Tema 8

Análisis Cluster

8.1. Introducción

El análisis cluster se conoce también como análisis de conglomerados. Se trata de una técnica estadística multivariante que se utiliza para agrupar objetos o individuos a modo de partición (cada objeto pertenece a un grupo y solamente a uno), de forma que los elementos más parecidos entre sí formen parte del mismo grupo y los elementos que más se diferencian queden en grupos separados. Cada grupo se denomina cluster o conglomerado.

Dentro del conjunto de técnicas multivariantes, el análisis cluster se engloba en el grupo de métodos de interdependencia, donde no se establecen diferencias de roles entre variables dependientes e independientes. El punto de partida del análisis cluster es una matriz que contiene medidas de varias variables en un conjunto de individuos. No se conoce la estructura de las categorías (a diferencia de lo que ocurre con el análisis discriminante) y el objetivo será precisamente descibrir esta estructura en la que encajan las observaciones.

En nuestro caso, toda la información disponible servirá para medir la proximidad de los objetos que se desea clasificar. En este sentido, será preciso hacer uso de medidas de distancia o de similaridad para conocer la situación relativa de los objetos a clasificar. Se podrán utilizar variables métricas, pero también variables no métricas, e incluso dicotómicas, siempre que esta medida quede bien definida. De esta forma, la medida que se utilice para cuantificar el grado de relación o la distancia entre las unidades a clasificar, juega un papel fundamental.

En este tema definiremos diferentes tipos de medidas así como diferentes metodologías para conseguir agrupar los objetos en clusters o conglomerados. Entre estas metodologías, diferenciaremos entre métodos jerárquicos y métodos no jerárquicos.

En este capítulo vamos a abordar las diferentes etapas necesarias para llevar a cabo un análisis cluster. En primer lugar establecemos el problema general que se desea resolver y el punto de partida para llevar a cabo la técnica estadística. En segundo término

observaremos las diferentes formas de medir distancias y similaridad entre individuos. A continuación estudiamos diferentes métodos jerárquicos para abordar el problema, así como diferentes métodos no jerárquicos. Para cerrar, se aborda la validación del análisis y la aplicación mediante el paquete R.

8.2. Punto de partida

Nuestro objetivo es conseguir una partición de m objetos o individuos a partir de la información de p variables estadísticas observadas en ellos, por lo que el punto de partida del problema es una matriz de datos X de dimensión $m \times p$ que contiene toda la información observada. Ahora bien, esta información puede estar armonizada, es decir, todas las p variables son comparables directamente porque estén medidas en las mismas unidades, o pueden no serlo. Por ello, además, tenemos que tener en cuenta el tipo de variable que es cada una de las p usadas. Así, habrá que diferenciar entre variables continuas, discretas, ordinales y nominales. En este sentido, tendremos que transformar los valores de las variables observadas para conseguir que sean comparables entre ellas y que estén medidas en los mismos rangos.

En términos generales, seguiremos las siguientes recomendaciones:

- Las variables dicotómicas (sí/no, presencia/ausencia) no suelen transformarse. Son las que menos grado de información proveen en una descripción, por su naturaleza.
- Las variables categóricas nominales suelen convertirse a dicotómicas (presencia/ausencia) para sus distintas modalidades.
- Las variables categóricas ordinales pueden mantenerse sin transformación, siempre que todas sean de esta naturaleza.
- Las variables cuantitativas se suelen normalizar para evitar diferentes unidades de medida en la misma matriz de datos. Esta normalización puede realizarse de varias formas:
  - Calculando las puntuaciones Z o valores tipificados; esto se consigue restando el valor medio y dividiendo entre el valor de la desviación típica todos los valores de la variable, es decir
    $$x_{ij}^* = \frac{x_{ij} - \overline{x}}{\sigma_j}$$
    donde i denota el individuo y j la variable.
  - Estandarizando a un rango entre 0 y 1; para ello, se resta a todos los valores de la variable el valor mínimo y después se divide entre el rango observado en la variable:
    $$x_{ij}^* = \frac{x_{ij} - min\{x_{ij}\}}{max\{x_{ij}\} - min\{x_{ij}\}}$$

- Estandarizando a un rango entre -1 y 1; esto se consigue realizando la operación

$$x_{ij}^{*} = \frac{2x_{ij} - (max\{x_{ij}\} + min\{x_{ij}\})}{max\{x_{ij}\} - min\{x_{ij}\}}$$

  para cada individuo i de la variable j-ésima.

- Magnitud con máximo en 1; para transformar los valores y conseguir que el máximo sea 1 se dividen todas las observaciones entre el máximo:

$$x_{ij}^{*} = \frac{x_{ij}}{max\{x_{ij}\}}.$$

- Estandarización para que las variables tengan media unitaria; se consigue dividiendo todos los valores observados entre el valor de la media (siempre que esta no sea 0):

$$x_{ij}^{*} = \frac{x_{ij}}{\overline{x}}.$$

- En el caso de que X contenga todas la variables medidas en la misma escala, no será necesaria la estandarización.

Si bien existen estas recomendaciones generales, el criterio y la experiencia de quien lleva a cabo el análisis resulta crucial a la hora de estandarizar o transformar las variables de la matriz X. En general es deseable considerar variables cuyos valores sean directamente comparables entre ellos, para evitar que las diferencias de unidades de medida tergiverse la medición del peso de las variables en la clasificación.

8.2.1. Agrupación de individuos vs. agrupación de variables

Una cuestión que se debe dilucidar antes de comenzar con el análisis es si se desea clasificar a los individuos o bien se desea una clasificación de las variables. Notemos que, transponiendo la matriz inicial, el análisis puede aplicarse para realizar una clasificación u otra.

En este sentido, una clasificación de individuos nos dará como resultado k grupos disjuntos donde cada individuo pertenece a un grupo y solamente a uno de ellos. Podríamos pensar que en este aspecto el análisis cluster es similar al discriminante. Sin embargo, el análisis discriminante parte de una matriz donde la clasificación ya está definida, y además esta debe haberse definido por otros medios; lo que se persigue es establecer la relación entre el grupo de pertenencia y las variables observables. Se parte de que la agrupación tiene una estructura definida y lo que establece es la relación de esta con las variables observadas. Por su parte, el análisis cluster parte de que no se conoce a priori la estructura de agrupación, y es esta la que se desea descubrir.

En cuanto a la agrupación de variables, podríamos poner en paralelo el análisis cluster para agrupación de variables con el análisis factorial, puesto que en ambos casos se agrupan variables. La diferencia fundamental entre ambas técnicas es que en el análisis factorial los factores se buscan para reducir la dimensión del problema estudiado, de forma que se seleccionan aquellas variables que mejor resuman la información de la matriz completa. No es así en el análisis cluster, donde los grupos de variables se forman por el parecido o la similaridad entre ellas, teniendo todas el mismo rol dentro del grupo formado.

Una vez determinado el tipo de análisis que se quiere llevar a cabo, se avanza hasta el siguiente paso que es la elección de la medida (distancia o similaridad) que se utilizará para establecer qué objetos deben situarse en el mismo grupo. Este paso, que se describe a continuación, dependerá así mismo de si el análisis que se desea realizar es sobre individuos o sobre variables, dado que las medidas más adecuadas no son las mismas en ambos casos.

8.3. Medidas: distancias y similaridades

Las medidas son necesarias para caracterizar la relación entre individuos o variables. Se trata de establecer la relación entre vectores; serán los vectores fila de la matriz inicial de observaciones X en el caso de pretender clasificar a los individuos o los vectores fila de la matriz traspuesta X' en el caso de querer clasificar las variables.

8.3.1. Definiciones de distancia y de similaridad

Definición de distancia métrica. Sea A un conjunto. Llamamos distancia métrica a una función $d : A \times A \to \mathbb{R}_0^+$ verificando $\forall x, y \in A$

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \Leftrightarrow x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$, $\forall z \in A$

Las dos primeras propiedades implican que d es una función definida positiva, la tercera es la propiedad de simetría y la cuarta es conocida como la desigualdad triangular.

Definición de similaridad. Sea A un conjunto finito o infinito de elementos, $s : U \times U \to \mathbb{R}$ una función, y $s_0 \in \mathbb{R}$ un número real finito arbitrario. Entonces, s se define como una similaridad si $\forall x, y \in U$ se verifica que

1. $s(x, x) = s_0$
2. $s(x, y) \leq s_0$

3. $s(x, y) = s(y, x)$

El valor s_0 normalmente es la unidad, que representa la similaridad máxima o del 100 % de los objetos x e y, la cual se alcanza cuando se mide la similaridad entre un objeto x y él mismo.

Habitualmente se miden las distancias en el caso de clasificación de individuos, mientras que se observan las similaridades en el caso de la clasificación de variables. Cualquiera de las medidas puede, sin embargo, aplicarse a cualquier tipo de análisis sin más que considerar la matriz traspuesta, sin embargo la interpretación que de ello pueda originarse no queda muy clara. En nuestro caso expondremos a continuación algunas de las medidas más habituales: de asociación para variables y de distancias para individuos.

8.4. Medidas habituales de distancia entre individuos

Considerando la matriz inicial de partida X, en la cual cada fila $\mathbf{x_h}$ representa un vector conteniendo la información de las p variables observadas en el mismo individuo, podemos definir diferentes distancias entre estos vectores, las cuales representan la cercanía o lejanía entre los individuos, que nos dará pie a realizar la clasificación más adelante. Las distancias más habituales son

Distancia euclídea. Se trata de la métrica más intuitiva y la más utilizada en la práctica. Así, la distancia entre los individuos r y s se expresará

$$d_2(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{\sum_{h=1}^{p} (x_{rh} - x_{sh})^2}.$$

Distancia de Minkowski. Es una generalización de la distancia euclídea, que a su vez dará lugar a diferentes definiciones. Se define la distancia de Minkowski como

$$d_q(\mathbf{x}_r, \mathbf{x}_s) = \left(\sum_{h=1}^{p} |x_{rh} - x_{sh}|^q \right)^{\frac{1}{q}}.$$

(Nótese que para $q = 2$ se tiene la distancia euclídea).

Distancia de bloque, ciudad o Manhattan. Es el caso particular de la distancia Minkowski cuando $q = 1$:

$$d_1(\mathbf{x}_r, \mathbf{x}_s) = \sum_{h=1}^{p} |x_{rh} - x_{sh}|.$$

Distancia de Chebychev o del máximo. Caso particular de la distancia Minkowski cuando $q = \infty$:

$$d_\infty(\mathbf{x}_r, \mathbf{x}_s) = \max_{h=1,\ldots,p} \{|x_{rh} - x_{sh}|\}.$$

Distancia Camberra. Es una versión ponderada de la distancia Manhattan, muy sensible para valores cercanos al 0:

$$d_C(\mathbf{x}_r, \mathbf{x}_s) = \sum_{h=1}^{p} \frac{|x_{rh} - x_{sh}|}{|x_{rh}| + |x_{sh}|}.$$

Distancia Pearson. Es una generalización de la distancia euclídea en la que se reescala cada variable en unidades de desviación típica:

$$d_P^2(\mathbf{x}_r, \mathbf{x}_s) = \sum_{h=1}^{p} \frac{(x_{rh} - x_{sh})^2}{\sigma_h^2}.$$

Distancia de Mahalanobis.

$$d_M(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{(\mathbf{x_r} - \mathbf{x_s})' \Sigma^{-1} (\mathbf{x_r} - \mathbf{x_s})}$$

con Σ la matriz (definida positiva) de varianzas-covarianzas.

Distancia de Gower. Se trata realmente de una medida de similaridad, aunque habitualmente nos referimos a ella como "distancia". Será útil cuando se quiera establecer la similaridad entre individuos y el conjunto de variables sea mixto, es decir, se tenga una mezcla de tipos de variables diferentes. El coeficiente se define:

$$d(\mathbf{x}_r, \mathbf{x}_s) = \frac{\sum_{h=1}^{p} \left(1 - \frac{|x_{rh} - x_{sh}|}{G_h} + a + \alpha\right)}{p_1 + (p_2 - d) + p_3}$$

con los siguientes valores:

- p_1 es el número de variables métricas no binarias
- p_2 es el número de variables métricas binarias
- p_3 es el número de variables no métricas
- a es el número de coincidencias $(1, 1)$ y d el número de coincidencias $(0, 0)$ en las variables binarias
- α es el número de coincidencias en las variables no métricas
- G_h es el rango o recorrido entre el máximo y el mínimo de la h-ésima variable métrica

8.5. Medidas habituales de asociación entre variables

En el caso en que se pretenda establecer la asociación entre variables, no entre individuos, es habitual utilizar medidas de similaridad basadas en la asociación o relación entre las mismas. A continuación se describen someramente las más utilizadas.

8.5.1. Medidas de asociación para variables binarias

Supongamos que queremos conocer la similaridad entre variables binarias, que vienen representadas con valores 1 para la presencia y 0 para la ausencia. En ese caso, se tiene que para cada pareja de variables, se puede obtener una tabla de doble entrada:

$\mathbf{x_r}$ / $\mathbf{x_s}$	Presencia (1)	Ausencia (0)	Total
Presencia (1)	a	b	$a+b$
Ausencia (0)	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=m$

A partir del conteo de la frecuencia con que ocurren las parejas de valores, se definen una serie de medidas para establecer la similaridad entre las variables r-ésima y s-ésima (si se utiliza la matriz traspuesta, se pueden observar las mismas medidas para establecer la similaridad entre los individuos r y s, en ese caso el total sumaría p en lugar de m):

Medida de Ochiai. Es una particularización del coseno del ángulo formado por los vectores, para variables binarias:

$$Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a}{\sqrt{(a+b)(a+c)}}.$$

Medida Φ. Es el coeficiente de correlación lineal de Pearson para variables dicotómicas:

$$Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}.$$

Medida de Russel y Rao. Es la probabilidad de coincidencia $(1,1)$ o presencia-presencia en las dos variables:

$$Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a}{m}.$$

Parejas simples. Es la probabilidad de coincidencia de cualquier tipo ($(1,1)$ y $(0,0)$, es decir presencia-presencia o ausencia-ausencia:

$$Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a+d}{m}.$$

Medida de Jaccard. Es la probabilidad de coincidencia de tipo $(1,1)$ condicionada a que no hay coincidencia $(0,0)$, es decir:

$$Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a}{a+b+c}.$$

Medida de Dice. Es una variación de la medida de Jaccard, donde la coincidencia $(1,1)$ se pondera doblemente:

$$Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{2a+d}{2a+b+c}.$$

Medida de Rogers Tanimoto. Mide la probabilidad de coincidencia entre las dos variables teniendo en cuenta la duplicación en la ponderación de las no coincidentes:

$$Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a+d}{a+d+2(b+c)}.$$

Medida de Kulcynski. Mide la proporción de coincidencias tipo $(1,1)$ entre las no coincidencias:

$$Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a}{b+c}.$$

8.5.2. Medidas de asociación para variables ordinales

Para definir algunas medidas de asociación entre variables ordinales, se tiene en cuenta la idea de parejas empatadas, parejas concordantes y parejas discordantes en tablas de doble entrada. Las parejas concordantes son aquellas parejas de individuos que presentan valores menores en una de las variables y mayores en otra (en las mismas). Las parejas discordantes sin embargo son aquellas en las que uno de los individuos presenta valores menores en una variable mientras que el otro individuo presenta valores mayores en esa variable. En los pares empatados, al menos una de las variables no presenta valores mayores ni menores.

Si tenemos una tabla de doble entrada para una pareja de variables, podemos esquematizar esta idea

$\mathbf{x}_r$ / $\mathbf{x}_s$	Modalidad 1	Modalidad 2	$\cdots$	Modalidad q_s
Modalidad 1	A		$\ldots$	
Modalidad 2	D		$\ldots$	C
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Modalidad q_r		B	$\ldots$	

de forma que podemos observar que los individuos A y B forman una pareja concordante, ya que para las dos variables se observan valores en A menores que en B. La pareja de individuos A y C son una pareja asimismo concordante. La pareja B y C serán discordante, puesto que en B la variable r-ésima es mayor que para C pero para la variable s-ésima ocurre al contrario. La pareja A-D será una pareja empatada, ya que se observa la misma modalidad en $\mathbf{x}_s$ para los dos individuos. La pareja C-D también es una pareja empatada. Resumiendo:

$$\begin{array}{ll} A-B & \text{concordante} \\ A-C & \text{concordante} \\ A-D & \text{empatada} \\ B-C & \text{discordante} \\ B-D & \text{concordante} \\ C-D & \text{empatada} \end{array}$$

Teniendo en cuenta estas definiciones, se definen las siguientes medidas:

Coeficiente τ de Kendall. Se define como

$$\tau(\mathbf{x}_r, \mathbf{x}_s) = \frac{n_c - n_d}{n_c + n_d}$$

donde

- n_c es el número de parejas concordantes
- n_d es el número de parejas discordantes

Coeficiente de concordancia $\tau - c$ de Kendall. La versión más utilizada para el cálculo del coeficiente de concordancia es la medida $\tau - c$ de Kendall, que se define como:

$$\tau - c(\mathbf{x}_r, \mathbf{x}_s) = \frac{2k(n_c - n_d)}{m^2(k-1)}$$

donde

- k es el mínimo entre el número de modalidades de la variable r y el número de modalidades de la variable s, es decir $k = \text{mín}\{q_r, q_s\}$ va a coincidir con el número mínimo de casos no empatados entre las variables
- n_c es el número de parejas concordantes
- n_d es el número de parejas discordantes

Se utiliza cuando el número de niveles no es igual en las dos variables $q_r \neq q_s$.

Coeficiente de correlación de Spearman. Mide el grado de asociación entre los rangos que se le asignan a los valores de las variables analizadas. Se definen los valores $x_{(il)}$ como los ordinales asociados a los correspondientes x_{il}, y a partir de ellos, el coeficiente de correlación de Spearman se define:

$$\rho(\mathbf{x}_r, \mathbf{x}_s) = 1 - \frac{6\sum_{i=1}^{m}(x_{(il)} - x_{(ik)})^2}{m(m^2-1)}.$$

8.5.3. Medidas de asociación para variables cuantitativas

Coseno del ángulo entre los vectores. El coseno es una buena medida para saber si dos vectores son o no paralelos, ya que una condición de paralelismos entre vectores es que el coseno del ángulo entre ellos toma el valor 0.

$$\cos(\widehat{\mathbf{x}_r, \mathbf{x}_s}) = \frac{\sum_{k=1}^{m} x_{kr}x_{ks}}{\sum_{k=1}^{m} x_{kr}^2 \sum_{k=1}^{m} x_{ks}^2}.$$

Coeficiente de correlación lineal de Pearson. Se define este coeficiente entre las variables r y s-ésimas como

$$\rho(\mathbf{x}_r, \mathbf{x}_s) = \frac{Cov(X_r, X_s)}{\sqrt{Var(X_r)Var(X_s)}} = \frac{\sum_{k=1}^{m}(x_{kr} - \overline{x}_r)(x_{ks} - \overline{x}_s)}{\sqrt{\sum_{k=1}^{m}(x_{kr} - \overline{x}_r)^2 \sum_{k=1}^{m}(x_{ks} - \overline{x}_s)^2}}.$$

8.5.4. Consideraciones finales sobre las distancias y las similaridades

La elección de una medida que represente bien la diferencia o similaridad entre individuos o variables es uno de los pasos principales para obtener una clasificación interpretable. Así, si bien algunas medidas son más apropiadas que otras, dependiendo del tipo de análisis y del tipo de variable que se tiene, se debe seleccionar una o varias medidas que aborden el problema coherentemente. Aunque cualquier medida es matemáticamente plausible para cualquier tipo de información, lo cierto es que la interpretabilidad de la información no siempre se consigue, por lo que habrá que realizar la selección cuidadosamente antes de proceder con la agrupación.

Habitualmente se seleccionan varias medidas y se realiza la agrupación bajo diferentes medidas, comparando después los resultados obtenidos. En el caso de tener resultados muy dispares, la medición estará jugando un papel muy relevante en la agrupación, por lo que será preciso analizar este hecho detenidamente. Si, en cambio, los resultados son iguales o muy similares, esto será un indicativo de una estructura en grupos bien definida, que bajo diferentes medidas permanece estable. Será un indicativo de que la agrupación es consistente.

Finalmente, debemos hacer notar que las distancias pueden ser convertidas en similaridades, y viceversa, utilizando la transformación $d(\mathbf{x}_r, \mathbf{x}_s) = 1 - Sim(\mathbf{x}_r, \mathbf{x}_s)$ o, más convenientemente: $d(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{1 - Sim(\mathbf{x}_r, \mathbf{x}_s)}$

8.6. Métodos jerárquicos

Los métodos jerárquicos toman la información de la matriz de datos y la convierten en una matriz de distancias o similaridades, cuadrada y simétrica, donde se establece la medida existente entre cada dos objetos. Esta matriz tendrá tamaño $m \times m$ si se trata de un análisis cluster para individuos y tamaño $p \times p$ si se trata de un análisis cluster para variables. En general, para tomar un caso que abarque ambos, consideraremos esa matriz de medidas de tamaño $n \times n$ para clasificar n objetos, sin determinar si se trata de variables o individuos.

En este punto, los métodos jerárquicos se dividen en dos tipos: los aglomerativos y los disociativos. En el primer caso, lo métodos jerárquicos aglomerativos, la forma de proceder es la siguiente:

1. Se parte de tantos grupos (clusters) como objetos se tenga que clasificar (n).

2. Observamos en la matriz de medidas aquellos dos clusters más similares o menos distantes.

3. Estos dos clusters se juntan para formar un solo cluster, pasando a tener $n - 1$ clusters.

4. Se recalcula la distancia o similaridad entre clusters en una nueva matriz de medidas, de tamaño $(n - 1) \times (n - 1)$.

5. Se vuelve al paso 2, para agrupar los clusters más parecidos entre sí y obtener un número de clusters $n - 2$.

6. Se repite este proceso hasta que quede un solo cluster conteniendo todos los objetos iniciales.

Los métodos jerárquicos disociativos realizan el procedimiento de forma inversa: partiendo de un solo cluster con todos los objetos, se llega a tantos clusters como objetos.

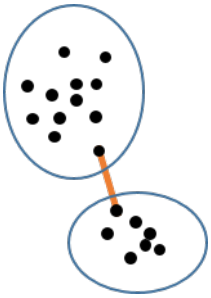
En ambos casos, se deben calcular las matrices con las medidas de distancia o similaridad en cada paso, una vez formado o deshecho un cluster. Este paso precisa de una metodología que indique cómo se va a medir la distancia o la similaridad entre clusters con varios objetos cada uno. Entre objetos es claro, pero entre grupos de objetos, existen diferentes alternativas, que dan pie a metodologías diferentes de análisis.

Vemos a continuación algunos de lo métodos más habituales.

8.6.1. Método del vecino más próximo (*Linkage simple*)

En este método se considera que la distancia entre dos grupos será la mínima que exista entre elementos de ambos grupos (en caso de similaridad, la máxima), es decir, se toma como distancia de todo el grupo la de aquellos objetos de los grupos que estén más cerca (los vecinos más cercanos).

Gráficamente, la idea es medir entre grupos de la forma más cercana



Ejemplo Como ejemplo, supongamos que se dispone de 5 objetos $(O_1, \ldots, O_5)$ para clasificar, y que se obtiene una matriz de similaridades para ellos:

$$\begin{pmatrix} & O_1 & O_2 & O_3 & O_4 & O_5 \\ O_1 & 1 & 0{,}59 & 0{,}44 & 0{,}61 & 0{,}93 \\ O_2 & & 1 & 0{,}21 & 0{,}83 & 0{,}49 \\ O_3 & & & 1 & 0{,}71 & 0{,}57 \\ O_4 & & & & 1 & 0{,}02 \\ O_5 & & & & & 1 \end{pmatrix}$$

Siguiendo los pasos descritos, en primer lugar consideramos que cada objeto es en sí mismo un cluster, y observamos los clusters más cercanos, que en este caso son los que tengan una similaridad mayor: O_1 y O_5. Entonces, estos dos clusters se unen formando

uno solo y volvemos a calcular la matriz, teniendo en cuenta que la similaridad entre un cluster y otro será la máxima entre elementos de los clusters. Así, la similaridad entre el nuevo cluster $O_1 - O_5$ y el cluster O_2 será la máxima entre la similaridad de la pareja $O_1 - O_2$ y la similaridad entre $O_5 - O_2$, en este caso la máxima es la similaridad entre O_1 y O_2. (En caso de que la matriz contuviera distancias, elegiríamos la distancia menor entre las dos posibles.) Llegamos entonces a una nueva matriz de similaridades:

$$\begin{pmatrix} & \{O_1, O_5\} & O_2 & O_3 & O_4 \\ \{O_1, O_5\} & 1 & 0{,}59 & 0{,}57 & 0{,}61 \\ O_2 & & 1 & 0{,}21 & 0{,}83 \\ O_3 & & & 1 & 0{,}71 \\ O_4 & & & & 1 \end{pmatrix}$$

En el siguiente paso, los clusters que se unen, por ser los más similares según la medida elegida, son O_2 y O_4, teniendo ahora que volver a calcular la matriz de similaridades, que quedará, con el método de *Linkage simple* como

$$\begin{pmatrix} & \{O_1, O_5\} & \{O_2, O_4\} & O_3 \\ \{O_1, O_5\} & 1 & 0{,}61 & 0{,}57 \\ \{O_2, O_4\} & & 1 & 0{,}71 \\ O_3 & & & 1 \end{pmatrix}$$

Siguiendo con este procedimiento, los clusters más similares son el $O_2 - O_4$ y O_3, por lo que forman un cluster nuevo, y se recalcula la matriz de similaridades eligiendo la máxima entre los elementos de los clusters que se acaban de unir:

$$\begin{pmatrix} & \{O_1, O_5\} & \{O_2, O_4, O_3\} \\ \{O_1, O_5\} & 1 & 0{,}61 \\ \{O_2, O_4, O_3\} & & 1 \end{pmatrix}$$

En el último paso, los dos clusters se unen formando uno solo con todos los objetos iniciales. De esta forma hemos realizado 4 uniones: la primera de los objetos 1 y 5, la segunda de los objetos 2 y 4, la tercera los objetos 2 y 4 se unen con el 3, y por último se unen todos los objetos en un solo cluster. El método del vecino más próximo indica que en cada paso la similaridad entre grupos sea la máxima entre los individuos, y en caso de tener distancias, elegiremos la mínima (la más *favorable*).

Cuando se lleva a cabo un método jerárquico los pasos seguidos se representan en un gráfico llamado dendrogama, similar a un árbol, cuyas ramas se unen (o separan si se trata de un método jerárquico disociativo) y que resume el procedimiento: pueden verse las uniones y la medida entre grupos que propicia cada una de las uniones de clusters.

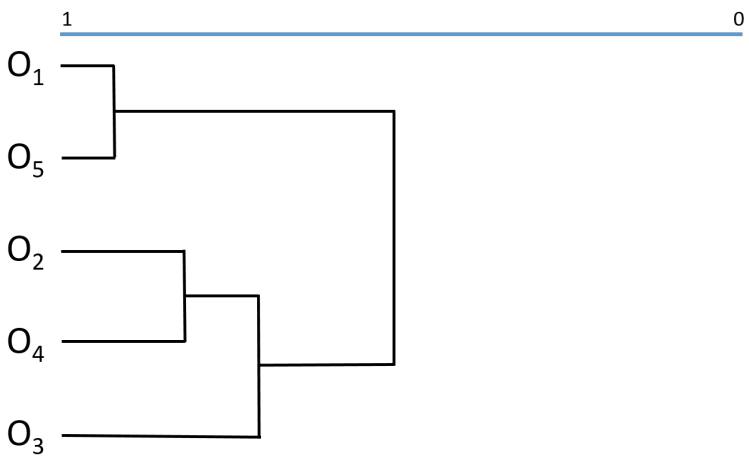
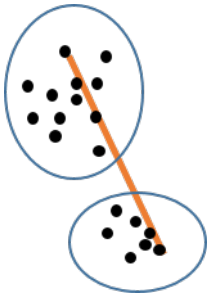


Figura 8.1: Dendrograma para el ejemplo

8.6.2. Método del vecino más lejano (*Linkage completo*)

En este método se considera que la distancia entre dos grupos será la máxima (si hablamos de similaridad, la mínima) entre los elementos de ambos grupos.



Así, si resolvemos el ejemplo anterior utilizando este método tendremos que, partiendo de la misma matriz de similaridades, el primer paso coincidirá, puesto que siempre se unen los objetos más similares, y por tanto se unen los objetos 1 y 5, pero al calcular de nuevo la matriz de similaridades se tiene que establecer la similaridad entre este nuevo cluster y los demás, tomando el caso más desfavorable (la menor similaridad) y por tanto quedará la matriz

$$\begin{pmatrix} & \{O_1,O_5\} & O_2 & O_3 & O_4 \\ \{O_1,O_5\} & 1 & 0{,}49 & 0{,}44 & 0{,}02 \\ O_2 & & 1 & 0{,}21 & 0{,}83 \\ O_3 & & & 1 & 0{,}71 \\ O_4 & & & & 1 \end{pmatrix}$$

En el siguiente paso se unen los clusters O_2 y O_4, teniendo ahora que volver a calcular la matriz de similaridades, que quedará, con el método de *Linkage completo*

$$\begin{pmatrix} & \{O_1,O_5\} & \{O_2,O_4\} & O_3 \\ \{O_1,O_5\} & 1 & 0{,}02 & 0{,}44 \\ \{O_2,O_4\} & & 1 & 0{,}21 \\ O_3 & & & 1 \end{pmatrix}$$

En el siguiente paso uniremos los clusters $\{O_1,O_5\}$ y O_3, por lo que forman un cluster nuevo, y con él se vuelve a calcular la matriz de similaridades:

$$\begin{pmatrix} & \{O_1,O_5,O_3\} & \{O_2,O_4\} \\ \{O_1,O_5,O_3\} & 1 & 0{,}02 \\ \{O_2,O_4\} & & 1 \end{pmatrix}$$

8.6.3. Otros métodos o linkages

Otros métodos no consideran la distancia entre clusters como la mínima o la máxima de entre las medidas de sus respectivos objetos, sino que pueden promediar dichas medidas, de forma ponderada según el número de objetos que contenga cada grupo o no ponderada, o establecer la medida entre clusters como la medida entres los centroides de cada cluster, lo cual da una medida más realista de la relación entre los grupos.

Mención especial debe hacerse del método de Ward, que se basa en la búsqueda de minimizar la varianza dentro de cada grupo a la hora de unir clusters, por lo tanto, en cada paso se unirán los clusters que incrementen menos el valor total de la suma de los cuadrados de las diferencias entre cada individuo del cluster al centroide del mismo, dentro de cada cluster, teniendo en cuenta que el menor incremento de los errores cuadráticos es proporcional a la distancia euclídea al cuadrado de los centroides de los clusters unidos.

El método de Wald es uno de los más utilizados ya que da lugar a clusters pequeños y de tamaños similares y es capaz de acercarse más que otros métodos a la clasificación óptima.

Una vez comprendido el mecanismo de las diferentes metodologías, se puede observar que los mecanismos definidos se pueden generalizar según la fórmula de recurrencia de Lance y Williams, que se enuncia a continuación.

Fórmula de recurrencia de Lance y Williams. Sean A, B y C clusters con n_A, n_B y n_C elementos respectivamente. Supongamos la unión en una etapa de los clusters B y C, entonces, la fórmula recurrente de Lance y Williams para calcular la distancia entre A y el nuevo cluster formado por B y C es

$$d(A,\{B,C\}) = \alpha_B d(A,B) + \alpha_C d(A,C) + \beta d(B,C) + \gamma |d(A,B) - d(A,C)|$$

De esta forma, tomando $\alpha_B = \alpha_C = \frac{1}{2}$; $\beta = 0$ y $\gamma = \frac{-1}{2}$ se tiene que la fórmula nos lleva al método del vecino más cercano lo *linkage simple*, mientras que para obtener el vecino más lejano tendríamso que tomar los valores $\alpha_B = \alpha_C = \frac{1}{2}$; $\beta = 0$ y $\gamma = \frac{1}{2}$.

Otros métodos se derivan de diferentes valores en la fórmula de Lance y Williams, como puede verse resumido en la tabla 8.1.

Método	α_B	α_C	β	γ
Vecino más próximo	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{-1}{2}$
Vecino más lejano	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Media no ponderada	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Media ponderada	$\frac{n_B}{n_B+n_C}$	$\frac{n_C}{n_B+n_C}$	0	0
Del centroide	$\frac{n_B}{n_B+n_C}$	$\frac{n_C}{n_B+n_C}$	$-\alpha_b\alpha_C$	0
De Ward	$\frac{n_B+n_A}{n_B+n_C+n_A}$	$\frac{n_C+n_A}{n_B+n_C+n_A}$	$\frac{-n_A}{n_B+n_C+n_A}$	0

Cuadro 8.1: Resumen de métodos de clustering jerárquico según la fórmula de Lance y Williams.

8.6.4. Coeficientes para valorar la idoneidad del método

Coeficiente de aglomeración. El coeficiente de aglomeración es una medida del ajuste de la aplicación de un método de clasificación jerárquico aglomerativo a los datos.

Para cada objeto i se denota $\mu(i)$ al cociente entre la distancia de i al primer grupo con el que se ha fusionado y la distancia de i al último grupo formado. El coeficiente de

aglomeración se calcula realizando la media aritmética de todos los valores $1 - \mu(i)$. Este coeficiente toma valores entre 0 y 1. Cuanto más cercano al 1, mejor será el ajuste del método escogido para nuestros datos.

Coeficiente de división o de disociación. Para métodos jerárquicos disociativos, se define $d(i)$ el cociente entre el diámetro del último grupo que se ha formado y el diámetro de todo el conjunto de datos. El coeficiente de disociación es la media aritmética de todos los valores $1 - d(i)$. Este coeficiente toma valores entre 0 y 1. Cuanto más cercano sea este valor a la unidad, mejor será el ajuste del método escogido a nuestros datos.

Coeficiente de correlación cofenético. Este coeficiente es un coeficente de correlación lineal de Pearson entre dos matrices; la de medidas entre objetos inicial y la matriz coefenética. Esta última es la matriz que se va utilizando en las etapas del análisis sin reducir la dimensionalidad de la matriz original.

Cuando se tengan valores del coeficiente de correlación cofenético ceracanos a 1 o -1, será indicativo de que la distribución en grupos de los datos es adecuada. Sin embargo, valores cercanos a 0 estarán indicando lo contrario, es decir, una estructura en grupos poco adecuada.

8.6.5. Elección del número de clusters

Cuando se lleva a cabo un método jerárquico, el número de grupos en que quedará la partición no tiene por qué estar establecido de antemano. Es habitual realizar esta selección una vez llevado a cabo alguno de los métodos jerárquicos, teniendo en cuenta algunas consideraciones.

Elección subjetiva. El procedimiento más básico consiste en cortar el dendrograma de forma subjetiva, en el lugar en que la experiencia en la investigación o el criterio experto considere más apropiado. Este criterio es poco satisfactorio, puesto que depende exclusivamente de la opinión del equipo experto. Algunas ideas intuitivas pueden apoyar la decisión, como por ejemplo, seleccionar en el dendrograma el paso en el que aumenta la distancia de fusión reflejada en el eje.

Método del codo. Este método consiste en minimizar la suma de las distancias al cuadrado de cada individuo con su centroide, lo que se conoce como WSS por ser las siglas de *within sum of squares*, es decir, la suma de las varianzas internas de los grupos. El algoritmo del codo pretende realizar una representación gráfica del WSS para diferentes valores del número de grupos k. La aparición de un *codo* en la gráfica indica una variación significativa en el WSS y señala el número idóneo de clusters.

Coeficiente de silueta. El método del coeficiente de silueta o sombra determina que el número óptimo de grupos es aquel valor que maximiza el coeficiente de sombra, que es una estimación de la distancia media entre grupos, representando cuán cerca están los

individuos de un grupo de las observaciones de los grupos vecinos. Para cada objeto i, el coeficiente de silueta S_i se calcula como

$$S_i = \frac{b_i - a_i}{\text{máx}\{a_i, b_i\}}$$

donde a_i representa la distancia media entre el objeto i y los demás objetos del mismo cluster y b_i se calcula como $b_i = \text{mín}\{d(\mathbf{x}_i, \mathbf{x}_s\}$, con $\mathbf{x}_s$ el resto de objetos del mismo cluster.

Cuando el valor del coeficiente de silueta se aproxima a la unidad, indica que el objeto i está bien clasificado. Valores cercanos a 0 indican que el objeto probablemente se encuentre entre dos grupos. Si el valor es negativo, es síntoma de que el objeto no ha quedado clasificado en el grupo que le corresponde.

El algoritmo entonces consiste en calcular, para cada valor de número de grupos k, el valor medio de los coeficientes de sombra y buscar qué valor de k lo hace máximo.

Estadístico GAP**.** Este estadístico se utiliza para comparar el valor de k obtenido con el que se obtendría bajo la distribución de una hipótesis nula apropiada. El objetivo es estar lejos de la distribución uniforme, ya que esta proporciona la colección de puntos más alegados los unos de los otros. El algoritmo concluye con la regla de seleccionar el valor de k para el cual el valor del estadístico sea mayor.

8.7. Métodos no jerárquicos

Los métodos no jerárquicos están diseñados para la clasificación de individuos y no de variables, en k grupos o clusters. Este valor de k debe ser conocido de antemano, ya que la metodología básicamente consiste en realizar una partición y a partir de ella hacer una reasignación de individuos a clusters para obtener una partición mejor. La forma de obtener la partición inicial y la forma de reasignar los individuos a los clusters son los que dan pie a los diferentes métodos no jerárquicos.

8.7.1. Método de las *k*-medias

El método de las k-medias consiste en partir de k semillas y asignar, según la proximidad de cada individuo a las semillas, cada individuo al cluster con semilla más cercana. Tras cada asignación, la semilla se recalcula como el centroide del cluster formado.

El cálculo de las semillas iniciales puede realizarse de varias formas:

1. Asignando los primeros k individuos como semillas.

2. Hacer un muestreo sistemático dentro de los individuos, seleccionando aquellos que estén en las posiciones $\left[\frac{im}{k}\right]$, $i = 1, \ldots, k$, donde $[x]$ representa la parte entera de x.

3. Realizar un muestreo aleatorio simple de tamaño k de entre los m individuos a clasificar.

4. Tomar una partición pre-establecida en k grupos. Esto por ejemplo puede realizarse tras un análisis jerárquico que indique una partición posible para los individuos. Las k semillas son entonces los k centroides de los grupos iniciales establecidos.

5. Seguir el algoritmo propuesto por Astrahan, que consiste en estudiar la densidad de cada individuo, entendida como el número de casos que los rodean hasta una distancia d_1, para ordenar las densidades y seleccionar como primera semilla el núcleo que posea mayor densidad; se continua escogiendo los núcleos según este criterio, con el añadido de que deben distar unos de otros una distancia d_2, hasta obtener las k semillas deseadas.

6. La opción propuesta por Ball y Hall es tomar como primer punto semilla el vector de medias de los datos y como subsiguientes semillas cualesquiera puntos que disten de los anteriores al menos una distancia d.

Una vez establecidas las semillas, se procede a la partición del grupo de individuos: se clasifican según su proximidad a los núcleos, clasificando dentro del mismo grupo aquellos con una menor distancia o con mayor proximidad al centroide del cluster.

Una vez se ha realizado la asignación, se repite de nuevo. Los individuos pueden cambiar de grupo, dado que los centroides de los mismos han cambiado tras la incorporación de nuevos individuos al cluster. El procedimiento es iterativo, y se repite hasta que los centroides no cambien en una iteración, o bien cuando se alcance un número de iteraciones pre-establecido.

Existen algunas variantes de este método, como la variante llamada **k-medoides** donde en lugar de utilizar los puntos medios (centroides) se utilizarán los medoides, que son los puntos pertenecientes al grupo de datos (individuos) que minimizan la distancia media entre los demás individuos, es decir, lo más centralizados, mientras que los centroides no tienen por qué pertenecer al conjunto. En el ámbito computacional se conoce como *PAM (Partitioning Around Medoids).*

Otra variante es el llamado método **k-medians**, que escoge las medianas como puntos semilla en lugar de los centroides. Una variante más es la conocida como **k-means++** el cual busca una optimización de las semillas iniciales, de entre todas las posibles combinaciones, para minimizar la varianza dentro de grupos en la elaboración de los clusters.

Para ello se eligen los núcleos en base a una distribución de probabilidad ponderada que determinarán con probabilidad máxima los puntos que minimizan esa varianza.

8.7.2. Otros métodos no jerárquicos

Dentro de los llamados *métodos de reasignación*, como es el propio método de las k-medias, un individuo asignado en un paso determinado a un cluster puede ser asignado a otro grupo en un paso posterior si con ello se consigue una mejor clasificación. El proceso termina cuando no quedan individuos cuya reasignación permita optimizar el resultado. Dentro de estos métodos, además del ya expuesto, se encuentran otros como **Quick-Cluster analysis**, método **de Forgy** o método **de las nubes dinámicas**.

Otros métodos, los de *búsqueda de la densidad*, tienen como objeto la formación de grupos buscando las zonas en las cuales se de una mayor concentración de individuos. Entre ellos tenemos el método **de Wishart**, el de **Taxmap** o el método **Fortin**. Otro método de búsqueda de densidad, el método **de Wolf** aborda el problema considerando que las variables siguen una ley de probabilidad, según la cual los parámetros varían de un grupo a otro y trata de encontrar los individuos que pertenecen a la misma distribución.

Los métodos llamados *directos* permiten clasificar simultáneamente a los individuos y a las variables. El algoritmo más conocido dentro de este grupo es el **Block-Clustering**.

Por su parte, los conocidos como *métodos de reducción de dimensiones* consisten en la búsqueda de factores en el espacio de los individuos; cada factor corresponde a un grupo. Se les conoce como **Análisis Factorial tipo Q**.

8.8. Validación del análisis

La aplicación de este tipo de análisis no requiere hipótesis de normalidad, homoscedasticidad o independencia, como ha ocurrido habitualmente en otras técnicas de análisis multivariante. Por lo tanto, la validación en estos caso se centra en validar de forma gráfica, y utilizando los coeficientes definidos anteriormente, la elección del número adecuado de grupos y de metodología.

Una de las técnicas más habituales para realizar la validación de resultados es la aplicación de un análisi multivariante de la varianza (MANOVA) o bien desarrollar varios análisis de la varianza (ANOVA) sobre cada variable en cada cluster, con el objetivo de ver cuáles difieren significativamente.

Estos métodos no son definitivos y pueden plantear problemas, por lo que no se consideran de forma única. Otros análisis que se pueden realizar con posterioridad es un Análisis Factorial o un Análisis de Componentes Principales para visualizar los clusters gráficamente y observar las diferencias existentes entre ellos.

El estudio del coeficiente de silueta también forma parte de la validación, ya que se puede obtener un gráfico silueta, que consiste simplemente en una representación de las siluetas para todas las observaciones puestos en orden descendente dentro de su cluster. La proporción de superficie contenida en las barras, respecto del área de ancho 1, corresponde al coeficiente de silueta, y mientras más largas sean las barras representadas, mayor validez tiene la estructura encontrada.

8.9. Aplicación con R

Para poder realizar un análisis cluster en R se pueden utilizar las funciones `hclust` y `kmeans` dentro de su paquete básico `stats`. Sin embargo, para un análisis más en profundidad existen diferentes paquetes como `bayesclust`, `clues`, `clustofvar`, `pvclust` o `cluster`.

8.9.1. Funciones de R en el paquete básico stats para análisis cluster jerárquico

Función «dist»

Esta función nos permite calcular una matriz de distancias o similaridades para los individuos o variables de la matriz de datos original. Devuelve el triángulo inferior de la matriz de distancias o similaridades. Su sintaxis es

```
dist(x, method = ‘‘euclidean’’, diag = FALSE, upper = FALSE, p = 2)
```

donde

- x puede ser una matriz o un objeto `data.frame`.
- *method* es la medida de asociación que se usará y podrá ser `euclidean`, `maximum`, `manhattan`, `canberra`, `binary` o `minkowski`.
- *diag* se iguala a un valor lógico para indicar si la diagonal de la matriz de distancia debe ser impreso por `print.dist`. Por defecto el valor es `FALSE`, por lo qeu no se imprime.
- *upper* se iguala a un valor lógico para indicar si el tirángulo superio de la matriz de distancias debe ser impreso por `print.dist`. Por defecto presenta el valor `FALSE`.
- p es la potencia de la distancia de Minkowski.

Función «hclust»

Esta función permite utilizar los distintos tipos de enlace entre clusters (limkage simple, completo, ...) pero solamente se puede trabajar con matrices de distancias o similaridades. Si partimos de una matriz de datos al uso, en primer lugar debemos transformarlos mediante la función `dist`. La sintaxis de esta función es

```
hclust(d,‘‘method=complete’’)
```

donde `d` es la matriz de distancias que se puede calcular utilizando la función `dist` y `method` contiene el método de enlace o unión que se utilizará: `ward.D`, `ward.D2`, `single`, `complete`, `average`, `median` o `centroid`.

Esta función devuelve un objeto de la clase `hclust` que describe los principales resultados producidos por el proceso aglomerativo. Este objeto nos proporciona los elementos siguientes

- *merge:* nos da una matriz de dimensión $(m-1) \times 2$ donde cada fila describe la fusión de los clusters en las sucesivas etapas. Si un elemento en la fila es negativo, entonces la observación se fusiona en esa etapa por primera vez. Si el elemento aparece positivo es porque ese elemento ya estaba en un cluster con más elementos; entradas negativas de `merge` indican aglomeraciones de grupos únicos, y entradas positivas indican las aglomeraciones de los no únicos.

- *height:* es un conjunto de $n-1$ valores reales que nos proporciona los valores de las alturas distancias de fusión según el método usado.

- *labels:* son las etiquetas para cada uno de los elementos que se agrupan.

- *call:* la llamada que produjo el resultado.

- *method:* el método cluster que se ha utilizado.

- *dist.method:* la distancia que se ha utilizado para crear `d`.

Un objeto de tipo `hclust` permite el uso de las funciones genéricas como `print`, `plot` y `rect.hclust`.

Función «plot»

La función `plot` se aplica directamente al objeto `hclust` y nos proporciona el dendrograma.

Función «cutree»

Esta función permite cortar los árboles a partir de objetos de la clase `hclust`, ya sea especificando el número deseado de grupos o especificando la altura de corte. La sintaxis de esta función es

```
cutree(tree, k=NULL, h=NULL)
```

donde los argumentos son

- *tree:* es un árbol obtenido de un objeto `hclust`. Solamente se puede expresar con los componentes `merge`, `height` y `labels` con su respectivo contenido cada uno.
- *k:* es un número entero con el número deseado de grupos (o un vector para explorar varias soluciones).
- *h:* es un entero o vector con las alturas donde el árbol debe ser cortado. Se debe especificar este valor o bien el anterior (`k`) para que se realice el corte.

La función devuelve un vector con la pertenencia a los grupos según se haya especificado con el argumento `k` o el argumento `h`. En caso de haber especificado un vector para explorar diferentes opciones, el resultado será una matriz de pertenencia de cada individuo a cada grupo en cada solución, y el nombre de cada columna indicará el número de grupos.

Función «rect.clust»

Esta función sirve para representar en el dendrograma, unos rectángulos alrededor de las ramas que nos permite diferenciar entre los grupos. En primer lugar el dendrograma se corta en un cierto nivel y entonces se dibuja un rectángulo alrededor de las ramas de cada grupo. La sintaxis de esta función es

```
rect.clust(tree, k=NULL, which=NULL,x=NULL, h=NULL,
    border=2, cluster=NULL)
```

con los siguientes argumentos

- *tree:* un objeto como los obtenidos de `hclust`.
- *k* y *h:* valores numéricos para indicar el corte en grupos, como en la función anterior `cutree`.
- *which* y *x:* vector que selecciona los grupos alrededor de los cuales se dibujará un rectángulo; `which` selecciona los grpos por número (de izquierda a derecha en el dendrogama) y `x` selecciona los grupos que contienen las respectivas coordenadas horizontales. Por defecto `which` es `1:k`.
- *border* es un vector con los colores de la frontera de los rectángulos

Funciones «cor» y «cophenetic»

La función `cor` permite calcular la correlación entre dos vectores o entre dos matrices, mientras que la funcion `cophenetic` nos permitirá hallar la matriz cofenética. Combinando ambos comandos podremos calcular el coeficiente correlación cofenético y que nos peritirá conocer en qué medida representa la estructura final obtenida las militudes o diferencias entre los distintos objetos de estudio. La sintaxis para la función `cor` es

```
cor(x, y=NULL, use=‘‘everything’’, method=‘‘pearson’’)
```

con los argumentos

- *x:* es un vector numérico o matriz (o bien un objeto `data.frame`)ç
- *y:* por defecto toma el valor `NULL`, es un vector, matriz u hoja de datos con dimensiones compatibles con `x`
- *use:* es opcional, sirve para seleccionar un método para calcular la covarianza cuando existan datos faltantes. Sus posibles valores son `everything` (por defecto), `all.obs`, `complete.obs`, `na.or.complete` o `pairwise.complete.obs`.
- *method:* indica qué tipo de coeficiente debe ser calculado. Se pueden elegir `pearson` (es el que se calculará por defecto), `kendall` o `spearman`.

La sintaxis de la función `cophenetic` es simplemente `cophenetic(x)`, donde `x` es un objeto `hclust`.

8.9.2. Funciones de R en el paquete básico stats para análisis no jerárquico

Función «kmeans»

La sintaxis de esta función es

```
kmeans(x, centers, iter.max=10, nstart=1,
    algorithm=‘‘Hartigan-Wong’’, trace=FALSE)
```

con los siguientes argumentos

- *x:* es la matriz de datos
- *centers:* se puede indicar un valor k de número de grupos o un conjunto de centros iniciales o semillas para la primera partición de grupos. Si se indica un valor k, se elegirán aleatoriamente mediante muestreo aleatorio simple un grupo de k individuos para que conformen las semillas y se conformen los grupos iniciales.

- *iter.max:* es el número máximo de iteraciones que se va a permitir.
- *nstart:* indica el número de veces que se realizará el proceso, cada vez con una asignación aleatoria diferente. Es recomendable que este valor sea elevado, entre 25 y 50, para eviar malos resultados que se deban a la mala selección inicial de centroides o semillas.
- *algorithm:* se puede seleccionar entre diferentes variantes del método de las k-medias. Las opciones disponibles son `Hartigan-Wong`, `Lloyd`, `Forgy` o `MacQueen`.
- *trace:* en caso de tomar el valor lógico `TRUE` se devuelve un seguimiento de las iteraciones del proceso.

La función devuelve un objeto de tipo `kmeans` que describe los principales resultados producidos por el proceso. Este objeto nos proporciona una lista de elementos:

- *clusters:* es un vector de enteros (de 1 a k) indicando a qué cluster es asignado cada punto.
- *centers:* una matriz con los centroides de los clusters.
- *totss:* la suma total de cuadrados.
- *withinss:* vector de suma de cuadrados dentro de los grupos, uno de los componentes por grupo.
- *tot.withinss:* suma total de cuadrados dentro de los grupos, es decir, la suma de los `withinss`.
- *betweens:* es la suma de cuadrados entre los grupos.
- *size:* el número de individuos o puntos en cada grupo, es decir, su tamaño.
- *iter:* el número de iteraciones realizadas.
- *infault:* indicador de un posible error en el algoritmo (es un valor entero).

8.9.3. Ejemplo de aplicación

En el ejemplo de aplicación utilizaremos unos datos referentes a los hábitos culturales de la población de las comunidades autónomas españolas, junto con Ceuta y Melilla. Los datos se refieren a los resultados obtenidos en que se pueden obtener en la Encuesta de hábitos y prácticas culturales 2018-2019 `https://www.culturaydeporte.gob.es/servicios-al-ciudadano/estadisticas/cultura/mc/ehc/2018-2019/presentacion.html`.

Para este ejemplo se seleccionan los datos acerca del porcentaje de habitantes que van a conciertos de música clásica (*MClasica*), porcentaje que acude a algún concierto de música actual (*MActual*), porcentaje de la población que compra música en algún formato (*CompraMusica*), porcentaje de la población que escucha la radio (*Radio*), porcentaje que acude a eventos deportivos (*Deportivos*), porcentaje que acude a ferias (*Ferias*), porcentaje que visita parques de atracciones (*Parques*) y porcentaje que visita parques acuáticos (*Acuaticos*).

Con estos datos agruparemos las comunidades utilizando el análisis cluster.

En primer lugar debemos cargar los datos

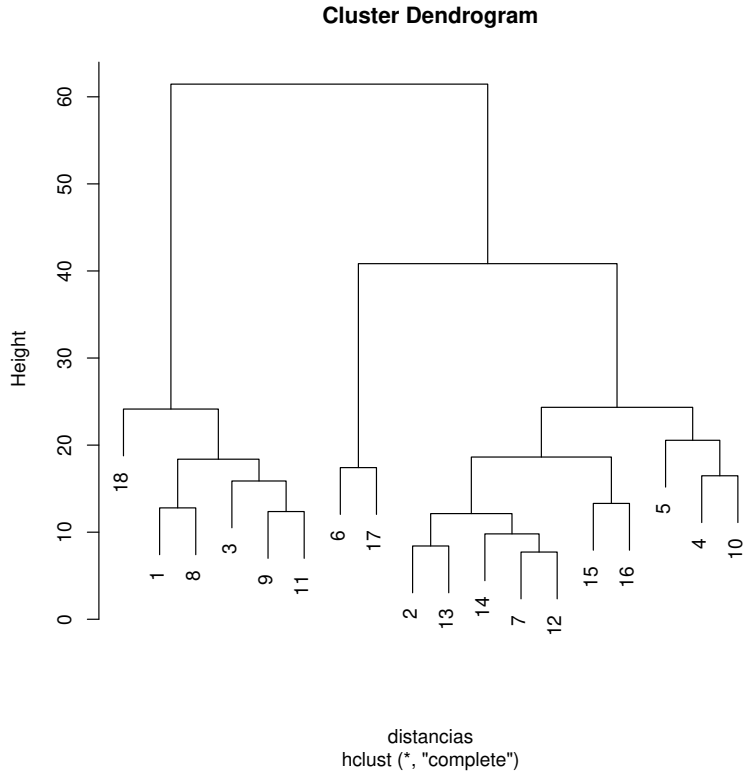
```
datos<-read.table("datoscluster1.txt", header=TRUE, sep="\t")
datos
```

	Comunidad_Autonoma	MClasica	MActual	CompraMusica	Radio	Deportivos	Ferias	Parques	Acuaticos
1	Andalucía	8.3	34.0	8.6	76.7	36.3	57.4	21.1	23.6
2	Aragón	9.7	30.1	7.6	82.8	33.3	43.1	18.0	16.3
3	Asturias (Principado de)	9.0	31.1	4.7	70.0	34.7	57.8	11.9	11.6
4	Baleares (Illes)	9.3	32.3	4.3	76.4	29.5	47.7	22.4	26.3
5	Canarias	6.6	23.1	3.4	84.3	26.4	39.3	13.0	19.1
6	Cantabria	8.3	36.8	5.6	77.3	19.4	30.4	12.1	7.0
7	Castilla y León	9.2	33.3	8.3	76.5	30.8	48.8	16.2	11.8
8	Castilla-La Mancha	11.4	37.4	6.0	76.6	32.3	60.2	18.1	14.4
9	Cataluña	7.0	24.1	4.9	72.9	36.7	62.4	23.6	12.4
10	Comunitat Valenciana	10.1	24.8	4.8	71.2	31.5	37.6	21.5	19.2
11	Extremadura	4.6	31.8	3.3	68.7	32.5	61.7	20.5	17.2
12	Galicia	10.0	33.9	7.0	77.8	33.5	44.1	13.7	15.4
13	Madrid (Comunidad de)	12.7	27.7	6.5	78.5	32.6	43.5	22.7	13.9
14	Murcia (Región de)	11.1	34.1	6.4	77.8	27.1	43.5	16.9	17.6
15	Navarra (Comunidad Foral de)	10.9	33.5	6.0	83.4	31.3	34.7	22.2	16.0
16	País Vasco	12.4	39.8	10.4	78.5	35.6	37.0	15.8	13.3
17	Rioja (La)	7.7	28.4	7.3	76.6	25.1	18.1	8.7	7.0
18	Ceuta y Melilla	7.4	33.7	5.6	65.8	35.7	67.6	22.9	28.1

Una vez cargados los datos, observamos que las unidades son comparables, ya que todas las variables representan porcentaje de población. Esto quiere decir que no es necesaria ninguna transformación de los datos previa a la aplicación del análisis.

Calculamos la matriz de distancias, utilizando la distancia euclídea y aplicamos el algoritmo de cluster jerárquico sobre ella para después utilizar la función `plot` y obtener el dendrograma correspondiente.

```
distancia<-dist(datos)
jerarquicobasico<-hclust(distancia)
plot(jerarquicobasico)
```



Según este primer resultado, donde hemos aplicado las opciones por defecto de las funciones `dist` y `hclust`, parece haber al menos tres grupos naturales; el primero de ellos agrupa Ceuta y Melilla con Andalucía, Castilla-La Mancha, Asturias, Cataluña y Extremadura. El segundo de los grupos contiene Cantabria y La Rioja, y el tercero de los grupos está formado por Aragón, Madrid, Murcia, Castilla y León, Galicia, Navarra, País Vasco, Canarias, Baleares y Comunidad Valenciana.

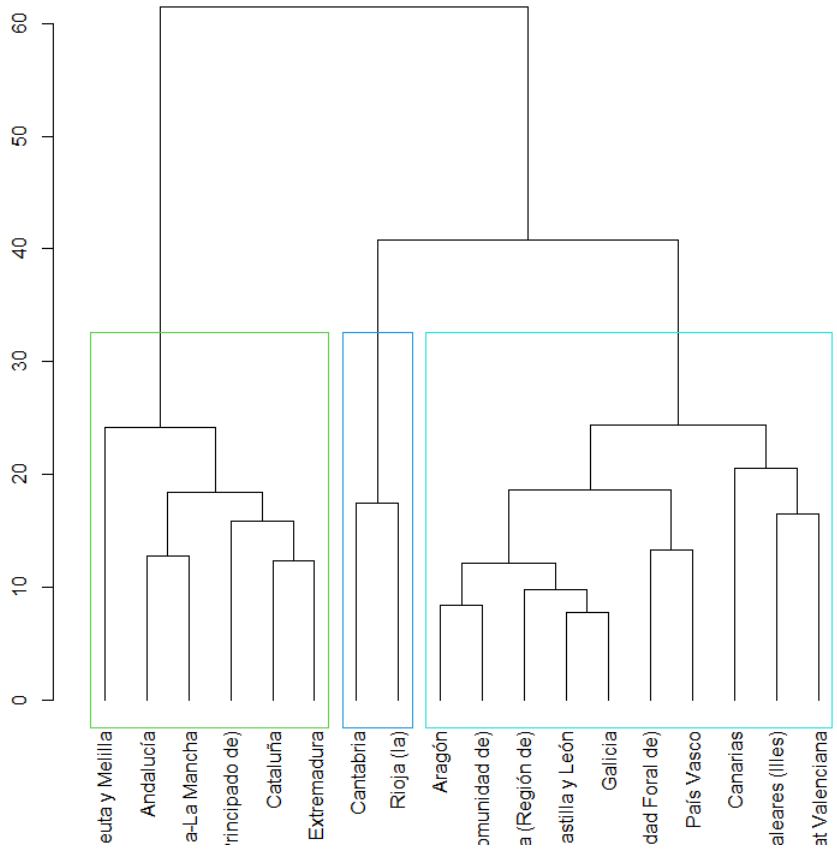
Si utilizamos la función `cutree` el árbol se corta para formar el número de grupos deseado, por ejemplo

```
solucion_basica<-cutree(tree, k=3, h=NULL)
solcion_basica
                  Andalucía                      Aragón     Asturias (Principado de)
                          1                          2                            1
          Baleares (Illes)                   Canarias                    Cantabria
                          2                          2                            3
           Castilla y León          Castilla-La Mancha                     Cataluña
                          2                          1                            1
       Comunitat Valenciana                 Extremadura                      Galicia
                          2                          1                            2
     Madrid (Comunidad de)          Murcia (Región de) Navarra (Comunidad Foral de)
                          2                          2                            2
               País Vasco                  Rioja (la)             Ceuta y Melilla
                          2                          3                            1
```

Para que el dendrograma tenga una forma más amigable se pueden utilizar las siguientes sentencias:

```
row.names(datos)<-as.character(datos$Comunidad_Autonoma)
distancia<-dist(datos)
jerarquicobasico<-hclust(distancia)
plot(as.dendrogram(jerarquicobasico))
rect.hclust(jerarquicobasico, k = 3, border = 3:5)
```

Con la primera orden lo que hacemos es ponerle a cada objeto o individuo como etiqueta su nombre, que está almacenado en la primera columna y, a continuación, repetimos el análisis, usando `as.dendrogram` para que el gráfico posicione todos los individuos en la misma horizontal, y usamos `rect.hclust` para enmarcar los tres grupos que obviamente se obtienen. Después de aplicar estas órdenes se obtiene el gráfico



Para calcular el coeficiente de correlación cofenético utilizamos

```
cor(x = distancia, cophenetic(jerarquicobasico))
[1] 0.6616181
```

Aplicamos ahora el método de las k-medias básico del paquete `stats` al mismo conjunto de datos, indicando que se hará una agrupación en 3 clusters.

```
datoscluster2<-datoscluster1[,-1]
kmeans(datoscluster2,centers=3,nstart=25)

K-means clustering with 3 clusters of sizes 2, 6, 10

Cluster means:
      X  Mactual CompraMusica    Radio Deportivos   Ferias  Parques Acuaticos
1  8.00 32.60000     6.450000 76.95000      22.25 24.25000 10.40000   7.00000
2  7.95 32.01667     5.516667 71.78333      34.70 61.18333 19.68333  17.88333
3 10.20 31.26000     6.470000 78.72000      31.16 41.93000 18.24000  16.89000

Clustering vector:
                    Andalucía                       Aragón     Asturias (Principado de)
                            2                           3                            2
            Baleares (Illes)                     Canarias                    Cantabria
                            3                           3                            1
             Castilla y León           Castilla-La Mancha                     Cataluña
                            3                           2                            2
        Comunitat Valenciana                 Extremadura                      Galicia
                            3                           2                            3
       Madrid (Comunidad de)           Murcia (Región de) Navarra (Comunidad Foral de)
                            3                           3                            3
                  País Vasco                    Rioja (la)              Ceuta y Melilla
                            3                           1                            2

Within cluster sum of squares by cluster:
[1] 134.820 636.145 965.519
 (between_SS / total_SS =  65.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

En este caso, la solución en 3 clusters es similar a la obtenida con el método jerárquico anterior, donde se utilizaba la distancia euclídea y el método del *linkage completo*. Además, tenemos información acerca de la suma de cuadrados en los clusters y el porcentaje que supone la suma de cuadrados entre clusters dividida entre la suma de cuadrados total. También se tiene información acerca de los centroides de los grupos observados, de forma que podemos ver que las comunidades del grupo 1 se diferencia sobre todo en la proporción de habitantes que acuden a eventos deportivos, ferias, parques de atracciones y parques acuáticos, ya que los porcentajes referidos a la asistencia a conciertos, compra de música y escucha de radio son más similares entre clusters diferentes.

Bibliografía

[1] Bridges CC. *Hierarchical Cluster Analysis*; Psychological Reports; 18(3):851-854, 1966.

[2] Carrasco, J.L.; Hernán, M.A. *Estadística multivariante en las ciencias de la vida. Fundamentos, métodos y aplicación*; Ciencia 3, D.L., 1993.

[3] Kassambara, A. *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol 1). Sthda, 2017.

[4] Cuadras, C.M. *Nuevos Métodos del Análisis Multivariante*; CMC, 2018.

[5] Hamerly, G., Elkan, C. *Learning the k in k-means*; Advances in neural information processing systems, 16:281-288, 2004.

[6] Gnanadeskian, R. *Methods for statistical data analysis of multivariate observations*; (Vol 321) John Wiley & Sons, 2011.

[7] Rencher, A.C. *Methods of Multivariate Analysis*; Wiley, N. York, 1995.

[8] Cluster Analysis in R, R-Bloggers `https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/`