# Modelos de Soporte No Supervisado

COMPARATIVO DE DENDOGRAMAS

# Cluster jerárquico aglomerativo

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:

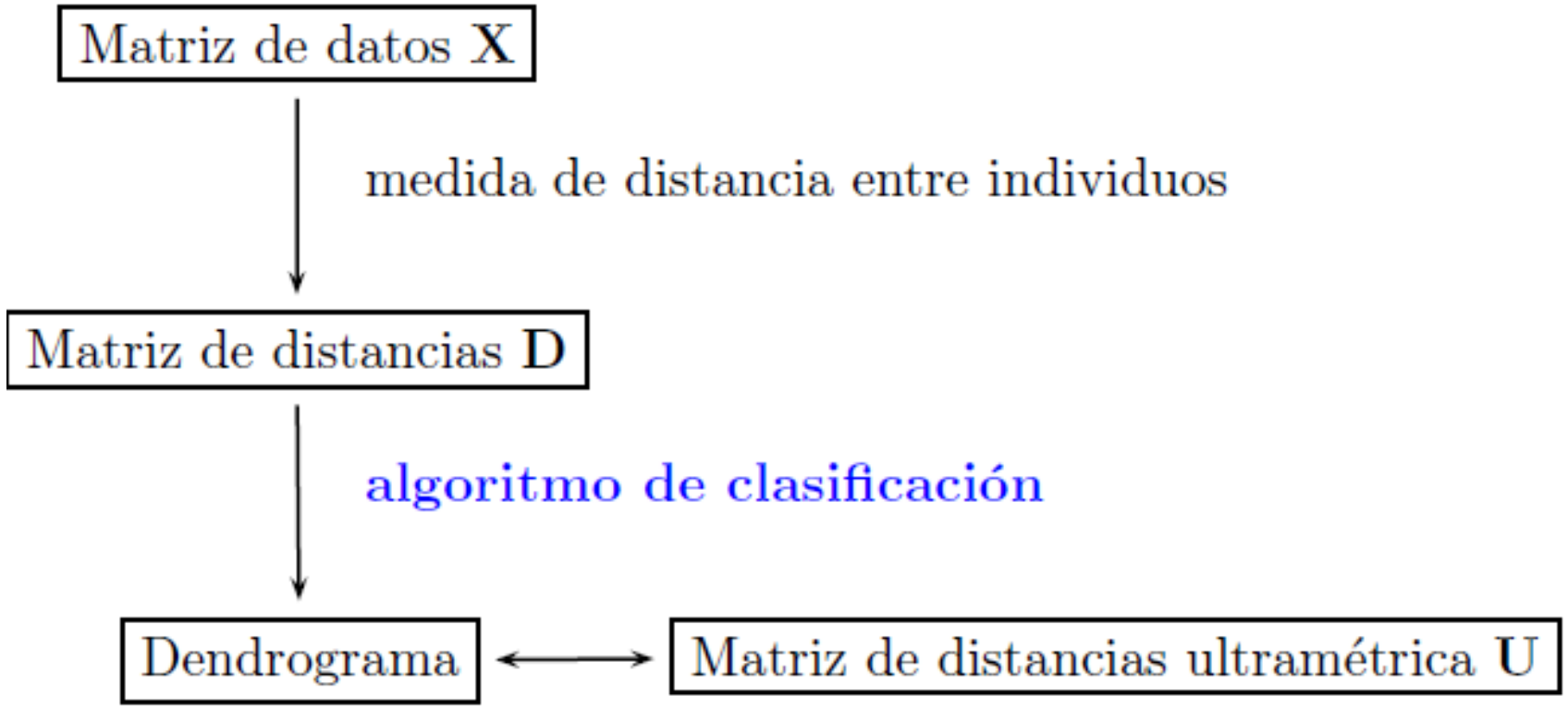
$$\left.\begin{array}{ll} \delta_{ij} = \delta_{ji}, & \text{para todo } i, j, \\ \delta_{ii} = 0, & \text{para todo } i, \end{array}\right\} \text{ disimilaridad o casi-métrica}$$

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.

# Cluster jerárquico aglomerativo



# Cluster jerárquico 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.

# Cluster jerárquico aglomerativo

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:

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

# Cluster jerárquico aglomerativo

- 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.

# Cluster jerárquico aglomerativo

**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.

# Cluster jerárquico aglomerativo

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

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.

# Cluster jerárquico aglomerativo

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

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

# Cluster jerárquico aglomerativo

Resumen del algoritmo de clasificación

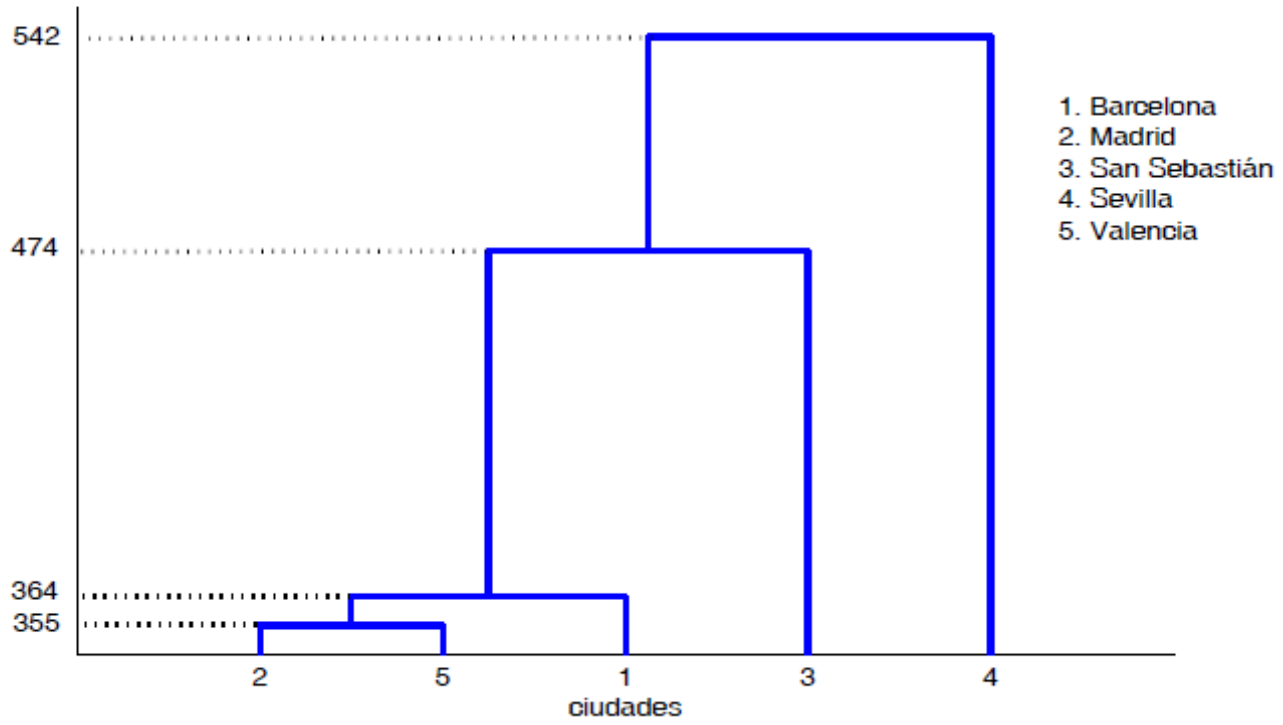
| Etapa | distancias | clasificación / conglomerados |
|---|---|---|
| 0 | - | $C_0 = \{B\} + \{M\} + \{SS\} + \{S\} + \{V\}$ |
| 1 | $\delta_{M,V} = \mathbf{355}$ | $C_1 = \{B\} + \{M, V\} + \{SS\} + \{S\}$ |
| 2 | $\delta_{B,MV} = \mathbf{364}$ | $C_2 = \{B, M, V\} + \{SS\} + \{S\}$ |
| 3 | $\delta_{BMV,SS}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

# Cluster jerárquico aglomerativo

Dendrograma (método del mínimo) de las ciudades.



# Cluster jerárquico aglomerativo: Correlación cophenetic

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 $(\delta_{ij}, u_{ij})$, 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.

# Cluster jerárquico aglomerativo: Correlación cophenetic

Sea D^2 la matriz de distancias para los siguientes datos:

| | 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 **D** ultramétrica?

# Cluster jerárquico aglomerativo: Correlación cophenetic

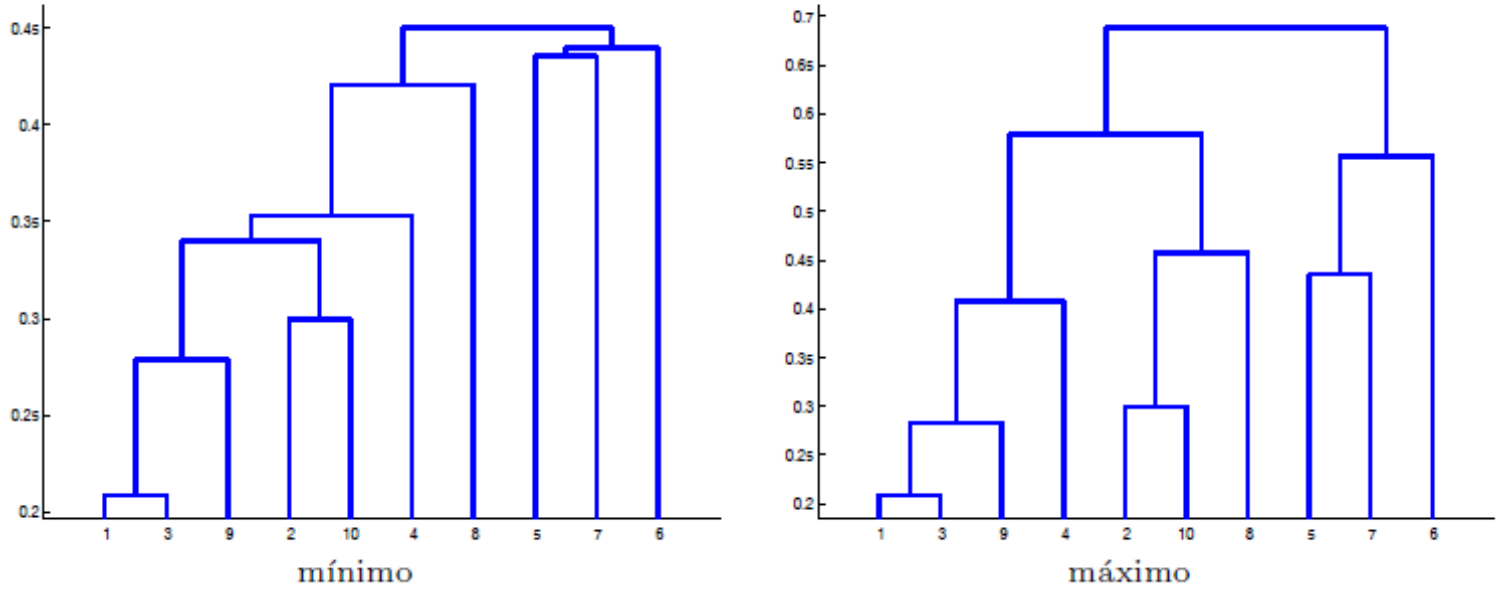
Calculamos préviamente la matriz de distancias $\mathbf{D}$:

```
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 ];
```

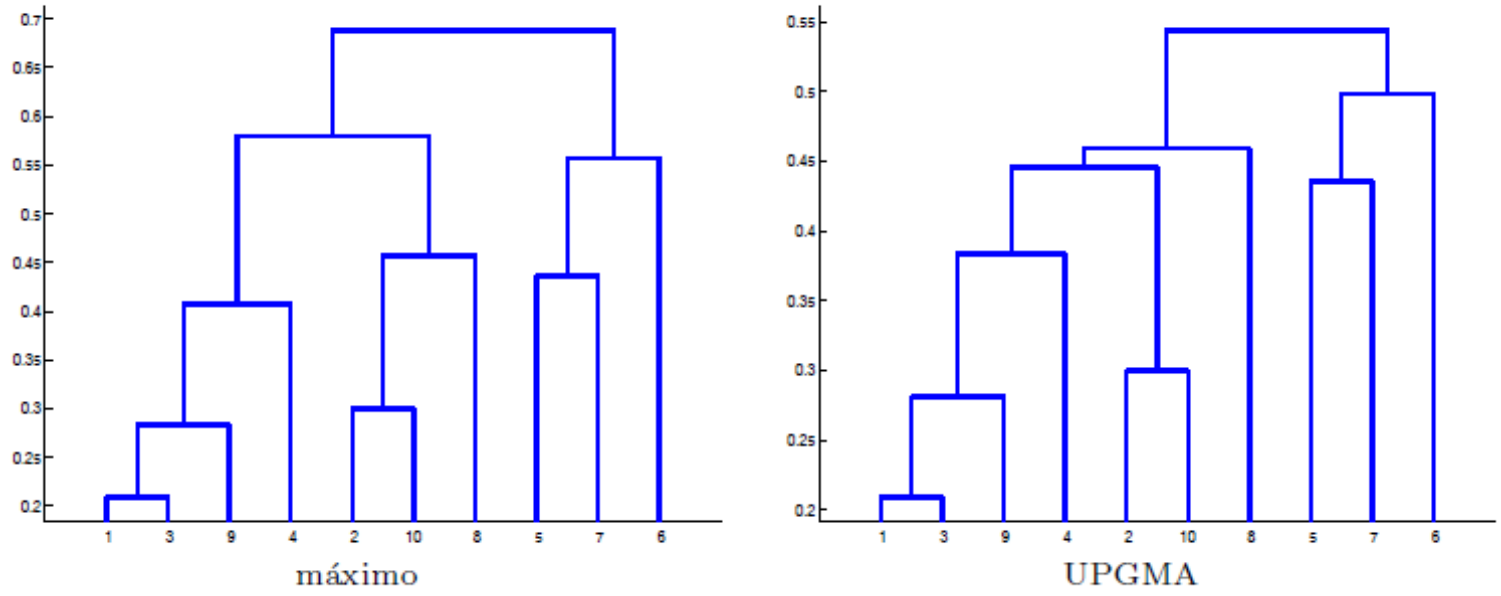
No es ultramétrica puesto que, por ejemplo,

$\delta_{1,6} = 0.6298 > \max\{\delta_{1,3}, \delta_{3,6}\} = \max\{0.2086, 0.6030\}.$

# Cluster jerárquico aglomerativo: Correlación cophenetic



# Cluster jerárquico aglomerativo: Correlación cophenetic



# Cluster jerárquico aglomerativo: Correlación cophenetic

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.

# Cluster jerárquico aglomerativo: Correlación cophenetic

Suppose that the original data $\{X_i\}$ have been modeled using a cluster method to produce a dendrogram $\{T_i\}$; that is, a simplified model in which data that are 'close' have been grouped into a hierarchical tree. Define the following distance measures. $x(i,j) = |X_i - X_j|$, the ordinary Euclidean distance between the $i$th and $j$th observations. $t(i,j)$ = the dendrogrammatic distance between the model points $T_i$ and $T_j$. This distance is the height of the node at which these two points are first joined together. Then, letting $x$ be the average of the $x(i,j)$, and letting $t$ be the average of the $t(i,j)$, the cophenetic correlation coefficient $c$ is defined as in (1) [9].

$$c = \frac{\sum_{i<j}(x(i,j) - x)(t(i,j) - t)}{\sqrt{[\sum_{i<j}(x(i,j) - x)^2][\sum_{i<j}(t(i,j) - t)^2]}}. \quad (1)$$

# Cluster jerárquico aglomerativo: Correlación cophenetic

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.

# Cluster jerárquico aglomerativo

- Una vez construido el dendograma podemos comparar diferentes diseños del dendograma
- Para ello utilizaremos dos funciones:

  ✓tanglegram (comparación visual)
  ✓cor.dendlist (correlación entre dendogramas)

# Cluster jerárquico aglomerativo

- df<- scale(USArrests)

Para una fácil comparación entre dendogramas seleccionamos una muestra aleatoria

- set.seed(234235)
- ss<-simple(1:50, 10)

# Cluster jerárquico aglomerativo

**Primero generamos dos análisis de cluster con dos métodos diferentes de Linkage**

```
library(dendextend)

# Estimamos la matriz de distancias

res.dist <- dist(df, method = "euclidean")

# Generamos los dos análisis de cluster jerárquico aglomerativo

hc1 <- hclust(res.dist, method = "average")
hc2 <- hclust(res.dist, method = "ward.D2")
```

# Cluster jerárquico aglomerativo

*# Generamos los respectivos dendogramas*

dend1 <- **as.dendrogram** (hc1)

dend2 <- **as.dendrogram** (hc2)

*# Generamos una lista que contenta los dos dendogramas*

dend_list <- **dendlist**(dend1, dend2)

# Cluster jerárquico aglomerativo

*Para comparar visualmente dos dendrogramas, usaremos la función tanglegram () [dendextend paquete], que traza los dos dendrogramas, uno al lado del otro, con sus etiquetas conectadas por líneas.*

*La calidad de la alineación de los dos árboles se puede medir utilizando la función entanglement(). Es una medida entre 1 y 0. Un coeficiente más bajo corresponde a una buena alineación.*

# Cluster jerárquico aglomerativo

**tanglegram**(dend1, dend2)

**tanglegram**(dend1, dend2,

highlight_distinct_edges = FALSE, *# Turn-off dashed lines*

common_subtrees_color_lines = FALSE, *# Turn-off line colors*

common_subtrees_color_branches = TRUE, *# Color common branches*

main = **paste**("entanglement =", **round**(**entanglement**(dend_list), 2))

)

# Cluster jerárquico aglomerativo

**La función cor.dendlist () se usa para calcular la  matriz de correlación "Baker" o "Cophenetic“ entre una lista de árboles.**

**El valor puede oscilar entre -1 a 1. Con valores cerca de 0 significa que los dos árboles no son estadísticamente similares.**

# Cluster jerárquico aglomerativo

*# Matriz de correlaciones Cophenetic*

**cor.dendlist**(dend_list, method = "cophenetic")

*# Matriz de correlación de Baker*

**cor.dendlist**(dend_list, method = "baker")

*# Coeficiente de correlación Cophenetic*

**cor_cophenetic**(dend1, dend2)

*# Coeficiente de correlación de Baker*

**cor_bakers_gamma**(dend1, dend2)

# Cluster jerárquico aglomerativo

*También es posible comparar simultáneamente dendrogramas múltiples. Un operador de encadenamiento*

*%>% se usa para ejecutar múltiples funciones al mismo tiempo. Es útil para simplificar el*

*código:*

*# Crear multiples dendogramas*

dend1 <- df %>% dist %>% **hclust**("complete") %>% as.dendrogram

dend2 <- df %>% dist %>% **hclust**("single") %>% as.dendrogram

dend3 <- df %>% dist %>% **hclust**("average") %>% as.dendrogram

dend4 <- df %>% dist %>% **hclust**("centroid") %>% as.dendrogram

# Cluster jerárquico aglomerativo

```
# Matriz de correlaciones

dend_list <- dendlist("Complete" = dend1, "Single" = dend2,
                      "Average" = dend3, "Centroid" = dend4)
                      cors <- cor.dendlist(dend_list)

# Analizamos matriz de correlaciones
round(cors, 2)
```