Análisis de Datos Multivariantes

3. ANÁLISIS DE CONGLOMERADOS

2016/17

Universidad Carlos III de Madrid

Contenido

1. Introducción
   - Objetivos
   - Distancias
   - Ejemplos
   - Descomposición de la variación total
2. Métodos jerárquicos
   - Componentes del algoritmo de asociación jerárquica
   - Monotonía
   - Ejemplo
3. Método de Ward
   - Motivación
   - Distancia de Ward
   - Ejemplo
4. Algoritmos no jerárquicos
   - Construcción
   - Ejemplo

Objetivos

Los **objetivos** de este tema son:

- Describir los fundamentos del **Análisis de Conglomerados** como técnica exploratoria de agrupación de n observaciones en k clases.
- Analizar **ejemplos** de aplicación.

Dada una **matriz de datos** $\mathcal{X} = (x_{ij})$ de $n \times p$:

- Existe un **gran número posible** de divisiones en k grupos disjuntos no vacíos de tamaños **arbitrarios**, $k = 1, \ldots, n$.
- Agrupación **con sentido** de proximidad basada en medidas numéricas de **distancia.**

Distancias entre individuos

- **Individuos.** Filas genéricas de $\mathcal{X} \to \mathbf{x}^T$, $\mathbf{y}^T$, donde:

$$\mathbf{x} = (x_1, \ldots, x_p)^T \quad , \quad \mathbf{y} = (y_1, \ldots, y_p)^T \ .$$

- **Ejemplos de distancias.** $(\mathbf{x}, \mathbf{y}) \to d(\mathbf{x}, \mathbf{y})$, donde:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \text{ (\textbf{Unidades homogéneas})}.$$

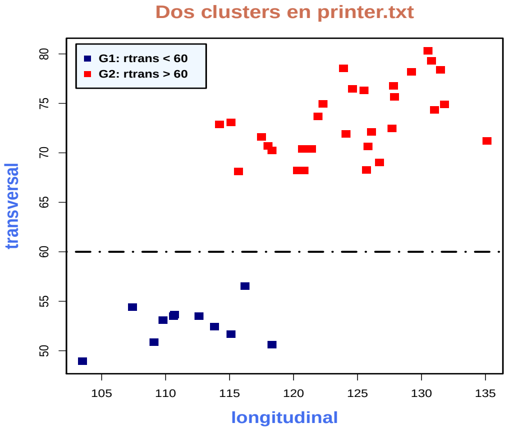
$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{D}^{-1/2}(\mathbf{x} - \mathbf{y})\| \text{ (\textbf{Tipificación})}.$$

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{S}^{-1/2}(\mathbf{x} - \mathbf{y})\| \text{ (\textbf{Mahalanobis})}.$$

- **Elección subjetiva**, naturaleza de datos, ...
- **Matriz** $n \times n$ de **distancias entre individuos:**

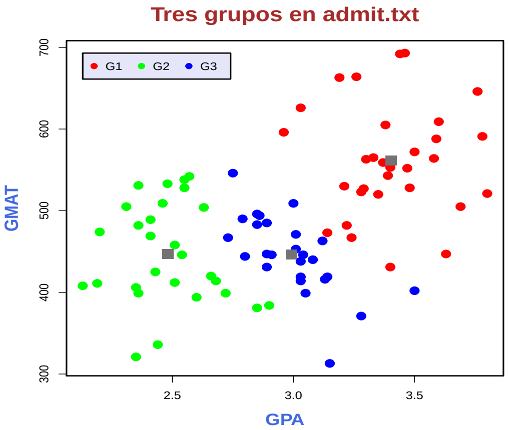
$$(d(\mathbf{x}_i, \mathbf{x}_j) : i, j = 1, \ldots, n) \ .$$

Ejemplo 1: Calidad del Papel de Impresora



Nube de puntos resistencia transversal frente a longitudinal

Ejemplo 2: Admisión



Nube de puntos GMAT frente a GPA

Descomposición de la variación total

Universidad Carlos III de Madrid

Dada una **partición** de $\mathcal{X} = (x_{ij})$ en k grupos $\{C_1, \ldots, C_k\}$, considerar:

- **Matriz** $p \times p$ **de variación total** en $\mathcal{X} = (x_{ij})$:

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T .$$

- **Matriz** $p \times p$ **de variación externa** (**entre grupos**):

$$\mathbf{B}_k = \sum_{l=1}^{k} \frac{n_l}{n} (\overline{\mathbf{x}}_l - \overline{\mathbf{x}})(\overline{\mathbf{x}}_l - \overline{\mathbf{x}})^T .$$

- **Matriz** $p \times p$ **de variación interna** (**dentro de los grupos**):

$$\mathbf{W}_k = \frac{1}{n} \sum_{l=1}^{k} \sum_{J=1}^{n_l} (\mathbf{x}_{lJ} - \overline{\mathbf{x}}_l)(\mathbf{x}_{lJ} - \overline{\mathbf{x}}_l)^T = \sum_{l=1}^{k} \frac{n_l}{n} \mathbf{S}_l .$$

Descomposición de la variación total

Descomposición de la variación total

$$\mathbf{S} = \mathbf{B}_k + \mathbf{W}_k \ .$$

Como **consecuencia:**

$$\boxed{\mathrm{VT} = \mathrm{VE}_k + \mathrm{VI}_k}$$

- $\mathrm{VT} = \mathrm{tr}(\mathbf{S}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \overline{\mathbf{x}}\|^2/n$.
- $\mathrm{VE}_k = \mathrm{tr}(\mathbf{B}_k) = \sum_{l=1}^{k}(n_l/n)\|\overline{\mathbf{x}}_l - \overline{\mathbf{x}}\|^2$.
- $\mathrm{VI}_k = \mathrm{tr}(\mathbf{W}_k) = \sum_{l=1}^{k}\sum_{J=1}^{n_l} \|\mathbf{x}_{lJ} - \overline{\mathbf{x}}_l\|^2/n = \sum_{l=1}^{k}(n_l/n)tr(\mathbf{S}_l)$.

Definición de grupos:

- **Precisa** en Ej. 1. $\mathrm{VE}_2/\mathrm{VT} = .7635 = 76.35\%$.
- **Imprecisa** en Ej. 2. $\mathrm{VE}_3/\mathrm{VT} = .4630 = 46.30\%$.

Métodos jerárquicos

Componentes del **algoritmo de asociación** en k grupos $\{C_1, \ldots, C_k\}$:

- Criterios de **distancia entre grupos.** Dados A y B:
  - **Simple.** $D(A,B) = \min_{\mathbf{x}\in A,\ \mathbf{y}\in B} d(\mathbf{x},\mathbf{y})$.
  - **Completa.** $D(A,B) = \max_{\mathbf{x}\in A,\ \mathbf{y}\in B} d(\mathbf{x},\mathbf{y})$.
  - **Promedio.** $D(A,B) = \sum_{\mathbf{x}\in A,\ \mathbf{y}\in B} d(\mathbf{x},\mathbf{y})/n_A n_B$.
- Proceder **secuencialmente** para $k = n,\ n-1,\ \ldots,$ **formando en cada paso** el grupo

$$AB = \arg \min_{i \neq j = 1, \ldots, k} D(C_i, C_j)\ .$$

- **Terminar** cuando $D_k = \min_{i\neq j = 1, \ldots, k} D(C_i, C_j)$ es **grande.**

Monotonía

Propiedad de monotonía

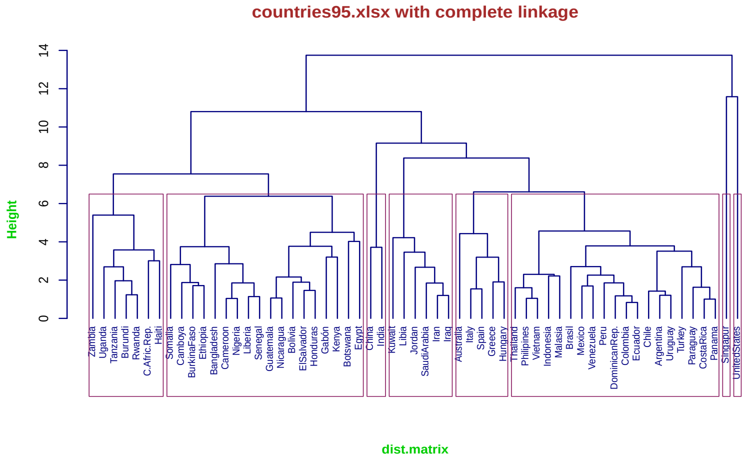
Para los criterios de asociación **simple**, **completa** y **promedio**, se tiene:

$$D_n \leq D_{n-1} \leq \cdots \leq D_2 \ .$$

- Los grupos **nunca disminuyen** de tamaño.
- **Estabilidad** en las asociaciones del algoritmo.
- **Demostración** en `lectura3_1.pdf`.

Ejemplo 3: Características de países

Universidad Carlos III de Madrid



Dendrograma con asociación completa (k = 8)

Método de Ward

- Considerar una **partición** en k grupos $\{C_1, \ldots, C_k\}$.

Lema 1

Al **asociar dos grupos** A y B, se tiene

$$\mathrm{VI}_{k-1} = \mathrm{VI}_k + \frac{1}{n}\frac{n_A n_B}{n_A + n_B}\left\|\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B\right\|^2 .$$

Como **consecuencia:**

- VI_k crece cuando el número de grupos k decrece.
- VE_k decrece cuando el número de grupos k decrece.
- **Demostración** en `lectura3_1.pdf`.

Ejemplo 4: districts.txt

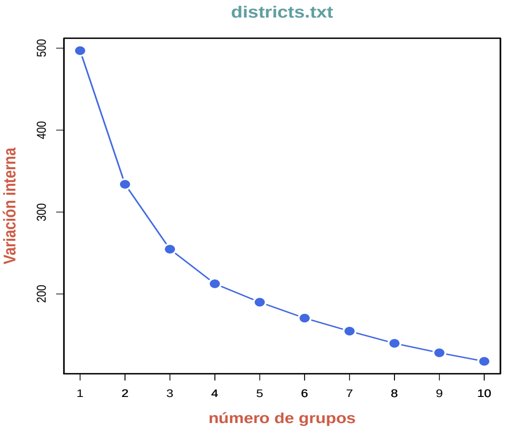


Gráfico de VI_k como función del número de grupos

Distancia de Ward

Universidad Carlos III de Madrid

- Dada $\{C_1, \ldots, C_k\}$, A y B se asocian cuando proporcionan el **menor incremento posible** de la variación interna. Es decir, si **minimizan**

$$W(A,B) = \frac{n_A n_B}{n_A + n_B} \|\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B\|^2 = \frac{\|\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B\|^2}{(1/n_A) + (1/n_B)} .$$

- $W_k = \min_{i \neq j = 1,\ldots,k} D(C_i, C_j)$. **Monotonía:**

$$W_n \leq W_{n-1} \leq \cdots \leq W_2 .$$

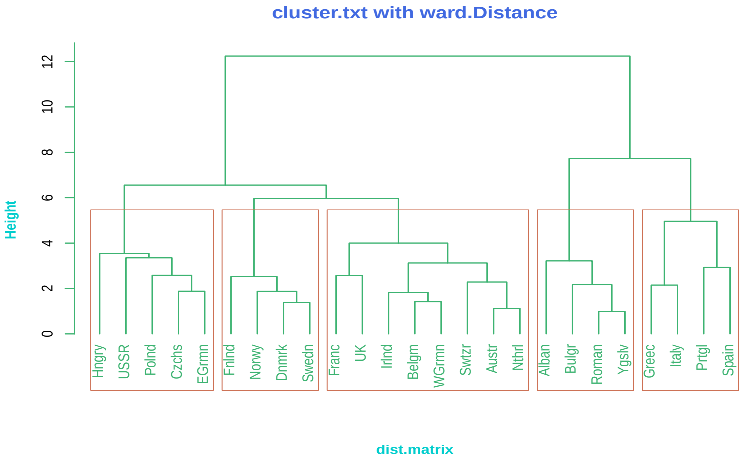
Lema 2

Para los grupos A, B, y C de la partición, se tiene:

$$W(AB, C) = \frac{(n_A + n_C)W(A,C) + (n_B + n_C)W(B,C) - n_C W(A,B)}{n_A + n_B + n_C}$$

- **Demostración** en `lectura3_1.pdf`.

Ejemplo 5: cluster.txt



Dendrograma con el método de Ward

Método de las $k-$medias

Universidad Carlos III de Madrid

Algoritmo de las $k-$medias

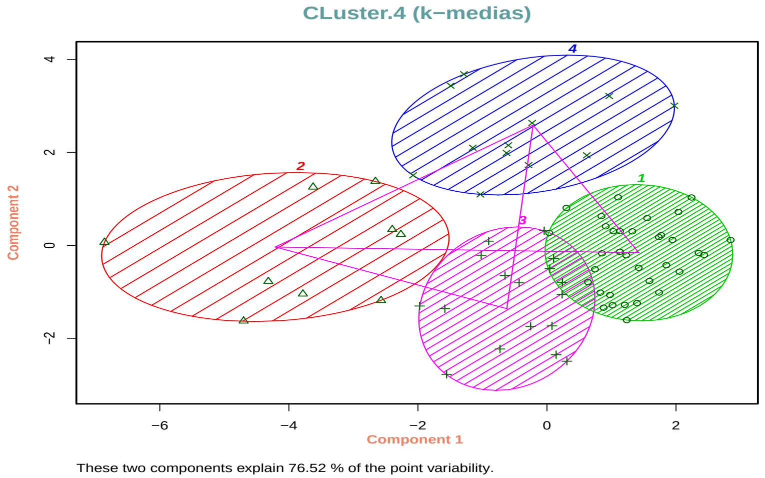
- Paso 1. Dividir la base de datos **aleatoriamente** en k grupos con **centros** $\bar{\mathbf{x}}_l$, $l = 1, \ldots, k$.
- Paso 2. Formar nuevos grupos mediante la regla

$$\textbf{Asignar } \mathbf{x}_i \rightarrow C_l \colon \arg\min_{1 \leq l \leq k} \left\| \mathbf{x}_i - \bar{\mathbf{x}}_l \right\|^2 .$$

- Paso 3. **Actualizar** los vectores de medias $\bar{\mathbf{x}}_l$, $l = 1, \ldots, k$.
- **Iterar** pasos $2-3$ hasta conseguir **estabilidad** en la **asignación.**

- **Selección** de k usando el **gráfico** de VI_k (**Ejemplo 4**).

Ejemplo 6: districts.txt



$k = 4$ **grupos** en el espacio (PC1, PC2)

Resumen

Universidad Carlos III de Madrid

1. Introducción
2. Métodos jerárquicos
3. Método de Ward
4. Algoritmos no jerárquicos

- **Referencias:** Johnson, R.A. y Wichern, D.W. (2007) [Cap. 12].