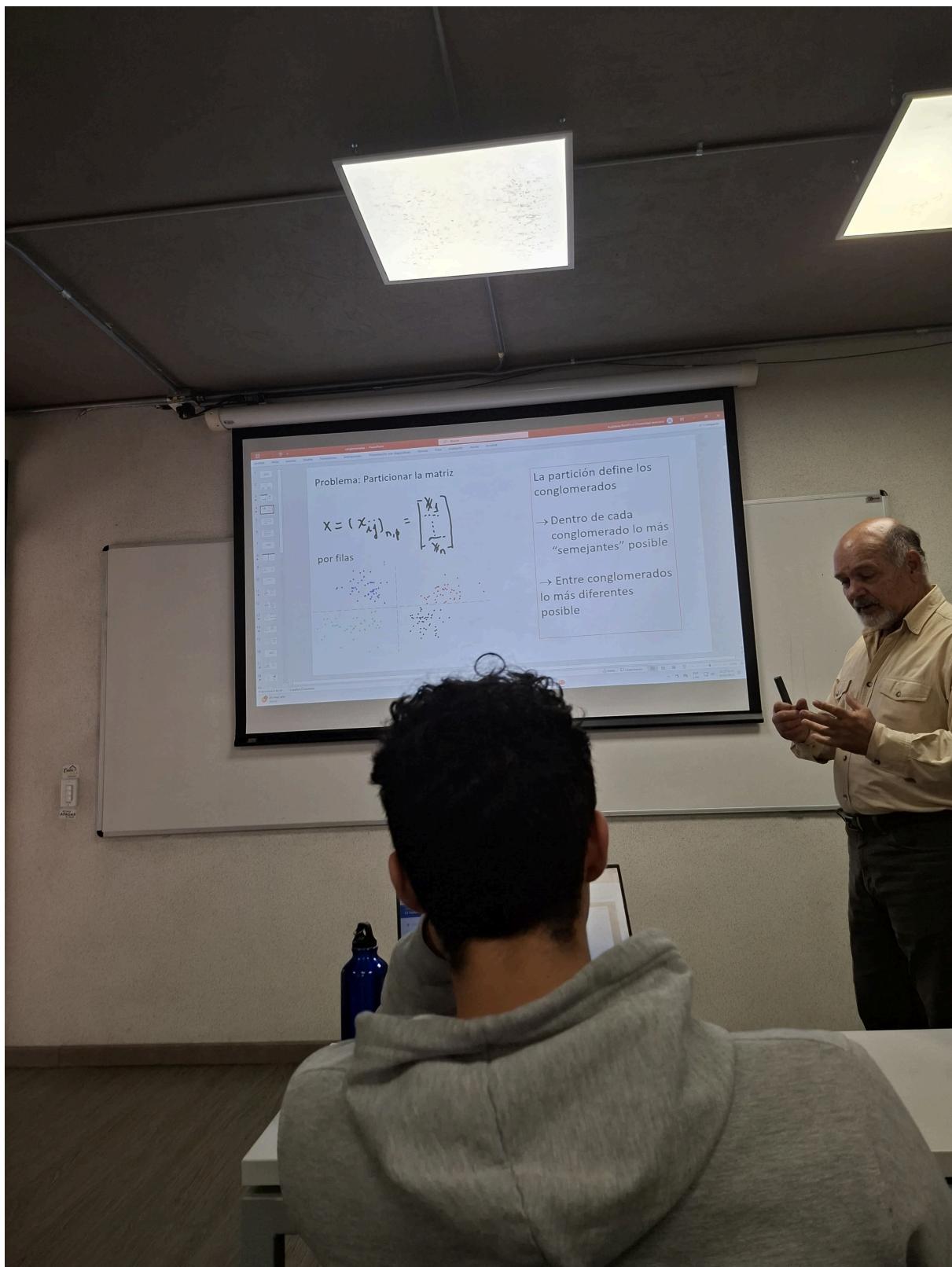




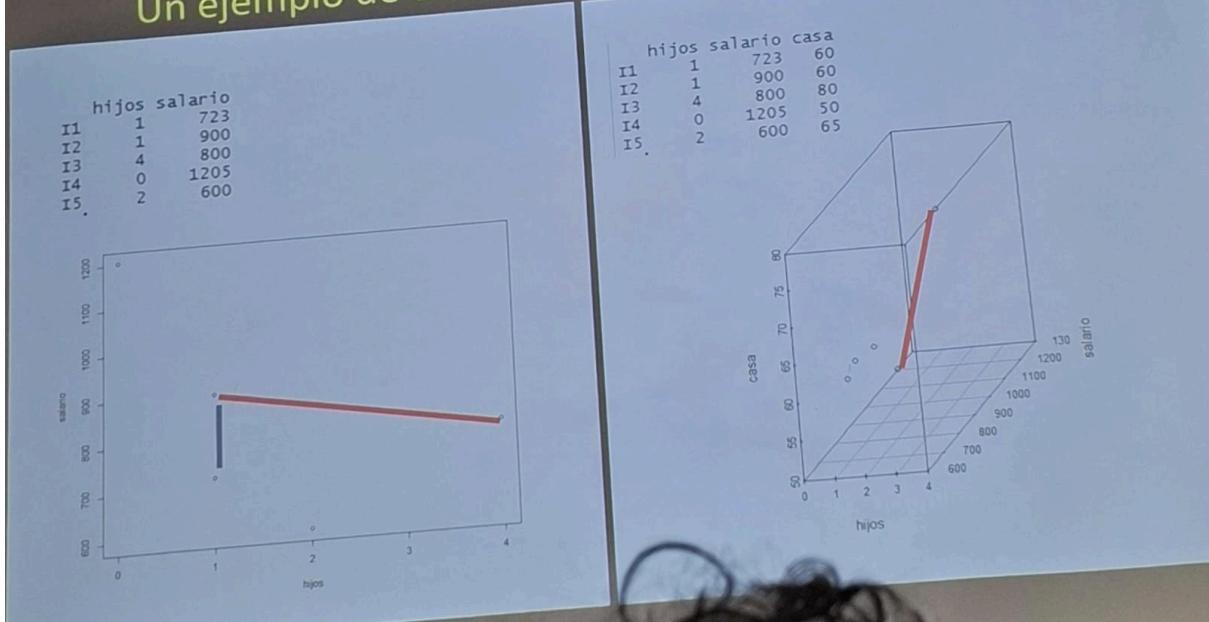
Analisis de conglomerados

⊕ Creadas	@6 de agosto de 2024 14:11
⊗ Asignatura	Analisis Multivariado





Un ejemplo de disimilaridad: la distancia euclídea



Hay que definir formalmente lo que se entiende por semejanza.

Criterios (medidas) de proximidad

- Disimilaridad
- Similaridad

Criterios para seleccionar la medida de proximidad más adecuada:

- ✓ El interés en el análisis
- ✓ Las escalas de medición de las variables
- ✓ El tipo de variable

Ejemplos de distancias

Valores de las variables en los individuos r y s

$$\mathbf{x}_r, \mathbf{x}_s$$

Distancia Euclídea

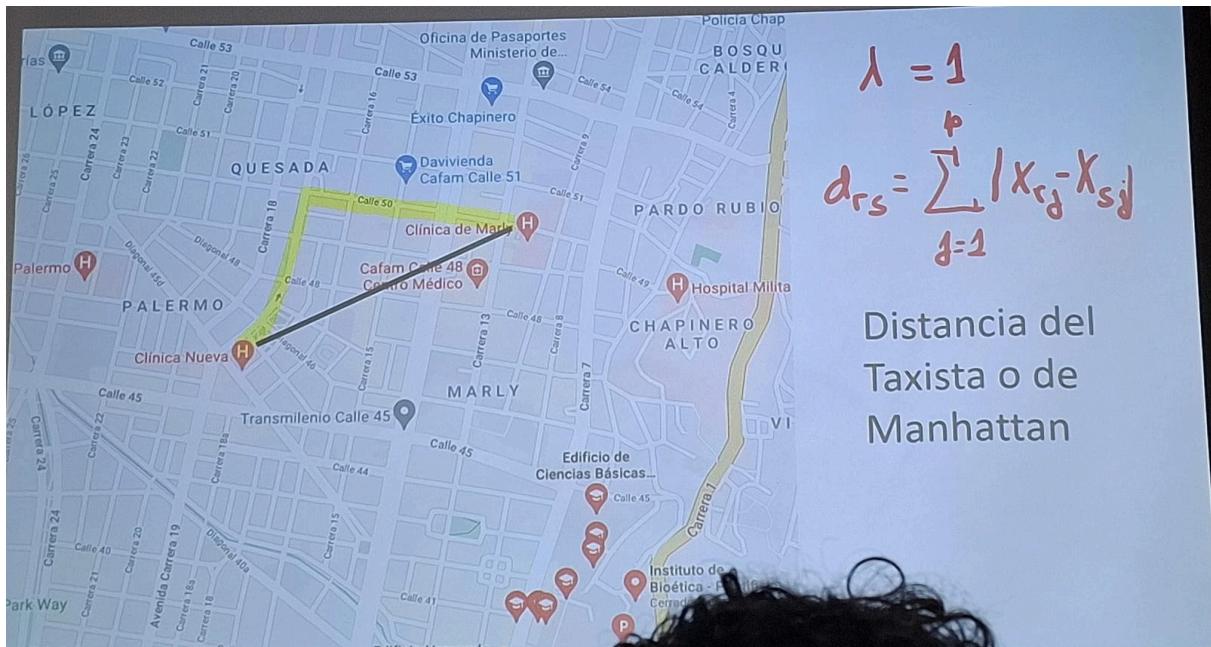
$$d_{r,s} = \left[(\mathbf{x}_r - \mathbf{x}_s)^T (\mathbf{x}_r - \mathbf{x}_s) \right]^{1/2}$$
$$= \| \mathbf{x}_r - \mathbf{x}_s \|$$

Esta medida puede verse afectada si hay muchas diferencias en las escalas de medición: Solución estandarizar o mejor

Distancia de Mahalanobis

$$d_{rs} = \left[(\mathbf{x}_r - \mathbf{x}_s)^T \mathbf{S}^{-1} (\mathbf{x}_r - \mathbf{x}_s) \right]^{1/2}$$

En todos estos casos hay una tendencia a enmascarar conglomerados



MEDIDAS DE SIMILARIDAD

Paso 1. Estandarizar

paso 2. hacer matriz de distancias

paso 3. vizualizar el dendograma

Ejemplos

Variables a analizar en el problema:

$$\mathbb{X} = (X_1, \dots X_p)^t$$

$$X_i = \begin{cases} 1 & A_i \\ 0 & A_i^c \end{cases}$$

Valores de las variables en los individuos r y s

$$X_r, X_s$$

Tabla de concordancias

		X_s	
	1	a	b
1	0	c	d

Índice de Similaridad de Jaccard

$$\frac{a}{a+b+c}$$

Valores de las variables en los individuos r y s

$$X_r, X_s$$

Partiendo de una medida de similaridad

$$\sigma_{rs}$$

Es posible construir una medida de disimilaridad

$$d_{rs} = 1 - \sigma_{rs}$$

Por ese motivo en todo lo que sigue trabajaremos con medidas de disimilaridad

Como

$$d_{rs} = 0 \text{ssi } X_r = X_s$$

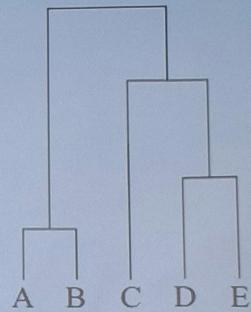
Mientras menor sea la medida de disimilaridad más "parecidos" son los individuos

Para la identificación de los conglomerados lo primero es construir una matriz de proximidades utilizando una medida de disimilaridad (por ejemplo, una distancia)

$$D_{n \times n} = (d_{rs})$$

Datos				Distancias					
	hijos	salario	casa	I1	I2	I3	I4	I5	
I1	1	723	60		0.0	177.0	79.6	482.1	123.1
I2	1	900	60		177.0	0.0	102.0	305.2	300.0
I3	4	800	80		79.6	102.0	0.0	406.1	200.6
I4	0	1205	50		482.1	305.2	406.1	0.0	605.2
I5	2	600	65		123.1	300.0	200.6	605.2	0.0

Para representar las agrupaciones con la que se van construyendo los conglomerados se utilizan

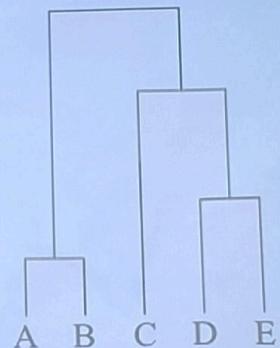


Arboles que permiten visualizar agrupaciones → Los dendogramas

Para construir las agrupaciones (conglomerados) hay dos estrategias:

- Métodos Jerárquicos (aglomerativos)
- Métodos no Jerárquicos (divisivos)

Divisivo



Aglomerativo

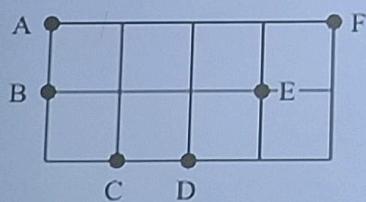
Se parte de una división inicial donde cada individuo es un conglomerado y usando la matriz de proximidades (disimilaridades) se van construyendo paso a paso las agrupaciones

A B C D E

Aglomerativo

Problema

Construya conglomerados donde agrupe individuos (6) ubicados en un plano según su cercanía. Utilice la distancia del taxista para representar las distancias entre nodos del plano



$$d(A, B) = 1$$



$$d(A, C) = 3$$



Matriz de distancias (disimilaridades)

	A	B	C	D	E	F
A	0					
B	1	0				
C	3	2	0			
D	4	3	1	0		
E	4	3	3	2	0	
F	4	5	5	4	2	0

Paso 1

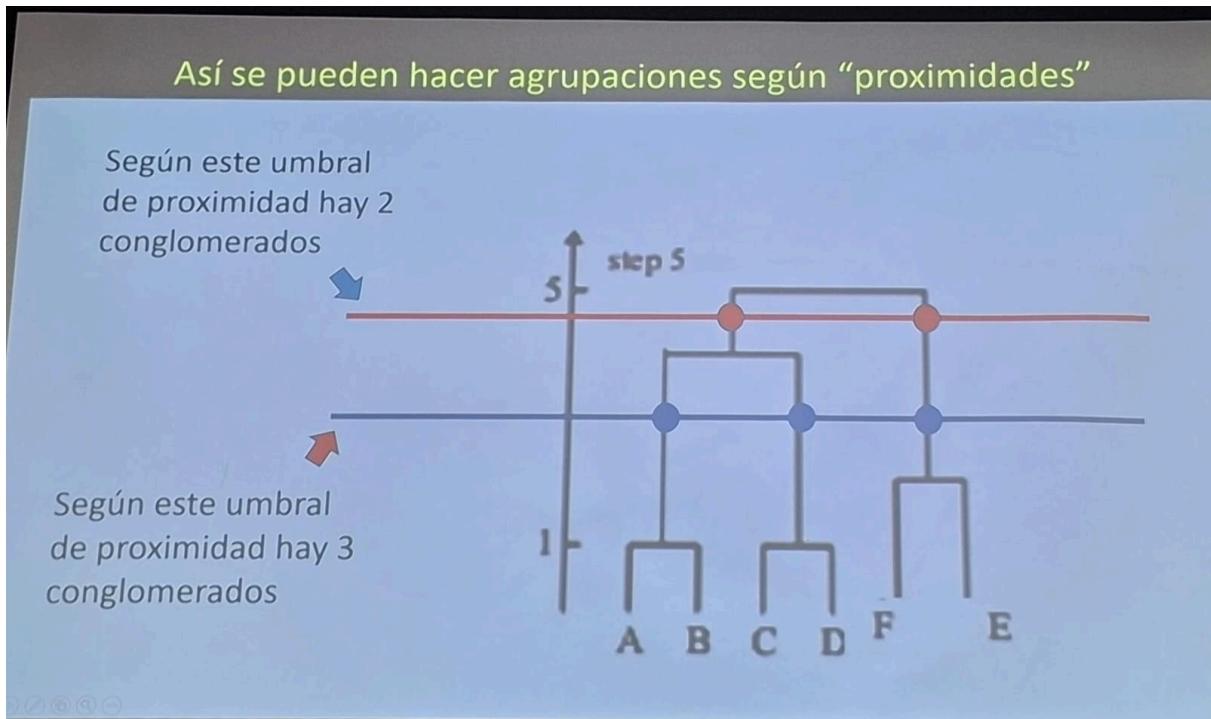
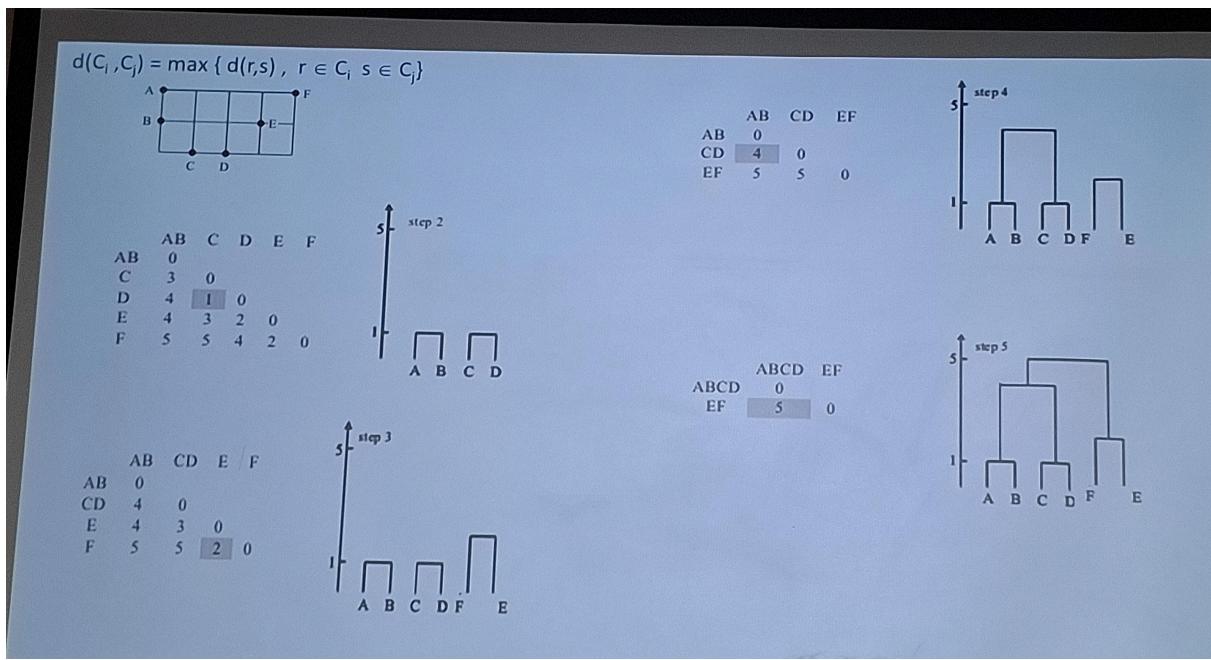
Eje para representar las distancias

	A	B	C	D	E	F
A	0					
B	1	0				
C	3	2	0			
D	4	3	1	0		
E	4	3	3	2	0	
F	4	5	5	4	2	0



¿Cómo calcular ahora las distancias donde uno de los componentes a considerar es un grupo?

Hay diferentes **métodos de aglomeración** para formar los grupos



Métodos de aglomeración para formar los grupos

Consideremos dos agrupaciones: C_i y C_j

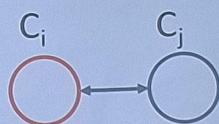
Método 1: Método del vecino más lejano

$$d(C_i, C_j) = \max \{ d(r, s) , r \in C_i, s \in C_j \}$$



Método 2: Método del vecino más cercano

$$d(C_i, C_j) = \min \{ d(r, s) , r \in C_i, s \in C_j \}$$



Método 3: Método de la media de grupos

$$d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{r \in C_i, s \in C_j} d(r, s)$$

Métodos de aglomeración para formar los grupos

Método 4: Método centroide

Valores de las variables en los individuos del conglomerado

C_i

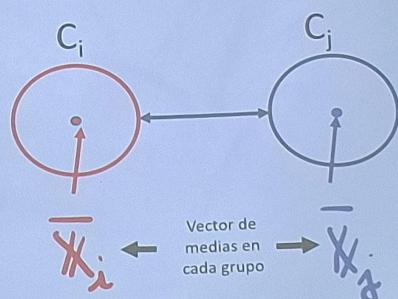
$$(x_r)_{r \in C_i}$$

Valores de las variables en los individuos del conglomerado

C_j

$$(x_s)_{s \in C_j}$$

$$d^2(C_i, C_j) = \| \bar{x}_i - \bar{x}_j \|^2$$



¿Cómo determinar el número de conglomerados?
El índice CH (Calinski–Harabasz)

$$k \rightarrow CH_k = \frac{\text{Variación entre}}{\text{Variación dentro}}$$

Número de conglomerados $\rightarrow \underset{k}{\operatorname{argmax}} CH_k$

Variacion entre alta

Variacion dentro baja

(es el mejor caso)

Base de datos arrestos
Cantidades de arrestos en las ciudades de EU según tipo de delito
Variables
Murder arrests (per 100000)
Assault arrests (per 100000)

Urban population (%)
Rape (per 100000)

Agrupar las ciudades según los tipos de arresto y caracterícelas

```

#paquetes
library(FactoMineR)
library(factoextra)

#Preparacion de los datos estandarizacion
Arr_E=scale(Arr)

#calcular la matriz de disimilaridades
#usemos la distancia euclidea es el implicito
D=dist(Arr_E,method="euclidean")

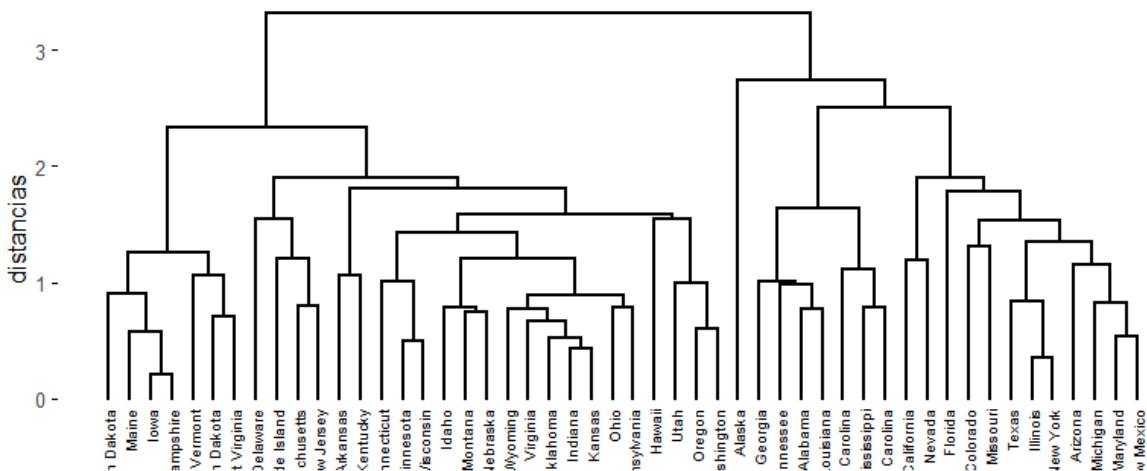
#elaboracion del cluster jerarquico
C_arrestos=hclust(D,method="average")
#metodos:
#cercano - single
#lejano - complete
#media - average
#centroid - centroid

#metodo de aglomeracion
#single: vecino mas proximo
#complete: vecino mas lejano
#average: media de grupos
#centroid: centroide

#elaboracion del dendograma
fviz_dend(C_arrestos,cex=0.5,ylab="distancias",main="dendograma iris")

```

dendograma iris

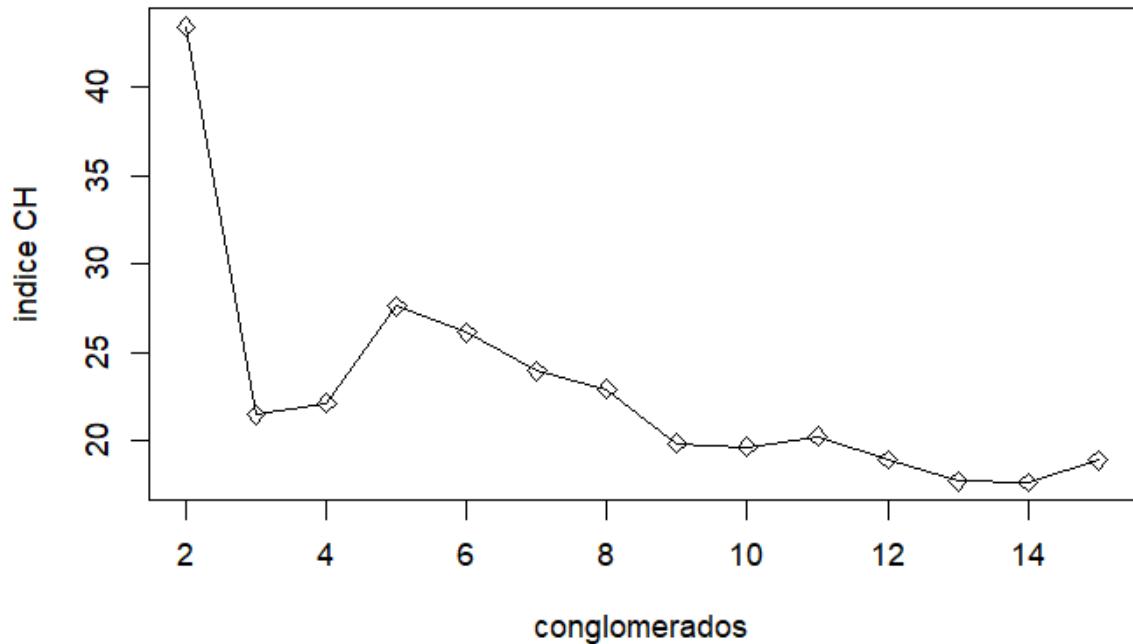


```

#numero de conglomerados a considerar
library(UniversalCVI)
I=CH.IDX(Arr_E,kmax=15,kmin=2,method="hclust_average")
plot(I$k,I$CH,pch=5,main="numero conglomerados",xlab="conglomerados",ylab="indice CH")
lines(I$k,I$CH, col = "black")

```

numero conglomerados

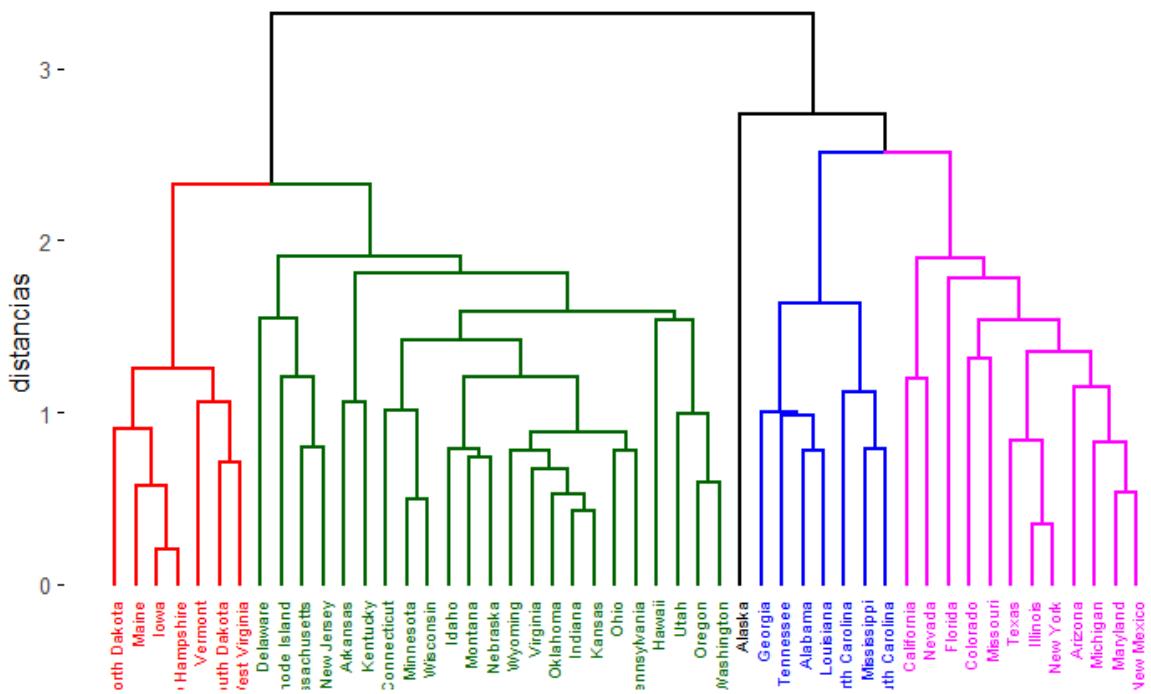


El Indice CH se explica antes para que es

```
#cortar el dendograma en diferentes grupos
#cortar en 5 grupos
grupos<-cutree(C_arrestos,k=5)
#nombres de las ciudades en cada conglomerado
table(grupos)
#nombres de las ciudades en los grupos, ejemplo en el 1
#para hacer el grafico con los grupos
fviz_dend(c_arrestos,k=5,cex=0.5,k_colors= c("red","darkgreen","black","blue","magenta"),color_labels_by_k = TRUE,
ylab="distancias",main="dendograma ciudades")
#se pueden utilizar las componentes principales para caracterizar los grupos
```

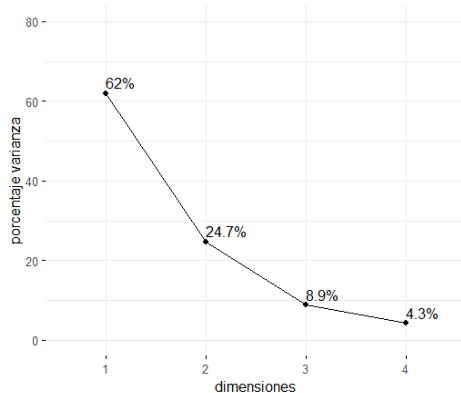
```
> table(grupos)
grupos
 1  2  3  4  5 
 1 12 23 7 
> #nombre de las ciudades en los grupos, ejemplo en el 1
> rownames(Arr)[grupos==1]
[1] "Alabama"      "Georgia"       "Louisiana"     "Mississippi"   "North Carolina"
[6] "South Carolina" "Tennessee"    +
```

dendograma ciudades



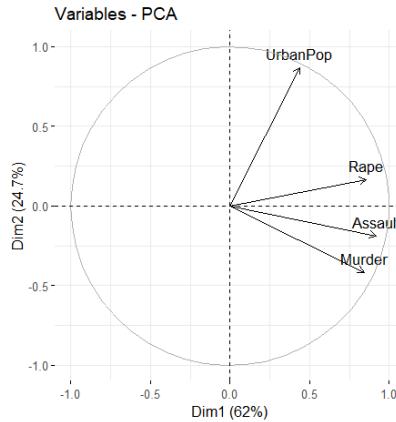
#obtencion de las componentes

```
CP_CC=PCA(Arr,scale.unit=TRUE,ncp=6,graph=FALSE)
fviz_screeplot(CP_CC,choise="eigenvalue",addlabels=TRUE,geom="line", ylim=c(0,80),
xlab="dimensiones",ylab="porcentaje varianza",main="",ncp=6)
```

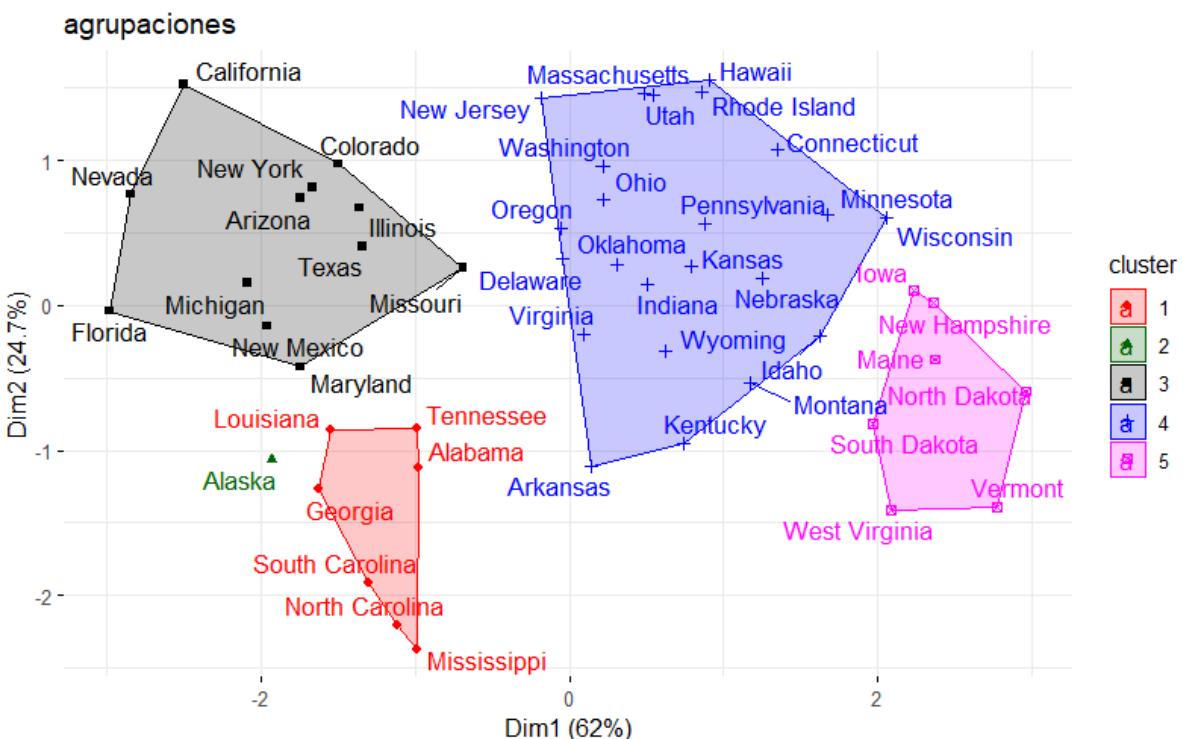


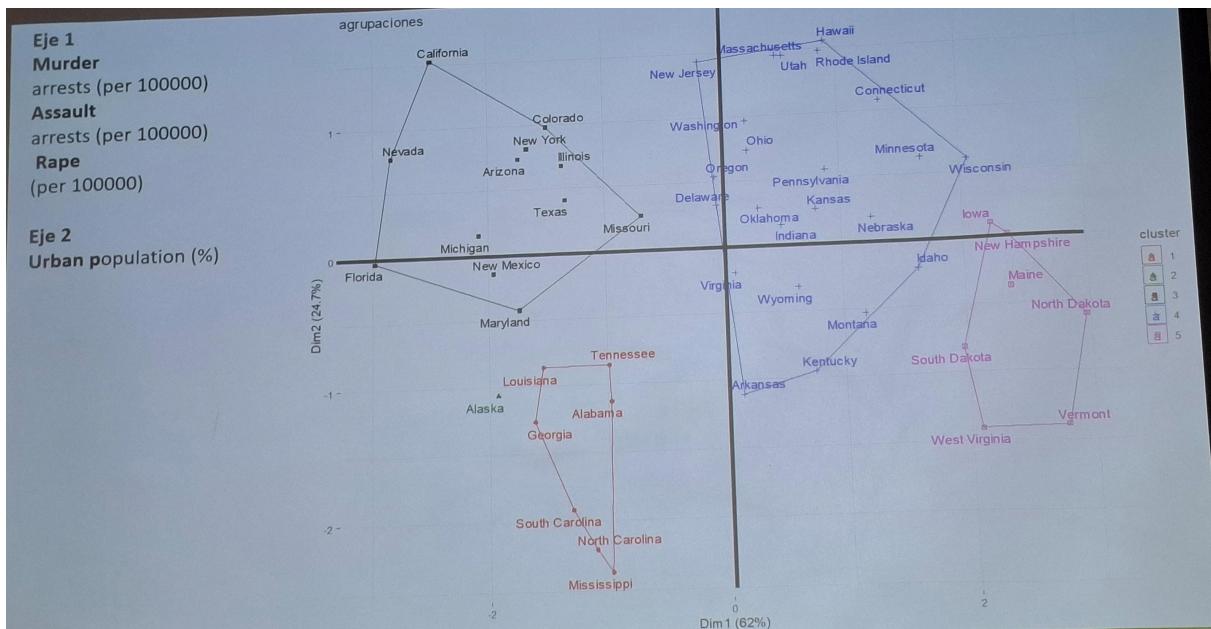
#interpretacion de los dos primeros ejes

```
fviz_pca_var(CP_CC,axes=c(1,2))
```



```
#representación de los conglomerados ejes 1 y 2
fviz_cluster(list(data=Arr,cluster=grupos),palette=c("red","darkgreen","black","blue","magenta"),
            ellipse.type="convex",repel=TRUE,
            show.clust.cent = FALSE,ggtheme = theme_minimal(),main="agrupaciones")
```





EJERCICIOS

Se dispone información de cinco familias. Datos ejercicio 1

	hijos	salario	casa
I1	1	723	60
I2	1	900	60
I3	4	800	80
I4	0	1205	50
I5	2	600	65

Utilizando la distancia euclídea y como método de aglomeración las medias, analice si es posible identificar algún tipo de agrupación para las familias

