Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 15: 28/10/24 – 02/11/24

Introducción

- En **estudios de marketing**, una de las estrategias para aplicar una determinada campaña, suele ser inicialmente la identificación de **grupos similares de clientes** para después establecer/trazar un plan específico de dicha campaña en cada grupo.
- En **estudios de mercado**, cuando se va lanzar un nuevo producto, se identifican los **productos relacionados que ya están en el mercado** y con base en eso determinar el desempeño que tendrá éste.
- En **Ecología**, se clasifica a un **grupo de plantas o animales** de acuerdo con un conjunto de características que tienen en común.

Introducción

- En **Psicología**, se colectan información sobre algunas características de las personas y con base en ellas se **agrupan algunas personas** que de alguna forma tienen **personalidades similares**.
- En **Ciencias Sociales**, las **personas** con **condiciones socioeconómicas homogeneas** se **agrupadas**/consideradas dentro de un **mismo estrato**.

Introducción

Observe que en todos los casos las **unidades muestrales o poblacionales son congregados/reunidos/clasificados en grupos**.

Introducción

Observe que en todos los casos las **unidades muestrales o poblacionales son congregados/reunidos/clasificados en grupos**. **El Análisis de Conglomerado (cluster o de clasificación)** tiene como propósito definir la estructura de los datos colocando las **unidades más parecidas en grupos**.

Introducción

Observe que en todos los casos las **unidades muestrales o poblacionales son congregados/reunidos/clasificados en grupos**. **El Análisis de Conglomerado (cluster o de clasificación)** tiene como propósito definir la estructura de los datos colocando las **unidades más parecidas en grupos**. El análisis de agrupamiento es bastante popular en **Data Mining (mineración de datos)**.

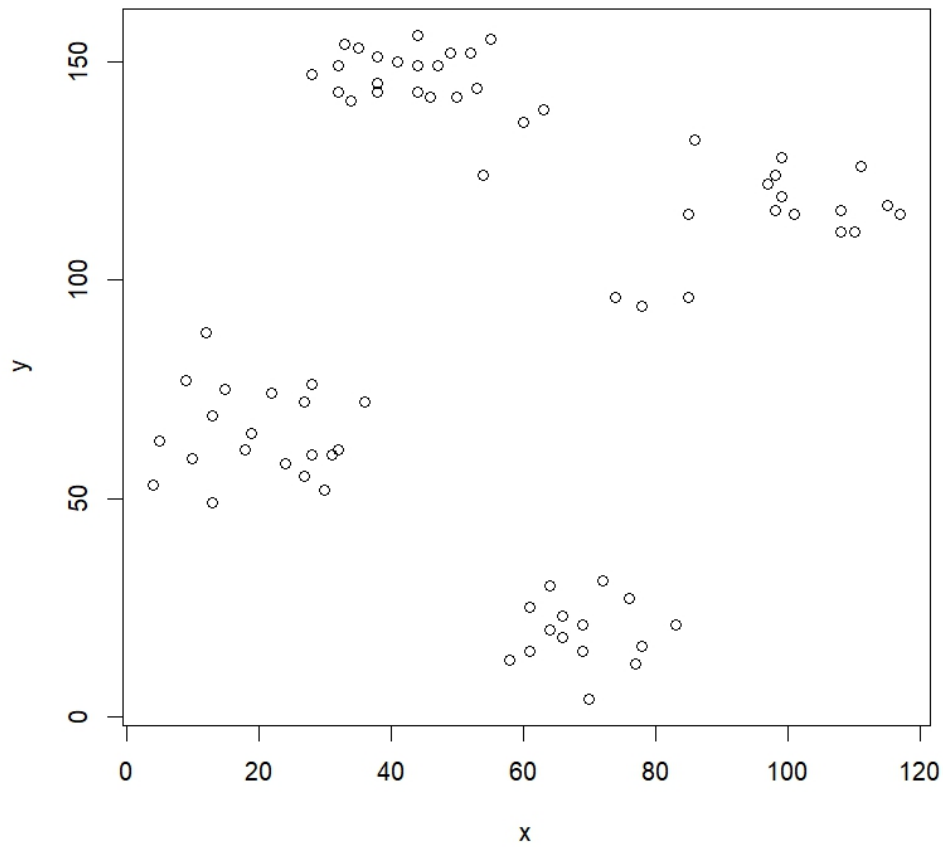
Introducción

Los datos `ruspini`[1] del paquete `cluster` contiene 75 observaciones sobre 2 variables que dan las coordenadas x e y de los puntos, respectivamente.

```
# ----- Datos Ruspini ----- #
require(cluster)
data("ruspini")
head(ruspini)
plot(ruspini)
```

[1] E. H. Ruspini (1970) Numerical methods for fuzzy clustering. Inform. Sci. 2, pp. 319–350.

Introducción



Se pueden observar que los puntos se aglomeran en **4 grupos**

Objetivos del Análisis de Conglomerados

El **Análisis de Conglomerados/Cluster/Clasificación** tiene como objetivo principal:

Dividir las unidades muestrales/poblacionales en grupos de tal forma que aquellas que pertenecen a un grupo específico sean similares (homogéneos) entre si de acuerdo con las características (variables) que fueron medidas en ellas, y aquellas unidades de grupos diferentes sean disimilares (heterogéneos) con relación a esas mismas características.

Objetivos del Análisis de Conglomerados

De acuerdo con los objetivos plantedos se deben considerar los siguientes aspectos:

1. ¿Cómo se mide la **similitud**?
   Se debe establecer algún **mecanismo/instrumento** que permita **comparar** las **unidades/observaciones** en términos de las variables medidas sobre ellos. Éste instrumento debe registrar la **proximidad entre pares de unidades/observaciones** de tal forma que las **distancias indiquen la similitud**.

Objetivos del Análisis de Conglomerados

2. ¿Cómo se forman los **conglomerados o cluster**?
   Se debe establecer el **pocedimiento** mediante el cual se **agrupan las observaciones** que son **más similares dentro** de un determinado **conglomerado/cluster**.
3. ¿**Cuántos grupos** se deben formar?
   El criterio debe tener en cuenta la homogeneidad media alcanzada dentro de los conglomerados.

Medidas de Similitud y Disimilitud

Medidas de Similitud y Disimilitud

Suponga que un conjunto de datos contiene n unidades muestrales u **observaciones** y p **variables** y que el objetivo es **agrupar** esas observaciones en g **grupos**.

Observaciones	X_1	X_2	$\cdots$	X_i	$\cdots$	X_p
1	X_{11}	X_{21}	$\cdots$	X_{i1}	$\cdots$	X_{p1}
2	X_{12}	X_{22}	$\cdots$	X_{i2}	$\cdots$	X_{p2}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
j	X_{1j}	X_{2j}	$\cdots$	X_{ij}	$\cdots$	X_{pj}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	X_{1n}	X_{2n}	$\cdots$	X_{in}	$\cdots$	X_{pn}

Medidas de Similitud y Disimilitud

Cada $j = 1, 2, \ldots, n$ es una **observación multivariada**

$$\mathbf{X}_j = [X_{1j} \quad X_{2j} \quad \ldots \quad X_{pj}]^\top$$

donde X_{ij} es el valor observado de la variable i $(i = 1, 2, \ldots, p)$ en la unidad/observación j.

Medidas de Similitud y Disimilitud

Cada $j = 1, 2, \ldots, n$ es una **observación multivariada**

$$\mathbf{X}_j = [X_{1j} \quad X_{2j} \quad \ldots \quad X_{pj}]^\top$$

donde X_{ij} es el valor observado de la variable i $(i = 1, 2, \ldots, p)$ en la unidad/observación j. Para **agrupar esas observaciones** es necesario que se defina apriori la **medidad de similaridad o disimilaridad** que será utilizada.

Medidas de Similitud y Disimilitud

Existen varias **medidas diferentes** que proporcionan un determinado tipo de **agrupamiento**:

1. Medidas de **Distancia:**
   Principalmente para **variables cuantitativas**.
   Cuanto menor sea la distancia, más parecidas son las unidades.
2. Coeficientes de **Concordancia/Asociación**:
   Principalmente para **variables cualitativas**
3. Coeficientes de **Correlación**:
   Se usa sólo para variables en la escala de intervalo.
4. Medidas **Probabilísticas de Similitud**:
   Se usa sólo para variables dicótomas.

Medidas de Distancia:

Sean $\mathbf{X}_k$ e $\mathbf{X}_l$ los vectores de las mediciones de las **unidades** k **e** l, respectivamente. A seguir se presentan las técnicas distancias entre $\mathbf{X}_k$ e $\mathbf{X}_l$.

- Distancia **Euclidiana**:

$$d_{kl} = \sqrt{(\mathbf{X}_k - \mathbf{X}_l)^\top (\mathbf{X}_k - \mathbf{X}_l)} = \sqrt{\sum_{i=1}^{p} (X_{ik} - X_{lk})^2}$$

Medidas de Similitud y Disimilitud

- Distancia **Generalizada o Ponderada**:

$$d_{kl} = \sqrt{(\mathbf{X}_k - \mathbf{X}_l)^\top \mathbf{A}(\mathbf{X}_k - \mathbf{X}_l)}$$

donde $\mathbf{A}$ es una matriz de ponderación:

(i) Si $\mathbf{A} = \mathbf{I}$ se tiene la distancia **Euclidiana**.
(ii) Si $\mathbf{A} = \mathbf{\Sigma}^{-1}$ se tiene la distancia de **Mahalanobis**.
(iii) $\mathbf{A} = \mathrm{diag}(1/p)$ se tiene la distancia **Euclidiana Media**.

Medidas de Similitud y Disimilitud

- Distancia de **Manhattan** o **city block**:

$$d_{kl} = \sum_{i=1}^{p} | \mathbf{X}_{ik} - \mathbf{X}_{il} |$$

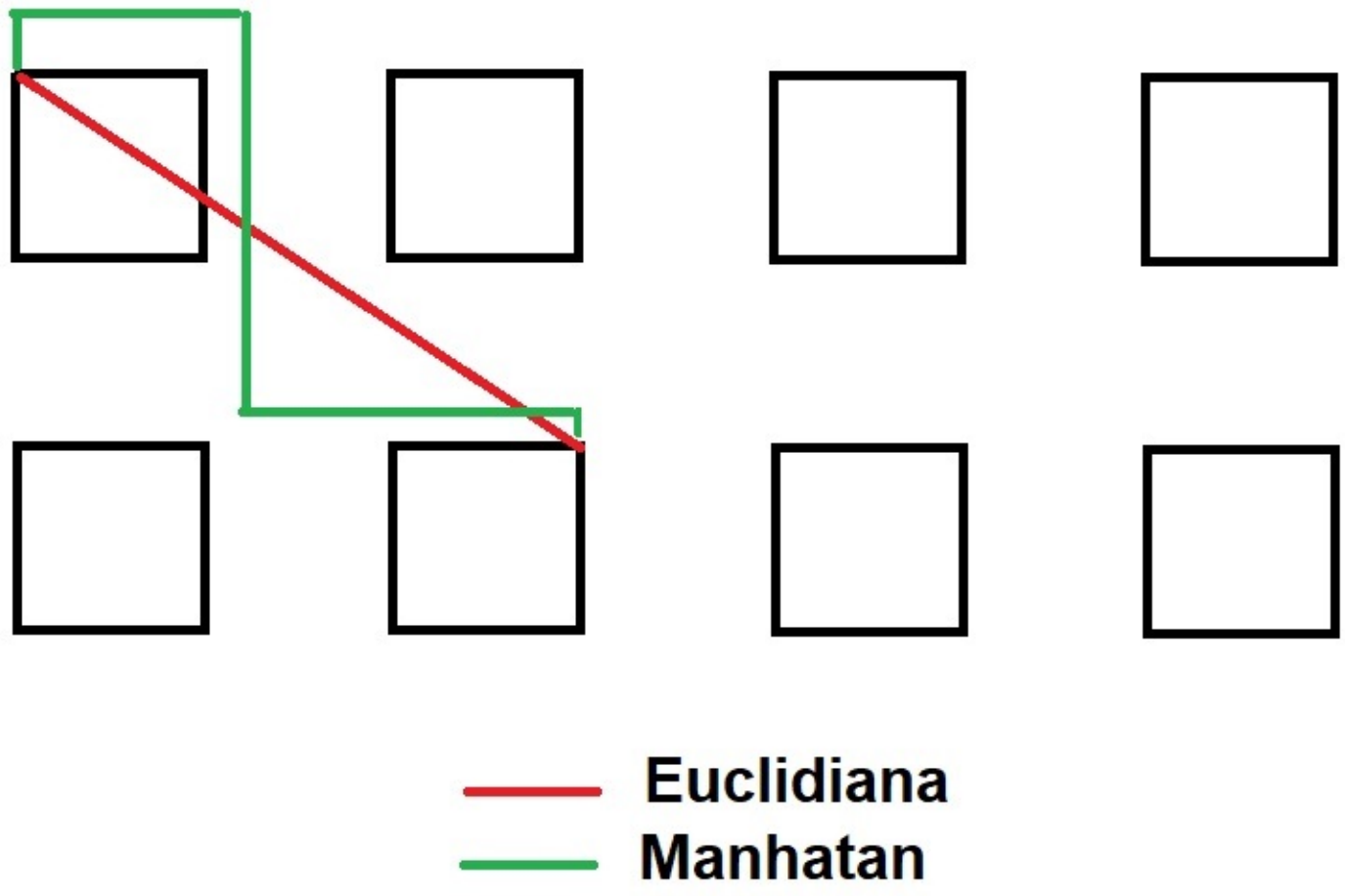
- Distancia de **Minkowski**:

$$d_{kl} = \left(\sum_{i=1}^{p} | \mathbf{X}_{ik} - \mathbf{X}_{il} |^{r} \right)^{1/r} \qquad r = 1, 2, \ldots$$

Note que si

- $r = 1$, se tiene la distancia de **Manhatan**.
- $r = 2$, se tiene la distancia **Euclidiana**.

Medidas de Similitud y Disimilitud



Medidas de Similitud y Disimilitud

Una vez se defina la distancia, se **organizan las distancias en una matriz**:

$$\begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ & 0 & d_{23} & \cdots & d_{2n} \\ & & & \ddots & \vdots \\ & & & & d_{n-1,n} \\ & & & & 0 \end{bmatrix}$$

Observe que ésta es una matriz **simétrica**.

Medidas de Similitud y Disimilitud

Example

Suponga que para $n = 4$ **personas** se tiene información sobre su edad $(X_1, \text{en años})$, su estatura $(X_2, \text{en metros})$ y su peso $(X_3, \text{en kilogramos})$:

Persona	X_1	X_2	X_3
A	23	1.69	61
B	40	1.70	72
C	26	1.65	68
D	38	1.68	70

Medidas de Similitud y Disimilitud

Tenemos

$$\overline{\mathbf{X}} = \begin{bmatrix} 31.75 \\ 1.68 \\ 67.75 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 72.25 & 0.08 & 35.58 \\ 0.08 & 0.00 & 0.00 \\ 35.58 & 0.00 & 22.92 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1.00 & 0.45 & 0.87 \\ 0.45 & 1.00 & 0.03 \\ 0.87 & 0.03 & 1.00 \end{bmatrix}$$

Las distancias Euclidianas:

$$d_{AB} = \sqrt{(23-40)^2 + (1.69-1.70)^2 + (61-72)^2} \approx 20.25$$

$$d_{AC} = \sqrt{(23-26)^2 + (1.69-1.65)^2 + (61-68)^2} \approx 7.62$$

$$d_{AD} = \sqrt{(23-38)^2 + (1.69-1.68)^2 + (61-70)^2} \approx 17.49$$

Medidas de Similitud y Disimilitud

$$d_{BC} = \sqrt{(40-26)^2 + (1.70-1.65)^2 + (72-68)^2} \approx 14.56$$

$$d_{BD} = \sqrt{(40-38)^2 + (1.70-1.68)^2 + (72-70)^2} \approx 2.83$$

$$d_{CD} = \sqrt{(26-38)^2 + (1.65-1.68)^2 + (68-70)^2} \approx 12.17$$

Por lo tanto, la matriz de **distancias Euclidiana**:

$$\begin{bmatrix} 0 & 20.25 & 7.62 & 17.49 \\ & 0 & 14.56 & 2.83 \\ & & 0 & 12.17 \\ & & & 0 \end{bmatrix}$$

Medidas de Similitud y Disimilitud

Observe que:

- La **menor distancia** (personas más cercanas) son B y D ($d_{BD} = 2.83$).
- La **segunda menor distancia** se observa entre las personas A y C ($d_{AC} = 7.62$).
- Las personas **más distantes** son A y B ($d_{AB} = 20.25$).

Medidas de Similitud y Disimilitud

Observe que:

- La **menor distancia** (personas más cercanas) son B y D ($d_{BD} = 2.83$).
- La **segunda menor distancia** se observa entre las personas A y C ($d_{AC} = 7.62$).
- Las personas **más distantes** son A y B ($d_{AB} = 20.25$).

Recuerde que las variables X_1, X_2, X_3 están **medidas en diferentes escalas**. Así que se deben **estandarizar**, por lo cuál calculamos la distancia de **Mahalanobis**:

Medidas de Similitud y Disimilitud

$$\begin{bmatrix} 0 & 7.21 & 6.36 & 10.01 \\ & 0 & 8.89 & 15.62 \\ & & 0 & 7.96 \\ & & & 0 \end{bmatrix}$$

Observe que:

- con la distancia de **Mahalanobis**, la **mayor distancia** es entre las personas B y D ($d_{BD} = 15.62$), mientras que con la distancia **Euclidiana**, estas dos personas presentaban las **menores distancias**.

Medidas de Similitud y Disimilitud

Observe que:

- con la distancia de **Mahalanobis**, la **menor distancia** se observa entre las personas A y C ($d_{AC} = 6.36$)
- con la distancia de **Mahalanobis**, la **segunda menor distancia** (7.21) se observa entre las personas A y B, las cuales eran **las más altas** en la **distancia Euclidiana**.

```
# ----- Distancia Euclidiana ----- #
 Datos<-data.frame(Persona = c("A","B","C","D"),
                   x1=c(23,40,26,38),
                   x2=c(1.69,1.70,1.65,1.68),
                   x3=c(61,72,68,70))
```

Medidas de Similitud y Disimilitud

```
X<-Datos[,-1]
Xbarra<-colMeans(X)
S<-round(cov(X),2)
R<-round(cor(X),2)

require(abdiv)
dAB<-round(euclidean(X[1,],X[2,]),2)
dAC<-round(euclidean(X[1,],X[3,]),2)
dAD<-round(euclidean(X[1,],X[4,]),2)
dBC<-round(euclidean(X[2,],X[3,]),2)
dBD<-round(euclidean(X[2,],X[4,]),2)
dCD<-round(euclidean(X[3,],X[4,]),2)
```

Coeficientes de Concordancia/Asociación:

Cuando las **variables** son **nominales**, usamos medidas de **similaridad**. En general, las **unidades/observaciones son comparadas** de acuerdo con la **presencia o ausencia de ciertas características**, éstas pueden ser representadas por una variable binária (dummy): **0 (de ausencia) y 1 (presencia)**. Las variables dummy son resumidas en una tabla de doble entrada, una para cada par de observaciones/unidades. Las observaciones "parecidas" deben tener en común más variables que concuerdan que no concuerdan.

Medidas de Similitud y Disimilitud

Example

Supponga que se tienen $n = 2$ unidades/observaciones, para los cuales se tiene la siguiente información:

Obs	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0	1	1	1	1	0	1	0	0	0
2	0	0	1	1	1	0	1	1	0	0

Al comparar estas **dos observaciones** se tiene:

- ambas tengan presente el carácter comparado: $(1, 1)$
- ambas tengan ausente el carácter comparado: $(0, 0)$

Medidas de Similitud y Disimilitud

- La primera tenga presente el carácter comparado y la segunda no lo tenga: $(1, 0)$.
- la prirmera no tenga presente el carácter comparado y la segunda lo tenga: $(0, 1)$.

Resumimos las frecuencias de estas características:

Obs 1	Obs 2	
	1	0
1	$a = 4$	$b = 1$
0	$c = 1$	$d = 4$

Medidas de Similitud y Disimilitud

De ahí tenemos que el **índice/coeficiente** de **concordancias/similitud** entre las observaciones $k = 1$ y $l = 2$ es:

$$S_{k,l} = \frac{a+d}{a+b+c+d} = \frac{4+4}{4+1+1+4} = \frac{8}{10} = 0.80$$

Ésta medida es conocida como

Coeficiente de Similitud Simple (S)

Claramente $0 \leq S_{k,l} \leq 1$. Cuanto **mayor sea el valor de** $S_{k,l}$, **mayor es la similaridad** entre las observaciones $k = 1$ y $l = 2$.

Medidas de Similitud y Disimilitud

Algunas veces, el **0 no es tan informativo como el 1**, en otras palabras el par **(0,0) puede no representar una concordancia** (aveces los individuos tienen características diferentes). Por ejemplo, si

$$X = \begin{cases} 1 & \text{, tiene ojos azules} \\ 0 & \text{, no tiene ojos azules} \end{cases}$$

Así que

$$(1,1) : \text{ambos individuos tienen ojos azules}$$
$$(0,0) : \text{puede ser (negro,café)}$$

Medidas de Similitud y Disimilitud

Omitiendo/**excluyendo el par** $(0,0)$ tenemos el **coeficiente de Jaccard**:

$$J_{kl} = \frac{a}{a+b+c} = \frac{4}{4+1+1} = \frac{4}{6} \approx 0.67$$

Claramente $0 \leq J_{k,l} \leq 1$. Cuanto mayor sea J_{kl}, mayor es la similaridad de las observaciones k y l. Este coeficiente es quizás el más utilizado.

Medidas de Similitud y Disimilitud

Existen otras medidas de similaridad:

- ✓ **Roger y Tanimoto:** considere más importante las diferencias.

$$RT_{kl} = \frac{a+d}{a+2b+2c+d}$$

Asume valores entre 0 y 1.

- ✓ **Sorense y Dice:** considera de mayor importancia a las coincidencias en estado de presencia.

$$SD_{kl} = \frac{2a}{2a+b+c}$$

Asume valores entre 0 y 1.

Medidas de Similitud y Disimilitud

- ✓ **Sokal y Sneath:** Tiene más en cuenta las coincidencias, tanto por presencia como por ausencia de las características:

$$SS_{kl} = \frac{2(a+d)}{2(a+d)+c+d}$$

Asume valores entre 0 y 1.

- ✓ **Hamann:** considera importante las diferencias entre coincidencias y no coincidencias.

$$H_{kl} = \frac{(a+d)-(c+b)}{a+b+c+d}$$

Asume valores entre -1 y 1.

Medidas de Similitud y Disimilitud

Cuando se tienen **variables cualitativas y cuantitativas al mismo tiempo**, se tienen las siguientes alternativas:

1. Transformar las variables cualitativas en cuantitativas por medio de atritutos numéricos a las categorías.
2. Transformar las variables cuantitativas en variables cualitativas a través de la categorizacion (usando algún criterio) de sus valores.
3. Construir medidas de similaridad mixtas y emplearlas para la comparación de las unidades/observaciones (combinación lineal de las variables cuantitativas (q) y cualitativas(p)):

$$c(A,B) = w_p \underbrace{c_p(A,B)}_{\text{Coef. Var. Cuant}} + w_q \underbrace{c_q(A,B)}_{\text{Coef. Var. Cual.}}$$

donde

$$w_p = \frac{p}{p+q} \qquad w_q = \frac{q}{p+q}$$

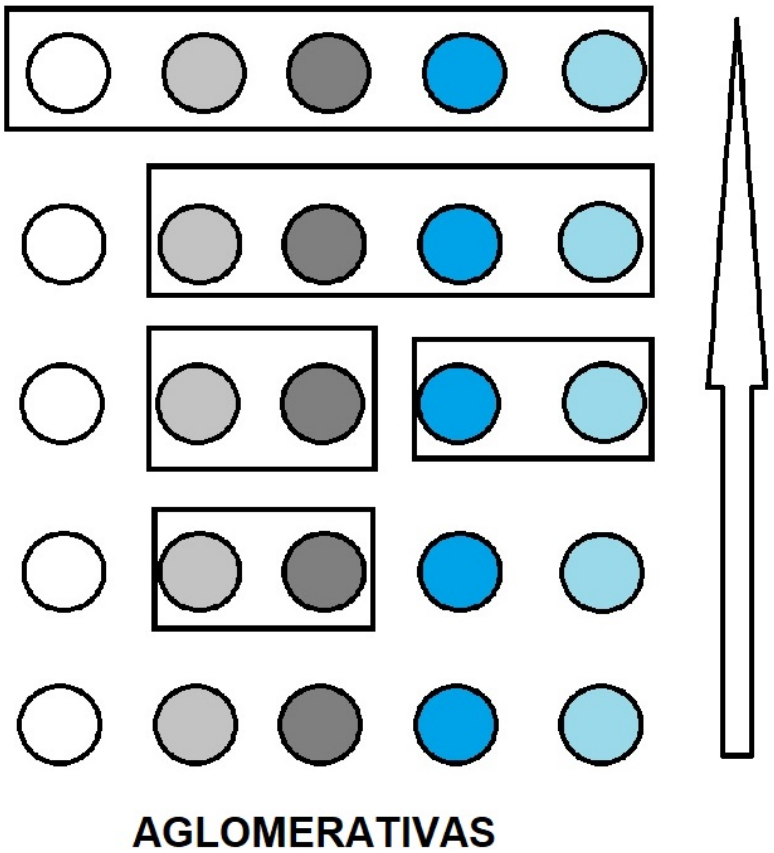
Técnicas/Métodos de Agrupamiento

Métodos de Agrupamiento

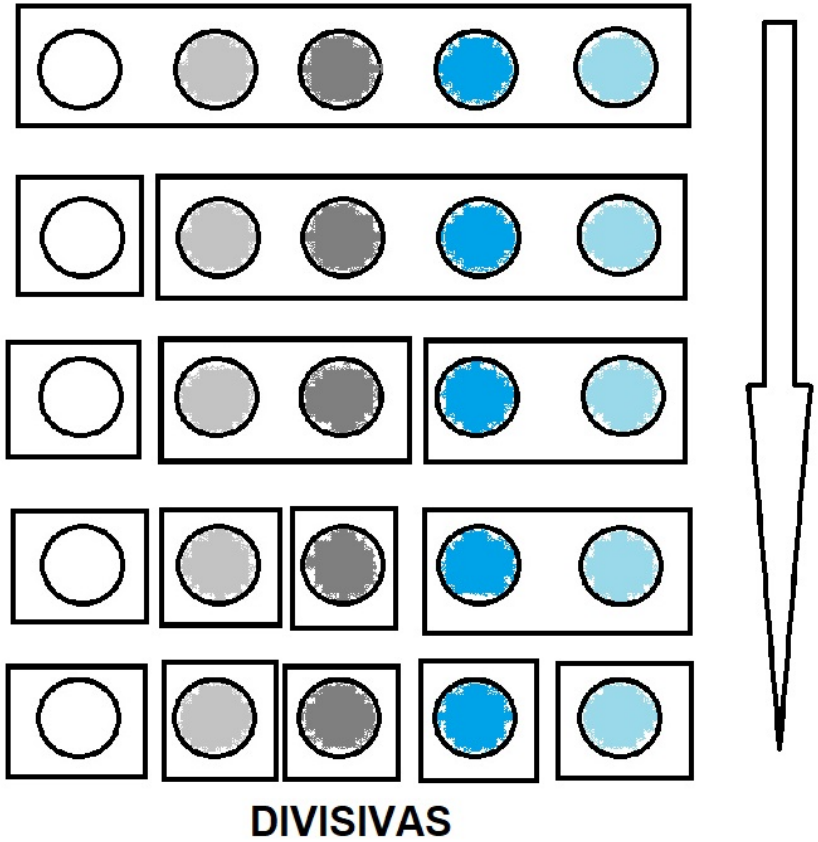
Las técnicas de para la construcción de conglomerados o cluster son clasificadas en:

1. **Jerarquicas:** usualmente usadas como exploratorias y su finalidad es identificar los posibles agrupamientos y el número probable de grupos (g). Existen técnicas jerarquicas agromerativas (cada unidad/observación forma un cluster y se empiezan aglomerar) y divisivas (el conjunto de datos es un cluster que se empieza a dividir).
2. **No-Jerarquicas:** el número de grupos (g) es predefinido por el investigador.

Métodos de Agrupamiento



AGLOMERATIVAS



DIVISIVAS

Técnicas Jerarquicas Aglomerativas:

- EL proceso incia considerando que se tienen n conglomerados, es decir, cada unidad/observación es considerada un cluster de tamaño 1.
- En cada paso del algoritmo, las unidades van siendo agrupadas, hasta que todas las unidades son consideradas como un único grupo.
- En el primer paso se tiene la menor dispersión interna posible, mientras que en el último se tienen la mayor.
- En cada etapa, los grupos son comparados a través de alguna medida de similaridad previamente definida.

Métodos de Agrupamiento

Básicamente el algoritmo es el siguiente:

- Cada observación constituye un cluster.
- En cada paso, los pares de conglomerados más similares son combinados y pasan a constituir un cluster. Se forma un nuevo conglomerado en cada paso, así que en cada paso del proceso el número de conglomerados va disminuyendo.

Note que en cada etapa del algoritmo, cada nuevo conglomerado formado es un agrupamiento de conglomerados formado en las etapas anteriores, ésta es la **jerarquía**. A partir de esta jerarquía, se puede construir un gráfico llamado **Dendrograma**, el cuál representa ese árbol de agrupamiento. Este gráfico permite definir el número de grupos (g).

Métodos de Agrupamiento

Existen varios **métodos de agrupamiento jerarquicos**:

- **Enlace Simple (Single Linkage) o vecino más cercano**: La similaridad entre dos conglomerados es definida por las dos unidades más parecidas entre sí.

  Suponga que en un determinado paso del algoritmo de agrupamiento, se tienen $g = 2$ grupos:

$$G_1 : \mathbf{X}_1, \mathbf{X}_2$$
$$G_2 : \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5$$

  La distancia entre los dos grupos es

$$d(G_1, G_2) = \text{mín}\{d(\mathbf{X}_l, \mathbf{X}_k)\} \;\; l \neq k, \; l = 1, 2, \;\; k = 3, 4, 5$$

Métodos de Agrupamiento

donde X_l son las observaciones de G_1 y X_k son las observaciones de G_2.

- **Enlace Completa (Complete Linkage) o vecino más lejano:**
  La similaridad entre dos conglomerados es definida por las dos unidades menos semejantes entre sí.

  Para nuestro ejemplo ilustrativo

$$d(G_1, G_2) = \text{máx}\{d(\mathbf{X}_l, \mathbf{X}_k)\} \;\; l \neq k, \; l = 1, 2 \;\; k = 3, 4, 5$$

Métodos de Agrupamiento

- **Distancias Medias (Average Linkage):**
  Trata las distancias entre dos conglomerados como la media de las distancias entre todos los pares de unidades que pueden ser formados con las unidades de los dos conglomerados que están siendo comparados:

$$d(G_1, G_2) = \sum_{l \in G_1} \sum_{k \in G_2} \frac{1}{n_1 n_2} d(\mathbf{X}_l, \mathbf{X}_k)$$

Para nuestro ejemplo ilustrativo

$$d(G_1, G_2) = \frac{1}{6} \left[d(\mathbf{X}_1, \mathbf{X}_3) + d(\mathbf{X}_1, \mathbf{X}_4) + d(\mathbf{X}_1, \mathbf{X}_5) + d(\mathbf{X}_2, \mathbf{X}_3) + d(\mathbf{X}_2, \mathbf{X}_4) + d(\mathbf{X}_2, \mathbf{X}_5) \right]$$

Métodos de Agrupamiento

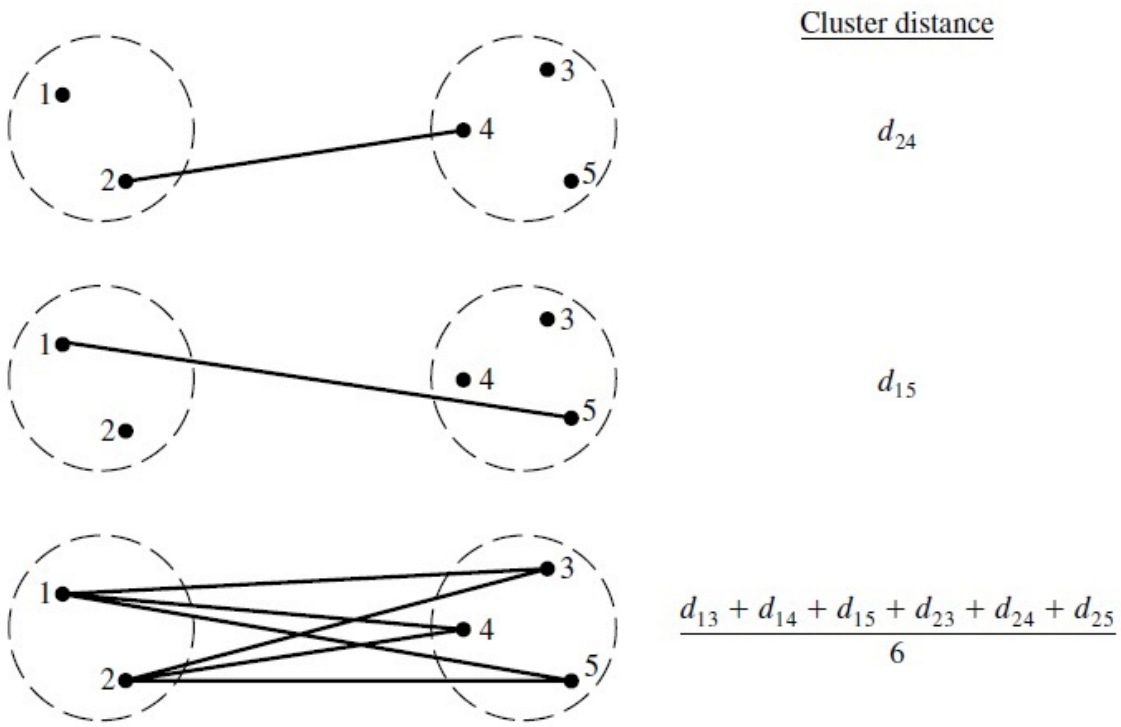


Figura: Figura 12.2 de Johnson and Wichern (2013). Applied Multivariate Statistical Analysis, pp. 681

Métodos de Agrupamiento

- **Método de Ward:**
  Busca que la partición produzca grupos lo más homogeneos posibles de tal forma que las unidades dentro de cada grupo sean homogeneos. Es decir, en cada paso se agrupan conglomerados que minimicen la varianza de los grupos.

En general:

- ✓ El método de ligación simple es incapaz de delinear grupos pocos separados
- ✓ El método de ligación completa tiende a producir conglomerados de aproximadamente el mismo diámetro, además que en los primeros pasos, tiende aislar datos discrepantes de la muestra.

Métodos de Agrupamiento

- ✓ El método de la media de las distancias, tienen a producir conglomerados de aproximadamente la misma varianza interna y produce mejores particiones que los métodos de ligación simple y completa.
- ✓ Los métodos de ligación simples, completa y de la media pueden ser utilizados para variables cualitativas y cuantitativas.
- ✓ El método de Ward es recomendado sólo para variables cuantitativas y tiende a producir grupos con aproximadamente el mismo número de unidades.

Número de Conglomerados

Número de Conglomerados

Criterios para determinar el **número de conglomerados/cluster**:

- **El gráfico de Dendrogramas:**
  Este gráfico es el más utilizado para describir los resultados de un cluster jerarquico.
  El eje vertical indica el nivel de similaridad (o disimilaridad) y en el eje horizontal se colocan las unidades/observaciones en un orden conveniente relativa a la historia de agrupamiento.
- **Análisis del Comportamiento del Coeficiente de Fusión**:
  Se grafica el número de conglomerados de un árbol jerárquico versus el nivel de la distancia (nivel de fusión) del agrupamiento de cada etapa del proceso, y se puede visualizar si hay puntos de salto relativamente grandes con relación a los demás valores de distancia.

Número de Conglomerados

Example

Considere que se tienen $n = 4$ unidades/observaciones y que se tiene la matriz de distancias simples (vecino más cercano)

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix} \end{array}$$

Número de Conglomerados

Observe que la menor distancia es

$$d(A, B) = 2$$

Así que el primer cluster es $\{C, E\}$ y

$$\begin{array}{c} \\ CE \\ A \\ B \\ D \end{array} \begin{array}{c} \begin{array}{cccc} CE & A & B & D \end{array} \\ \begin{pmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{pmatrix} \end{array}$$

pues

$$\text{mín}\{d(A, CE)\} = \text{mín}\, d(A, C), d(A, E) = \text{mín}\{3, 11\} = 3$$
$$\text{mín}\{d(B, CE)\} = \text{mín}\, d(B, C), d(B, E) = \text{mín}\{7, 10\} = 7$$
$$\text{mín}\{d(D, CE)\} = \text{mín}\, d(D, C), d(D, E) = \text{mín}\{9, 8\} = 8$$

Número de Conglomerados

La menor distancia es

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

Así que el segundo cluster es $\{CE, A\}$ y

$$\begin{array}{c} \\ CEA \\ B \\ D \end{array} \begin{array}{c} \begin{array}{ccc} CEA & B & D \end{array} \\ \begin{pmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & \end{pmatrix} \end{array}$$

pues

$$\text{mín}\{d(CEA, B)\} = \text{mín}\, d(C, B), d(E, B), d(A, B) = \text{mín}\{7, 10, 9\} = 7$$
$$\text{mín}\{d(CEA, D)\} = \text{mín}\, d(C, D), d(E, D), d(A, D) = \text{mín}\{9, 8, 6\} = 6$$

Número de Conglomerados

Note que la menor distancia es

$$d(B, D) = 5$$

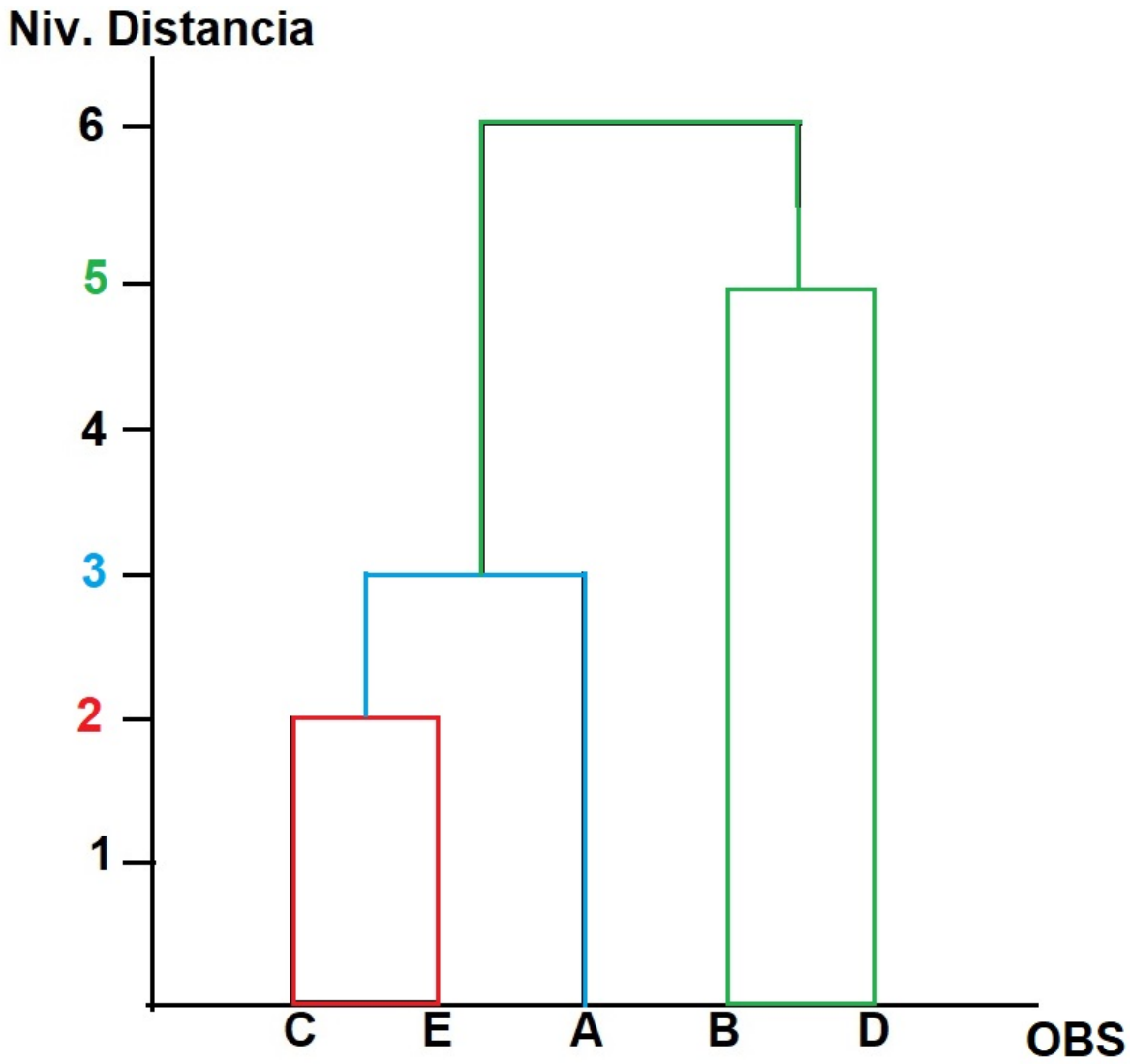
y por lo tanto el tercer y último cluster es $\{B, D\}$.

Resumiendo

Pasos	No. Grupos Formados	Obs. Fusionadas	Nivel de Distancia
1	4	$\{C\}$ y $\{E\}$	2
2	3	$\{CE\}$ y $\{A\}$	3
3	2	$\{CEA\}$ y $\{BD\}$	6

Por lo tanto, el **Dendrograma** es

Número de Conglomerados



Número de Conglomerados

```
x<-matrix(c(0, 9, 3, 6, 11,
            9, 0, 7, 5, 10,
            3, 7, 0, 9, 2,
            6, 5, 9, 0, 8,
            11,10,2, 8, 0),ncol=5)
dimnames(x)<-list(paste("Obs",1:5,sep=""),
                  paste("Obs",1:5,sep="") )
y<-as.dist(x)

# Enlace simple (vecino mas cercano):
cl<-hclust(y,
           method = "single",
           members = NULL)
plot(cl,hang = -1)
```

Número de Conglomerados

```
data("USArrests")
head(USArrests)
Datos<-scale(USArrests)

D<-dist(Datos)
cl<-hclust(D)
plot(cl)

g<-cutree(cl,4)
table(g)

Data_cluster<-cbind(USArrests,g)
aggregate(USArrests,by=list(cluster=g),mean)
```

Número de Conglomerados

```
require(factoextra)
fviz_nbclust(Datos,
             FUN = hcut,
             method = "silhouette")
fviz_nbclust(Datos,
             FUN = hcut,
             method = "wss")
```