Análisis de Conglomerados

(Cluster analysis)

Aurea Grané
Departamento de Estadística
Universidad Carlos III de Madrid

Antecedente histórico

Las representaciones que se estudian en este tema tienen su antecedente histórico en el sistema taxonómico de los seres vivos debido al naturalista Carl von Linné.

También conocido como Carlos Linneo o Carolus Linnaeus, es llamado con frecuencia el Padre de la Taxonomía. Todavía se usa (aunque con muchos cambios) su sistema para nombrar, ordenar y clasificar los organismos vivos.

Sus ideas sobre la clasificación han influenciado a generaciones de biólogos mientras vivía y mucho después de su muerte, aún a aquellos que se oponían a los fundamentos filosóficos y teológicos de su trabajo.

Ejemplo ilustrativo

Algunas frutas comunes (ciruela, cereza, melocotón, albaricoque, pera, manzana, níspero) pertenecen a la familia de las Rosáceas,

género *Prunus* (P. domestica = ciruela, P. avium = cereza, P. persica = melocotón, P. ameniaca = albaricoque),

género *Pyrus* (P. communis = pera, P. malus = manzana)

y al género *Mespilus* (M. germanica = níspero).

Los botánicos, siguiendo criterios naturalistas, clasifican estas frutas de la forma siguiente:

Familia	**G**énero	**E**specie
Rosáceas	*Prunus*	*P.domestica*(ciruela)
		P.avium(cereza)
		P.persica(melocotón)
		P.ameniaca(albaricoque)
	Pyrus	*P.communis*(pera)
		P.malus(manzana)
	Mespilus	{*M.germanica*(níspero)

Objetivo

Sea $\mathcal{E}$ un conjunto de n objetos o individuos sobre los que se ha calculado alguna medida de distancia.
Sea $\mathbf{D} = (\delta_{ij})_{1 \leq i,j \leq n}$ la matriz de distancias entre estos n individuos.

El objetivo del análisis de conglomerados (o *cluster analysis*) es la **clasificación** (no supervisada) de los elementos de $\mathcal{E}$, es decir, su agrupación en clases disjuntas, que se denominan **conglomerados** (o *clusters*).

Si estas clases se agrupan sucesivamente en clases de un nivel superior, el resultado es una estructura jerárquica de conglomerados, que puede representarse gráficamente mediante un árbol, llamado **dendrograma**.

Ultramétricas

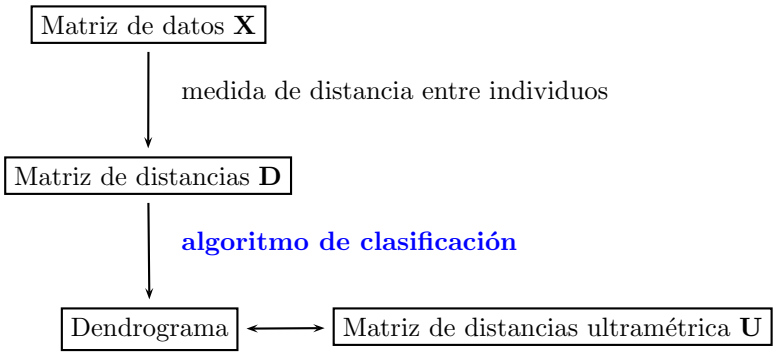
Se dice que una matriz de distancias $\mathbf{D}$ es **ultramétrica** si para todos los elementos del conjunto $\mathcal{E}$ se verifica que:

$$\left.\begin{array}{ll} \delta_{ij} = \delta_{ji}, & \text{para todo } i,j, \\ \delta_{ii} = 0, & \text{para todo } i, \end{array}\right\} \text{ disimilaridad o casi-métrica}$$

y además verifican la **desigualdad ultramétrica**:

$$\delta_{ij} \leq \max\{\delta_{ik}, \delta_{kj}\}, \quad \text{para todo } i,j,k.$$

Puede demostrarse que a cada dendrograma le corresponde una matriz de distancias ultramétrica y viceversa.



Clasificación jerárquica

Algoritmos de tipo divisivo: Se parte de un único conglomerado, formado por el conjunto de todos los objectos a clasificar, y se subdivide en particiones cada vez más finas.

Algortimos de tipo aglomerativo: Se parte de n conglomerados, de un solo objeto cada uno, y se van agrupando en conjuntos sucesivamente mayores.

Algoritmos de tipo aglomerativo

Se dispone de un conjunto $\mathcal{E}$ de n elementos u objetos y de una matriz de distancias $\mathbf{D} = (\delta_{ij})_{1 \leq i,j \leq n}$ entre ellos.

Idea: se juntan los elementos o conglomerados más próximos, y se procura obtener distancias ultramétricas.

1. Se empieza con la partición: $\mathcal{E} = \{1\} + \{2\} + \ldots + \{n\}$.
2. Sean i, j los dos elementos más próximos, es decir, $\delta_{ij} = \min \delta_{kl}$. Éstos se unen dando lugar a un nuevo conglomerado:

   $$\{i\} \cup \{j\} = \{i, j\}$$

   y se define la distancia del conglomerado $\{i, j\}$ al resto de elementos del conjunto $\mathcal{E}$:

   $$\delta'_{k,(ij)} = f(\delta_{ik}, \delta_{jk}), \quad k \neq i, j,$$

   donde f es una función adecuada.

3. Se considera la nueva partición: $\mathcal{E} = \{1\} + \ldots + \{i, j\} + \ldots + \{n\}$ y se repiten los pasos 2 y 3, hasta que todos los elementos estén contenidos en un único conglomerado.

La función f (paso 2) se define adecuadamente de manera que se cumpla la propiedad ultramétrica. Los distintos **métodos de clasificación jerárquica** dependen de la elección de la función f:

- método del mínimo (o *single linkage*). Se toma f igual al mínimo:

  $$\delta'_{k,(ij)} = \min(\delta_{ik}, \delta_{jk}), \quad k \neq i, j,$$

- método del máximo (o *complete linkage*). Se toma f igual al máximo:

  $$\delta'_{k,(ij)} = \max(\delta_{ik}, \delta_{jk}), \quad k \neq i, j,$$

- método de la media.

  $$\delta'_{k,(ij)} = \frac{1}{2}(\delta_{ik} + \delta_{jk}), \quad k \neq i, j,$$

- UPGMA (*Unweighted Pair Group Method using arithmetic Averages*), que utiliza medias ponderadas según el número de elementos que hay en cada conglomerado. Si E_i, E_j, E_k son conglomerados de n_i, n_j, n_k elementos, respectivamente y E_i, E_j son los conglomerados más próximos, entonces

  $$\delta'(E_k, E_i \cup E_j) = \frac{n_i}{n_i + n_j}\,\delta(E_i, E_k) + \frac{n_j}{n_i + n_j}\,\delta(E_j, E_k)$$

Si la matriz de distancias original $\mathbf{D}$ no cumple la propiedad ultramétrica, los distintos métodos de clasificación darán lugar a matrices ultramétricas distintas y, por tanto, a representaciones jerárquicas distintas.

Ejemplo 1: Problema 6.2

Distancias por carretera (en km) entre ciudades.

	Barcelona	Madrid	San Sebastián	Sevilla	Valencia
Barcelona	0	639	606	1181	364
Madrid	639	0	474	542	**355**
San Sebastián	606	474	0	908	597
Sevilla	1181	542	908	0	679
Valencia	364	**355**	597	679	0

Etapa cero: $C_0 = \{B\} + \{M\} + \{SS\} + \{S\} + \{V\}$

Etapa uno: $C_1 = \{B\} + \{M, V\} + \{SS\} + \{S\}$ y se recalculan las distancias (por ejemplo, mediante el método del mínimo) del conglomerado {M,V} al resto.

$$\delta_{(MV),B} = \min\{\delta_{M,B}, \delta_{V,B}\} = \min\{639, 364\} = 364,$$
$$\delta_{(MV),SS} = \min\{\delta_{M,SS}, \delta_{V,SS}\} = \min\{474, 597\} = 474,$$
$$\delta_{(MV),S} = \min\{\delta_{M,S}, \delta_{V,S}\} = \min\{542, 679\} = 542,$$

de manera que la matriz de distancias ha quedado:

Paso 0	B	M	SS	S	V
B	0	639	606	1181	364
M		0	474	542	**355**
SS			0	908	597
S				0	679
V					0

$\rightarrow$

Paso 1	B	(M,V)	SS	S
B	0	**364**	606	1181
(M,V)		0	474	542
SS			0	908
S				0

Etapa dos: $C_2 = \{B, M, V\} + \{SS\} + \{S\}$ y se recalculan las distancias del conglomerado {B,M,V} al resto de individuos.

$$\delta_{(BMV),SS} = \min\{\delta_{B,SS}, \delta_{(MV),SS}\} = \min\{606, 474\} = 474,$$
$$\delta_{(BMV),S} = \min\{\delta_{B,S}, \delta_{(MV),S}\} = \min\{1181, 542\} = 542,$$

y la matriz de distancias ha quedado:

Paso 2	(B,MV)	SS	S
(B,MV)	0	**474**	542
SS		0	908
S			0

$\rightarrow$

Paso 3	(BMV,SS)	S
(BMV,SS)	0	542
S		0

Etapa tres: $C_3 = \{B, M, V, SS\} + \{S\}$ y se recalculan las distancias del conglomerado {B,M,V,SS} al resto de individuos.

$$\delta_{(BMVSS),S} = \min\{\delta_{(BMV),S}, \delta_{SS,S}\} = \min\{542, 908\} = 542,$$

Etapa cuatro: $C_4 = \{B, M, V, SS, S\}$

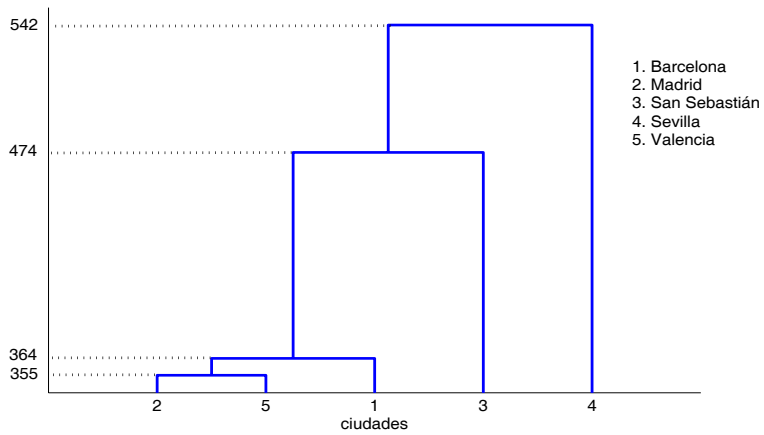
Resumen del algoritmo de clasificación

Etapa	distancias	clasificación / conglomerados
0	-	$C_0 = \{B\} + \{M\} + \{SS\} + \{S\} + \{V\}$
1	$\delta_{M,V} = \mathbf{355}$	$C_1 = \{B\} + \{M, V\} + \{SS\} + \{S\}$
2	$\delta_{B,MV} = \mathbf{364}$	$C_2 = \{B, M, V\} + \{SS\} + \{S\}$
3	$\delta_{BMV,SS} = \mathbf{474}$	$C_3 = \{B, M, V, SS\} + \{S\}$
4	$\delta_{BMVSS,S} = \mathbf{542}$	$C_4 = \{B, M, V, SS, S\}$

Utilizando las distancias a las que se forman los conglomerados se reconstruye la **matriz de distancias ultramétrica**:

	Barcelona	Madrid	San Sebastián	Sevilla	Valencia
Barcelona	0	364	474	542	364
Madrid		0	474	542	355
San Sebastián			0	542	474
Sevilla				0	542
Valencia					0

Dendrograma (método del mínimo) de las ciudades.



Correlación cofenética

Como ocurría en el caso euclídeo, en general, una matriz de distancias $\mathbf{D}$, obtenida a partir de una matriz de datos multivariantes $\mathbf{X}$, no cumple la propiedad ultramétrica.

Esto da lugar al problema de aproximar la matriz de distancias $\mathbf{D} = (\delta_{ij})$ con una matriz ultramétrica $\mathbf{U} = (u_{ij})$ según algún criterio de proximidad adecuado.

La medida de proximidad que se utiliza es la **correlación cofenética**, que es el coeficiente de correlación lineal (de Pearson) entre los $n\,(n-1)/2$ pares de distancias (δ_{ij}, u_{ij}), para $1 \leq i < j \leq n$.

Este coeficiente vale 1 cuando ambas matrices son proporcionales (iguales). Esto equivale a decir que la matriz $\mathbf{D}$ ya cumple la propiedad ultramétrica y, por tanto, la clasificación es exacta.

Ejemplo 2: Problema 6.3

Sea $\mathbf{D}^{(2)}$ la matriz de distancias de Bhattacharyya obtenida para los datos siguientes:

	Población	grupo A	grupo AB	grupo B	grupo O
1.	francesa	0.21	0.06	0.06	0.67
2.	checa	0.25	0.04	0.14	0.57
3.	germánica	0.22	0.06	0.08	0.64
4.	vasca	0.19	0.04	0.02	0.75
5.	china	0.18	0.00	0.15	0.67
6.	ainu	0.23	0.00	0.28	0.49
7.	esquimal	0.30	0.00	0.06	0.64
8.	negra USA	0.10	0.06	0.13	0.71
9.	española	0.27	0.04	0.06	0.63
10.	egipcia	0.21	0.05	0.20	0.54

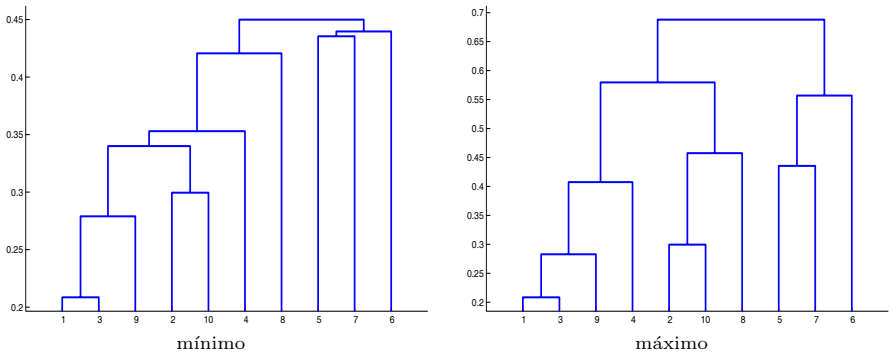
a) ¿Es $\mathbf{D}$ ultramétrica?

Calculamos préviamente la matriz de distancias $\mathbf{D}$:

```
D=[  0  0.3959  0.2086  0.3530  0.5351  0.6298  0.5121  0.4301  0.2828  0.4695
0.3959       0  0.3400  0.5162  0.4733  0.5104  0.4976  0.4575  0.3693  0.2995
0.2086  0.3400       0  0.4074  0.5211  0.6030  0.5107  0.4206  0.2789  0.4227
0.3530  0.5162  0.4074       0  0.5675  0.6879  0.5106  0.5055  0.3895  0.5796
0.5351  0.4733  0.5211  0.5675       0  0.4397  0.4354  0.5206  0.5151  0.4991
0.6298  0.5104  0.6030  0.6879  0.4397       0  0.5569  0.6084  0.6035  0.4921
0.5121  0.4976  0.5107  0.5106  0.4354  0.5569       0  0.6007  0.4499  0.5680
0.4301  0.4575  0.4206  0.5055  0.5206  0.6084  0.6007       0  0.4938  0.4469
0.2828  0.3693  0.2789  0.3895  0.5151  0.6035  0.4499  0.4938       0  0.4702
0.4695  0.2995  0.4227  0.5796  0.4991  0.4921  0.5680  0.4469  0.4702  0 ];
```

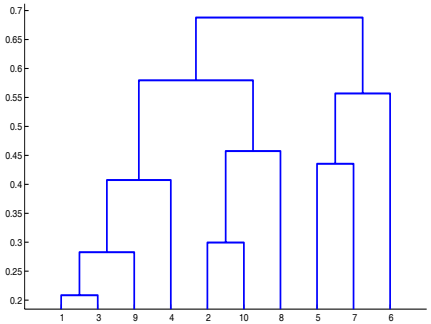
No es ultramétrica puesto que, por ejemplo,
$\delta_{1,6} = 0.6298 > \max\{\delta_{1,3}, \delta_{3,6}\} = \max\{0.2086, 0.6030\}$.

b) Clasificación jerárquica con los métodos del mínimo, del máximo y UPGMA

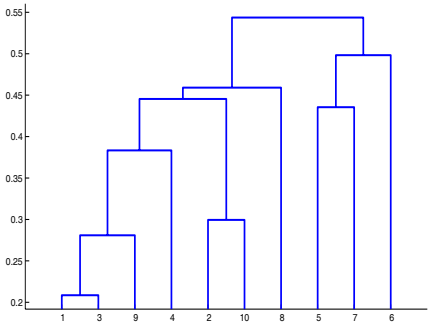


mínimo máximo

El método del mínimo tiende a contraer el espacio (observad los valores del índice de la jerarquía, que se encuentran representados en el eje vertical del gráfico), mientras que el método de máximo tiende a dilatar el espacio.



máximo



UPGMA

Las clasificaciones que se obtienen mediante los métodos del máximo y UPGMA son muy parecidas.

c) Calcular la correlación cofenética en cada caso.

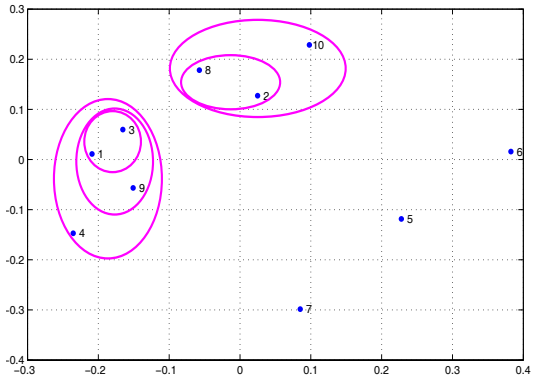
La correlación cofenética es el coeficiente de correlación lineal de Pearson entre los elementos de la matriz de distancias original y los elementos de la matriz de distancias ultramétrica.

En este caso, obtenemos:

`c_min=0.7910`, `c_max=0.8132` y `c_UPGMA=0.8413`,

indicando que el método UPGMA es el que menos distorsiona (de los tres que hemos visto) la matriz de distancias original.

d) Comparar los dendrogramas con la representación en coordenadas principales.



{1,3,9,4}: poblaciones europeas {*francesa, germánica, española, vasca*}
{2,8,10}: poblaciones {*checa, negra USA, egipcia*}
{5,6,7}: poblaciones {*china, ainu, esquimal*}.

Los dendrogramas obtenidos con el método del máximo y con el método UPGMA son los que mejor reflejan estas agrupaciones.

Clasificación no jerárquica

Una clasificación no jerárquica de n objetos en relación a una matriz de datos $\mathbf{X}$, consiste en obtener g grupos homogéneos y excluyentes (conglomerados).

Si tenemos g conglomerados, estamos en la misma situación del Tema 5 (MANOVA), y podemos considerar la descomposición de la variabilidad total $\mathbf{T} = \mathbf{B} + \mathbf{W}$.

Una partición en g conglomerados que maximice $\mathbf{B}$ o minimice $\mathbf{W}$, en relación a algún criterio, dará una solución al problema, puesto que tendremos máxima dispersión entre conglomerados. Algunos criterios, justificados por el MANOVA son:

a) minimizar $\mathsf{tr}(\mathbf{W})$,
b) minimizar $|\mathbf{W}|$,
c) minimizar $\Lambda = |\mathbf{W}|/|\mathbf{T}|$,
d) maximizar $\mathsf{tr}(\mathbf{W}^{-1}\,\mathbf{B})$.

La cantidad de maneras diferentes de agrupar n objetos en g grupos es del orden de $g^n/g!$, número muy grande incluso para valores moderados de n y g.

Por ejemplo, si $n = 50$ y $g = 3$, se formarían más de 10^{23} conglomerados !!! Es necesario seguir algún algoritmo de agrupación.

Algoritmo de k-medias (*k-means*):

1. Seleccionar g puntos del espacio $\mathbb{R}^p$ como centros (centroides) de los grupos iniciales:
   a) asignando aleatoriamente los elementos a los grupos y calculando así los centroides,
   b) tomando como centroides los g puntos más alejados entre sí,
   c) construyendo unos grupos iniciales con información *a priori* y calculando sus centroides, o bien seleccionando los centroides *a priori*.

2. Calcular las distancias euclídeas (δ_E) de cada elemento a los g centroides, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las cordenadas del nuevo centroide.
3. Repetir el paso 2, calculando cada vez la cantidad $|\mathbf{W}|$, o el criterio de optimización escogido.
4. Parar cuando $|\mathbf{W}|$ ya no disminuye, o el criterio de optimización escogido ya no mejora.

Se demuestra que la suma de cuadrados de las distancias euclídeas de los elementos de cada grupo al centroide disminuye a cada paso:

$$\sum_{k=1}^{g} \sum_{i=1}^{n_k} \delta_E^2(\mathbf{x}_{ki}, \bar{\mathbf{x}}_k).$$