha}(X)\,\Phi(d(X,M))\,dP \\
&\quad + \int \tau_{M,\alpha}(X)\,dP \int (\tau_{M,\beta}(X) - \tau_{M,\alpha}(X))\,\Phi(d(X,M))\,dP \\
&\geq \int \tau_{M,\alpha}(X)\,dP \int \tau_{M,\alpha}(X)\,\Phi(d(X,M))\,dP \\
&\quad + \int \tau_{M,\alpha}(X)\,\Phi(d(X,M))\,dP \int (\tau_{M,\beta}(X) - \tau_{M,\alpha}(X))\;dP \\
&= \int \tau_{M,\beta}(X)\,dP \int \tau_{M,\alpha}(X)\,\Phi(d(X,M))\,dP
\end{aligned}$$

es decir,

$$(1-\alpha)\int \tau_{M,\beta}(X)\,\Phi(d(X,M))\,dP \geq (1-\beta)\int \tau_{M,\alpha}(X)\,\Phi(d(X,M))\,dP.$$

Lo que implica que $V_{\Phi,\alpha}(M) \leq V_{\Phi,\beta}(M)$, justo como queríamos probar.

Ahora bien, la igualdad en (a) se cumple si y solo si (1.13) y (1.14) son igualdades. Veamos cuando se dan ambas.

La igualdad en (1.13) se cumple si y solo si

$$\int_{\overline{B}^C} (\tau_{M,\beta}(X) - \tau_{M,\alpha}(X))\;dP = 0$$

(donde $B = B(M, r_\alpha(M))$ es la bola generalizada de centro M y radio $r_\alpha(M)$).
Esta igualdad que se cumple si y solo si $\int_{\overline{B}^C} \tau_{M,\beta}(X) = 0$, lo cual ocurre si y solo si $r_\alpha(M) = r_\beta(M)$.

Análogamente, la igualdad en (1.14) se da si y solo si

$$\int_B \tau_{M,\alpha}(x)\,dP = 0,$$

o equivalentemente cuando,

$$P(B) = 0$$

□

Proposición 1.1. *Manteniendo la notación de los lemas 1.1 y 1.2, se tiene que*

$$V_{k,\ \Phi,\ \alpha} = \inf_{\substack{M \subset \Re^{\ p} \\ \#M = k}} V_{\Phi,\ \alpha}(M)$$

Demostración. Sea $\tau \in \tau_{\alpha^-}$ y $M \subset \mathbb{R}^p$, con $\#M = k$. Aplicando sucesivamente los lemas 1.2 y 1.1 se tiene

$$\frac{1}{1-\alpha}\int \tau_{M,\ \alpha}(X)\ \Phi(d(X,M))\ dP \leq \frac{1}{\int \tau(X)\ dP}\int \tau(X)\ \Phi(d(X,M))\ dP$$

es decir,

$$V_{\Phi,\ \alpha}(M) \leq V_{\Phi}^{\tau}(M), \qquad \textit{para cada } \tau \in \tau_{\alpha^-}$$

y a partir de esto se deduce el resultado.

□

Este resultado nos permite reducir el problema de k-medias recortadas por funciones de recorte a la búsqueda de un conjunto de k elementos, $M_0 = \{m_1, \ldots, m_k\} \subset \mathbb{R}^p$, tal que

$$V_{\Phi,\ \alpha}(M_0) = V_{k,\ \phi,\ \alpha} \tag{1.15}$$

Recuérdese que, en virtud del Lema 1.1, la variación β-recortada respecto un conjunto M, $V_{\Phi,\ \beta}(M)$, se minimiza tomando cualquier función de $\tau_{M,\ \beta}$, que es en esencia la función indicatriz de una bola generalizada centrada en M. Por lo tanto, en caso de existir, la solución del problema de k-medias recortadas por funciones de recorte es la solución del problema de k-medias recortadas básico, es decir, por conjuntos de recorte.

Por otro lado, el Lema 1.2 muestra que para minimizar la variación β-recortada sobre M, es estrictamente mejor recortar la cantidad exacta α, salvo el caso en el que la masa de probabilidad de $\overline{B}(M, r_\alpha(M))$ esté concentrada en la frontera de la bola.

1.4. Existencia de k-medias recortadas

Una vez presentado el problema de k-medias recortadas, nos centraremos en analizar algunas de sus propiedades básicas. Comenzamos estudiando el problema de existencia, esto es, nos preguntamos si siempre existe un conjunto de k elementos que minimice la variación de X, es decir, que verifique (1.15). Para estudiar esta cuestión seguiremos la serie de resultados planteados en [4], que aseguran la existencia de k-medias recortadas para cualquier vector aleatorio X cuando $\alpha \in (0,1)$.

En el caso $\alpha = 0$, que corresponde al método de k-medias clásico, la función objetivo a minimizar puede escribirse como

$$V_{\phi}^{Sop(X)}(M) \coloneqq \int_{Sop(X)} \phi\left(\inf_{i=1,\ldots,k} \|X - m_i\|\right)\ dP_X$$

luego es necesario imponer ciertas restricciones que aseguren que la integral es finita para algún $M \subset \mathbb{R}^p$.

En el caso $\alpha > 0$, este inconveniente desaparece, pues siempre nos podemos restringir a un compacto en el que dicha integral será finita. Basta por ejemplo considerar una bola $B = B(0, r)$ tal que $P_X(B) \geq 1 - \alpha$, y se tiene

$$V_{k,\ \phi,\ \alpha}(X) \leq V_\phi^B(0) = \frac{1}{P_X(B)} \int I_B(X)\ \phi(d(X,0))\ dP_X \leq \phi(r) < \infty \tag{1.16}$$

Gracias al recorte, podemos entonces asegurar la existencia de una solución óptima, resultado que recogemos en el siguiente teorema, y que supone el objetivo principal de esta sección.

Teorema 1.1. ***Existencia de k-medias recortadas***
Sea $X : \Omega \longrightarrow \mathbb{R}^p$ un vector aleatorio con probabilidad asociada P_X. Sean $\alpha \in (0, 1)$, $k \in \mathbb{N}$ y $\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función contínua, no decreciente y tal que $\phi(0) = 0$ y $\phi(x) < \phi(\infty)$ para todo x. Entonces, existe una k-media recortada de X.

Antes de demostrar este teorema, presentamos dos resultados claves para su prueba. Primero veremos la continuidad de la variación recortada $V_{\Phi,\ \alpha}(M)$ con respecto a M, lo cual será también esencial en la consistencia del método. Posteriormente, comprobaremos la idea natural de que esta variación mejora al añadir más clusters, pues es evidente que un mayor número de representantes aproximará mejor la estructura de la distribución.

Proposición 1.2. *Sea $M_n = \{m_1^n, \ldots, m_k^n\},\ n = 0, 1, 2, \ldots$, una sucesión de conjuntos de k elementos en $\mathbb{R}^p$ que satisface*

$$M_n \to M_0 \quad \textit{en la distancia de Hausdorff cuando} \quad n \to \infty$$

Entonces se tiene

$$V_{\Phi,\alpha}(M_n) \to V_{\Phi,\alpha}(M_0) \quad \textit{cuando } n \to \infty.$$

Demostración. Con el fin de hacer la demostración más clara denotaremos $r_n = r_\alpha(M_n)$ y $\tau_n = \tau_{M_n,\,\alpha}$, para $n = 0,\ 1, \ldots$. Es sencillo probar que $\lim_{n\to\infty} r_n = r_0$.

Denotemos por $D_n = |V_{\Phi,\,\alpha}(M_n) - V_{\Phi,\,\alpha}(M_0)|$. Entonces, se tiene que

$$\begin{aligned}
(1-\alpha)\, D_n &= \left| \int \tau_n(X)\, \Phi\big(d(X, M_n)\big)\, dP - \int \tau_0(X)\, \Phi\big(d(X, M_0)\big)\, dP \right| \\
&\leq \left| \int \tau_n(X) \big(\Phi\big(d(X, M_n)\big) - \Phi\big(d(X, M_0)\big) \big)\, dP \right| \\
&\quad + \left| \int \big(\tau_n(X) - \tau_0(X)\big) \Phi\big(d(X, M_0)\big)\, dP \right| \\
&=: G_n + H_n
\end{aligned}$$

Donde se ha aplicado la desigualdad triangular. Basta ver entonces que $lim_{n\to\infty} G_n = lim_{n\to\infty} H_n = 0$ para deducir el resultado.

Considerando la distancia de Haussdorff, sabemos que $d(X, M_n) - d(X, M_0) \to 0$ cuando $n \to \infty$. Además, en virtud del teorema de Heine-Cantor, Φ es uniformemente contínua en todo conjunto compacto, por lo que

$$G_n \leq \int \tau_n(X) |\Phi\big(d(X, M_n)\big) - \Phi\big(d(X, M_0)\big)| \, dP$$

$$\leq (1-\alpha)\Big(\sup_{x \in \overline{B}(M_n, r_n)} |\Phi\big(d(x, M_n)\big) - \Phi\big(d(x, M_0)\big)| \Big) \to 0 \quad \text{cuando } n \to \infty.$$

Para demostrar que $H_n \to 0$ cuando $n \to \infty$, denotaremos

$$E_n := \{x \in \mathbb{R}^p : \tau_n(x) > \tau_0(x)\} \quad \text{y} \quad F_n := \{x \in \mathbb{R}^p : \tau_n(x) < \tau_0(x)\}.$$

Como $\tau_n \in \tau_\alpha$ para $n = 0, 1, 2, \ldots$, entonces se tiene

$$0 = \int \big(\tau_n(x) - \tau_0(x)\big) \, dP_X$$

$$= \int_{E_n} \big(\tau_n(x) - \tau_0(x)\big) \, dP_X + \int_{F_n} \big(\tau_n(x) - \tau_0(x)\big) \, dP_X$$

de donde se deduce que

$$\int_{E_n} \big(\tau_n(x) - \tau_0(x)\big) \, dP_x = \int_{F_n} \big(\tau_0(x) - \tau_n(x)\big) \, dP_x.$$

Además, como $E_n \subset B^c(M_0, r_0) \ \cap \ \overline{B}(M_n, \ r_n)$, para todo $x \in E_n$ se cumple

$$\Phi\big(d(x, M_0)\big) \leq \Phi\big(d(x, M_n) + d(M_n, M_0)\big) \leq \Phi\big(r_n + d(M_n, M_0)\big)$$

De manera análoga, teniendo en cuenta que $F_n \subset \overline{B}(M_0, r_0) \ \cap \ B^c(M_n, \ r_n)$, para todo $x \in F_n$,

$$\Phi(d(x, \ M_0)) \geq \Phi(d(x, \ M_n) - d(M_n, \ M_0)) \geq \Phi(r_n - d(M_n, \ M_0))$$

Por otro lado, en virtud de la definición de τ_0

$$\int (\tau_n(X) - \tau_0(X)) \ \Phi(d(X, M_0)) \ dP \geq 0,$$

pues se trata de la función de recorte óptima para M_0. A partir de estas desigualdades podemos acabar deduciendo:

$$H_n = \int (\tau_n(X) - \tau_0(X)) \, \Phi(d(X, M_0)) \, dP$$

$$= \int_{E_n} (\tau_n(X) - \tau_0(X)) \, \Phi(d(X, M_0)) \, dP$$

$$- \int_{F_n} (\tau_0(X) - \tau_n(X)) \, \Phi(d(X, M_0)) \, dP \leq$$

$$
\begin{aligned}
&\leq \Phi(r_n + d(M_n, M_0)) \int_{E_n} (\tau_n(X) - \tau_0(X))\, dP \\
&\quad - \Phi(r_n - d(M_n, M_0)) \int_{F_n} (\tau_0(X) - \tau_n(X))\, dP \\
&\leq \Phi(r_n + d(M_n, M_0)) - \Phi(r_n - d(M_n, M_0)) \to 0 \quad \text{cuando } n \to \infty.
\end{aligned}
$$

En consecuencia, $D_n = |V_{\Phi,\,\alpha}(M_n) - V_{\Phi,\,\alpha}(M_0)| \longrightarrow 0$, y hemos probado la continuidad de la variación α-recortada.

□

Proposición 1.3. *Sean $M = \{m_1, \ldots, m_k\} \subset \mathbb{R}^p$ y $\alpha \in (0,1)$. Entonces las siguientes afirmaciones son equivalentes:*

(a) $V_{\Phi,\alpha}(M) > 0$*;*

(b) existe $m_0 \in \mathbb{R}^p$ *tal que* $V_{\Phi,\alpha}(M \cup \{m_0\}) < V_{\Phi,\alpha}(M)$.

Demostración. Solo probamos que (a) implica (b), la otra implicación es trivial, pues la variación α-recortada es mayor o igual que 0.

Supongamos que $V_{\Phi,\,\alpha}(M) > 0$. En consecuencia, se tiene que $r_\alpha(M) > 0$ y $P_X(M) < 1 - \alpha$. Además, para todo $r < r_\alpha(M)$, tenemos que $P_X(\overline{B}(M, r)) < 1 - \alpha$ y por tanto se puede encontrar un $m_0 \in \mathbb{R}^p$ y $r_0 > 0$ tal que la bola $B_0 = B(m_0, r_0)$ satisface:

(i) $\int_{B_0} \tau_{\alpha,M}(X)\, dP > 0$;

(ii) $\min_{i=1,\ldots,k} \|m_i - m_0\| > \frac{2}{3} r_\alpha(M)$;

(iii) $r_0 < \frac{1}{3} r_\alpha(M)$.

Y en consecuencia,

$$
\begin{aligned}
V_{\Phi,\alpha}(M) =& \frac{1}{1-\alpha} \int \tau_{M,\alpha}(X) \Phi(d(X, M))\, dP \\
&= \frac{1}{1-\alpha} \int_{B_0} \tau_{M,\alpha}(X) \Phi(d(X, M))\, dP \\
&\quad + \frac{1}{1-\alpha} \int_{B_0^c} \tau_{M,\alpha}(X) \Phi(d(X, M))\, dP \\
&> \frac{1}{1-\alpha} \int_{B_0} \tau_{M,\alpha}(X) \Phi(d(X, m_0))\, dP \\
&\quad + \frac{1}{1-\alpha} \int_{B_0^c} \tau_{M,\alpha}(X) \Phi(d(X, M))\, dP
\end{aligned}
$$

$$\geq \frac{1}{1-\alpha}\int \tau_{M,\alpha}(X)\,\text{mín}\{\Phi(d(X,M)),\Phi(d(X,m_0))\}\,dP$$
$$\geq \frac{1}{1-\alpha}\int \tau_{M\cup\{m_0\},\alpha}(X)\Phi(d(X,M\cup\{m_0\}))\,dP$$
$$= V_{\Phi,\alpha}(M\cup\{m_0\}).$$

Y obtenemos la desigualdad deseada. □

En virtud de la Proposición 1.1, sabemos que existe una sucesión de conjuntos de k elementos $M_n = \{m_1^n,\ \ldots,\ m_k^n\} \subset \mathbb{R}^p,\ n = 0,1,2,\ldots$ tal que:

$$V_{\Phi,\ \alpha}(M_n) \downarrow V_{k,\ \Phi,\ \alpha}(X) \quad si \quad n \to \infty \tag{1.17}$$

Con el fin de probar el Teorema 1, veremos primero la existencia de subsucesiones convergentes de $\{M_n\}_n$, y posteriormente comprobaremos que los conjuntos límite son k-medias recortadas del vector aleatorio X. Comenzamos con un lema previo:

Lema 1.3. *Sea $\{M_n\}_n$ la sucesión descrita en* (1.17)*. Sean $a_n = \text{mín}_{i=1,\ldots,k}\, d(m_i^n, 0)$ y $r_n = r_\alpha(M_n)$. Entonces, las sucesiones $\{a_n\}_n$ y $\{r_n\}_n$ están acotadas.*

Demostración. Sea $\gamma < \infty$ tal que $P_X(B(0,\gamma)) > 1-\alpha$. Entonces, para todo $n = 1,2,\ldots$, se tiene que

$$a_n - \gamma \leq r_n \leq a_n + \gamma.$$

Luego basta con probar que una de las sucesiones mencionadas está acotada.

Primero, obsérvese que en virtud de (1.16) y (1.17):

$$V_{\Phi,\alpha}(M_n) \downarrow V_{k,\Phi,\alpha}(X) \leq \Phi(\gamma) < \Phi(\infty). \tag{1.18}$$

Sean $\{\varepsilon_n\}_n$ y $\{\gamma_n\}_n$ dos sucesiones de números positivos tales que $\varepsilon_n \downarrow 0$, $\gamma_n \uparrow \infty$, y $P[X \in B(0,\gamma_n)] \geq 1-\varepsilon_n$. Si $\{a_n\}_n$ no estuviera acotada, podríamos encontrar una subsucesión (que denotamos como la inicial) tal que $a_n > 2\gamma_n$ para todo $n = 1,2,\ldots$ y entonces, si denotamos $B_n = B(0,\gamma)$, para cada $x \in B_n$ se tiene que

$$d(x,M_n) = \min_{i=1,\ldots,k} d(x,m_i^n) \geq \min_{i=1,\ldots,k}(d(m_i^n,0) - d(x,0)) = a_n - \gamma > \gamma$$

.

En consecuencia:

$$V_{\Phi,\alpha}(M_n) \geq \frac{1}{1-\alpha}\int_{B_n} \tau_n(X)\,\Phi(d(X,M_n))\,dP$$

$$\geq \frac{1}{1-\alpha}\int_{B_n} \tau_n(X)\,\Phi(\gamma_n)\,dP$$

$$\geq \Phi(\gamma_n) \cdot \frac{1-\alpha-\varepsilon_n}{1-\alpha} \uparrow \Phi(\infty),$$

lo cual contradice (1.18).

□

Podemos ahora demostrar el teorema angular de esta sección, que garantiza la existencia de k-medias recortadas, cuya prueba exponemos a continuación:

Demostración Teorema 1.1- Existencia de k-medias recortadas. En virtud del Lema 1.3, existe un conjunto no vacío $I \subseteq \{1,\ldots,k\}$ y una subsucesión de $\{M_n\}_n = \{\{m_1^n,\ldots,m_k^n\}\}_n$ (que denotamos como la inicial) tal que:

$$\text{si } i \notin I, \text{ entonces } d(m_i^n, 0) \to \infty \text{ cuando } n \to \infty, \tag{1.19}$$

$$\text{si } i \in I, \text{ existe } m_i^0 \in \mathbb{R}^p \text{ tal que } m_i^n \to m_i^0 \text{ cuando } n \to \infty. \tag{1.20}$$

A partir de una reordenación correspondiente de los índices se puede suponer, sin pérdida de generalidad, que $I = \{1,\ldots,h\}$ con $1 \leq h \leq k$. Si denotamos por $M_n^h = \{m_1^n,\ldots,m_h^n\}$ y $r'_n = r_\alpha(M_n^h)$, para $n = 1,2,\ldots$, resulta trivial que $r'_n \geq r_n$, $n = 1,2,\ldots$, y que la sucesión $\{r'_n\}_n$ está acotada.

Veamos que la variación de la sucesión $\{M_n^h\}_n$ tiende a la α-variación mínima, es decir

$$V_{\Phi,\alpha}(M_n^h) \to V_{h,\Phi,\alpha} \quad \text{cuando } n \to \infty \quad \text{y} \quad V_{h,\Phi,\alpha} = V_{k,\Phi,\alpha} \tag{1.21}$$

Tomemos $\{\varepsilon_n\}_n$ y $\{\gamma_n\}_n$ tales que $\varepsilon_n \downarrow 0$, $\gamma_n \uparrow \infty$, y $P[X \in B(0,\gamma_n)] \geq 1-\varepsilon_n$. Por (1.19) y (1.20), podemos suponer, sin pérdida de generalidad, que para todo $n \in \mathbb{N}$,

$$d(m_i^n, 0) > 2\gamma_n \quad \text{para } i = h+1,\ldots,k,$$

$$\left(\bigcup_{i=1}^{h} \overline{B}(m_i^n, r_n)\right) \cap \left(\bigcup_{i=h+1}^{k} \overline{B}(m_i^n, r_n)\right) = \emptyset.$$

y

$$P_X\left(\bigcup_{i=h+1}^{k} \overline{B}(m_i^n, r_n)\right) \leq \varepsilon_n.$$

A partir de esto, se pueden acotar las variaciones por

$$V_{\Phi,\alpha}(M_n^h) \leq \frac{1}{1-\alpha}\left[\int_{B(M_n^h, r_n)} \tau_n(X)\, \Phi(d(X, M_n^h))\, dP + \Phi(r'_n)\, \varepsilon_n\right]$$

y deducir que

$$\begin{aligned}(1-\alpha)V_{\Phi,\alpha}(M_n) &\geq \int_{B(M_n^h, r_n)} \tau_n(X)\, \Phi(d(X, M_n^h))\, dP \\ &\geq (1-\alpha)V_{\Phi,\alpha}(M_n^h) - \Phi(r'_n)\, \varepsilon_n \\ &\geq (1-\alpha)V_{h,\Phi,\alpha}(X) - \Phi(r'_n)\, \varepsilon_n.\end{aligned}$$

Ahora tomando límites en la expresión anterior, y teniendo en cuenta que $\lim_{n\to\infty} \Phi(r'_n)\varepsilon_n = 0$ pues $(r'_n)_n$ está acotada, se tiene

$$\lim_{n\to\infty} V_{\Phi,\alpha}(M_n) \geq \lim_{n\to\infty} V_{\Phi,\alpha}(M_n^h) \geq V_{h,\Phi,\alpha}(X), \tag{1.22}$$

y a partir de estas desigualdades y de (1.17) se obtiene

$$V_{k,\Phi,\alpha} = \lim_{n\to\infty} V_{\Phi,\alpha}(M_n) \geq V_{h,\Phi,\alpha}.$$

Entonces, necesariamente $V_{k,\Phi,\alpha} = V_{h,\Phi,\alpha}$ y (1.21) se cumple. Además, por la Proposición 1.2, tenemos

$$V_{\Phi,\alpha}(M_n^h) \to V_{\Phi,\alpha}(M_0^h) \quad \text{cuando } n \to \infty \tag{1.23}$$

y de (1.22) y (1.23) se deduce que

$$V_{\Phi,\alpha}(M_0^h) = V_{h,\Phi,\alpha}(X),$$

y entonces $M_0^h = \{m_1^0, \ldots, m_h^0\}$ es una h-media recortada de X.

Ahora, si $h = k$, M_0^h es una k-media y la demostración está completa. Si $h < k$, la igualdad $V_{h,\phi,\alpha} = V_{k,\phi,\alpha}$ en (1.21) y la Proposición 1.3 implican que $V_{\Phi,\alpha}(M_0^h) = 0$ y entonces la existencia está obviamente garantizada para todo $k \geq h$, pues basta añadir puntos hasta llegar a k elementos. □

Como ya adelantamos previamente, el lema 1.1 establece una relación entre las k-medias recortadas y las funciones de recorte óptimas. Así pues, la función de recorte óptima coincide con la función indicatriz de un conjunto medible, lo que asegura la existencia de k-medias recortadas. Esta relación se recoge en el siguiente corolario:

Corolario 1.1. *Bajo las hipótesis del teorema 1.1, si* Φ *es estrictamente creciente y* τ_0 *y* M_0 *son soluciones de* (1.7)*, entonces*

$$I_{B(M_0,\ r_\alpha(M_0))} \leq \tau_0 \leq I_{\overline{B}(M_0,\ r_\alpha(M_0))}\ , \qquad P_x - c.t.p$$

De hecho, si P_x *es absolutamente contínua respecto a la medida de Lebesgue en* $\mathbb{R}^p$*, se tiene,*

$$I_{B(M_0,\ r_\alpha(M_0))} = \tau_0\ , \qquad P_x - c.t.p$$

1.5. Unicidad de k-medias recortadas y puntos frontera entre clusters

Una vez se ha garantizado la existencia de soluciones óptimas para el problema de k-medias recortadas, surge de forma natural preguntarse si se da la unicidad, o si, en cambio, podemos encontrar distintos conjuntos de k elementos que minimicen la variación. Este factor tiene una gran influencia no solo a nivel teórico, sino también en

la práctica, pues en caso de existir varias soluciones óptimas, podríamos llegar a una u otra solución en función de la inicialización que tomemos.

Consideramos una k-media de X, $M_0 = \{m_1^0, \ \ldots, \ m_k^0\}$, con función de recorte óptima asociada τ_0 y radio óptimo r_0, esto es,

$$I_{B(M_0,\ r_0)} \leq \tau_0 \leq I_{\overline{B}(M_0,\ r_0)},$$

donde $I_{\overline{B}(M_0,\ r_0)}$ es el conjunto de recorte óptimo salvo por quizá parte de la frontera. Sabemos que M_0 induce una partición en $\overline{B}(M_0,\ r_0)$ en k clusters de la siguiente manera: el cluster A_i está formado por los puntos que están más cerca de m_i^0 que de los $k-1$ puntos restantes de M_0. Los puntos de la frontera entre clusters pueden ser asignados a cualquiera de estos, pues la k-variación recortada no varía.

De hecho, el conjunto M_0 también induce una partición de la k-variación recortada en variaciones correspondientes a cada cluster:

$$\begin{aligned} V_{k,\ \phi,\ \alpha} &= \frac{1}{1-\alpha} \int \tau_0(X)\Phi(d(X,\ M_0))\ dP \\ &= \frac{1}{1-\alpha} \sum_{i=1}^{k} \int_{A_i} \tau_0(X)\Phi(d(X,\ m_i^0))\ dP \end{aligned}$$

A su vez, para la partición obtenida, m_i^0 debe ser una ϕ-media del correspondiente cluster A_i para cada $i = 1, \ldots, k$, o más precisamente, una ϕ-media de X dado A_i; es decir, una solución de

$$\inf_{m \in \mathbb{R}^p} \int_{A_i} \tau_0(X)\Phi(d(X,\ m))\ dP$$

De lo contrario podríamos disminuir la variación en algunos clusters remplazando m_i^0, $i = 1, \ldots, k$, por ϕ-medias de los clusters correspondientes, y M_0 no sería una k-media recortada de X.

Como consecuencia, concluímos que la unicidad de las k-medias recortadas depende no solo de la unicidad del conjunto de recorte, sino también de la unicidad de las ϕ-medias dado cada cluster. Es sencillo ver que la unicidad del conjunto de recorte no se da en general, y no resulta complicado encontrar algún caso que lo ilustre. Por ejemplo, tomemos $\alpha = \frac{1}{2}$, k=1, y supongamos que P es una distribución uniforme en la bola unidad. Es evidente que existen infinitas posibilidades de recorte de los datos que nos llevan a soluciones óptimas de nuestro problema. La figura 1.6 muestra 3 de las posibles soluciones que minimizan la variación de la distribución.

En cuanto a la unicidad de las k-ϕ-medias no existen resultados generales, no obstante, a continuación recogemos un resultado para el caso k=1.

Proposición 1.4. *Si Φ es estrictamente convexa, entonces la Φ-media de P es única.*

Demostración. Supongamos que m, $m^* \in \mathbb{R}^p$ son dos Φ-medias distintas de P. Sea $\lambda \in (0,1)$, consideramos $m_\lambda := \lambda m + (1-\lambda)m^*$. Teniendo en cuenta que Φ es creciente

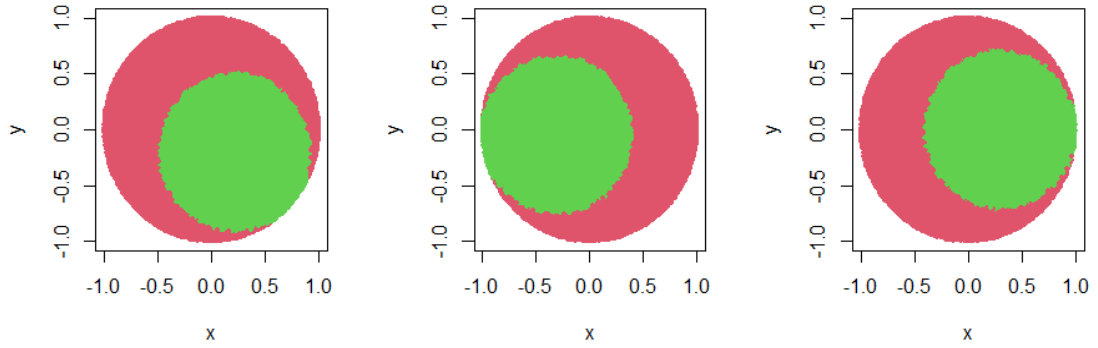


Figura 1.6: Soluciones óptimas por recorte de 1-media de una distribución uniforme en la bola unidad.

y convexa, se tienen las siguientes desigualdades

$$\begin{aligned}\int \Phi[\|x - m_\lambda\|]P(dx) &\leq \int \Phi[\lambda \|x - m\| + (1-\lambda)\|x - m^*\|]P(dx) \\ &\leq \lambda \int \Phi[\|x - m\|]P(dx) + (1-\lambda)\int \Phi[\|x - m^*\|]P(dx) \\ &= \int \Phi[\|x - m\|]P(dx) \qquad (1.24)\end{aligned}$$

Si Φ es estrictamente convexa, la última desigualdad es estricta salvo que $\|x - m\| = \|x - m^*\|$ $P - c.t.p$, lo cual solo ocurre si

$$P\left\{x + \frac{m + m^*}{2} : \langle x, m - m^*\rangle = 0\right\} = 1 \qquad (1.25)$$

donde $\langle \cdot, \cdot \rangle$ denota el producto escalar usual en $\mathbb{R}^p$.

Por otro lado, si x verifica $\langle x, m - m^* = 0\rangle$, entonces

$$\left\|\left(x + \frac{m + m^*}{2}\right) - m_\lambda\right\| < \left\|\left(x + \frac{m + m^*}{2}\right) - m\right\|$$

La convexidad estricta de Φ implica que Φ es estrictamente creciente. Por lo tanto, si se cumple (1.25), entonces la primera desigualdad en (1.24) es estricta, lo cual es absurdo pues m es una Φ-media de P. □

Hemos probado por lo tanto no solo que una k-media recortada induce una partición de $\overline{B}(M_0, r_0)$ en k clusters, sino también que, si Φ es estrictamente convexa y excluímos casos degenerados en los que algún cluster tiene probabilidad cero, la partición determina la k-media recortada. Resumimos estas consideraciones en la siguiente proposición.

Proposición 1.5. *Con la notación previa, m_i^0 es una Φ-media de P dada τ_0 y A_i, para $i = 1, \ldots, k$. De hecho, si Φ es estrictamente convexa y $P(A_i) > 0$ entonces m_i^0 es la única Φ-media de P dada τ_0 y A_i, para $i = 1, \ldots, k$.*

Estos resultados han sido sacados de [3], donde se analiza la forma en la que P puede asignar probabilidad a la frontera de los clusters. Podríamos pensar que, para los puntos que se encuentran en la frontera entre dos clusters, el método "reparta" su masa de probabilidad entre ambos grupos, sin embargo, veremos que bajo ciertas condiciones de la función Φ, la frontera de un cluster tendrá probabilidad cero. Antes de llegar al teorema que recoge este resultado, enunciamos otra propiedad importante de las Φ-medias, consecuencia directa del teorema de separación por hiperplanos.

Lema 1.4. *Supongamos que Φ es estrictamente convexa y que A es un conjunto convexo. Entonces la Φ-media de P dado A está en la adherencia de A.*

Teorema 1.2. *Usando la misma notación introducida en los resultados previos. Sea $M_0 = \{m_1, \ldots, m_k\}$ una k-media recortada de P dado $\alpha \in (0, 1)$ y $B = B(M_0, r_\alpha(M_0))$. Si ϕ es estrictamente convexa y su primera derivada existe y es contínua, entonces:*

1. *$P\{x \in B : \|x - m_i\| = \|x - m_j\|\} = 0$ si $i \neq j$*
2. *Sea D la frontera topológica de $B(M_0, r_0)$. Entonces la función de recorte óptima τ_0 verifica una de las siguientes condiciones:*
   a) *$\int_D \tau_0(x) dP_x = 0$ (es decir, D no se encuentra en el conjunto óptimo de recorte, $P_X[B(M_0, r_0)] = 1 - \alpha$)*
   b) *$\int_D \tau_0(x) dP_x = P(D)$ (es decir, D está contenido en el conjunto óptimo de recorte, $P_X[\overline{B}(M_0, r_0)] = 1 - \alpha$)*
   c) *Existe un punto x_0 en D tal que toda la probabilidad de D está concentrada en x_0 (es decir, $P(D) = P(\{x_0\})$*

La demostración de este teorema se incluye en el apéndice.

1.6. Consistencia de las k-medias recortadas

Una vez presentado el problema de k-medias recortadas y estudiado alguna de sus propiedades básicas, nos preguntamos acerca de la consistencia del método. Generalmente, trabajaremos sobre un conjunto de datos de n individuos, y encontraremos la mejor representación de este conjunto mediante k-medias. A menudo, este conjunto de datos es una muestra de una distribución de probabilidad, y cabe preguntarse si los centros óptimos obtenidos para el caso muestral se aproximan a los de la distribución teórica.

Estudiaremos por lo tanto el comportamiento asintótico de este método, comprobando la convergencia de las k-medias recortadas empíricas a las k-medias recortadas teóricas, siguiendo el estudio realizado en [4].

En lo que sigue, $\{X_n\}_n$ es una sucesión de vectores aleatorios en $\mathbb{R}^p$ definidos en el espacio probabilístico $(\Omega,\ \sigma,\ P)$ y $M_n = \{m_1^n, \ldots, m_k^n\}$, n=0,1,2,..., es una k-media recortada de X_n con función de recorte óptima asociada τ_n y radio óptimo r_n. Además, denotamos por $V_n[= V_{k,\ \Phi,\ \alpha}(X_n)]$, n=0,1,2, ..., a la k-variación recortada de X_n.

Comenzaremos con un pequeño lema técnico necesario para la posterior prueba de la consistencia.

Lema 1.5. *Si $X_n \xrightarrow{c.s} X_0$, y denotamos por $a_n = min_{i=1,\ldots,k}\ d(m_i^n, 0)$ para $n = 1, 2, \ldots,$ entonces $\{a_n\}_n$ y $\{r_n\}_n$ son sucesiones acotadas.*

Demostración. Sabemos que existe una bola $B(0,\gamma)$, con $\gamma < \infty$, tal que $P_{X_n}[B(0,\gamma)] > 1-\alpha$ para todo $n = 1, 2, \ldots$ Entonces, para todo $n = 1, 2, \ldots$, tenemos

$$a_n - \gamma \leq r_n \leq a_n + \gamma,$$

por lo que es suficiente con probar que una de las sucesiones está acotada. Para empezar, nótese que

$$V_n \leq \frac{1}{P_{X_n}(B(0,\gamma))} \int I_{B(0,\gamma)}(X_n)\Phi(d(X_n,0))dP \leq \Phi(\gamma) < \Phi(\infty) \quad \text{para } n = 1, 2, \ldots \tag{1.26}$$

Ahora, sean $\{\varepsilon_n\}_n$ y $\{\gamma_n\}_n$ sucesiones tales que $\varepsilon_n \downarrow 0$, $\gamma_n \uparrow \infty$ y $P[X_n \in B(0,\gamma_n)] \geq 1-\varepsilon_n$.

Siguiendo un razonamiento análogo al del lema 1.3, si $\{a_n\}_n$ no estuviera acotada podríamos obtener una subsucesión (que denotaremos como la inicial para facilitar la notación) tal que $a_n > 2\gamma_n$ para todo $n = 1, 2, \ldots$ y entonces tendríamos

$$\begin{aligned} V_n &\geq \frac{1}{1-\alpha} \int_{B(0,\gamma_n)} \tau_n(X_n)\Phi(d(X_n, M_n))\, dP_{X_n} \\ &> \frac{1}{1-\alpha} \int_{B(0,\gamma_n)} \tau_n(X_n)\Phi(\gamma_n)\, dP_{X_n} \\ &\geq \Phi(\gamma_n)\frac{1-\alpha-\varepsilon_n}{1-\alpha} \uparrow \Phi(\infty) \end{aligned}$$

lo cual es contradice (1.26). □

A continuación, se expone un resultado previo a la consistencia, que demuestra que bajo ciertas condiciones, las k-medias recortadas de una sucesión de distribuciones que converge casi seguro a otra distribución, convergen a la k-media recortada de esta última. Para facilitar la visualización y el seguimiento del desarrollo, la prueba de este teorema se incluirá en el apéndice.

Teorema 1.3. *Manteniendo la notación previa, asumimos que:*

a) $X_n \xrightarrow{c.s} X_0$

b) P_{X_0} *es absolutamente contínua;*

c) $M_0 = \{m_1^0, \ldots, m_k^0\}$ *es la única k-media recortada de* X_0

Entonces

$$M_n \to M_0 \quad \textit{(en la distancia de Haussdorff) cuando } n \to \infty$$

y

$$V_n \to V_0 \quad \textit{cuando } n \to \infty$$

Nótese que, en el teorema anterior, se pide que $\{X_n\}_n$ converga casi seguro a X. No obstante, el teorema de representación de Skorohod permite extender el resultado incluyendo la convergencia en distribución. Este resultado es clave para poder asegurar la consistencia del método, a través de la convergencia de probabilidades empíricas a la probabilidad teórica.

Corolario 1.2. *Si suponemos que todas las hipótesis del Teorema 1.3 se satisfacen y (a) se remplaza por:*

- *(a*)* $X_n \to X_0$ *converge en distribución*

Entonces

$$M_n \to M_0 \textit{ (en la distancia de Haussdorff) cuando } n \to \infty$$

y

$$V_n \to V_0 \quad \textit{cuando } n \to \infty$$

Demostración. Aplicando el teorema de representación de Skorohod, existe una sucesión $\{Y_n\}_n$ de vectores aleatorios en $\mathbb{R}^p$ tales que $P_{Y_0} = P_X$, $P_{Y_n} = P_{X_n}$ e $Y_n \to Y_0$ c.s. Por tanto, el resultado se deduce aplicando el teorema anterior a la sucesión $\{Y_n\}_n$. □

Como consecuencia de este corolario, llegamos al teorema objetivo de esta sección, que demuestra la consistencia del método de k-medias recortadas.

Teorema 1.4. ***Consistencia de las k-medias recortadas.*** *Sea* $\{X_n\}_n$ *una sucesión de vectores aleatorios independientes e igualmente distribuídos, con probabilidad asociada* P_X*. Sea* $\{P_n^\omega\}$ *una sucesión de medidas de probabilidad empíricas (es decir,* $P_n^\omega(A) = \frac{1}{n}\sum_{1\le i\le n} I_A[X_i(\omega)]$*. Asumimos que* P_X *es contínua y que existe una única k-media recortada para* P_X*, que denotamos por* M_0*. Si* $\{M_n^\omega\}_n$ *es una sucesión de k-medias recortadas empíricas, entonces:*

- *(a)* $d(M_n^\omega,\ M_0) \to 0$ *, para P-casi todo* ω
- *(b)* $V_{k,\ \Phi,\ \alpha}(P_n^\omega) \to V_{k,\ \Phi,\ \alpha}(P_X)$*, para P-casi todo* ω

Demostración. Sea $A := \{\omega \in \Omega : P_n^\omega \to_d P_X\}$. Es bien sabido que $P(A) = 1$, luego el resultado se sigue del corolario 1.2. □

Capítulo 2

Clustering sobre formas elípticas

A menudo se tiende a pensar en el análisis de datos como una ciencia exacta, que a través de distintos procedimientos y técnicas sistemáticos siempre devuelve resultados objetivos. No obstante, esta percepción está lejos de la realidad, y los resultados de nuestro análisis cluster suelen estar en gran parte influenciados por el método utilizado, así como por las suposiciones realizadas sobre el modelo probabilístico que genera los datos. Por ello, debemos conocer bien las limitaciones que presentan los métodos que usamos para un análisis cluster, y procurar usar un procedimiento que se adapte correctamente a la estructura de los datos.

El algoritmo de k-medias clásico, introducido en el primer capítulo, presenta una conocida preferencia por la formación de clusters esféricos. Esta característica es causa de la propia definición de los clusters (ver (1.2)), los cuales se forman asignando cada punto al centroide más cercano mediante la distancia usual en $\mathbb{R}^p$. Esta falta de flexibilidad se induce también en el método de k-medias recortadas, donde de nuevo se emplea la distancia usual en la asignación de puntos.

Otra limitación del método de k-medias recortadas está relacionada con la suposición de igualdad del tamaño de los clusters. Este método asume que los grupos poseen una varianza similar, lo que puede generar problemas ante la presencia de grupos con tamaños o densidades muy distintas. Por ejemplo, si un grupo pequeño se encuentra próximo a un grupo grande, el algoritmo tiende a "absorber" el grupo de tamaño más reducido dentro del de mayor tamaño. De la misma manera, un grupo de gran extensión puede ser dividido en clusters más pequeños, pues el método favorece la formación de particiones de dispersión similar.

En este capítulo presentaremos algunos de los métodos de clustering robusto basados en modelos, surgidos como una extensión de las k-medias recortadas con el objetivo de solucionar estas limitaciones. Analizaremos la evolución de estos métodos, así como las motivaciones que condujeron a su desarrollo. La base de estos procedimientos es la metodología TCLUST, introducida en García-Escudero, L. A., Gordaliza, A., Matran, C., & Mayo-Iscar, A. (2008)[12], que constituye un marco general para el tratamiento de los distintos problemas que se presentarán, y que ha ido evolucionando durante los años según estas extensiones del modelo han sido desarrolladas. Por último, describiremos un método de clustering basado en modelos de mezcla, que permite establecer probabilidades de pertenencia a los clusters para cada individuo, proporcionando así una mayor interpretabilidad a los resultados.

2.1. Extensión del modelo: metodología TCLUST

En aplicaciones reales de clustering es frecuente tratar con conjuntos de datos de estructura variada y heterogénea, en los que podemos encontrar grupos de formas alargadas y muy diversas, lejanas de la esfericidad.

Las k-medias recortadas, tratadas en el capítulo anterior, surgieron como una robustificación del método de k-medias clásico, solucionando su sensibilidad a la presencia de datos átipicos mediante el uso del recorte. Sin embargo, conservan muchos de los principales inconvenientes y problemas del método original, entre los que destaca la falta de flexibilidad, pues, al tomar la distancia euclídea como medida de similaridad, se establece una preferencia por la búsqueda de clusters con forma aproximadamente esférica y de tamaño similar (véase la figura 2.1).

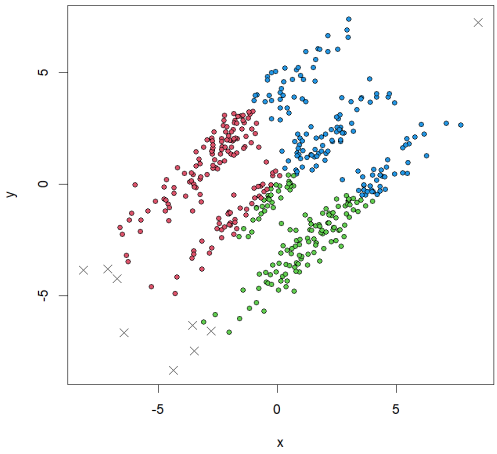


Figura 2.1: Resultado de aplicar 3-medias recortadas sobre un conjunto de datos formado por muestras de 3 normales multivariantes.

Esta limitación justifica la elección de otra medida de similaridad de los datos que permita identificar clusters de forma diversa. La suposición de normalidad de los clusters ofrece un marco de trabajo flexible, permitiendo el uso de las distancias de Mahalanobis asociadas a cada cluster como medida de proximidad de las observaciones, definida como

$$d_\Sigma(x;\mu) = \sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

siendo μ el centroide del cluster y Σ la matriz de covarianzas asociada.

Esta medida tiene en cuenta la correlación entre las variables y las escalas de cada dimensión. Gallegos y Ritter(2005)[11] proponen una extensión de las k-medias recortadas basada en esta distancia. Aparece así el denominado "modelo de outliers espurios", que modela la estructura de los datos, asumiendo la generación de las observaciones regulares por k distribuciones normales multivariantes, más la presencia de contaminación generada por una distribución indefinida.

Consideremos un conjunto de datos $\{x_1,\ldots,x_n\} \subset \mathbb{R}^p$. Sea $f(x,\mu,\Sigma)$ la función de densidad de una normal multivariante de dimensión p con vector de medias μ y matriz

de covarianzas Σ. La función de verosimilitud del "modelo de outliers espurios"[11] se describe como:

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j} f(x_i;\mu_j,\Sigma)\right]\left[\prod_{i\in R_0} g_i(x_i)\right] \tag{2.1}$$

donde $\{R_0,\ldots,R_k\}$ es una partición de $\{1,\ldots,n\}$ con $\#R_0 = \lceil n\alpha\rceil$. El conjunto de índices R_0 hace referencia a las observación "no regulares", generadas por distribuciones de probabilidad g_i desconocidas.

Este modelo introduce una matriz de dispersión no necesariamente esférica para los clusters, permitiendo así identificar grupos "alargados", como los de la figura 2.1. No obstante, mantiene la asunción de la igualdad de variabilidad de los mismos. Esto lo convierte en un modelo adecuado cuando se quieren obtener clusters con formas semejantes, lo cuál puede resultar interesante en ciertos contextos y problemas. Sin embargo, queda limitado cuando los grupos a identificar presentan formas y tamaños disparejos.

Ante esta nueva limitación, la siguiente extensión natural es considerar distintas matrices de dispersión Σ_j para los clusters en lugar de una matriz común Σ. Bajo esta consideración, se puede generalizar el módelo de "outliers espurios" (2.1), mediante la función de verosimilitud

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j} f(x_i;\mu_j,\Sigma_j)\right]\left[\prod_{i\in R_0} g_i(x_i)\right] \tag{2.2}$$

donde se mantienen como parámetros los centros μ_j's y la partición $R_0,\ldots,R_k$.

Los modelos que se acaban de exponer asumen que los datos regulares vienen generados por distribuciones normales multivariantes. Si bien esta suposición puede ser errónea y no adaptarse correctamente a determinados casos, suele dar buenos resultados en general, siendo un enfoque mucho más flexible que el de las k-medias recortadas.

Las matrices de dispersión introducidas en (2.2) controlan la variabilidad de cada cluster, es decir, determinan tanto su forma como la amplitud del mismo, pero no su densidad. En consecuencia, los algoritmos presentados para este modelo tienden a encontrar clusters con un número similar de observaciones, lo cual si bien puede resultar interesante en algunos tipos de estudios, no supone un modelo idóneo en el caso general.

La figura 2.2(a) muestra el resultado de aplicar el algoritmo TCLUST para el modelo (2.2) sobre un conjunto de datos generado artificialmente con grupos de 300, 300 y 50 observaciones respectivamente (y 20 observaciones no regulares). La preferencia intrínseca de este método por grupos igualmente densos provoca la partición de los clusters grandes en dos, asignando parte de las observaciones al cluster de menor tamaño.

Una idea sencilla para enfrentar este problema es añadir pesos a los grupos, incluyendo nuevos parámetros π_j, con $\sum_{j=1}^{k}\pi_j = 1$, a la función de verosimilitud a maximizar. La figura 2.2(b) ilustra la clasificación obtenida tras considerar estos pesos en el algoritmo TCLUST, que describiremos en una sección posterior. Este enfoque permite

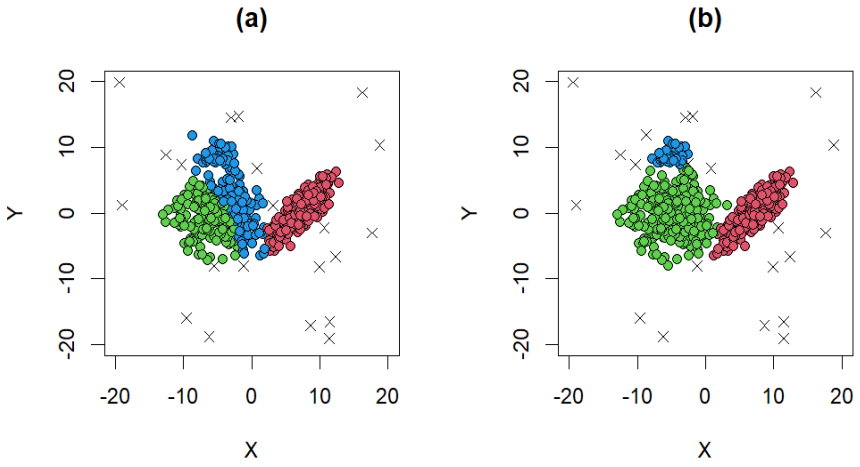


Figura 2.2: Resultado de aplicar TCLUST: (a)Sin pesos (b)Con pesos.

identificar grupos conformados por diversas cantidades de observaciones, manteniendo la flexibilidad en las formas y tamaños.

Esta es la propuesta presentada en García-Escudero, L. A., Gordaliza, A., Matran, C., & Mayo-Iscar, A. (2008)[12], que generaliza el módelo de "outliers espurios", asumiendo matrices de dispersión Σ_j's y pesos π_j's (con $\sum_{j=1}^{k} \pi_j = 1$) distintos para cada grupo, lo que lleva a la maximización de la función de verosimilitud

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j} \pi_j f(x_i;\mu_j,\Sigma_j)\right]\left[\prod_{i\in R_0} g_i(x_i)\right] \tag{2.3}$$

Según los intereses de nuestro análisis, deberemos emplear uno u otro de los métodos presentados. No obstante, debido a su mayor flexibilidad y adaptabilidad a distintos contextos, durante este capítulo nos centraremos en el modelo con pesos (2.3), y analizaremos el problema de maximización siguiendo el desarrollo presentado en [12].

Las observaciones "no regulares", correpondientes al segundo factor en (2.3), son de naturaleza desconocida, lo que impide en la práctica maximizar esta función, pues no conocemos la expresión de las g_i's. No obstante, estas observaciones pueden ser vistas como "outliers" asumiendo que

$$\underset{\mathcal{R}}{arg\,max} \quad \max_{\mu_j,\,\Sigma_j} \prod_{j=1}^{k}\prod_{i\in R_j} \pi_j f(x_i;\mu_j,\Sigma_j) \subset \underset{\mathcal{R}}{arg\,max} \prod_{i\in R_0} g_i(x_i). \tag{2.4}$$

donde $\mathcal{R}$ denota el conjunto de particiones de los índices $\{1,2,\ldots,n\}$ en $k+1$ grupos $\{R_0,R_1,\ldots,R_k\}$ con $\#R_0 = [n\alpha]$.

Esta condición se cumple bajo supuestos razonables para las $g_i's$, como se analiza en [11]. Se trata de una suposición sensata, que permite evitar la contribución no regular.

Bajo esta asunción, podemos entonces ignorar en (2.3) el factor producido por las observaciones contaminantes, simplificando el problema a la búsqueda de k centros $\mu_1, \ldots, \mu_k$ en $\mathbb{R}^p$, k matrices $p \times p$ simétricas definidas positivas $\Sigma_1, \ldots, \Sigma_k$, pesos $\pi_1, \ldots, \pi_k$ con $\sum_{i=1}^{k} \pi_i = 1$, y una partición $\{R_0, R_1, \ldots, R_k\}$ de $\{1, \ldots, n\}$ tal que R_0 incluye una proporción $[n\alpha]$ de los índices, maximizando

$$\sum_{j=1}^{k} \sum_{i \in R_j} log(\pi_j f(x_i; \mu_j, \Sigma_j)) \tag{2.5}$$

Nótese que ahora las observaciones pertenecientes a R_0 no se tienen en cuenta, es decir, son recortadas. A la expresión (2.5) la denominaremos *función de verosimilitud de clasificación recortada*. La maximización directa de esta función no es un problema bien definido, y, como veremos en la siguiente sección, será necesaria la imposición de alguna restricción sobre las matrices de dispersión $\Sigma_j's$ que controle su tamaño relativo, garantizando así la existencia de un máximo.

Con el propósito de estudiar las propiedades matemáticas de este problema, y de manera análoga a lo realizado para las k-medias recortadas, generalizamos el modelo a un marco teórico y probabilístico, para una distribución P.

Para ello, se introducen las llamadas *funciones de asignación,* z_j, $j = 1, \ldots, k$. Dada una observación $x \in \mathbb{R}^p$, definimos $z_j(x) = 1$ si x es asignada a la clase R_j, $j = 1, \ldots, k$ o $z_0(x) = 1$ si ha sido recortada (es decir, si $x \in R_0$). Esta noción permite rescribir el problema formulado en (2.5) como la maximización de:

$$\prod_{i=1}^{n} \left[\prod_{j=1}^{k} \pi_j^{z_j(x_i)} f(x_i; \mu_j, \Sigma_j)^{z_j(x_i)} \right], \tag{2.6}$$

donde z_j son funciones con llegada en $\{0, 1\}$, definidas en todo el espacio muestral, que verifican $\sum_{j=0}^{k} z_j(x_i) = 1$ y $\sum_{i=1}^{n} z_0(x_i) = [n\alpha]$.

Ahora, de nuevo mediante el uso de la esperanza matemática, se puede generalizar el modelo para una distribución de probabilidad. Dado un vector aleatorio X, con distribución de probabilidad asociada P_X, queremos maximizar:

$$E_{P_X} \left[\sum_{j=1}^{k} z_j(\cdot) \left(log\ \pi_j + log\ f(\cdot; \mu_j, \Sigma_j) \right) \right] \tag{2.7}$$

en términos de las funciones de asignación:

$$z_j : \mathbb{R}^p \longrightarrow \{0, 1\} \qquad \text{tales que } \sum_{j=0}^{k} z_j = 1 \text{ y } E_{P_X}(z_0(\cdot)) = \alpha$$

y los parámetros $\theta = (\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k)$ correspondiendo a los pesos $\pi_j \in [0, 1]$, con $\sum_{j=1}^{k} \pi_j = 1$, vectores de medias $\mu_j \in \mathbb{R}^p$ y matrices $p \times p$ simétricas y definidas positivas Σ_j, $j = 1, \ldots, k$.

Nótese que si P_n representa la medida de probabilidad empírica, $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$, remplazando P por P_n, recuperamos el problema muestral original (tal vez, $E_{P_n} z_0(\cdot) = \alpha$ no pueda cumplirse).

2.1.1. Restricciones de las matrices de dispersión

La maximización de (2.5) y (2.7) sin imponer restricciones no constituye un problema bien definido desde el punto de vista matemático. La elección de una matriz Σ_j con un determinante arbitrariamente cercano a 0 y un centro μ_j igual a una de las observaciones, hace tender las funciones de verosimilitud recortadas (2.5) y (2.7) hacia infinito, conduciendo a agrupaciones espurias lejos de la naturaleza de la estructura de los datos. Para evitar este tipo de soluciones degeneradas es esencial establecer alguna restricción sobre las matrices de covarianza $\Sigma_j' s$. Existen en la literatura numerosas restricciones a este problema. A continuación, se presentan alguna de las más habituales. Cabe resaltar que los resultados obtenidos dependerán fuertemente de la restricción establecida, luego debemos elegirla cuidadosamente en función de la estructura de los datos y el propósito de nuestro estudio, conociendo las implicaciones de cada una.

Restricción de autovalores (RA)

El enfoque más utilizado consiste en establecer una restricción de similitud de las matrices de dispersión basada en sus autovalores. En [12] se introduce la llamada "restricción de autovalores", que se detalla a continuación.

Sea $c \geq 1$ una constante prefijada, se exige que:

$$\frac{M_n}{m_n} \leq c \tag{2.8}$$

donde

$$M_n = \max_{j=1,\ldots,k}\left(\max_{l=1,\ldots,p}\lambda_l(\Sigma_j)\right) \quad \text{y} \quad m_n = \min_{j=1,\ldots,k}\left(\min_{l=1,\ldots,p}\lambda_l(\Sigma_j)\right),$$

y $\{\lambda_j(\Sigma)\}_{j=1}^{p}$ denota el conjunto de autovalores de una matriz de dispersión Σ. El conjunto de parámetros θ que satisfacen esta restricción se denota por Θ_c.

Esta constante controla simultáneamente el tamaño relativo de los grupos y la desviación de la esfericidad de cada cluster, lo que puede ayudarnos a manejar el resultado del método según nuestros intereses. Obsérvese que, si tomamos $c = 1$ (la restricción más estricta posible) entonces $\Sigma_1 = \ldots = \Sigma_k = r \times I_p$, donde I_p denota la matriz identidad $p \times p$ y $r > 0$. Este caso se reduce por lo tanto al método de k-medias recortadas pero con pesos.

Ejemplo: necesidad de restricción

Veamos ahora un ejemplo que ilustre la utilidad del parámetro c para adaptar el método a las características y propósitos de nuestro estudio. Supongamos que la

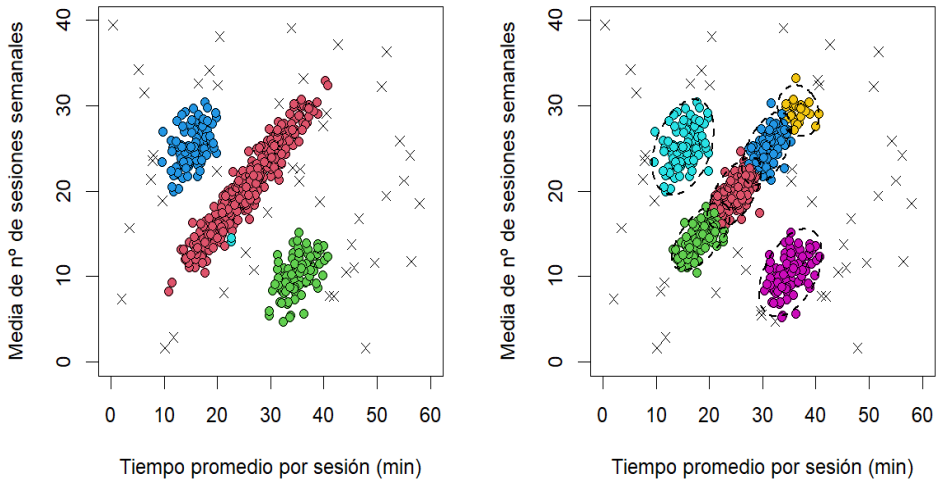


Figura 2.3: Resultados de aplicar TCLUST con $k = 6$, $\alpha = 0.07$, $c = 100$ (izq.) y $k = 6$, $\alpha = 0.07$, $c = 2$ (dcha.).

empresa dueña de una app quiere segmentar a sus usuarios en distintos grupos de comportamiento para personalizar notificaciones. Para cada usuario se dispone de la información correspondiente a dos variables: el promedio de número de sesiones por semana y el tiempo promedio por sesión (en minutos). La figura 2.3 muestra la representación de estos usuarios, así como los resultados de aplicar el método TCLUST con $k = 6$, $\alpha = 0.07$, $c = 100$ (izquierda) y $k = 6$, $\alpha = 0.07$, $c = 2$ (derecha).

Se puede observar como un valor grande de c provoca la determinación de un cluster muy alargado, agrupando usuarios con comportamientos totalmente distintos, lo que carece de sentido para el objetivo planteado. En cambio, para valores bajos como $c = 2$, se evita que un clúster pueda tener variabilidad excesiva en una sola variable, determinando grupos más "esféricos", enfoque que se adapta mejor a este caso concreto.

En ocasiones, el objetivo del análisis cluster no es obtener la mejor clasificación posible en función de la estructura subyacente de los datos, sino que la clasificación deseada puede estar determinada por ciertas características particulares del estudio. En este ejemplo podemos ver cómo el manejo de la constante de restricción c proporciona al usuario otra herramienta para ajustar el modelo.

Restricción de determinantes

Otra alternativa para controlar la relación entre las matrices de dispersión es mediante sus determinantes. Denotemos por

$$\tilde{M}_n = \max_{j=1,\dots,k} |\Sigma_j| \quad \text{y} \quad \tilde{m}_n = \min_{j=1,\dots,k} |\Sigma_j|$$

los determinantes máximos y mínimos de las matrices Σ_j's. Sea $c \geq 1$ una constante, la restricción de determinantes se define como:

$$\frac{\tilde{M}_n}{\tilde{m}_n} \leq c. \tag{2.9}$$

Esta restricción limita los volumenes relativos de los elipsoides, pero no establece ninguna restricción sobre su forma. Por tanto, resulta adecuada cuando se busca invarianza afín, es decir, que la agrupación obtenida no se vea afectada por transformaciones lineales del espacio como giros o escalados. Esta restricción generaliza la impuesta en [10], donde se asumía implícitamente que $|\Sigma_1| = \ldots = |\Sigma_k|$, caso correspondiente a tomar $c = 1$ en (2.9).

2.1.2. Propiedades matemáticas

A continuación, nos proponemos analizar alguna de las propiedades fundamentales de este método, como la existencia de soluciones y la consistencia. Nos centraremos en probar estos resultados para el modelo descrito en (2.7) bajo la restricción de los autovalores (RA) presentada en (2.8), ya que se trata del caso más flexible e interesante en la práctica. Para este objetivo seguiremos esencialmente el desarrollo en [12].

Con el fín de excluir aquellas distribuciones de probabilidad claramente inapropiadas para el enfoque de nuestro método, incluiremos otra restricción leve sobre la distribución P. Esta condición será necesaria para garantizar la existencia de soluciones de (2.7).

(PR) La distribución P no está concentrada en k puntos tras eliminar una masa de probabilidad igual a α.

Bajo estas consideraciones, simplificaremos el problema en (2.7) mediante una reformulación sensata del mismo, expresando las funciones de asignación z_j en términos de los parámetros θ. Esta reformulación no solo facilitará el desarrollo de la prueba de la existencia, sino la descripción del algoritmo que resuelva la parte muestral de nuestro problema, el cuál expondremos en la siguiente sección.

Definición 5. *Dados $\theta \in \Theta_c$ y una medida de probabilidad P, consideramos las funciones discriminantes definidas por:*

$$D_j(x;\theta) = \pi_j \ f(x;\mu_j,\Sigma_j) \quad y$$
$$D(x;\theta) = max\{D_1(x;\theta),\ldots,D_k(x;\theta)\}$$

Y consideramos la función de distribución de $D(\cdot,\theta)$ y su correspondiente α-cuantil:

$$G(u;\theta,P) \coloneqq P(D(\cdot;\theta) \leq u) \quad y \quad R(\theta,P) \coloneqq \inf_u \{G(u;\theta,P) \geq \alpha\} \tag{2.10}$$

El lector familiarizado con la materia se habrá percatado que estas funciones son ya conocidas de la regla de Bayes del análisis discriminante, de ahí el nombre que reciben. Además, permiten establecer una medida de "atipicidad" de las observaciones, que ayudará a determinar cuales de ellas recortar como aplicación del algoritmo del método.

Una vez introducida esta noción, tenemos una caracterización directa de las funciones de asignación, y podemos formalizar el problema como sigue:

Proposición 2.1. *Sean k un número natural, $\alpha \in (0,1)$, y $c \geq 1$ una constante. El problema de clustering robusto se puede simplificar, usando las funciones discriminantes, a la maximización en $\theta \in \Theta_c$ de:*

$$\theta \longmapsto L(\theta, P) \coloneqq E_{P_X}\left[\sum_{j=1}^{k} z_j(\cdot;\theta)\ log\ D_j(\cdot;\theta)\right], \tag{2.11}$$

donde las funciones de asignación se obtienen de θ como

$$z_j(x;\theta) = I\left\{x\ /\ \{D(x;\theta) = D_j(x;\theta)\} \cap \{D_j(x;\theta) \geq R(\theta,P)\}\right\}$$

y

$$z_0(x;\theta) = 1 - \sum_{j=1}^{k} z_j(x;\theta)$$

Es decir, se asigna x a la clase j con el mayor valor de la función discriminante $D_j(x;\theta)$ o x es recortada cuando todos los $D_j(x;\theta)$'s (y en consecuencia $D(x;\theta)$) son más pequeños que $R(\theta,P)$. En caso de empate entre distintas clases se puede establecer una regla de desempate como el orden lexicográfico.

Para probar la existencia de soluciones del problema (2.11), consideremos una sucesión de parámetros $\{\theta_n\}_{n=1}^{\infty} = \{(\pi_1^n,\ldots,\pi_k^n,\mu_1^n,\ldots,\mu_k^n,\Sigma_1^n,\ldots,\Sigma_k^n)\}_{n=1}^{\infty}$ tal que

$$\lim_{n\to\infty} L(\theta_n,P) = \sup_{\theta\in\Theta_c} L(\theta,P) = M > -\infty \tag{E.1}$$

(para ver la acotación inferior basta considerar $\pi_1 = 1$, $\mu_1 = 0$ y $\Sigma_1 = I$, y establecer el valor 0 para el resto de pesos). Como $[0,1]^k$ es un conjunto compacto, podemos extraer una subsucesión de $\{\theta_n\}_{n=1}^{\infty}$ (que denotaremos como la original para no complicar la notación) tal que:

$$\pi_j^n \to \pi_j \in [0,1] \quad \text{para } 1 \leq j \leq k, \tag{E.2}$$

y satisfaciendo para algún $g \in \{0,1,\ldots,k\}$ (una reordenación puede ser necesaria) que

$$\mu_j^n \to \mu_j \in \mathbb{R}^p \quad \text{para } 1 \leq j \leq g \quad \text{y} \quad \min_{j>g} \left\|\mu_j^n\right\| \to \infty \tag{E.3}$$

En cuanto a las matrices de dispersión, bajo la restricción (RA), se puede considerar una nueva subsucesión verificando una (y solo una) de las siguientes afirmaciones:

$$\Sigma_j^n \to \Sigma_j \quad \text{para } 1 \leq j \leq k \tag{E.4}$$

$$M_n = \max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l(\Sigma_j) \to \infty \tag{E.5}$$

o

$$m_n = \min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l(\Sigma_j) \to 0 \tag{E.6}$$

Veamos ahora una serie de lemas claves que nos ayudarán a completar la prueba. El primero de ellos nos señala la necesidad de las restricciones (RA) y (PR) para garantizar la existencia de una solución óptima.

Lema 2.1. *Si se mantiene la restricción (RA) y P satisface (PR), entonces solo la convergencia* (E.4) *es posible.*

Demostración. Veamos que (E.5) o (E.6) implican que $\lim_{n\to\infty} L(\theta_n, P) = -\infty$. Sean $\lambda_{l,j}^n := \lambda_l(\Sigma_j^n)$ para $j = 1, \ldots, k$ y $l = 1, \ldots, p$, los autovalores de las matrices de covarianza de los grupos, y $\|v_{l,j}^n\| = 1$ sus autovectores unitarios asociados. Entonces tenemos que:

$$\begin{aligned} L(\theta_n, P) = E_P \Bigg[\sum_{j=1}^{k} z_j(\cdot;\theta_n) \Bigg(\log \pi_j^n - \frac{p}{2}\log 2\pi - \frac{1}{2}\sum_{l=1}^{p} \log \lambda_{l,j}^n \\ - \frac{1}{2}\sum_{l=1}^{p} (\lambda_{l,j}^n)^{-1} (\cdot - \mu_j^n)' v_{l,j}^n (v_{l,j}^n)' (\cdot - \mu_j^n) \Bigg) \Bigg] \end{aligned} \tag{E.7}$$

$$\leq E_P \left[\sum_{j=1}^{k} z_j(\cdot;\theta_n) \left(\log \pi_j^n - \frac{p}{2}\log 2\pi - \frac{p}{2}\log m_n - \frac{1}{2} M_n^{-1} \| \cdot - \mu_j^n \|^2 \right) \right]$$

Por lo tanto, si suponemos (E.5), es decir, que $M_n \to \infty$, por la restricción (RA) necesariamente $m_n \to \infty$, y en virtud de la desigualdad anterior, tendríamos que $L(\theta_n, P) \to -\infty$, lo que contradice (E.1).

Supongamos ahora que se satisface (E.6). Podemos garantizar mediante el lema 2.2. (que probaremos a continuación) que si P satisface (PR), entonces existe una constante h tal que

$$E_P \left[\sum_{j=1}^{k} z_j(\cdot;\theta_n) \| \cdot - \mu_j^n \|^2 \right] \geq h > 0. \tag{E.8}$$

Dado que $\log \pi_j^n \leq 0$, y $P[z_1(\cdot) + \cdots + z_k(\cdot)] = 1 - \alpha$, podemos acotar la expresión en (E.7) y deducir que:

$$L(\theta_n, P) \leq (1-\alpha)\left(-\frac{p}{2}\log 2\pi - \frac{p}{2}\log m_n \right) - \frac{1}{2} M_n^{-1} E_P \left[\sum_{j=1}^{k} z_j(\cdot;\theta_n) \| \cdot - \mu_j^n \|^2 \right].$$

Y ahora, en virtud de (RA) y (E.8) tenemos que:

$$L(\theta_n, P) \leq (1-\alpha)\left(-\frac{p}{2}\log 2\pi - \frac{p}{2}\log m_n \right) - \frac{1}{2}(c m_n)^{-1} h. \tag{E.9}$$

Pero esta cota superior tiende a $-\infty$ cuando $m_n \to 0$, lo que contradice (E.1). En consecuencia, no se pueden dar (E.5) ni (E.6), luego necesariamente se cumple (E.4). □

Probaremos a continuación un pequeño lema técnico que ha sido utilizado en la prueba del lema 2.1.

Lema 2.2. *Si P satisface la condición (PR), entonces existe una constante $h > 0$ tal que la desigualdad* (E.8) *se cumple.*

Demostración. El problema de k-medias recortadas, al cual dedicamos el capítulo 1, fue introducido como la búsqueda de un conjunto de k puntos $M = \{m_1, \ldots, m_k\}$ en $\mathbb{R}^p$ y un conjunto de Borel A_0 que minimice:

$$\min_{A_0:P(A_0)\geq 1-\alpha} \min_{\mu_1,\ldots,\mu_k} \frac{1}{P(A_0)} \int_{A_0} \inf_{1\leq j\leq k} \|x-\mu_j\|^2 \, dP(x). \tag{E.10}$$

El Teorema 1.1 garantiza la existencia de soluciones para este problema. Denotando por $V_{\alpha,\,k}$ el valor mínimo alcanzado en la expresión (E.10), se tiene para toda elección de θ:

$$E_{P_X}\left[\sum_{j=1}^{k} z_j(\cdot;\theta)\|\cdot-\mu_j\|^2\right] \geq E_{P_X}\left[\sum_{j=1}^{k} z_j(\cdot;\theta) \inf_{1\leq l\leq k} \|\cdot-\mu_l\|^2\right] \geq (1-\alpha)V_{\alpha,\,k}$$

pues $\bigcup_{j=1}^{k}\{x : z_j(x;\theta) = 1\}$ es un conjunto de Borel cuya probabilidad es mayor o igual que $1-\alpha$. Luego, basta tomar $h := (1-\alpha)V_{\alpha,k} > 0$ siempre que la condición (PR) se cumpla para P. □

El siguiente paso es demostrar que, siempre que las clases en la partición óptima tengan probabilidades estrictamente positivas, podemos garantizar la convergencia de los centros μ_j^n. Este resultado también es importante para entender el papel que juegan los pesos π_j^n en este enfoque.

Lema 2.3 (A3)**.** *Cuando se satisfacen (PR) y (RA), si todos los pesos π_j en* (E.2) *son estrictamente positivos, es decir, $\pi_j > 0$ para $j = 1, \ldots, k$, entonces $g = k$ en* (E.3)*.*

Demostración. Si $g = 0$, podemos tomar una bola con centro en 0 y radio suficientemente grande, $B(0,R)$, tal que $P[B(0,R)] > \alpha$. Se comprueba entonces fácilmente que

$$E_{P_X}\left[\sum_{j=1}^{k} z_j(\cdot;\theta_n)\left\|\cdot-\mu_j^n\right\|^2\right] \to \infty.$$

pues siempre existe algún $x \in B(0,R)$ con $z_j(x,\theta_n) = 1$ para algún $j \in \{1,2,\ldots,k\}$. A partir de esto, la restricción (RA) y de la desigualdad en (E.7), podemos deducir que $L(\theta_n, P) \to -\infty$, lo cual contradice (E.1).

Por otro lado, si $g > 0$, probamos primero que

$$E_P\left[\sum_{j=g+1}^{k} z_j(\cdot;\theta_n)\right] \to 0. \tag{E.11}$$

Esto se deduce del teorema de convergencia dominada, teniendo en cuenta que la sucesión está obviamente acotada por $1-\alpha$, y que

$$\{x : z_j(x;\theta_n) = 1\} \subseteq \left\{x : \max_{j=g+1,\ldots,k} D_j(x;\theta_n) \geq D_1(x;\theta_n)\right\} \tag{E.12}$$

para $j = g+1,\ldots,k$, donde el lado derecho converge hacia el conjunto vacío cuando $n \to \infty$, debido a (E.3) y (E.4).

Ahora, a partir de (E.11) se deduce:

$$\lim_{n\to\infty} \sup L(\theta_n, P) \leq \lim_{n\to\infty} E_P\left[\sum_{j=1}^{g} z_j(\cdot;\theta_n)\left(\log \pi_j^n - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_j^n| \right.\right.$$

$$\left.\left. -\frac{1}{2}(\cdot-\mu_j^n)'(\Sigma_j^n)^{-1}(\cdot-\mu_j^n)\right)\right]$$

$$= E_P\left[\sum_{j=1}^{g} z_j(\cdot;\tilde{\theta})\left(\log \pi_j - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(\cdot-\mu_j)'\Sigma_j^{-1}(\cdot-\mu_j)\right)\right],$$

donde $x \mapsto z_j(x;\tilde{\theta})$ son las funciones de asignación que se obtendrían al trabajar con g (en lugar de k) poblaciones y $\tilde{\theta}$ es igual a un límite de la subsucesión $\{\theta_n\}_{n=1}^{\infty} = \{(\pi_1^n,\ldots,\pi_g^n,\mu_1^n,\ldots,\mu_g^n,\Sigma_1^n,\ldots,\Sigma_g^n)\}_{n=1}^{\infty}$.

Como $\sum_{j=1}^{g}\pi_j < 1$, la prueba finaliza mostrando que podemos cambiar los pesos $\pi_1,\ldots,\pi_k$ por

$$\pi_j^* = \frac{\pi_j}{\sum_{j=1}^{g}\pi_j} \quad \text{para } 1 \leq i \leq g \text{ y } \pi_{g+1}^* = \cdots = \pi_k^* = 0 \tag{E.13}$$

(modificando apropiadamente las funciones de asignación z_j). Este cambio provoca una disminución estricta en la función objetivo, lo que contradice la optimalidad declarada en (2.11). Por lo tanto, necesariamente $g = k$, y se obtiene el resultado deseado. □

Concluímos ahora la prueba de la existencia de soluciones mediante el siguiente razonamiento.

Proposición 2.2 (Existencia)**.** *Si la distribución P satisface (PR), entonces existe algún $\theta \in \Theta_c$ tal que el máximo en* (2.7) *sujeto a la restricción (RA) es alcanzado.*

Demostración. Teniendo en cuenta los lemas previos, debe cumplirse una de las siguientes posibilidades para la sucesión de parámetros descrita anteriormente:

- (i) Si $\pi_j^n \to \pi_j > 0$ para $1 \leq j \leq k$, entonces la elección de θ es clara
- (ii) Si $\pi_j^n \to \pi_j > 0$ para $j \leq g$ y $\pi_j = 0$ para $g < j \leq k$, podemos definir los pesos π_j como $\pi_j = lim_{n\to\infty}\pi_j^n$ para $j = 1, \ldots, g$ y $\pi_{g+1} = \ldots = \pi_k = 0$, y tomar $\mu_j = lim_{n\to\infty}\mu_j^n$ y $\Sigma_j = lim_{n\to\infty}\Sigma_j^n$ para $j \leq g$. El resto de μ_j's y Σ_j's pueden ser elegidos arbitrariamente (por supuesto satisfaciendo (RA)).

□

Observación 2.1. *Nótese que en el resultado anterior hemos admitido pesos $\pi_j = 0$. Sin embargo, esto no representa una desventaja al tomar $\log \pi_j$, ya que en este caso $z_j(\cdot;\theta) \equiv 0$ y el conjunto $\{x : z_j(x;\theta) = 1\}$ es vacío. La presencia de grupos con peso cero ocurre a menudo en la práctica. Por ejemplo, cuando $k = 2$, $c = 1$, $\alpha = 0$ y P es la distribución $\mathcal{N}(0,1)$ en la recta real, podemos ver que $\theta = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_1^2) = (1, 0, 0, \mu_2, 1, 1)$ es la solución óptima para todo $\mu_2 \in \mathbb{R}$.*

Observación 2.2. *Recordamos que para el problema de k-medias recortadas, la función objetivo siempre mejoraba cuando se aumentaba el número de clusters k, como vimos en la proposición 1.3. Sin embargo, con este nuevo enfoque, la aparición de grupos con peso $\pi_j = 0$ implica que la función objetivo no mejora necesariamente cuando aumentamos k. Esto puede resultar interesante a la hora de desarrollar métodos para determinar un k óptimo, pues la presencia de un peso π_j cercano a 0 sugiere tomar un k más pequeño.*

Una vez probada la existencia de una solución óptima, cabe preguntarse por la consistencia del modelo en el sentido de la convergencia de las soluciones empíricas a la solución óptima del caso poblacional. Debido a la necesidad de técnicas y herramientas que quedan fuera del nivel de un traajo de fin de grado, nos limitaremos a enunciar el resultado de consistencia recogido en la siguiente proposición. Las demostraciones correspondientes pueden consultarse en [12].

Proposición 2.3 (Consistencia)**.** *Supongamos que la distribución P tiene una función de densidad estrictamente positiva y que θ_0 es la única solución que maximiza* (2.7) *bajo la restricción (RA). Si $\theta_n \in \Theta_c$ denota un estimador basado en la medida empírica P_n, entonces $\theta_n \to \theta_0$ casi seguro.*

2.1.3. Algoritmo TCLUST

El problema empírico de clustering robusto posee una complejidad computacional enorme, y encontrar una solución exacta resulta inviable incluso para tamaños de muestra moderados. Es esencial por tanto el desarrollo de algoritmos que puedan aproximar la solución óptima del problema muestral. En esta sección, nos centraremos en el algoritmo TCLUST introducido en [12], el cual aproxima una solución de (2.5) sujeto a la restricción (RA).

La base de este algoritmo es el principio EM, muy utilizado en la búsqueda de soluciones de problemas de máxima verosimilitud, como se analiza en [5]. Una pequeña modificación permite adaptar esta idea al clustering robusto, enfoque conocido como clasificación EM, introducido en [2].

Los pasos del algoritmo TCLUST son los siguientes:

1. Se seleccionan aleatoriamente valores iniciales para los centros m_j^0's, las matrices de covarianza ${S_j}^0$'s y los pesos de los grupos p_j^0's para $j = 1, \ldots, k$.

2. A partir del parámetro $\theta^l = (p_1^l, \ldots, p_k^l, m_1^l, \ldots, m_k^l, S_1^l, \ldots, S_k^l)$ devuelto por la iteración anterior:

   a) Se obtiene $d_i = D(x_i, \theta^l)$ para las observaciones $\{x_1, \ldots, x_n\}$ y se guarda el conjunto H con las $[n(1-\alpha)]$ observaciones con los mayores d_i's.

   b) Se genera una partición de H, denotada por $\{H_1, \ldots, H_k\}$, con $H_j = \{x_i \in H : D_j(x_i, \theta^l) = D(x_i, \theta^l)\}$.

   c) Se toma el número de datos n_j en H_j, su media muestral m_j y su matriz de covarianza muestral S_j, para $j = 1, \ldots, k$.

   d) Se considera la descomposición en valores singulares de $S_j = U_j' D_j U_j$ donde U_j es una matriz ortogonal y $D_j = \text{diag}(\Lambda_j)$ es una matriz diagonal (con elementos diagonales dados por el vector Λ_j).

   Si el vector completo de autovalores $\Lambda = (\Lambda_1, \ldots, \Lambda_k)$ no satisface la restricción de razón de autovalores (RA), se obtiene mediante el algoritmo de Dykstra[7] un nuevo vector $\tilde{\Lambda} = (\tilde{\Lambda}_1, \ldots, \tilde{\Lambda}_k)$ que cumpla con la restricción y tal que $\|\tilde{\Lambda} - \Lambda^{-1}\|^2$ sea lo más pequeño posible. (Λ^{-1} denota el vector compuesto por los inversos de los elementos de Λ). Nótese que la restricción (RA) para Λ es la misma que para Λ^{-1}.

   e) Se actualiza θ^{l+1} mediante:

   1) $p_j^{l+1} \leftarrow n_j/[n(1-\alpha)]$
   2) $m_j^{l+1} \leftarrow m_j$
   3) $S_j^{l+1} \leftarrow U_j' \widetilde{D}_j U_j$ y $\widetilde{D}_j = diag(\widetilde{\Lambda}_j)^{-1}$

3. Se realizan F iteraciones del proceso descrito en el paso 2 (un número moderado de iteraciones suele ser suficiente) y se evalúa la función $L(\theta^F, P_n)$ mediante la expresión (2.11).

4. Se comienza desde el paso 1 varias veces, guardando las soluciones que lleven a valores maximales de la función $L(\theta^F, P_n)$ y se escoge la mejor de ellas.

Analicemos ahora la idea detrás de este algoritmo. Cada iteración del paso 2 consta del cálculo de una esperanza (E-step) y una fase de maximización (M-step). En la primera fase, se calculan las probabilidades "a posteriori" $D_j(x_i, \theta^l) = p_j \; f(x_i; m_j, S_j)$, las cuales inducen una clasificación discreta dejando una proporción α de observaciones sin asignar, que son las más difíciles de clasificar basándonos en estas probabilidades. Más tarde, en la fase de maximización, se obtienen nuevos parámetros θ^{l+1} maximizando la esperanza condicional.

La obtención de las matrices de dispersión en cada iteración se descompone en la búsqueda de los autovalores y los autovectores correspondientes. Para cada elección de autovalores, los autovectores se derivan directamente de las matrices de covarianza muestrales de las observaciones en cada grupo. Esta descomposición es similar a la presentada en Gallegos [10], donde las "formas" y "tamaños" de las matrices de dispersión se trataban por separado.

Tanto la inicialización aleatoria (paso 1) como el refinamiento final (paso 4) son esenciales en este algoritmo. Para el paso 1, se ha observado que simplemente tomar k puntos aleatorios del conjunto de datos como centros, k matrices identidades para las covarianzas y el mismo peso para todos los grupos (igual a $\frac{1}{k}$) proporciona buenos resultados en la mayoría de casos.

En cuanto a la restricción de autovalores, necesitamos que $\Lambda = (\Lambda_1, \ldots, \Lambda_k)$ con $\Lambda_j = (\lambda_{1,j}, \ldots, \lambda_{p,j})$ pertenezca al cono:

$$\mathcal{C} = \left\{(\Lambda_1, \ldots, \Lambda_k) \in \mathbb{R}^{p\times k} : \lambda_{u,v} - c \cdot \lambda_{r,s} \leq 0 \text{ para todo } (u,v) \neq (r,s)\right\} \tag{2.12}$$

Si $\Lambda \notin \mathcal{C}$, reemplazaremos Λ^{-1} por $\tilde{\Lambda} \in \mathcal{C}$ con el mínimo $\|\tilde{\Lambda} - \Lambda^{-1}\|^2$. El algoritmo de Dykstra, presentado en [7], resuelve aproximadamente ese problema restringido de mínimos cuadrados, siempre que $\mathcal{C}$ sea la intersección de varios conos convexos cerrados

$$\mathcal{C}_h = \left\{(\Lambda_1, \ldots, \Lambda_k) \in \mathbb{R}^{p\times k} : \lambda_{u,v} - c \cdot \lambda_{r,s} \leq 0\right\} \quad \text{para } h = (u,v,r,s),$$

recurriendo a proyecciones iterativas sobre los conos individuales $\mathcal{C}_h$.

El algoritmo TCLUST puede adaptarse a otros modelos y restricciones vistos durante este capítulo mediante modificaciones sencillas. Por ejemplo, si queremos aplicar un algoritmo para aproximar la mejor solución bajo la restricción de los determinantes (2.9) basta considerar el cono

$$\mathcal{C}' = \{(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^k : \sigma_u - c \cdot \sigma_v \leq 0 \text{ para todos } u \neq v\},$$

donde la factorización del paso 2.4 es de la forma $S_j = \sigma_j \cdot U_j$ con $|U_j| = 1$ y $\sigma_j = |S_j|^{\frac{1}{p}}$.

En la aplicación del algoritmo expuesto, cada observación x_i es asignada al cluster más probable en función de las probabilidades "a posteriori", sin embargo, no todas las asignaciones son igualmente fiables. Las funciones discriminantes $D_j(x, \theta)$ permiten establecer una medida de credibilidad de una asignación, como se describe en Van Aeslt et al. [25]. En el capítulo 3 veremos más en profundidad como emplear estas funciones para obtener información de la calidad de una asignación, lo que permitirá evaluar la adecuación del modelo.

2.2. Modelo de mezcla con clasificación blanda

Los modelos descritos hasta ahora establecen una clasificación total de las observaciones, es decir, asignan cada observación no recortada a un determinado cluster. Esto

es lo que se conoce como clasificación "dura". Sin embargo, existen otro género de modelos de clustering que tratan el problema desde un enfoque más probabilístico, no solo indentificando los posibles datos atípicos, sino tratando de cuantificar la incertidumbre en la asignación de una observación a un cluster mediante probabilidades.

Este enfoque basado en probabilidades de pertenencia a cada cluster es conocido como clasificación blanda. Esta asignación suave de las observaciones resulta especialmente interesante cuando hay muchas asignaciones dudosas, ofreciendo una interpretabilidad más flexible de los datos.

Además, en muchos estudios, un enfoque de clasificación dura puede resultar demasiado estricto, y asignar cada observación totalmente a un cluster puede suponer una pérdida de información necesaria. Veamos esto con un ejemplo sencillo.

Supongamos que un banco quiere entrenar un detector de billetes fraudulentos. Para ello dispone de 200 billetes, entre los que se encuentran tanto billetes válidos como falsificados. Se pretende así realizar una clasificación de los mismos para poder detectar futuros billetes falsos, en función de dos medidas: las distancias del marco interior al borde superior y al borde inferior respectivamente. El conjunto de datos descrito corresponde al dataset 'swissbank', del paquete 'tclust' de R, del que se han tomado las variables "IF Lower" y "IF Upper".

Si se realiza una clasificación dura, cada observación es asignada al cluster de mayor probabilidad de pertenencia, sin tener en cuenta el valor de esta probabilidad. En este caso particular, conocer las probabilidades de pertenencia a cada grupo puede resultar de gran ayuda para la interpretación del resultado. Si un billete tiene un 40 % de probabilidad de ser falso, asignarlo completamente al grupo de billetes verdaderos no parece una buena idea, pues este procedimiento puede conducir a la aprobación de un número considerable de billetes fraudulentos, y, en consecuencia, una gran pérdida económica. Una clasificación blanda, en cambio, permitiría medir la credibilidad de cada billete, y establecer un margen más estricto para su aprobación.

La figura 2.4 muestra los resultados de realizar una clasificación dura o una clasificación blanda para el conjunto de billetes descrito. En la gráfica central se han representado en azul, las observaciones catalogadas como asignaciones dudosas, correspondientes a los billetes con menos de un 99 % de probabilidad de pertenencia a ambos

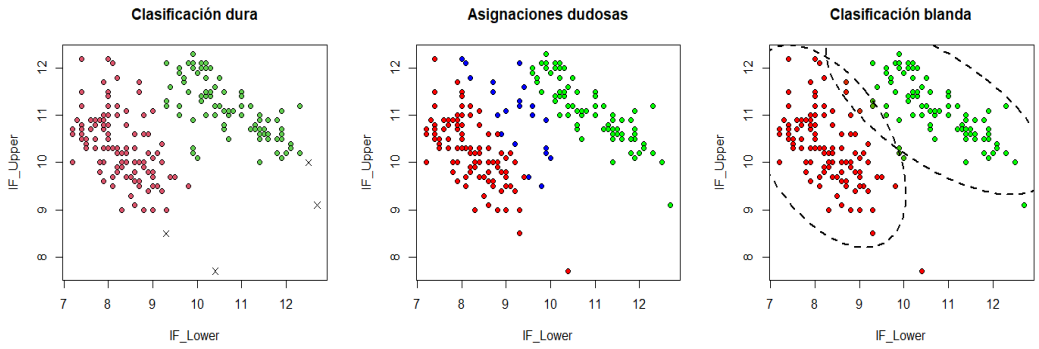


Figura 2.4: Resultados de aplicar una clasificación dura (TCLUST) y una clasificación blanda sobre las variables 'IF_Lower' e 'IF_Upper' del conjunto de billetes 'swissbank'.

clusters. La clasificación blanda permite establecer una medida de la veracidad de cada billete, pudiendo recurrir a otras medidas o herramientas de identificación para las asignaciones dudosas. Este ejemplo muestra que, si bien la interpretación de la clasificación blanda es más compleja que la de la clasificación usual, en determinados contexto aporta una información adicional muy beneficiosa o incluso imprescindible.

Al igual que ocurría con el método de k-medias clásico, los procedimientos de clasificación blanda convencionales pueden verse afectados gravemente por la presencia de observaciones irregulares. Esto conduce a la necesidad del desarrollo de modelos que robustifiquen esta métodología. Existen en la literatura diversas propuestas basadas en estadística robusta, como el conocido MCLUST, que modela una componente adicional de ruido, o la mezcla de distribuciones t de Student, tratada en [21].

No obstante, como nos ocupa en este trabajo, nos centraremos en los modelos de robustificación basados en el recorte de los datos, los cuales se ha comprobado que ofrecen un método flexible y computacionalmente viable, adaptable a situaciones de contaminación muy variadas. En esta sección seguiremos el desarrollo de [16], que introduce una extensión de la metodología TCLUST adaptada al contexto de la clasificación blanda.

Los modelos de mezcla gaussianos ofrecen una base adecuada para este enfoque, permitiendo modelar estructuras de datos compuestas por la combinación de varias distribuciones. Es decir, se asume que los datos observados siguen una distribución conjunta de densidad

$$g(x) = \sum_{j=1}^{k} \pi_j f(x, \mu_j, \Sigma_j)$$

donde $f(x, \mu_j, \Sigma_j)$ es la función de densidad de una normal multivariada con vector de medias μ_j y matriz de covarianzas Σ_j, y se buscan los parámetros $\{\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k\}$ que mejor representen la nube de datos, a través de la maximización de una función de log-verosimilitud completa.

La base de estos modelos es el problema de ajustar una mezcla de k distribuciones normales a un conjunto de datos $\{x_1, \ldots, x_n\} \subset \mathbb{R}^p$. Este enfoque, busca combinar la flexibilidad del modelo de mezclas gaussianas con la potencia y robustez del recorte, proporcionando así un marco ideal para el análisis cluster. Esta idea se formaliza mediante la maximización de la función de verosimilitud de mezcla recortada

$$\sum_{i=1}^{n} z(x_i) \left[\sum_{j=1}^{k} \pi_j f(x_i, \mu_j, \Sigma_j) \right] \tag{2.13}$$

donde z es la función indicadora de recorte, que devuelve $z(x_i) = 0$ si la observación x_i es recortada o $z(x_i) = 1$ si no. Al igual que en los modelos de recorte vistos previamente, se permite recortar una proporción α de las observaciones, luego se impone la condición $\sum_{i=1}^{n} z(x_i) = [n(1-\alpha)]$ a nuestro problema.

Este problema soluciona la falta de robustez de los modelos de mezcla ante la presencia de contaminación o ruido. Sin embargo, la maximización de (2.13) sigue llevando frecuentemente a soluciones espurias, incluso cuando los datos provienen realmente de una mezcla de k distribuciones normales, debido a que no es un problema bien definido,

como explicamos en la sección anterior.

Bajo el mismo razonamiento, surge de forma natural imponer una restricción en las matrices de dispersión. Se emplea así un estimador que combina las restricciones vistas para TCLUST con la metodología de la función de log-verosimilitud (2.13) presentada en [22]. Debido a la mayor flexibilidad de la misma, nos centraremos en el caso de tomar la restricción de autovalores (2.8), que denotamos por (RA).

Por tanto, sea un conjunto de datos $\{x_1, \ldots, x_n\} \subset \mathbb{R}^p$, el modelo de mezcla de clustering robusto consiste en la maximización de (2.13) bajo la restricción de autovalores:

$$\frac{M_n}{m_n} \leq c,$$

donde

$$M_n = \underset{j=1,\ldots,k}{max} \left(\underset{l=1,\ldots,p}{max} \lambda_l(\Sigma_j) \right) \quad \text{y} \quad m_n = \underset{j=1,\ldots,k}{min} \left(\underset{l=1,\ldots,p}{min} \lambda_l(\Sigma_j) \right),$$

y $\{\lambda_j(\Sigma)\}_{j=1}^p$ denota el conjunto de autovalores de una matriz de dispersión Σ.

Esta restricción permite controlar la relación de las formas y tamaños de las distintas matrices de dispersión, como ya analizamos para el TCLUST.

Al igual que hemos hecho con otros modelos, el problema de mezcla puede generalizarse a un marco teórico o probabilístico a través de la esperanza matemática. De este modo, sea P una medida de probabilidad, el problema restringido del ajuste de mezclas recortado se define como la maximización de

$$E_P \left[z(\cdot)\ log \left(\sum_{j=1}^{k} \pi_j f(\cdot; \mu_j, \Sigma_j) \right) \right] \tag{2.14}$$

en términos de la función indicadora de recorte

$$z : \mathbb{R}^p \longrightarrow \{0, 1\} \quad \text{tal que} \quad E_P\ z(\cdot) = 1 - \alpha$$

y los parámetros $\theta = \{\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k\}$ correspondientes a los pesos $\pi_j \in [0, 1]$, con $\sum_{j=1}^k \pi_j = 1$, centros $\mu_j \in \mathbb{R}^p$ y matrices $(p \times p)$ definidas positivas Σ_j verificando la restricción (RA) para una constante fija $c \geq 1$, es decir, $c \in \Theta_c$.

De forma análoga al desarrollo realizado en la sección dedicada al TCLUST, dado un parámetro $\theta \in \Theta_c$ podemos definir las funciones discriminantes $D_j(x; \theta) = \pi_j\ f(x; \mu_j, \Sigma_j)$ y $D(x; \theta) = \sum_{j=1}^k D_j(x; \theta)$, y escribir la función z en términos de dicho parámetro, como

$$z(x) \coloneqq z(x, \theta) = I\{x : D(x; \theta) \geq R(\theta, P)\},$$

con

$$H(u; \theta, P) = P(D(\cdot; \theta) \leq u) \quad \text{y} \quad R(\theta; P) = \underset{u}{inf}\{H(u; \theta, P) \geq \alpha\}.$$

Bajo ciertas condiciones suaves sobre la distribución P, podemos asegurar la existencia de soluciones del problema (2.14), así como un resultado de consistencia del

método. Basta imponer que la distribución P no esté concentrada en k puntos.

(PR) La distribución P no está concentrada en k puntos tras eliminar una masa de probabilidad igual a α.

En este trabajo nos limitaremos a enunciar los resultados principales de existencia y consistencia, análogos a los vistos para TCLUST. Estos resultados han sido tomados de [16], donde se pueden encontrar las demostraciones correspondientes, las cuales siguen la misma idea que las vistas en la sección anterior.

Proposición 2.4 (Existencia)**.** *Si la distribución P satisface (PR), entonces existe algún $\theta \in \Theta_c$ tal que el máximo en* (2.14) *sujeto a la restricción (RA) es alcanzado.*

Proposición 2.5 (Consistencia)**.** *Supongamos que se verifica (PR) para la distribución P que tiene una función de densidad estrictamente positiva y que θ_0 es la única solución que maximiza* (2.14) *bajo la restricción (RA). Si $\theta_n \in \Theta_c$ denota un estimador basado en la medida empírica P_n, entonces $\theta_n \to \theta_0$ casi seguro.*

A continuación presentamos un algoritmo adecuado para resolver el problema (2.14). Este algoritmo supone una mejora del algoritmo TCLUST desde el punto de vista computacional, y una adaptación al enfoque de la clasificación blanda.

Algoritmo de clasificación suave:

1. *Inicialización:* El algoritmo se inicializa N veces para distintos $\theta^{(0)} = (\pi_1^{(0)}, \ldots, \pi_k^{(0)}, \mu_1^{(0)}, \ldots, \mu_k^{(0)}, \Sigma_1^{(0)}, \ldots, \Sigma_k^{(0)})$. Para cada inicialización se seleccionan aleatoriamente $k \times (p+1)$ observaciones y se calculan k centros $\mu_j^{(0)}$ y k matrices de dispersión $\Sigma_j^{(0)}$ a partir de estas. Las restricciones sobre las matrices de dispersión (descritas en el paso 2.2) se aplican a las matrices $\Sigma_j^{(0)}$ si es necesario. Los pesos $\pi_1^{(0)}, \ldots, \pi_k^{(0)}$ en el intervalo $(0, 1)$ y tales que $\sum_{j=1}^{k} \pi_j = 1$ se eligen aleatoriamente.

2. *Pasos EM recortados:* Los siguientes pasos se realizan alternativamente hasta la convergencia ($\theta^{(l+1)} = \theta^{(l)}$) o hasta que un máximo de iteraciones M sea alcanzado.

   2.1 *Pasos E y C:* En esta iteración se descartan las observaciones en $I = \{i : D(x_i; \theta^{(l)}) \leq R(\theta^{(l)}; P_n)\}$. Dicho de otra manera, si

   $$D(x_{(1)}; \theta^{(l)}) \leq D(x_{(2)}; \theta^{(l)}) \leq \ldots \leq D(x_{(n)}; \theta^{(l)})$$

   es una reordenación de las funciones discriminantes, entonces $R(\theta^{(l)}; P_n) = D(x_{([n\alpha])}; \theta^{(l)})$, y descartamos la proporción α de las observaciones con valores $D(x_i; \theta^{(l)})$ más pequeños. Calculamos las probabilidades a posteriori como

   $$\tau_j(x_i; \theta^{(l)}) = \frac{D_j(x_i; \theta^{(l)})}{D(x_i; \theta^{(l)})}, \quad \text{para } i = 1, \ldots, n.$$

y para las observaciones descartadas se modifican los valores $\tau_j(x_i;\theta^{(l)})$ por

$$\tau_j(x_i;\theta^{(l)}) = 0 \quad \text{para todo } j = 1,\ldots,k, \text{ cuando } j \in I.$$

2.2 *Paso M:* Se actualizan los parámetros como

$$\pi_j^{(l+1)} = \sum_{i=1}^{n} \frac{\tau_j(x_i;\theta^{(l)})}{[n(1-\alpha)]}$$

y

$$\mu_j^{(l+1)} = \frac{\sum_{i=1}^{n} \tau_j(x_i;\theta^{(l)}) x_i}{\sum_{i=1}^{n} \tau_j(x_i;\theta^{(l)})}$$

Inicialmente, se actualizan las matrices de dispersión como

$$T_j = \frac{\sum_{i=1}^{n} \tau_j(x_i;\theta^{l})(x_i - \mu_j^{(l+1)})(x_i - \mu_j^{(l+1)})'}{\sum_{i=1}^{n} \tau_j(x_i;\theta^{l})}$$

Sin embargo, estas matrices pueden no satisfacer la restricción de autovalores (RA). En este caso, consideramos la descomposición en valores singulares $T_j = U_j' D_j U_j$, siendo U_j una matriz ortogonal y $D_j = diag(d_{j1}, d_{j2}, \ldots, d_{jp})$ una matriz diagonal. Se definen los autovalores truncados por

$$[d_{gl}]_t = \begin{cases} d_{js} & \text{si } d_{js} \in [t, ct] \\ t & \text{si } d_{js} < t \\ ct & \text{si } d_{js} > ct \end{cases}$$

para un valor umbral t. Entonces, se actualizan las matrices de dispersión finalmente como

$$\Sigma_j^{(l+1)} = U_j' D_j^* U_j,$$

con $D_j^* = diag([d_{j1}]_{t_{opt}}, [d_{j2}]_{t_{opt}}, \ldots, [d_{jk}]_{t_{opt}})$ y t_{opt} minimizando

$$t \longrightarrow \sum_{j=1}^{k} \pi_j^{(l+1)} \sum_{s=1}^{p} \left(log([d_{js}]_t) + \frac{d_{js}}{[d_{js}]_t} \right). \tag{2.15}$$

3. *Evaluación de la función objetivo:* Tras efectuar los pasos EM recortados, se calcula el valor de la función objetivo (asignando $z(x_i) = 0$ si $i \in I$ y $z(x_i) = 1$ si $i \in I$). El conjunto de parámetros que conduce al valor más alto de la función objetivo, y la función indicadora de recorte asociada z se devuelven como resultado del algoritmo.

Ahora, se pueden computar las probabilidades posteriores τ_j en función de los parámetros devueltos por el algoritmo. Estas probabilidades constituyen una clasificación blanda de las observaciones, lo que ofrece mayor riqueza e información a nuestro análisis. No obstante, este resultado posee una interpretabilidad más compleja que la clasificación dura, por lo que se debe elegir el método en función de los objetivos del estudio realizado, y de las características de la solución deseada.

Capítulo 3

Elección de parámetros y evaluación de la clasificación

A lo largo de este trabajo se ha expuesto en repetidas ocasiones que, el análisis cluster, lejos de ser una cuestión completamente definida, está sujeta a las suposiciones realizadas para cada estudio. La clasificación obtenida depende no solamente del modelo escogido para el análisis, sino también de los parámetros tomados para dicho modelo. Una de las cuestiones más debatidas en este contexto es la elección del número de clusters k a identificar. Este dilema constituye el problema fundamental del análisis cluster, ocupando numerosos estudios en busca de métodos objetivos para tratar esta decisión.

En el análisis cluster basado en el recorte de los datos, la elección de k debe tomarse conjuntamente con la del parámetro α, que indica la proporción de observaciones descartadas en el análisis. Ambos parámetros mantienen una estrecha relación entre sí, y la determinación de uno condiciona directamente la decisión del otro, lo que aumenta el nivel de complejidad de nuestro estudio.

En este capítulo se presentarán alguna de las herramientas más utilizadas en la toma de estas decisiones, y se discutirán claves y consideraciones a tener en cuenta en el ajuste del modelo a un conjunto de datos concreto. Además, se abordarán criterios para evaluar la calidad de las clasificaciones obtenidas, lo que resultará fundamental a la hora de identificar una elección errónea de los parámetros del modelo.

3.1. Selección del número de clusters (k) y nivel de recorte (α)

Los métodos presentados durante el desarrollo de este trabajo requieren de la elección del número de clusters k a identificar, y del nivel de recorte α, de manera previa a la búsqueda de la mejor agrupación posible para estos valores. A menudo, el parámetro k viene determinado por la finalidad de nuestro análisis, que precisa de la obtención de un número concreto de grupos. No obstante, esta no es la situación general, y, en la gran mayoría de casos, el parámetro k se determina por la propia estructura de los datos y las relaciones que guardan los individuos entre sí. Esta dependencia de la estructura subyacente de los datos convierte a la elección de k en el problema fundamental

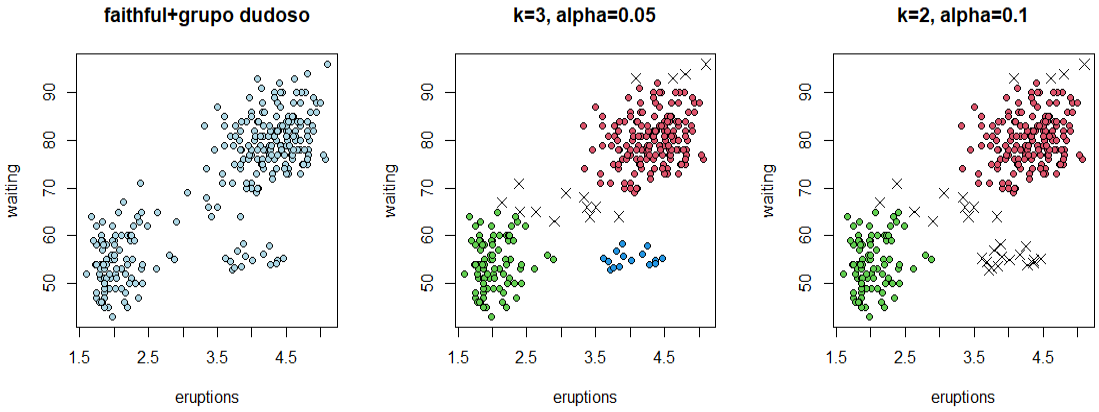


Figura 3.1: Resultados de aplicar TCLUST sobre el conjunto 'faithful' con un grupo adicional de pocas observaciones.

del análisis cluster, pues imposibilita el desarrollo de un método sistemático para su obtención.

En el análisis cluster basado en el recorte este problema es aún más complejo, pues la elección de k depende estrechamente del valor de α tomado. Por ejemplo, la elección de un α muy grande puede provocar el recorte de clusters pequeños, con pocas observaciones, precisando así de un k más pequeño. Por el contrario, un α demasiado pequeño puede conllevar la formación de clusters degenerados por agrupación de datos atípicos, afectando así la clasificación resultante. Una manera de combatir esta alteración es precisamente aumentando el valor de k, identificando artificialemente este tipo de agrupaciones espurias.

Esta relación se ilustra en la figura 3.1, donde se ha considerado el conjunto de datos "faithful" de R, que recoge el tiempo de duración de las erupciones y el tiempo de espera entre erupciones del géiser Old Faithful, a los que se ha añadido un pequeño grupo artificial de apenas 15 observaciones. Si fijamos un total de tres grupos y un nivel bajo de recorte, el algoritmo detecta esta pequeña agrupación como un cluster. Sin embargo, al aumentar el nivel de recorte al 10 %, estas observaciones son descartadas y consideradas como ruido, y el resultado óptimo pasa a ser una partición en dos grupos. Es decir, la determinación de un nivel de recorte α condiciona el valor óptimo del número de clusters k y viceversa, por lo que la elección de estos dos parámetros debe estudiarse de manera conjunta.

En la práctica, la naturaleza de los datos observados es desconocida, lo que dificulta discernir cuando este tipo de pequeñas agrupaciones corresponden realmente a un cluster, o, por el contrario, se trata de un conjunto de datos atípicos casualmente agrupados. Como siempre, este tipo de dilemas deben ser tratados en función del objetivo de nuestro análisis y de la posible interpretación de estos grupos en el estudio.

En general, la determinación de los parámetros k y α no tiene una solución óptima única, si no que debe estudiarse su comportamiento conjunto en busca de una elección que proporcione una partición coherente del conjunto de datos. Como se mencionó en el

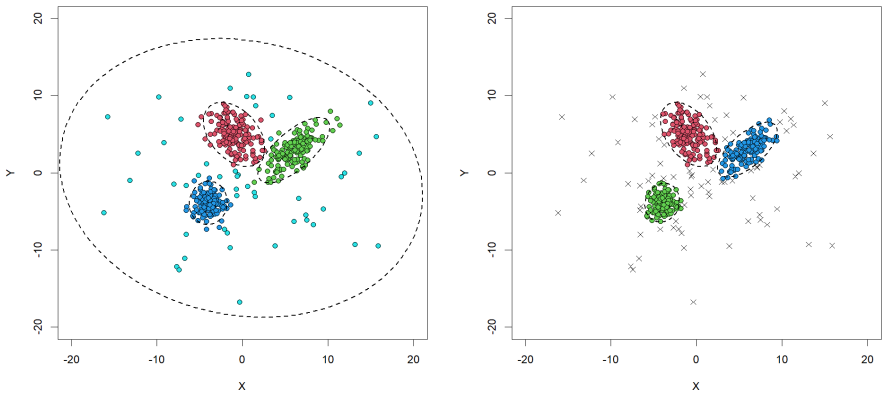


Figura 3.2: Resultados de aplicar TCLUST con $k = 4$, $\alpha = 0$, $c = 50$ (izq.) y $k = 3$, $\alpha = 0.15$, $c = 50$ (dcha).

primer capítulo, una posible estrategia contra la contaminación en el clustering consiste en modelar el ruido mediante la inclusión de un cluster adicional de gran dispersión que recoja estas observaciones atípicas. Si bien este enfoque no resulta generalmente válido, existen situaciones a las que se adapta correctamente, y permite incluso obtener información de estos datos irregulares.

Esto se ilustra en la figura 3.2, donde se parte de un conjunto formado por muestras de tres distribuciones normales, al que se ha añadido contaminación procedente de una cuarta distribución normal de mayor dispersión. La gráfica de la izquierda muestra el resultado de aplicar TCLUST tomando $k = 4$, $\alpha = 0$ y $c = 50$, de modo que el algoritmo identifica las observaciones contaminantes como un cluster adicional. Por otro lado, en la gráfica de la derecha se ha tomado $k = 3$, $\alpha = 0.15$ y $c = 50$, conduciendo al descarte de las observaciones contaminantes mediante el recorte. Ambos enfoques se adaptan correctamente a la estructura de los datos, y deben ser las consideraciones particulares de nuestro estudio las que determinen cuál de estas soluciones debe considerarse.

Recuérdese de nuevo que, el primer método se ajusta a este conjunto debido a la naturaleza particular de la contaminación, no siendo este el caso general. El objetivo de este ejemplo es únicamente mostrar la posible existencia de varias soluciones óptimas factibles.

En los métodos de clustering basados en modelos gaussianos, presentados en el capítulo anterior, entra en juego un tercer parámetro de ajuste, la constante c de la restricción impuesta sobre las matrices de dispersión. La elección conjunta de k y α está a su vez estrechamente relacionada con esta constante. Basta observar de nuevo la gráfica de la izquierda de la figura 3.2 para percatarse que esa elección de k y α no resultaría apropiada para valores pequeños de c, pues el algoritmo no podría detectar el grupo de mayor dispersión.

En general, valores pequeños del parámetro c, pueden provocar la partición de clusters de gran magnitud en clusters más pequeños. Esta idea se ilustra en la figura 2.3 del capítulo anterior, donde al reducir el valor de c se requiere de un aumento en

el número de clusters, evitando así la asignación de parte del grupo más extenso a los clusters más pequeños.

Si bien no existe un método sistemático que garantice una determinación óptima de estos parámetros, se pueden encontrar en la literatura numerosas herramientas para orientar este proceso de elección, facilitando así la discusión. En este capítulo presentamos dos de las más comunes.

3.1.1. Curvas CTL

Una vez fijada la constante de restricción c de acuerdo con los própositos de nuestro estudio, la atención debe centrarse en la elección de los parámetros k y α. Como consecuencia de la fuerte dependencia entre ambos, la idea más sensata consiste en realizar un estudio conjunto de ambos parámetros, monitorizando como varía la función objetivo del modelo según se modifican estos valores. Esta es la idea propuesta en García-Escudero et al. (2011)[15] que introduce las denominadas curvas CTL (Classification Trimmed Likelihood curves) como herramienta para asistir al usuario en la elección de los parámetros k y α.

Definición 6. *Bajo las hipótesis del problema de clustering, sea c una constante fija para la restricción de autovalores* (2.8)*, denotamos* $\mathcal{L}_c^{\Pi}(\alpha, k)$ *al valor máximo de* (2.5)*. Se define entonces la k-curva de verosimilitud recortada o curva ctl como la función:*

$$\alpha \longrightarrow \mathcal{L}_c^{\Pi}(\alpha, k) \quad \textit{para} \quad \alpha \in [0, 1)$$

Estas curvas miden la evolución de la calidad de la clasificación obtenida mediante la metodología TCLUST para un k determinado en función del parámetro α. El interés detrás de esta herramienta consiste en la comparación de distintas k-curvas, identificando así el valor de k a partir del cual no se observa una mejora significativa en la clasificación para un mínimo valor de recorte α. Consideramos la función $\Delta_c^{\Pi}(\alpha, k) := \mathcal{L}_c^{\Pi}(\alpha, k+1) - \mathcal{L}_c^{\Pi}(\alpha, k)$, que cuantifica la "ganancia" conseguida en la función objetivo al aumentar el número de clusters a considerar, pasando de k a $k+1$ grupos para un nivel de recorte α dado.

Nótese que, el valor $\Delta_c^{\Pi}(\alpha, k)$ puede ser igual a 0 cuando tomamos un número de grupos k mayor o igual que el "adecuado", pues la función objetivo no mejora al considerar más grupos. No obstante, esto representa un caso muy particular, y la función objetivo tiende a mejorar ligeramente al considerar más grupos aún cuando los datos originales están generados por k distribuciones, especialmente ante la presencia de ruido.

El criterio propuesto en [15] consiste en seleccionar el número de grupos k más pequeño tal que $\Delta_c^{\Pi}(\alpha, k)$ sea siempre cercano a 0 salvo para valores reducidos de α. Una vez fijado el número de clusters, una elección sensata del nivel de recorte será el valor α_0 más bajo para el que $\Delta_c^{\Pi}(\alpha, k)$ es cercano a 0 para todo $\alpha \geq \alpha_0$.

La complejidad computacional del problema TCLUST impide representar completamente estas curvas, realizándose así su evaluación únicamente para una cuadrícula

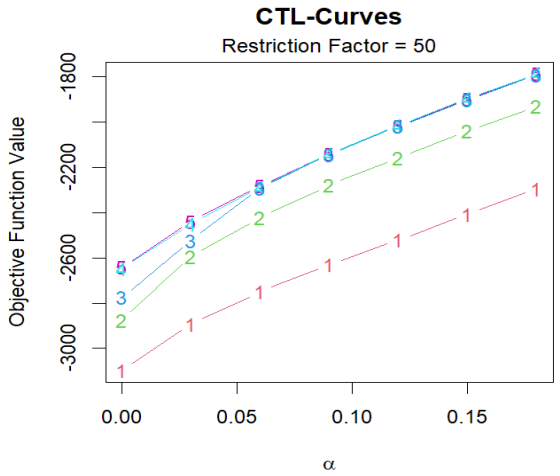


Figura 3.3: Curvas CTL para el conjunto de datos de la figura 4.2 con $c = 50$.

de valores de α. La figura 3.3 representa las curvas ctl $\mathcal{L}^{\Pi}_{50}(\alpha, k)$ para $k = 1, 2, 3, 4, 5$, α en $[0, 0.2]$ y $c = 50$ para el conjunto de datos de la figura 3.2. Se observa que, para $\alpha = 0$, la clasificación no mejora al incrementar el valor k de 4 a 5. De igual manera, una vez descartada al menos una proporción $\alpha_0 = 0.07$ de los datos, no se aprecia una mejora significativa al pasar de 3 a 4 grupos. Por lo tanto, este criterio sugiere tomar $k = 4$ y $\alpha = 0$ o bien $k = 3$ y $\alpha = 0.07$ como elecciones sensatas de los parámetros del modelo TCLUST cuando $c = 50$.

Estas curvas ofrecen al usuario información útil en la elección de los parámetros k y α, sin embargo, consitituyen una herramienta gráfica más que un criterio formal. Para proporcionar mayor objetividad a la decisión tomada por el usuario, sería necesario definir con precisión cuando $\Delta^{\Pi}_{c}(\alpha, k)$ es suficientemente cercano a 0.

Además, conviene recordar que esta representación se realiza para un valor de c preestablecido, el cual debe ser elegido por el usuario en función de otros criterios previos. Como ya se ha señalado, los valores óptimos de k y α están completemante relacionados con la elección de c, y esta dependencia se refleja también en las curvas CTL.

La figura 3.4 muestra la representación de estas curvas para el conjunto de datos de la figura 2.3, que utilizamos en el capítulo anterior como ejemplo de la utilidad del parámetro c en el ajuste del modelo. En la figura de la izquierda, correspondiente a la elección $c = 50$, se observa que, tras descartar una proporción $\alpha_0 = 0.05$ de las observaciones más atípicas, la elección adecuada es tomar $k = 3$ grupos, pues un número mayor de clusters no se traduce en una mejor clasificación. En cambio, para $c = 2$, caso correspondiente a la figura de la derecha, aumentar el parámetro k mejora significativamente la clasificación hasta llegar a $k = 5$, siempre que se recorte una proporción $\alpha_0 = 0.15$ o mayor de las observaciones.

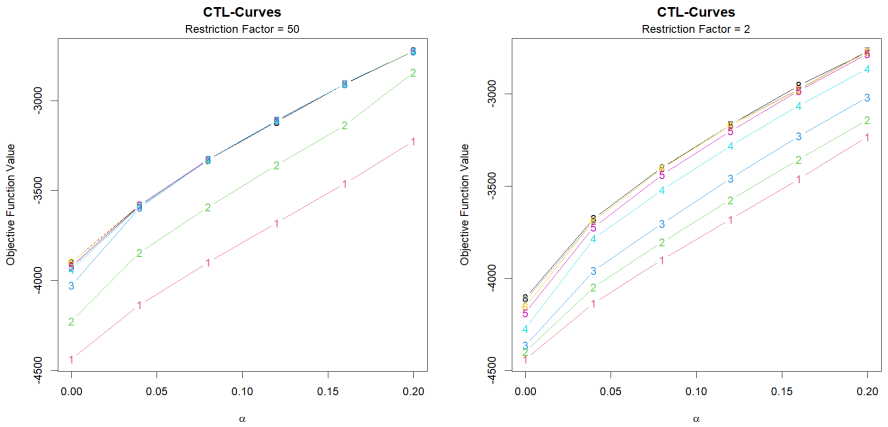


Figura 3.4: Curvas CTL para el conjunto de datos de la figura 4.2 con: (a)$c = 50$ (b)$c = 2$.

3.1.2. TBIC: Trimmed Bayesian Information Criterion

Uno de los métodos más utilizados en la comparación y selección de modelos estadísticos es el criterio BIC (Bayesian Information Criterion). Este criterio penaliza la función de log-verosimilitud en función del número de parámetros del modelo, buscando un equilibrio entre la calidad del ajuste y la complejidad del modelo. De esta manera, se evita el sobreajuste castigando modelos demasiado complejos, favoreciendo así la elección de modelos más sencillos siempre que sea posible.

Una extensión robusta de este criterio aparece en Neykov. et.al (2007)[22], adaptando el método a la comparación de modelos que incluyen recorte. Esta extensión recibe el nombre de TBIC (Trimmed Bayesian Information Criterion), y se basa en la medida de verosimilitud penalizada definida por:

$$\text{TBIC} \coloneqq -2\ log(TL(\tilde{\Theta}_k)) + m\ log(\lfloor n(1-\alpha) \rfloor) \tag{3.1}$$

donde $TL(\tilde{\Theta}_k)$ es la función de verosimilitud maximizada de (2.6) o (2.13), y m es el número de parámetros del modelo.

Nótese que para un modelo en el que se buscan k grupos en un conjunto de puntos con p variables, la función de verosimilitud requiere de $(k-1)$ pesos π_j (el último se obtiene por la restricción $\sum_{j=1}^{k} \pi_j = 1$), k vectores de medias en $\mathbb{R}^p$ y k matrices de dispersión simétricas, lo que hace un total de parámetros:

$$m = (k-1) + k \cdot p + k \cdot \frac{p(p+1)}{2}$$

El criterio TBIC tiene únicamente en cuenta las observaciones no recortadas también en el término de penalización, lo que permite comparar modelos ajustados con distintos niveles de recorte. Como es lógico, el caso $\alpha = 0$, se reduce al criterio BIC ya conocido.

Para ilustrar la efectividad del método, se ha utilizado nuevamente el conjunto de datos de la figura 3.2, ajustando el modelo TCLUST con $c = 50$ para los valores de $k = 1, 2, 3, 4, 5$, y variando en cada caso el nivel de recorte desde el 0 % al 30 % en incrementos del 5 %. Los resultados de los valores TBIC para estos modelos se muestran en el cuadro 3.1. Además, para cada nivel de recorte, se ha señalado en negrita el correspondiente valor mínimo (excepto para el caso $\alpha = 0.05$, donde se han señalado resaltado dos de estos valores por su proximidad). Observando la tabla, el valor mínimo de cada columna corresponde al modelo con $k = 3$ cuando se recorta parte de los datos, lo que confirma que se trata del modelo que mejor se ajusta a este conjunto de datos, tal como se había deducido anteriormente.

	0 %	**0.05 %**	**0.1 %**	**0.15 %**	**0.2 %**	**0.25 %**	**0.3 %**
k=1	6223.12	5611.19	5214.65	4844.05	4479.37	4089.38	3715.68
k=2	5852.68	5021.32	4559.47	4168.58	3819.78	3471.71	3147.53
k=3	5728.58	**4805.40**	**4303.75**	**3904.50**	**3550.37**	**3212.02**	**2899.07**
k=4	5616.14	**4807.63**	4333.61	3941.10	3586.70	3247.87	2932.07
k=5	**5544.37**	4841.75	4367.26	3977.76	3622.86	3284.77	2966.30

Cuadro 3.1: Tabla de valores de TBIC para el conjunto de datos de la figura 3.2.

Al igual que las curvas CTL, este estudio se realiza una vez se ha fijado el parámetro c de la restricción correspondiente en los métodos de TCLUST y modelos de mezcla con clasificación blanda. Como siempre este parámetro debe ser escogido por el usuario de manera previa, en función de las características del problema.

3.2. Evaluación de la clasificación

Una vez obtenida una solución de un análisis cluster, cabe preguntarse por la calidad de la representación lograda. No todos los conjuntos de datos permiten obtener clasificaciones igualmente coherentes, por lo que es necesario recurrir a técnicas posteriores que permitan valorar la fiabilidad de las agrupaciones.

Una estrategia efectiva es evaluar la calidad de las asignaciones realizadas de las observaciones a los distintos grupos. No todas las asignaciones son igualmente fiables. Los datos más próximos al centro de un cluster obtenido serán más probables de pertenecer al mismo, sin embargo, aquellos situados en los bordes son más susceptibles de ser mal clasificados, especialmente en zonas con clusters solapados.

Supongamos que se obtiene una solución óptima $\hat{R} = \{\hat{R}_0, \ldots, \hat{R}_k\}$, $\hat{\theta} = \{\hat{\theta}_0, \ldots, \hat{\theta}_k\}$ y $\hat{\pi} = \{\hat{\pi}_0, \ldots, \hat{\pi}_k\}$ mediante el algoritmo TCLUST para determinados valores de k, α y c. Las funciones discriminantes $D_j(x, \theta) = \pi_j \ f(x; \mu_j, \Sigma_j)$, descritas en el capítulo anterior, permiten establecer una medida de credibilidad para la asignación de cada observación, tal y como se describe en Van Aelst et al. [25].

Sea x_i una observación. Si denotamos por $D_{(1)}(x_i, \theta) \leq \ldots \leq D_{(k)}(x_i, \theta)$ la ordenación de los valores de las funciones discriminantes, el factor de Bayes para una observación no recortada x_i se define como

$$\mathrm{BF}(x_i) \coloneqq log\left(\frac{D_{(k-1)}(x_i, \theta)}{D_{(k)}(x_i, \theta)}\right)$$

Es claro que, cuanto menor es el factor de Bayes, más fiable resulta la asignación de x_i a su cluster correspondiente. Por tanto, estos valores permiten cuantificar la fiabilidad de la clasificación obtenida. La presencia de clusters con numerosas observaciones que presenten valores elevados (próximos a cero) del factor de Bayes, indican una posible elección errónea de los parámetros del modelo.

Esta noción puede extenderse también a las observaciones recortadas, como se expone en [15]. Se considera para cada observación de la muestra el valor $d_i = D_{(k)}(x_i; \theta)$, y se realiza la ordenación $d_{(1)} \leq \ldots \leq d_{(n)}$. El algoritmo TCLUST descarta las observaciones con los valores d_i más bajos, de modo que $\hat{R}_0 = \{i \in \{1, \ldots, n\} : d_{(i)} \leq d_{([n\alpha])}\}$. Se define entonces el factor de Bayes para una observación recortada como

$$\mathrm{BF}(x_i) \coloneqq log\left(\frac{D_{(k)}(x_i, \theta)}{d_{([n\alpha]+1)}}\right),$$

que cuantifica la credibilidad de considerar x_i como dato atípico. Valores altos de $\mathrm{BF}(x_i)$ indican decisiones poco claras en el recorte de la observación x_i. Así, al igual que ocurría para los clusters, un número elevado de observaciones recortadas con valores grandes del $\mathrm{BF}(x_i)$ sugieren un exceso en el nivel de recorte, y por tanto, una posible elección incorrecta de los parámetros.

Los valores $\mathrm{BF}(x_i)$ permiten construir resúmenes gráficos adicionales sobre la calidad de la clasificación, como los denominados "silhouette plots" o gráficos de silueta, que ayudan a identificar grupos con numerosas observaciones con factores de Bayes elevados. Esto indica poca confianza en la formación de dicho grupo, lo que puede deberse a la presencia de clusters solapados o a una elección errónea de los parámetros del modelo.

Cuando se trabaja con dos variables, es posible representar directamente las observaciones con valores altos de $\mathrm{BF}(x_i)$, las cuales catalogamos como asignaciones o recortes dudosos.

Como es de esperar, estas observaciones dudosas suelen encontrarse en la frontera de los clusters, como se observa en la figura 3.5(b), donde un único cluster ha sido separado artificialmente en dos al considerar un número de clusters mayor del necesario. Este tipo de situaciones, originadas por una mala elección de parámetros, genera numerosas observaciones con valores elevados de $\mathrm{BF}(x_i)$, como se aprecia en la figura 3.5(c), correspondiente al gráfico de siluetas asociado a esa clasificación. En dicho gráfico, el valor $\log(1/8)$ aparece representado mediante una línea discontínua, siendo un umbral comunmente establecido como límite a partir del cual se considera una asignación como dudosa.

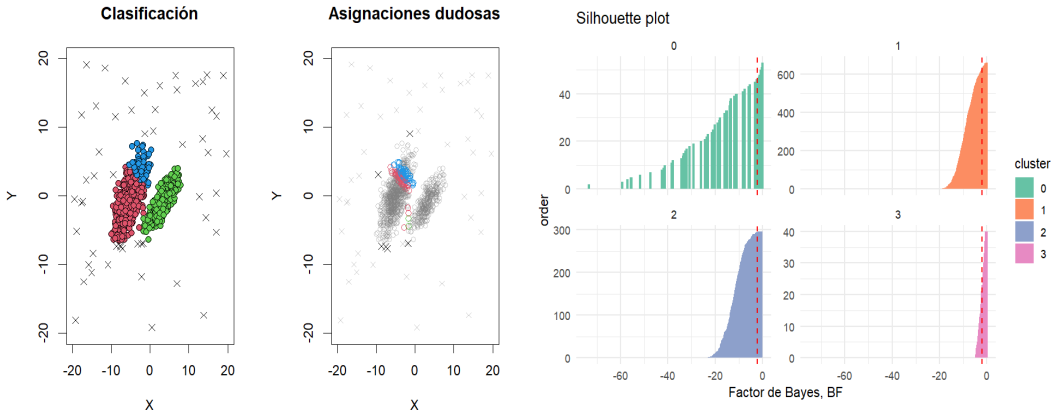


Figura 3.5: (a) Solución por TCLUST con $k = 3$, $\alpha = 0.05$ y $c = 2$. (b) Observaciones con factor de Bayes mayor que $log(1/8)$. (c) Silhoutte plot.

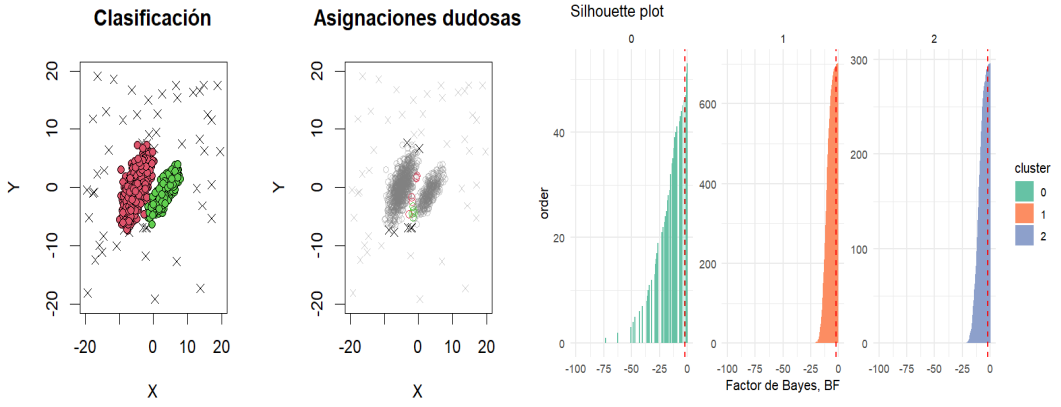


Figura 3.6: (a) Solución por TCLUST con $k = 2$, $\alpha = 0.05$ y $c = 2$. (b) Observaciones con factor de Bayes mayor que $log(1/8)$. (c) Silhoutte plot.

La figura 3.6 muestra las representaciones gráficas de la clasificación para k=2. En este caso, el número de asignaciones dudosas es mucho menor que para k=3, lo que sugiere que esta partición es más adecuada para el conjunto de datos. Esto también se refleja en el gráfico de siluetas (figura 3.6(c)), donde a penas unas pocas observaciones superan el umbral de log(1/8). Por tanto, atendiendo a ambos gráficos, la clasificación en dos grupos parece más razonable que la de tres.

Nótese que en ambas clasificaciones aparecen algunas observaciones recortadas con valores altos del factor de Bayes cerca de los bordes de los grupos obtenidos. Esto sugiere que dichas observaciones podrían haber sido recortadas de forma errónea y que, en realidad, podrían pertenecer a alguno de los clusters contiguos. Esta situación se vuelve aún más evidente en la clasificación mostrada en la figura 3.7, donde se ha utilizado un nivel de recorte $\alpha = 0.15$ considerablemente mayor al nivel de contaminación real. Como consecuencia, se observa la presencia de numerosas observaciones dudosamente recortadas, lo que indica este exceso en el recorte.

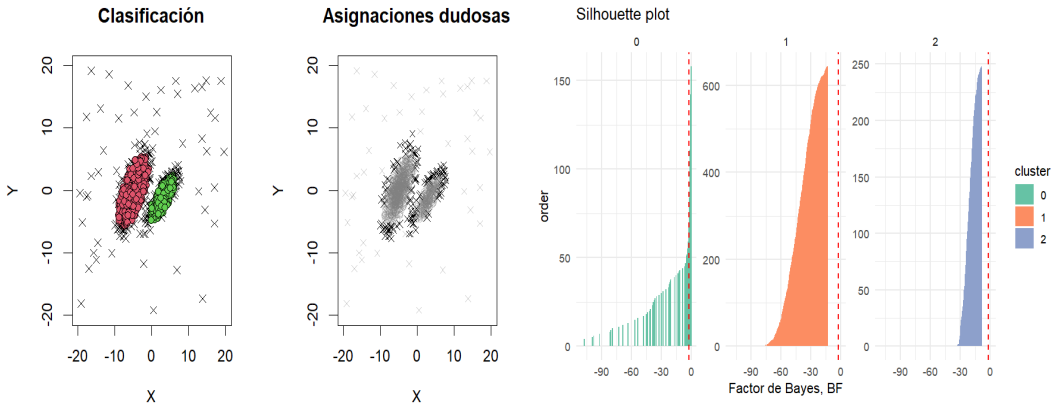


Figura 3.7: (a) Solución por TCLUST con $k = 3$, $\alpha = 0.15$ y $c = 2$. (b) Observaciones con factor de Bayes mayor que $log(1/8)$. (c) Silhoutte plot.

No obstante, tomar un valor de α ligeramente superior al necesario no suele afectar gravemente la clasificación, siempre que no existan clusters muy pequeños que puedan ser "absorbidos" por el recorte. Tal como se observa en la figura 3.7(a), aunque algunas observaciones de la frontera de los clusters han sido erróneamente recortadas, la estructura propia de los datos ha sido correctamente identificada.

Además, al conocer que observaciones se han recortado de forma dudosa, se podría reasignarlas al cluster más cercano, mejorando así la clasificación. Esta es la idea detrás de los métodos de reponderación o re-weighting, como el descrito en [6], que permiten reducir progresivamente el nivel de recorte hasta el nivel óptimo, recuperando así las observaciones erróneamente descartadas. Un procedimiento de este tipo puede complementar la serie de estrategias presentadas en este capítulo para la elección de los parámetros y valoración del modelo.

Se recomienda por tanto ejecutar el algoritmo TCLUST para los valores determinados mediante alguna de las herramientas presentadas anteriormente, y, a continuación, utilizar las representaciones descritas al comienzo de esta sección como método de evaluación de la clasificación obtenida. Por último, según las conclusiones obtenidas tras esta evaluación, se puede llevar a cabo una reponderación o una nueva clasificación con distintos parámetros si es necesario.

Capítulo 4

Extensiones y variantes

A lo largo de los capítulos anteriores se han presentado y discutido métodos de clustering robusto basados en el recorte, así como analizado su evolución adaptápdose a distintos contextos y estructuras de datos. Estos modelos parten de ciertos supuestos generales como la distribución normal de los datos, que establece un marco de trabajo flexible, adaptándose correctamente a casos muy diversos. No obstante, en la práctica es frecuente encontrar conjuntos de datos que distan considerablemente de estas suposiciones, requiriendo otro tipo de enfoques para obtener resultados adecuados.

Otro inconveniente de los métodos de clustering usuales es el tratamiento de conjuntos de datos de alta dimensionalidad, especialmente ante la presencia de contaminación en las medidas realizadas sobre las varibles. La alta dimensionalidad acentúa el problema de ruido, provocando un gran aumento en el número de datos atípicos, provocando una gran pérdida de información al aplicar los métodos de recorte conocidos hasta el momento.

En este capítulo se presenta un breve resumen de algunas extensiones de los métodos clásicos de clustering robusto, adaptadas a situaciones específicas frecuentes en la práctica y para las cuáles los métodos vistos hasta ahora no ofrecen soluciones adecuadas. Así, el objetivo del presente capítulo no es ofrecer una descripción detallada de estos modelos, sino exponer la motivación detrás del desarrollo de estos procedimientos y dar a conocer distintas ramas del análisis cluster robusto, las cuales se encuentran en plena expansión.

4.1. Clustering de subespacios afines

El objetivo principal del análisis cluster es, como se ha señalado en repetidas ocasiones, obtener agrupaciones en un conjunto de datos que permitan identificar relaciones entre las observaciones. Estas relaciones suelen estar motivadas por la similitud entre los elementos, la cual se cuantifica normalmente mediante alguna distancia. No obstante, existen otro tipo de vínculos naturales entre las observaciones que provocan la agrupación de las mismas en estructuras más complejas. Entre estas, destacan especialmente las agrupaciones en torno a subespacios lineales, presentes en numerosas aplicaciones prácticas.

El clustering en torno a subespacios lineales constituye una de las variantes más estudiadas, debido a su aplicación en campos como el reconocimiento de patrones o la tomografía. A continuación, presentamos un ejemplo de aplicación de este tipo de clustering.

Supongamos que se ha plantado un determinado vegetal en 75 parcelas de distintas regiones. En cada una de estas parcelas se ha utilizado una cantidad distinta de un mismo fertilizante, con el objetivo de analizar su efecto en la producción obtenida. No obstante, se sabe que este fertilizante tiene un impacto distinto en función del tipo de suelo cultivado. Entre las parcelas plantadas, habrá suelos deficientes en nutrientes, que responderán bien al fertilizante; suelos saturados de nutrientes, donde el uso de fertilizante puede ser perjudicial, y otros donde la producción no se vea prácticamente alterada. En cualquier caso, se espera que para un determinado tipo de suelo, la relación entre la producción y la cantidad de fertilizante utilizado siga una relación lineal.

Ante estas suposiciones, se pretende realizar una clasificación de las parcelas en función de esta relación, con el objetivo de predecir el tipo de suelo de cada una, y establecer una estrategia óptima en el uso de fertilizante en próximos periodos de cultivo.

El conjunto de parcelas descrito puede visualizarse en la figura 4.1(izda.), donde se representa una gráfica que recoge la cantidad de fertilizante utilizado, medido en kilogramos por hertárea, y la producción en cada parcela, medida en toneladas por hectárea. Es necesario mencionar que estos datos no pertenecen a ningún conjunto conocido, sino que han sido generados artificialmente con el propósito de representar una situación realista del problema descrito.

Los métodos de clustering robusto descritos hasta el momento no ofrecen una solución adecuada para este problema, pues la formación de los clusters se basa en la similaridad de las observaciones medida mediante una distancia, y no en la relación lineal entre las variables. La figura 4.1(dcha) muestra el resultado obtenido tras aplicar TCLUST con parámetros $k = 3$, $\alpha = 0.2$, $c = 500$ para el conjunto de cultivos descrito. La clasificación obtenida no parece recoger ningún tipo de relación entre las observaciones, y los clusters obtenidos no tienen ninguna interpretación práctica.

Ante esta incapacidad de los métodos clásicos de detectar relaciones lineales de los datos surge el clustering sobre espacios lineales. Este enfoque busca encontrar k espacios lineales que recojan lo mejor posible la estructura de los datos, medida en función de las proyecciones ortogonales de cada observación al subespacio más cercano.

No obstante, la presencia de datos contaminados tiene un gran impacto negativo en la formación de estos subespacios, desviando completamente la estructura determinada por este tipo de procedimientos. Esta falta de robustez pone de manifiesto la necesidad del desarrollo de métodos robustos para este tipo de clustering.

En García-Escudero et al.(2009)[13] se propone un primer método robusto de clustering sobre subespacios lineales, basado en la minimización de una suma recortada de residuos ortogonales. Este modelo, conocido como RLG (Robust Linear Grouping), permite no solo realizar una clasificación robusta en base a las relaciones lineales de los datos, sino también obtener una reducción de la dimensionalidad, extendiendo el Análisis de Componentes Principales(PCA) al contexto del análisis cluster.

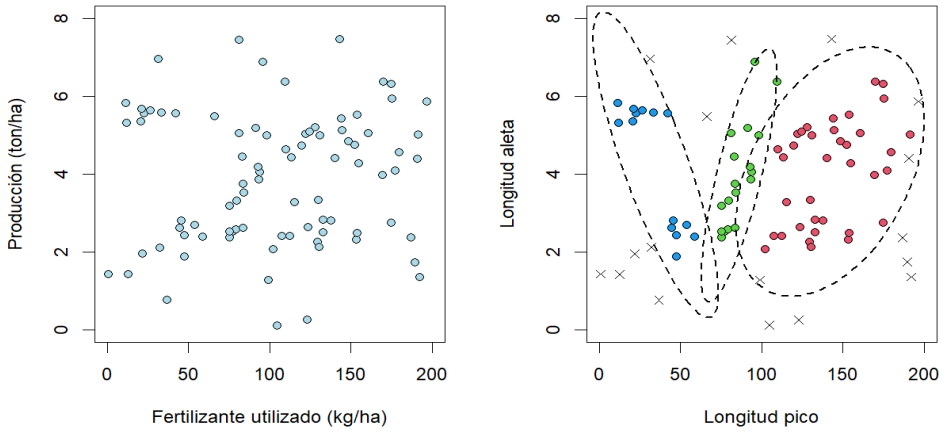


Figura 4.1: (a)Conjunto de cultivos (b)Resultado de aplicar TCLUST con $k = 3$, $\alpha = 0.2$ y $c = 500$ sobre el conjunto de cultivos.

El algoritmo RLG (Robust Linear Grouping), propuesto en [13], se basa en la búsqueda de k subespacios afines $h_1, \ldots, h_k$ de dimensiones intrínsecas $q_1, \ldots, q_k$ ($q_j < p$), junto con una partición $\{R_0, R_1, \ldots, R_k\}$ de $\{1, \ldots, n\}$, donde R_0 recoge una proporción $[n\alpha]$ de los índices, minimizando la expresión

$$\sum_{j=1}^{k} \sum_{i \in R_j} \left\| x_i - Pr_{h_j}(x_i) \right\|^2, \tag{4.1}$$

donde $Pr_h(x)$ denota la proyección ortogonal de x sobre el subespacio h. Cabe descacar que, en el caso particular $K = 1$ y $\alpha = 0$, este modelo se reduce al ya conocido PCA.

La figura 4.2 muestra el resultado de aplicar esta metodología al conjunto de datos de la figura 4.1, para $k = 3$ y dimensiones $q_1 = q_2 = q_3 = 1$. La gráfica de la izquierda muestra la clasificación obtenida para un nivel de recorte $\alpha = 0$, es decir, sin descartar ninguna observación. La presencia de datos atípicos dificulta la identificación de las relaciones lineales subyacentes entre las observaciones, lo que justifica la necesidad de robustificación del método. La gráfica de la derecha representa la clasificación obtenida para un nivel de recorte $\alpha = 0.2$. Este método detecta las relaciones lineales de las variables, identificando 3 grupos distinos con diferentes vínculos lineales entre estas medidas.

Volviendo al ejemplo de los cultivos, la clasificación obtenida permite identificar los tipos de suelo y las relaciones entre fertilizante y producción para cada uno de los grupos, pudiendo predecir la cantidad óptima de fertilizante a utilizar en cada parcela. Para las parcelas correspondientes al cluster de mayor pendiente se utilizará mucho fertilizante, maximizando así la producción. Las parcelas del grupo de pendiente negativa no requerirán de fertilizante, pues parece que disminuye la cantidad de vegetal cosechado. En cuanto al grupo de parcelas con pendiente baja en la relación producción-fertilizante, habría que estudiar los factores las ganancias o pérdidas al usar más o menos fertilizante en función de su precio y el beneficio de la venta de la producción a

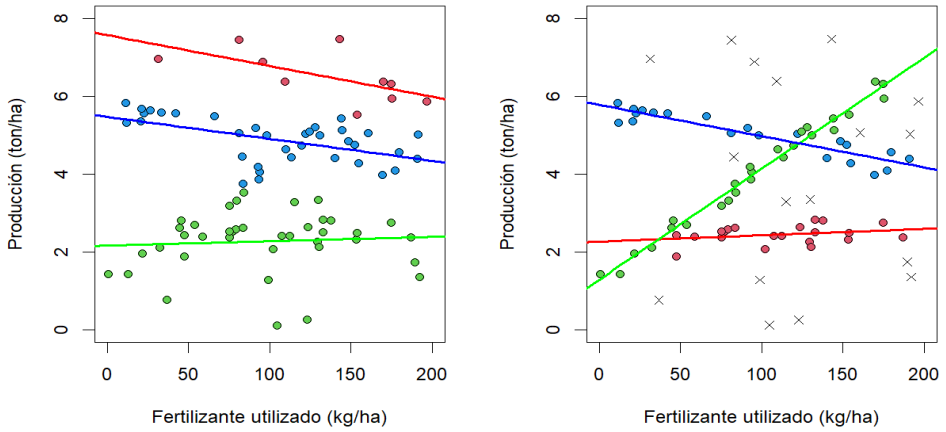


Figura 4.2: Resultados de aplicar RLG sobre el conjunto de datos de la figura 4.1 con (a) $\alpha = 0$ (b) $\alpha = 0.2$.

mayores. Por último, las observaciones descartadas, las cuales corresponden a los datos más atípicos, pueden asignarse al grupo más cercano o analizarse individualmente.

La elección ortogonal de los residuos en (4.1) implica que no existe una variable privilegiada usada como variable respuesta. Sin embargo, en numerosas aplicaciones resulta de interés explicar una variable concreta en función del resto. En estos casos, es más apropiado emplear los residuos de la regresión lineal clásica.

Este enfoque, que denominaremos "regresión lineal por clusters", permite además aprovechar propiedades fundamentales de la regresión lineal, como el análisis del comportamiento de cada variable explicativa, la capacidad de predicción de la variable respuesta o la validación del modelo mediante el estudio de los residuos.

Esta es la idea presentada en García-Escudero et al.(2010)[14], que extiende la metodología del TCLUST al contexto de la regresión lineal por clusters robusta. La flexibilidad de este enfoque permite trabajar con distintos pesos asignados a los grupos y considerar distintas dispersiones para los errores en cada cluster. Para evitar la aparición de soluciones degeneradas, es necesario imponer alguna restricción sobre la dispersión de los errores, de manera análoga a como ocurría en la metodología TCLUST. De este modo, se asegura la existencia y consistencia de los estimadores, permitiendo una aplicación fiable del método.

Supongamos que el número total de grupos k y el nivel de recorte α son fijados previamente. Sea un conjunto de datos $\{X_i\}_{i=1}^n$ donde cada observación $X_i = (y_i, x_i')'$ está compuesta por una variable respuesta $y_i \in \mathbb{R}$ y un vector de variables explicativas $x_i \in \mathbb{R}^p$. Si la observación X_i es asignada al cluster j, entonces se asume que el valor de la variable respuesta y_i está generada por una distribución $N(x_i'\beta_j\ , \sigma_j^2)$, para unos parámetros $\beta_j \in \mathbb{R}^p$ y $\sigma^2 > 0$ desconocidos. Además, se consideran pesos π_j asociados a cada cluster, sujetos a la restricción $\sum_{j=1}^k \pi_j = 1$.

De forma análoga a la construcción del modelo TCLUST, una modificación del modelo de outliers-espurios propuesto en [10] y [11] permite formular el problema de regresión lineal por clusters a través de la siguiente función de verosimilitud:

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j}\pi_j\ f(y_i;x_i'\beta_j,\sigma_j)\right]\left[\prod_{i\notin R}g_i(y_i;x_i')\right] \tag{4.2}$$

donde

$$f(y;x'\beta,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-(y-x'\beta)^2}{(2\sigma^2)}),$$

y $\{R_0,R_1\ldots,R_k\}$ es una partición de los índices $\{1,\ldots,n\}$, con $\#R_0=[n\alpha]$.

Las funciones $g_i(\cdot,x_i')$ representan densidades desconocidas (condicionadas al valor x_i') que modelan la generación de los datos contaminantes. Tal y como se discutió en el capítulo dedicado al TCLUST, bajo ciertas suposiciones sensatas se puede evitar su aportación en la práctica.

Estas suposiciones permiten reducir el problema de regresión lineal por clusters robusta a la búsqueda de una partición $R_0,R_1\ldots,R_k$ de los índices $\{1,\ldots,n\}$, con $\#R_0=[n\alpha]$, y de valores para los parámetros π_j, β_j y σ_j que maximicen la expresión

$$\sum_{j=1}^{k}\sum_{i\in R_j}log\left(\pi_j\ f(y_i;x_i'\beta_j,\sigma_j)\right) \tag{4.3}$$

De nuevo, tal y como ocurre en la metodología TCLUST, es necesario imponer ciertas restricciones sobre los parámetros de dispersión del modelo para evitar la aparición de soluciones degeneradas. Concretamente, se establece una restricción de similaridad de las dispersiones de los clusters, de la forma

$$\frac{M_n}{m_n}\leq c\quad\text{para}\quad M_n=\max_{j=1,\ldots,k}\sigma_j^2\quad\text{y}\quad m_n=\min_{j=1,\ldots,k}\sigma_j^2$$

Esta restricción garantiza que el problema de maximización de (4.3) esté bien definido, y evita la aparición de grupos con varianza nula, los cuales conducen a soluciones no interpretables.

Una modificación adecuada del algoritmo TCLUST, adaptada al contexto de la regresión lineal, permite encontrar una aproximación de la solución óptima del problema (4.3). Esta modificación introduce un segundo recorte de los datos, ya que, si bien es sabido que este tipo de algoritmos son efectivos ante la presencia de datos atípicos de la varible y, también es conocida la falta de robustez de la regresión lineal por mínimos cuadrados frente a la aparición de outliers en las variables explicativas x. Este procedimiento de doble recorte, proporciona una metodología robusta para la regresión lineal por clusters.

4.2. Clustering con recorte por celdas

En el análisis cluster con recorte, el enfoque tradicional consiste en descartar una proporción de las observaciones $x_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$ más atípicas, lo que resulta razonable en contextos de baja dimensión. Sin embargo, este procedimiento puede resultar extremo para valores elevados de p, ya que implica el descarte completo de cualquier de fila x_i que contenga al menos una celda x_{ij} irregular. En situaciones de alta dimensionalidad, incluso un pequeño porcentaje de celdas contaminadas puede llevar al descarte de una proporción considerable de observaciones, lo que conlleva una pérdida de información inmensa, y, en consecuencia, una disminución de la efectividad del método.

Por ejemplo, consideremos el conjunto de datos del cuadro 5.1, que recoge las respuestas de 10 individuos sobre la frecuencia o duración de ciertos hábitos cotidianos, y supongamos que queremos realizar una clasificación de los individuos según sus perfiles de comportamiento en base a estas medidas. Dado que no se disponía de una conjunto de datos real adecuado, se ha optado por construir un conjunto de datos ficticio, diseñado para representar una situación realista dentro del contexto de la contaminación por celdas. Las celdas marcadas en rojo representan los datos atípicos, que se salen de la norma. En esta tabla puede observarse que la presencia de "solamente" 7 celdas atípicas (un 10 % de los datos totales) podría forzar el descarte de hasta 5 individuos, es decir, la mitad de la muestra, si se realiza un recorte clásico. Esto supone una inmensa pérdida de información, lo que reduce drásticamente la efectividad del estudio, comprometiendo su utilidad.

Cuadro 4.1: Frecuencia diaria(d)/semanal(s) o duración de ciertas actividades.

Persona	Agua(d)	Comidas(d)	Móvil(d)	Ejercicio(s)	Horas de Sueño(d)	Horas de lectura(s)	Horas de TV(d)
1	8	3	25	3	7	1	3
2	**23**	4	20	0	8	0	7
3	9	3	30	**28**	6	5	15
4	8	4	22	5	7	2	5
5	9	3	25	2	8	1	**35**
6	8	3	**250**	1	7	0	0
7	11	5	15	7	8	6	9
8	5	5	45	0	**2**	**22**	12
9	2	2	35	2	6	0	0
10	7	3	**0**	7	7	3	2

Aunque en este ejemplo estamos trabajando un conjunto de datos de dimensión moderada ($p = 7$), la carencia expuesta aumenta según p crece. La extensa cantidad de datos generada por las nuevas tecnologías ha incitado al desarrollo de metodologías robustas capaces de solucionar esta pérdida de información, manteniendo así la efectividad del método en contextos de alta dimensionalidad.

Uno de los recursos más utilizados para abordar este problema es el denominado recorte por celdas o "cellwise trimming". Esta estrategia consiste en descartar únicamente algunas variables de ciertas observaciones, manteniendo el resto de la información, y minimizando así la pérdida de datos regulares.

En esta sección se presenta una de las primeras propuestas de clustering basadas en el recorte por celdas, intoducida en [8], bajo el nombre de snipped k-means. Este método consitituye una extensión robusta de las k-medias, diseñada específicamente para el descarte de celdas contaminadas.

Supongamos que se tiene una muestra de n observaciones $Y_1, \ldots, Y_n \in \mathbb{R}^p$, agrupadas en k clusters $J_1, \ldots, J_k$, donde $J = \bigcup_{j=1}^{k} J_j \subset \{1, \ldots, n\}$ y $J_t \cap J_s = \emptyset$ para $t \neq s$. Además, se considera contaminación por componentes, es decir, cada entrada de una observación puede estar contaminada de manera independiente con probabilidad ϵ.

Así, se asume que $np\epsilon$ entradas de la matriz de datos Y están contaminadas según el modelo que se detallará posteriormente. También se supone que $n(1-\epsilon) \leq |J| \leq n$, donde $|J|$ es el número de observaciones no completamente contaminadas. Nótese que si $|J| = \lfloor n(1-\epsilon) \rfloor$, toda la contaminación está formada por datos con todas las variables afectadas, mientras que si $|J| = n$, no hay ninguna observación completamente contaminada.

Se asume un modelo de contaminación a nivel de celda definido de la siguiente manera: sean b_{ij}, para $i = 1, \ldots, n$, variables indicadoras con $b_{ij} \in \{0, 1\}$ y $\sum_{ij} b_{ij}$ igual al entero más cercano a $np(1-\epsilon)$. El valor de b_{ij} indica si la entrada (i, j) de la matriz Y es regular ($b_{ij} = 1$) o está contaminada ($b_{ij} = 0$). Bajo este esquema se modelan las observaciones como:

$$Y_i = b_i X_i + (1 - b_i) Z_i \quad \text{para} \quad 1, \ldots, n \tag{4.4}$$

donde $Z_i \sim g_i(\cdot)$, siendo g una función de densidad en $\mathbb{R}^p$. Así, para cada componente j de la observación Y_i, si $b_{ij} = 1$, la entrada mantiene su valor original, mientras que, si $b_{ij} = 0$, será sustituída por un outlier. Además, se asume que cuando $i \in J_j$, $X_i \sim N(\mu_j, \sigma^2 Id)$, es decir, las observaciones regulares de cada grupos siguen una distribución normal. Por otro lado, las observaciones para las cuales $b_{ij} = 0$ para todo $j = 1, \ldots, p$ están fuera de J, y reciben el nombre de outliers estructurales. Nótese que, el modelo presentado en (4.4) sigue una contaminación heterogénea, en el sentido de que no todas las variables están igualmente contaminadas.

Bajo las consideraciones realizadas en el modelo de contaminación, cabe adoptar un enfoque de clasificación dura, es decir, cada observación es asignada a un único cluster, y se impone además una restricción de esfericidad de los grupos, como en el método de k-medias.

Partiendo de nuevo de las ideas propuestas en Gallegos y Ritter (2005)[11], se asume que al estimar μ, es posible ignorar las observaciones irregulares. Esta afirmación se fundamenta en el supuesto formal de que, para cualquier partición óptima que recorte una proporción ϵ de las celdas, las observaciones irregulares podrían identificarse al maximizar la verosimilitud asociada a las funciones $g_i(\cdot)$.

Bajo estos supuestos, la estimación de los centroides mediante el método de "snipped k-means" se obtiene resolviendo el problema:

$$\min_{V \in \mathcal{V}} \inf_{\{\mu_1, \ldots, \mu_k\} \in \mathbb{R}^p} \sum_{i=1}^{n} \min_{1 \leq t \leq k} \sum_{j=1}^{p} v_{ij} (Y_{ij} - \mu_{sj})^2 \tag{4.5}$$

donde $\mathcal{V}$ denota el subconjunto de $\mathcal{M}_{\{0,1\}}(n,p)$ (el conjunto de matrices binarias de dimensión $n \times p$) con exactamente $\lceil np\epsilon \rceil$ entradas iguales a 0.

Un cero en una entrada de V indica que ese valor ha sido descartado, por lo que una fila completa de ceros corresponde a una observación Y_i totalmente recortada. Conviene destacar que, aunque este método descarte el mismo número de entradas que el método de k-medias recortadas, en este último las entradas descartadas están todas alineadas en filas, mientras que el recorte por celdas distribuye los descartes a lo largo de toda la matriz, conservando la información relevante de cada observación.

El conjunto de todas la matrices de recorte por celdas $\mathcal{V}$ tiene un cardinal extremadamente elevado, lo que imposibilita tratar computacionalmente el problema (4.5), por lo que se debe recurrir a la aplicación de logaritmo que aproximen su solución.

El algoritmo propuesto en [8] se basa en la estrategia de aceptación-rechazo, actualizando de manera iterativa una solución inicial, identificada por $V(0) \in \mathcal{V}$. Este enfoque permite encontrar un equilibrio entre la exploración del espacio de soluciónes y la convergencia a la solución óptima.

El recorte por celdas supone así un remedio eficaz contra la pérdida de información provocada por el recorte usual, consolidándose como una herramienta cada vez más utilizada en el clustering sobre conjuntos de datos de alta dimensión.

En los últimos años, se han desarrollado métodos más sofisticados de recorte por celdas, el cual continúa siendo un campo de investigación activo, en búsqueda de nuevas estrategias que mejoren las prestaciones de los algoritmos actuales.

4.3. Co-clustering recortado

Los métodos de clustering clásicos tienen como objetivo la búsqueda de agrupaciones de n individuos (filas) en función de sus similitudes respecto a p características (columnas), identificando así posibles relaciones y patrones dentro de los datos. Sin embargo, en numerosas aplicaciones resulta de interés realizar una doble clasificación, es decir, agrupar estos individuos tanto por filas como por columnas.

Por ejemplo, si se dispone de una matriz de datos X en la que las filas representan clientes y las columnas una serie de productos, esta clasificación conjunta de filas y columnas permite no solo agrupar los clientes con perfiles de compra similares, sino descubrir la relación de estos grupos con respecto a clases específicas de productos. Esta técnica de doble agrupación recibe el nombre de co-clustering, y presenta aplicaciones en numerosos campos como la biología, la psicología o la sociología, donde es frecuente analizar tanto las relaciones de los individuos como las de sus características.

Un posible enfoque para obtener agrupaciones por filas y por columnas hubiera sido realizar estas clasificaciones por separado, sin embargo, esta estrategia ignora las relaciones presentes entre ambos ejes de la matriz de datos. El co-clustering no busca

solamente clasificar ambos ejes de forma simultánea, sino también ofrecer una representación comprimida de las estructura de los datos, lo que lo convierte en una herramienta especialemente útil en el estudio de datos masivos.

Esta técnica suele por tanto emplearse sobre matrices de datos de grandes dimensiones, permitiendo reducir su estructura a un tamaño mucho más manejable mediante la síntesis de las características del conjunto de datos en distintos grupos. Comenzamos exponiendo un ejemplo de co-clustering sin recorte para ilustrar la utilidad de este enfoque.

Consideremos el conjunto de datos "USArrests" de la librería de R, que contiene el número de arrestos por cada 100.000 habitantes por asalto, asesinato y violación en cada uno de los 50 estados de EEUU en 1973, además del porcentaje de población urbana en cada uno de los estados. Se quiere realizar una clasificación tanto de los estados como de las variables mencionadas (los 3 tipos de delitos y el porcentaje de población urbana) en función de estos valores, de forma que se puedan establecer relaciones entre los distintos grupos de estados y los conjuntos de variables. Un enfoque de clustering usual por filas y por columnas respectivamente ofrece una clasificación individual adecuada para cada una de estas medidas, pero no tiene en cuenta la relación subyacente entre ellas.

Bajo esta motivación por obtener una representación de la estructura de correspondencias entre ambos ejes nace el enfoque del co-clustering, que realiza una clasificación simultánea de individuos y variables. Las figuras 4.3 y 4.4 muestran distintas representaciones de la clasificación obtenida por la función 'cocluster', del paquete 'blockcluster'[23] de R, para la identificación de $k = 2 \times 3$ grupos sobre el cojunto 'UsArrests' tras realizar una escala de los datos.

Esta función implementa un clustering doble basado en modelos de bloques latentes (LBMs), introducidos en Govaert and Nadif(2003)[19]. Este tipo de modelos, que describiremos más adelante, extienden los modelos de mezcla finita al caso bidireccional, suponiendo un marco probabilístico de trabajo idóneo para esta modalidad de clustering.

La figura 4.3 muestra el mapa de calor del conjunto de datos 'USArrests', y la correspondiente reordenación y clasificación en $k = 2 \times 3$ grupos. Esta representación por intensidad de color facilita la interpretación de la clasificación por correspondencias obtenida, la cual se expone en el cuadro 4.2. El conjunto de variables se ha clasificado en un grupo de tres (asaltos, asesinatos y violaciones) frente a un grupo de una única variable: el porcentaje de población urbana. Esta separación sugiere que la proporción de habitantes en suelo urbano para cada estado no guarda una relación demasiado estrecha con la tasa de delitos cometidos. En cambio, si hay una fuerte relación entre los distintos tipos de delitos, es decir, en los estados que presentan una conducta más violenta, tienden a cometerse todo tipo de delitos.

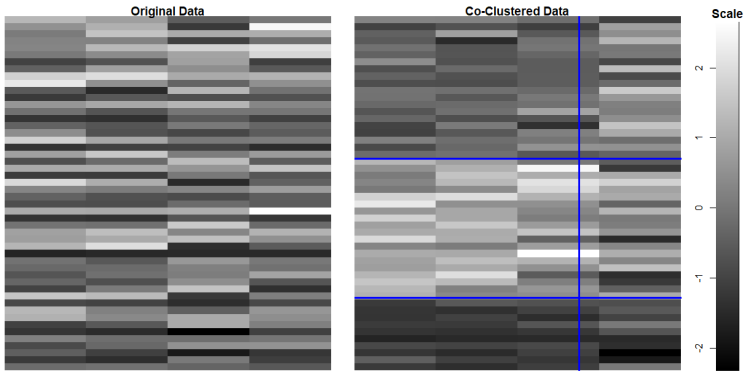


Figura 4.3: Representación por mapa de calor de la clasificación obtenida tras realizar co-clustering sobre el conjunto de datos 'USArrests' de R para k=2x3 grupos.

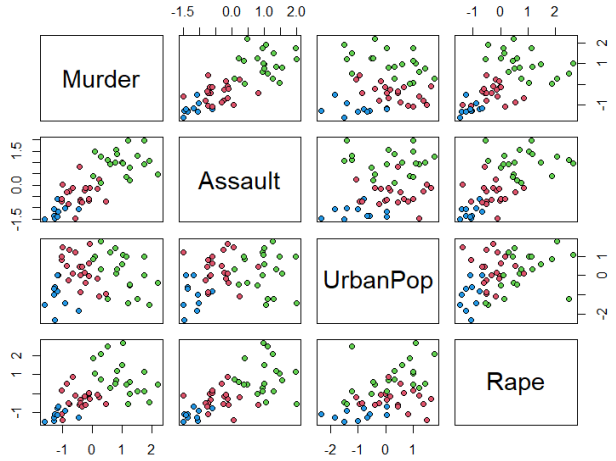


Figura 4.4: Gráfica de diagramas de dispersión de la clasificación obtenida tras realizar co-clustering sobre el conjunto de datos 'USArrests' de R para k=2x3 grupos.

En cuanto a la clasificación de los estados, el mapa de calor permite interpretar la clasificación realizada. El primero de los grupos está constituído por estados con tasas medias de criminalidad. El segundo de ellos está formado por los estados de mayor criminalidad, es decir, con altas tasas de delitos violentos. Por último, el tercero de los grupos lo conforman estados con valores bajos de criminalidad y de población urbana.

Un análisis conjunto de los grupos determinados permite deducir que, si bien la variable de población urbana es la que menos relación guarda con el resto, sí existe una correspondencia entre tasas bajas de criminalidad y estados con gran porcentaje de población rural. Estas interpretaciones se deducen también de la figura 4.4, donde se ha representado la dispersión de los grupos de estados para cada par de variables.

El ejemplo que acabamos de analizar consituye una aplicación sencilla del coclustering, con un número reducido de variables para facilitar la implementación e interpretación del método. No obstante, este tipo de análisis suele aplicarse sobre matrices de gran tamaño, permitiendo obtener relaciones entre un número mucho mayor de variables que resultarían imposibles de identificar a simple vista.

Cuadro 4.2: Resultado de aplicar co-clustering sobre el conjunto de datos 'USArrests' de R para k=2x3 grupos.

Estado	Murder	Rape	Assault	UrbanPop
Grupo filas 1				
Arkansas	8.8	190	19.5	50
Connecticut	3.3	110	11.1	77
Delaware	5.9	238	15.8	72
Hawaii	5.3	46	20.2	83
Indiana	7.2	113	21.0	65
Kansas	6.0	115	18.0	66
Kentucky	9.7	109	16.3	52
Massachusetts	4.4	149	16.3	85
Montana	6.0	109	16.4	53
Nebraska	4.3	102	16.5	62
New Jersey	7.4	159	18.8	89
Ohio	7.3	120	21.4	75
Oklahoma	6.6	151	20.0	68
Oregon	4.9	159	29.3	67
Pennsylvania	6.3	106	14.9	72
Rhode Island	3.4	174	8.3	87
Utah	3.2	120	22.9	80
Virginia	8.5	156	20.7	63
Washington	4.0	145	26.2	73
Wyoming	6.8	161	15.6	60
Grupo filas 2				
Alabama	13.2	236	21.2	58
Alaska	10.0	263	44.5	48
Arizona	8.1	294	31.0	80
California	9.0	276	40.6	91
Colorado	7.9	204	38.7	78
Florida	15.4	335	31.9	80
Georgia	17.4	211	25.8	60
Illinois	10.4	249	24.0	83
Louisiana	15.4	249	22.2	66
Maryland	11.3	300	27.8	67
Michigan	12.1	255	35.1	74
Mississippi	16.1	259	17.1	44
Missouri	9.0	178	28.2	70
Nevada	12.2	252	46.0	81
New Mexico	11.4	285	32.1	70
New York	11.1	254	26.1	86
North Carolina	13.0	337	16.1	45
South Carolina	14.4	279	22.5	48
Tennessee	13.2	188	26.9	59
Texas	12.7	201	25.5	80
Grupo filas 3				
Idaho	2.6	120	14.2	54
Iowa	2.2	56	11.3	57
Maine	2.1	83	7.8	51
Minnesota	2.7	72	14.9	66
New Hampshire	2.1	57	9.5	56
North Dakota	0.8	45	7.3	44
South Dakota	3.8	86	12.8	45
Vermont	2.2	48	11.2	32
West Virginia	5.7	81	9.3	39
Wisconsin	2.6	53	10.8	66

En estos problemas de gran dimensionalidad es frecuente la presencia de contaminación alterando las relaciones grupales subyacentes. Al igual que en otros contextos del análisis cluster, el recorte imparcial constituye una solución efectiva ante la aparición de datos atípicos de distinta índole, permitiendo obtener clasificaciones dobles más robustas y fiables.

Existen numerosos enfoques distintos en el contexto del co-clustering, constituyendo un campo no completamente definido, que se encuentra en plena expansión y contínuo desarrollo. En esta sección, seguiremos el planteamiento propuesto en Fibbi et al.(2023)[9], que introduce un recorte imparcial tanto por filas como por columnas sobre la estimación de Modelos de Bloques Latentes (LBMs) como el del ejemplo que acabamos de presentar.

A continuación, se exponen las hipótesis realizadas en este modelo, introduciendo además la notación necesaria para facilitar la posterior descripción del problema.

Modelos de Bloques Latentes (LBMs)

Los LBMs se basan en las siguientes suposiciones:

- Los co-clusters (o bloques) están definidos por el producto cartesiano de las particiones de filas y columanas. Por lo tanto, cada elemento X_{ij} de la matriz de datos X pertenece exactamente a unos de esos clusters.
- Se asume que el conjunto de etiquetas de fila $\{Z_i\}_{i=1}^n$ y etiquetas de columna $\{W_j\}_{j=1}^p$ son independientes entre sí.
- Las etiquetas de fila (resp. de columna) también son independientes e igualmente distribuídas, siguiendo una distribución categórica.
- Condicionalmente a las etiquetas $\{Z_i\}_{i=1}^n$ y $\{W_j\}_{j=1}^p$, los elementos X_{ij} de la matriz de datos son independientes y siguen una distribución condicional definida por $f(x\,|\lambda_{k,l})$, donde $\lambda_{k,l}$ son parámetros específicos del co-cluster identificado por el par de índices (k,l). La función f es una función de densidad de probabilidad que puede ser discreta o absolutamente contínua.

La elección de la función f depende del tipo de data que se pretende modelar. En la práctica se suelen considerar distribuciones normales o de Poisson para datos contínuos, distribuciones multinomiales para datos categóricos o distribuciones de Bernoulli para datos binarios.

Los LBMs constituyen un marco de trabajo muy flexible, y bajo las suposiciones presentadas previamente, podemos escribir una verosimilitud incompleta de los datos de la forma:

$$L(\vartheta|X) = \sum_{(Z,W)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,l} \rho_l^{w_{jl}} \prod_{i,j,k,l} f(x_{ij}|\lambda_{kl})^{z_{ik}\ w_{jl}}, \tag{4.6}$$

donde:

- X es una matriz $n \times p$ cuyos elementos x_{ij} toman valores en un conjunto $S \subseteq \mathbb{R}$, por ejemplo, $S = \mathbb{R}$, $S = \mathbb{N}$, o $S = \{0, 1\}$.
- Los índices i y j hacen referencia a las filas y columnas de X, respectivamente, por lo tanto $i \in \{1, 2, ..., n\}$ y $j \in \{1, \ldots, p\}$.
- Los índices k y l se refieren a las particiones de filas y columnas de tamaño g y m, respectivamente, por lo tanto $k \in \{1, 2, ..., g\}$, $l \in \{1, 2, ..., m\}$.
- $\mathcal{Z}$ es el conjunto de todas las matrices binarias $n \times g$ con suma por filas igual a 1:

$$\mathcal{Z} = \left\{ Z \in \{0,1\}^{n\times g} : \sum_k z_{i,k} = 1,\ \forall i \right\}.$$

  Para indicar que la fila i pertenece al clúster de fila k, se escribe $z_{i,k} = 1$. Si no pertenece, entonces $z_{i,k} = 0$. De forma análoga, W representa la matriz de etiquetas binarias de columnas, reemplazando n por p y g por m.

- $\boldsymbol{\pi} = \{\pi_k\}_k$ es el vector de proporciones para la partición de filas, con:

$$\pi_k \in [0,1], \quad \sum_k \pi_k = 1.$$

  De forma análoga, $\boldsymbol{\rho} = \{\rho_l\}_l$ representa las proporciones para las columnas.

- $\Lambda = \{\lambda_{k\ l}\}_{k\ l}$ es el conjunto de los parámetros de bloque.

- $\vartheta = (\text{vec}(\Lambda), \boldsymbol{\pi}, \boldsymbol{\rho})$ es el vector de todos los parámetros del modelo.

- f es una función de densidad de probabilidad, discreta o absolutamente continua, con soporte en S.

Estos modelos suponen la base del plantemiento introducido en [9], donde se considera una extensión de los modelos de bloques latentes añadiendo contaminación. Sean S y T los conjuntos de índices de filas y columnas de las celdas no contaminadas, una modificación de la función (4.6), modelando el ruido mediante la inclusión de un factor contaminante (enfoque similar al de [12] pero adaptado al co-clustering), permite expresar la función de verosimilitud de nuestro modelo contaminado de la forma:

$$L'(\vartheta|X) = \left[\sum_{(Z,W)\in\mathcal{Z}\times\mathcal{W}} \prod_{i\in S,k} \pi_k^{z_{ik}} \prod_{j\in T,l} \rho_l^{w_{jl}} \prod_{i\in S,\ j,\in T,\ k,\ l} f(x_{ij}|\lambda_{kl})^{z_{ik}\ w_{jl}}\right] \cdot \left[\prod_{(i,j)\notin S\times T} g_{ij}(x_{ij})\right] \tag{4.7}$$

donde las funciones g_{ij} denotan las funciones de densidad de los elementos contaminados, las cuales son desconocidas.

Dada esta función de verosimilitud y dos niveles de recorte α_1 y α_2, queremos resolver el problema de maximización:

$$\max_{S,\ T} \max_{\vartheta\in\Theta} L'(\vartheta|X)$$

donde S es un subconjunto de $\{1,\dots,n\}$ y T un subconjunto de $\{1,\dots,p\}$, con cardinales $S = n - \lceil \alpha_1 n \rceil$ y $T = p - \lceil \alpha_2 p \rceil$ respectivamente. Nótese que, si $\alpha_1 = \alpha_2 = 0$, recuperamos el LBM original.

Obtener una solución óptima para este problema de maximización es imposible en la práctica, pues las distribuciones g_{ij} que generan el ruido son desconocidas. No obstante, como se ha visto en otros contextos, bajo ciertas suposiciones sensatas sobre la naturaleza de la contaminación es posible reducir el problema a la maximización de la función verosimilitud de clasificación recortada:

$$\max_{S,\ T} \max_{Z,\ W,\ \vartheta} \prod_{i\in S,k} \pi_k^{z_{ik}} \prod_{j\in T,l} \rho_l^{w_{jl}} \prod_{i\in S,\ j,\in T,\ k,\ l} f(x_{ij}|\lambda_{kl})^{z_{ik}\ w_{jl}},$$

donde los conjuntos S y T están sujetos a las condiciones impuestas anteriormente.

El algoritmo propuesto en [9] permite aproximar una solución óptima de este modelo, y constituye una extensión de algoritmos de Clasificación EM adaptada a los modelos de bloques latentes. Siguiendo un enfoque de recorte imparcial basado en las probabilidades a posteriori, se realizan en cada iteración dos recortes consecutivos, uno por filas y otro por columnas, para niveles de recorte α_1 y α_2 respectivamente.

Este doble recorte permite descartar las observaciones fila y columna más atípicas de manera simultánea, sin embargo, requiere determinar dos parámetros de recorte distintos, aumentando así el grado de complejidad del modelo. No obstante, una adaptación de las técnicas presentadas en el capítulo 3 permite monitorizar la calidad de la clasificación obtenida en función de la variación de estas dos proporciones y del valor de k. Así, el método expuesto constituye una solución de recorte imparcial para el problema de co-clustering bajo elecciones adecuadas de α_1, α_2 y k.

Apéndice:

Demostración del Teorema 1.2. .

Demostración 1ª afirmación

Sean $i, j \in \{1, \ldots, k\}$, $i \neq j$. Sabemos por la definición de las k-medias recortadas que se evita el caso degenerado $m_i = m_j$. Podemos asumir que es posible construir A_i y A_j (según se han definido anteriomente) tales que $\int_{A_i} \tau_\alpha(x) dP_x$ y $\int_{A_j} \tau_\alpha(x) dP_x$ sean estrictamente positivos. En otro caso el resultado sería trivial.

Por la proposición 1.5 sabemos que m_i y m_j son las Φ-medias de P dados τ_α y los clústeres A_i y A_j respectivamente. Denotemos $B_{i,j} := \{x \in B : \|x - m_i\| = \|x - m_j\|\}$. Nótese que

$$\int_{A_i \cup A_j} \tau_\alpha(x) \inf_{l=1,\ldots,k} \Phi(\|x - m_l\|) dP_x$$

$$= \int_{A_i \cup A_j} \tau_\alpha(x) \operatorname{ínf} \left[\Phi(\|x - m_i\|), \Phi(\|x - m_j\|)\right] dP_x$$

$$= \int_{A_i \cup B_{i,j}} \tau_\alpha(x) \Phi(\|x - m_i\|) dP_x + \int_{A_j - B_{i,j}} \tau_\alpha(x) \Phi(\|x - m_j\|) dP_x.$$

La proposición 1.4 implica que la Φ-media de P dados τ_α y $(A_i \cup B_{i,j})$, la Φ-media de P dados τ_α y $(A_i - B_{i,j})$, y m_i, coinciden, pues, de no ser así, $\{m_1, \ldots, m_k\}$ no sería una k-media recortada de P.

Probemos ahora que si $P(B_{i,j}) > 0$ entonces las Φ-medias de P dados τ_α y los conjuntos $A_i - B_{i,j}$ y $A_i \cup B_{i,j}$ no son iguales, de donde se deducirá el resultado.

Sean m y m^1 las Φ-medias de P dados τ_α y los conjuntos $A_i - B_{i,j}$ y $B_{i,j}$ respectivamente. Tenemos que:

$$\int_{A_i - B_{i,j}} \tau_\alpha(x) \, dP_x > 0$$

porque, de otro modo, por el lema 1.4, m_i pertenecería a $\overline{B}_{i,j}$, y $B_{i,j}$ no sería la frontera entre A_i y A_j.

Como τ_α es estrictamente positivo en B (y $B_{i,j} \subset B$), tenemos que m y m^1 están bien definidos y $m \neq m^1$, porque, de otro modo, tendríamos que $m_i = m^1$ y entonces $m_i \in B_{i,j}$.

Consideremos una base ortonormal $\{e_1, \ldots, e_p\}$ en $\mathbb{R}^p$, con $e_1 = m^1 - m$, y supongamos que fijamos el origen en m, de modo que $m^1 = (1, 0, \ldots, 0)$. Sea $m^t = (t, 0, \ldots, 0)$,

definimos

$$\Phi'(\|x - m^{t_0}\|) := \frac{d}{dt}\Phi(\|x - m^t\|)\Big|_{t=t_0}.$$

Dado que $A_i \cup B_{i,j}$ es un conjunto acotado, la continuidad de la derivada primera de Φ nos permite concluir que las funciones

$$H_1(t) := \int_{A_i - B_{i,j}} \tau_\alpha(x)\Phi(\|x - m^t\|)P(dx)$$

y

$$H_2(t) := \int_{B_{i,j}} \tau_\alpha(x)\Phi(\|x - m^t\|)P(dx)$$

son diferenciables con derivadas contínuas dadas por

$$H_1'(t) = \int_{A_i - B_{i,j}} \tau_\alpha(x)\Phi'(\|x - m^t\|)P(dx)$$

y

$$H_2'(t) = \int_{B_{i,j}} \tau_\alpha(x)\Phi'(\|x - m^t\|)P(dx).$$

respectivamente.

Por definición de m, $H_1'(0) = 0$, pues al ser la Φ-media de P dados τ_α y $A_i - B_{i,j}$, el mínimo de H_1 se alcanza en $m = m^0$. Como $B_{i,j}$ está contenido en B, el corolario 1.1 implica que $\tau_\alpha(x) = 1$ para todo $x \in B_{i,j}$, y una prueba similar a la de la proposición 1.4 demuestra que H_2 es estrictamente convexa en $[0, 1]$. Teniendo en cuenta que H_2 alcanza su mínimo en $t = 1$ (por un razonamiento análogo al realizado para H_1), obtenemos que $H_2'(0) < 0$. Finalmente, consideremos la función:

$$H(t) \coloneqq \int_{A_i \cup B_{i,j}} \tau_\alpha(x)\Phi(\left\|x - m^t\right\|)dP_x = H_1(t) + H_2(t)$$

Las afirmaciones previas nos indican que H es decreciente en un intervalo $[0, t_0]$ con $t_0 > 0$. Por lo tanto, m no sería la Φ-media de P dado $A_i \cup B_{i,j}$, luego $P[B_{i,j}] = 0$.

Demostración de la 2ª afirmación

Para evitar trivialidades, asumimos que $P[D] > 0$. Sea S el soporte de la distribución condicional de P dado D, supongamos que

$$P\{x \in S : \tau_\alpha(x) \in (0, 1)\} > 0.$$

En este caso, existe $\varepsilon > 0$ tal que $P(S_\varepsilon) > 0$, donde

$$S_\varepsilon := \{x \in S : 0 < \tau_\alpha(x) < 1 - \varepsilon\}.$$

Supongamos que S_ε es un conjunto infinito. Entonces, existe i tal que $S_i := S_\varepsilon \cap A_i$ también es un conjunto infinito con $P(S_i) > 0$.

Consideremos una base ortonormal en $\mathbb{R}^p$. Al menos una proyección de S_i es infinita.

Por lo tanto, existe un hiperplano H tal que si denotamos por H_+ y H_- los dos semiespacios en los que divide $\mathbb{R}^p$, entonces

$$p := P(S_i \cap H_+) > 0 \quad \text{y} \quad q := P(S_i \cap H_-) > 0.$$

Supongamos que $p \geq q$. Sea β un número real tal que $1 < \beta < \text{ínf}\left(\frac{1}{1-\varepsilon}, \frac{p+q}{q}\right)$.

Un cálculo sencillo sirve para ver que existe $\gamma \in (0,1)$ tal que, si definimos

$$\tau_1(x) = \begin{cases} \tau_\alpha(x), & \text{si } x \notin S_i, \\ \beta\tau_\alpha(x), & \text{si } x \in S_i \cap H_-, \\ \gamma\tau_\alpha(x), & \text{si } x \in S_i \cap H_+, \end{cases} \qquad (*)$$

entonces $\int \tau_1(x)\, dP_x = \int \tau_\alpha(x)\, dP_x$ y τ_1 es una función de recorte de nivel α.

La Φ-media de P dado un conjunto depende de la función de recorte que estamos considerando, pero en cualquier caso, de manera similar a la afirmación 1, la Φ-media de P dados τ_α y $B(m_i, r)$ coincide con la Φ-media de P dada τ_1 y $B(m_i, r)$; pero esto no es posible por el mismo argumento presentado en la parte final de la demostración de la afirmación 1.

Supongamos que S_ε es finito y contiene al menos dos puntos. En este caso, existe un hiperplano H tal que, si denotamos H_+ y H_- a los semi-espacios asociados, entonces existen $x_+ \in S_\varepsilon \cap H_+$ y $x_- \in S_\varepsilon \cap H_-$ tales que $P(x_+) > 0$ y $P(x_-) > 0$, y podemos construir una función de recorte de nivel α mediante una construcción similar a la de (*) para obtener una contradicción.

Por lo tanto, hemos probado que, si el soporte de P dada la frontera de $B(M, r)$ contiene más de un punto, la función de recorte τ_α solo puede tomar los valores 0 y 1 salvo en un punto como máximo.

Si estamos en el caso a) o b), la demostración ha terminado. En otro caso, existe un hiperplano H que verifica que:

$$P[H_+ \cap \{x \in D : \tau_\alpha(x) = 1\}] > 0 \quad \text{y} \quad P[H_- \cap \{x \in D : \tau_\alpha(x) < 1\}] > 0$$

o

$$P[H_+ \cap \{x \in D : \tau_\alpha(x) = 0\}] > 0 \quad \text{y} \quad P[H_- \cap \{x \in D : \tau_\alpha(x) > 0\}] > 0$$

y podemos hacer una construcción similar a la de la primera parte de la prueba para obtener nuevamente una contradicción.

□

Demostración del Teorema 1.3. .

Basta con probar que toda subsucesión de $\{M_n\}_n$ (resp. $\{V_n\}_n$) admite una nueva subsucesión que converge a M_0 (resp. a V_0).

Para cada $n = 1, 2, \ldots$, designemos por τ'_n una función de recorte arbitraria en el conjunto $\tau_{M_0,\alpha}(X_n)$, y por r'_n, $n = 1, 2, \ldots$, el radio asociado a τ'_n, es decir:

$$I_{B(M_0,r_n)} \leq \tau'_n \leq I_{B(M_0,r'_n)}$$

Obviamente, $\{r'_n\}_n$ es una sucesión acotada, y podemos asumir, sin pérdida de generalidad, que $\lim_{n\to\infty} r'_n = r'_0$ para algún $r'_0 \in \mathbb{R}$. Entonces, por la continuidad de P_{X_0}, tenemos que

$$\tau'_n(X_n) \to I_{B(M_0,r'_0)}(X_0), \quad \text{P-c.s}$$

y entonces, tomando en cuenta que $|\tau'_n| \leq 1$ para todo $n \in \mathbb{N}$, podemos escribir:

$$1 - \alpha = \int \tau'_n(X_n)\, dP \to \int I_{B(M_0,r'_0)}(X_0)\, dP \quad \text{cuando } n \to \infty$$

De ahí,

$$\int I_{B(M_0,r'_0)}(X_0)\, dP = 1 - \alpha$$

y

$$I_{B(M_0,r'_0)} = \tau_0, \quad \text{P}_{X_0}\text{-c.s}$$

Además, tenemos

$$\tau'_n(X_n)\Phi(d(X_n, M_0)) \to \tau_0(X_0)\Phi(d(X_0, M_0)), \quad \text{P-c.s}$$

y $\{\tau'_n(X_n)\Phi(d(X_n, M_0))\}_n$ es uniformemente acotada. Así,

$$V_n \leq \frac{1}{1-\alpha} \int \tau'_n(X_n)\Phi(d(X_n, M_0))\, dP$$

$$\to \frac{1}{1-\alpha} \int \tau_0(X_0)\Phi(d(X_0, M_0))\, dP \quad \text{cuando } n \to \infty$$

y, en consecuencia,

$$\limsup_n V_n \leq V_0 \tag{4.8}$$

Por el lema 1.5, existe un subconjunto no vacío $I \subset \{1, \ldots, k\}$ y una subsucesión de $\{M_n\}_n$ (que denotamos como la inicial) tal que:

$$\text{si } i \notin I, \text{ entonces } d(m_n^i, 0) \to \infty \text{cuando } n \to \infty, \tag{4.9}$$

$$\text{si } i \in I, \text{ entonces existe } m^i \in \mathbb{R}^p \text{ tal que } m_n^i \to m^i \text{cuando } n \to \infty \tag{4.10}$$

Podemos asumir, sin pérdida de generalidad, que $I = \{1, \ldots, h\}$ con $1 \leq h \leq k$. Usemos la notación $M^{(h)} = \{m_1, \ldots, m_h\}$ y $M_n^{(h)} = \{m_n^1, \ldots, m_n^h\}$. También podemos

asumir que $\{r_n\}_n$ es una sucesión convergente, con límite r. Entonces, para n suficientemente grande:

$$I_{B(M^{(h)},r)}(X_n)+I_{B(M_n-M_n^{(h)},r)}(X_n)\leq \tau_n(X_n)\leq I_{\overline{B}(M_n^{(h)},r)}(X_n)+I_{\overline{B}(M_n-M_n^{(h)},r)}(X_n). \tag{4.11}$$

Además,

$$I_{B(M_n-M_n^{(h)},r)}(X_n)\to 0, \quad \text{P-c.s}$$

De (4.11) obtenemos que:

$$\lim_n \tau_n(X_n)=I_{B(M^{(h)},r)}(X_0), \quad \text{P-c.s}$$

Luego, considerando que $|\tau_n|\leq 1$ para todo $n\in\mathbb{N}$, tenemos:

$$1-\alpha=\int \tau_n(X_n)\,dP\to\int I_{B(M^{(h)},r)}(X_0)\,dP, \quad \text{cuando } n\to\infty,$$

por lo tanto $I_{B(M^{(h)},r)}$ es una función de recorte de nivel α para X_0. Además, en virtud del lema de Fatou:

$$\begin{aligned}\liminf_n V_n &= \frac{1}{1-\alpha}\liminf_n \int \tau_n(X_n)\Phi(d(X_n,M_n))\,dP\\ &\geq \frac{1}{1-\alpha}\int \liminf_n \left(\tau_n(X_n)I_{B(M_n^{(h)},r)}(X_n)\Phi(d(X_n,M_n))\right)dP\\ &+\frac{1}{1-\alpha}\int \liminf_n \left(\tau_n(X_n)I_{B(M_n-M_n^{(h)},r)}(X_n)\Phi(d(X_n,M_n))\right)dP.\end{aligned}$$

Y, en consecuencia de que

$$\frac{1}{1-\alpha}\int \liminf_n \left(\tau_n(X_n)I_{B(M_n-M_n^{(h)},r)}(X_n)\Phi(d(X_n,M_n))\right)dP=0,$$

obtenemos:

$$\liminf_n V_n\geq \frac{1}{1-\alpha}\int I_{B(M^{(h)},r)}(X_0)\Phi(d(X_0,M^{(h)}))\,dP\geq V_{h,\Phi,\alpha}(X_0).$$

Esto y (4.8) implican:

$$V_{h,\phi,\alpha}(X_0)=V_{k,\phi,\alpha}(X_0)=\lim_n V_n,$$

y la continuidad de P_{X_0}, junto con la unicidad de M_0, muestran que $I=\{1,\ldots,k\}$ y, en consecuencia, $\{m_1,\ldots,m_h\}=M_0$. □

Bibliografía

[1] Allison Marie Horst and Alison Presmanes Hill and Kristen B Gorman(2020) *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*, doi = 10.5281/zenodo.3960218, url = https://allisonhorst.github.io/palmerpenguins/

[2] Celeux, G. and Govaert, A. *A classification EM algorithm for clustering and two stochastic versions.* Computational Statistics & Data Analysis 14, pp.315-332 (1992)

[3] Cuesta-Albertos, J. A., Gordaliza, A., Matrán, C.. *Trimmed k-means and the Cauchy mean-value property.* Multivariate Statistics. Proceedings of the Fifth Tartu Conference on Multivariate Statistics 247-265 (1995)

[4] Cuesta-Albertos, J. A., Gordaliza, A., Matrán, C.. *Trimmed k-means: an attempt to robustify quantizers.* The Annals of Statistics 25 (2), pp 553–576 (1997).

[5] Dempster, A., Laird, N. and Rubin, D. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society. Ser.B (Methodological), Vol.39, No.1. , pp.1-38. (1977)

[6] Dotto, F., Farcomeni, A., García-Escudero, L. A., y Mayo-Iscar, A. (2018). *A reweighting approach to robust clustering.* Statistics and Computing, 28(2), 477–493.

[7] Dykstra, R. L. *An algorithm for restricted least squares regression.* J. Amer. Statist. Assoc. 78, no. 384, pp.837–842 (1983)

[8] Farcomeni, A. (2014). *Snipping for robust k-means clustering under component-wise contamination.* Statistics and Computing, 24(6), 907–919.

[9] Fibbi, E., Perrotta, D., Torti, F., van Aelst, S., y Verdonck, T. (2023). *Co-clustering contaminated data: A robust model-based approach.* Advances in Data Analysis and Classification, 18, 121–161.

[10] Gallegos, M. T. *Maximum likelihood clustering with outliers. In Classification, Clustering and Data Analysis: Recent Advances and Applications* (K. Jajuga, A. Sokolowski and H.-H. Bock, eds.) pp. 247–255. Springer, New York. (2002)

[11] Gallegos, M.T., Ritter, G. *A robust method for cluster analysis* The Annals of Statistics, Vol. 36, No. 3, p 1324–1345 (2005)

[12] García-Escudero, L. A., Gordaliza, A., Matran, C., Mayo-Iscar, A. *A general trimming approach to robust cluster analysis.* The Annals of Statistics, 36 (3), pp 1324–1345 (2008).

[13] García-Escudero, L. A., Gordaliza, A., San Martín, R., van Aelst, S., y Zamar, R. (2009). *Robust linear clustering.* Journal of the Royal Statistical Society. Series B: Statistical Methodology, 71(1), 301–318.

[14] García-Escudero, L. A., Gordaliza, A., Mayo-Iscar, A., y San Martín, R. (2010). *Robust clusterwise linear regression through trimming.* Computational Statistics and Data Analysis, 54(12), 3057–3069.

[15] García-Escudero, L. A., Gordaliza, A., Matran, C., y Mayo-Iscar, A.(2011) *Exploring the number of groups in robust model-based clustering.* Statistics and Computing, 21(4), pp 585–599.

[16] García-Escudero, L. A., Gordaliza, A., y Mayo-Iscar, A. (2014). *A constrained robust proposal for mixture modeling avoiding spurious solutions.* Advances in Data Analysis and Classification, 8(1), 27–43.

[17] García-Escudero, L. A., y Mayo-Iscar, A. (2024). *Robust clustering based on trimming.* WIREs Computational Statistics, 16(4), e1658

[18] Gordaliza, A. *Best aproximations to random variables based on trimming procedures.* Journal of Approximation Theory 64, 162-180 (1991).

[19] Govaert G, Nadif M (2003) *Clustering with block mixture models.* Pattern Recogn 36(2):463–473

[20] Heinrich Fritz and Luis A. Garcia-Escudero and Agustin Mayo-Iscar(2012) *tclust: An R Package for a Trimming Approach to Cluster Analysis* Journal of Statistical Software, 47, 12, 1-26, doi=10.18637/jss.v047.i12

[21] McLachlan, G., y Peel, D. A. (2000). *Finite mixture models.* Wiley Series in Probability and Statistics

[22] Neykov, N., Filzmoser, P., Dimova, R., y Neytchev, P. (2007). *Robust fitting of mixtures using the trimmed likelihood estimator.* Computational Statistics & Data Analysis, 4(1), 299–308.

[23] Parmeet (Singh Bhatia) and Serge Iovleff and Gérard Govaert(2017) *blockcluster: An R Package for Model-Based Co-Clustering*, Journal of Statistical Software, 76, 9, 1–24, doi = 10.18637/jss.v076.i09

[24] R Core Team (2025). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`

[25] Van Aelst, S., Wang, X., Zamar, R.H. y Zhu, R. (2006), *Linear grouping using orthogonal regression* Comput. Statit. Data Anal., 50, 1287-1312.