

CC0293 - Análise Multivariada
Notas de aula
Prof. Gualberto Agamez Montalvo

1 Análise de Agrupamento

A análise de agrupamento tem como objetivo agrupar n objetos (ou indivíduos) em um número desconhecido de grupos (g).

1.1 Medidas de similaridade ou dissimilaridade

A análise de agrupamento tenta identificar os vetores de observações que são semelhantes e agrupá-los em *clusters* (grupos), muitas técnicas usam um índice de proximidade (similaridade ou dissimilaridade) entre cada par de observações. Algumas medidas de proximidade entre dois vetores são:

- a distância Euclidiana;
- a distância Euclidiana padronizada;
- a distância generalizada de Mahalanobis;
- a métrica de Canberra dada por

$$d_{Ca}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^p \frac{|x_j - y_j|}{x_j + y_j};$$

- a métrica de Czekanowski dada por

$$d_{Cz}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{j=1}^p |x_j - y_j|}{\sum_{j=1}^p (x_j + y_j)};$$

- a métrica de Minkowski dada por

$$d_{Mi}(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^p |x_j - y_j|^r \right)^{\frac{1}{r}}$$

em que r é um número inteiro positivo.

Observação 1: a métrica de Minkowski é igual á distância euclidiana quando $r = 2$.

Observação 2: a escala de medição das variáveis é um fator importante quando se utiliza como medida a distância Euclidiana. Alterar a escala pode afetar as distâncias relativas entre os indivíduos.

Exemplo 1.1. Considere os vetores $\mathbf{y}_1 = (2, 5)$, $\mathbf{y}_2 = (4, 2)$ e $\mathbf{y}_3 = (7, 9)$. Portanto, a matriz de distâncias euclidianas $\mathbf{D} = (d_{ij})$ é

$$\mathbf{D} = \begin{bmatrix} 0 & 3,6056 & 6,4031 \\ 3,6056 & 0 & 7,6158 \\ 6,4031 & 7,6158 & 0 \end{bmatrix}.$$

Agora, suponha que $\mathbf{y}_1 = (200, 5)$, $\mathbf{y}_2 = (400, 2)$ e $\mathbf{y}_3 = (700, 9)$. Então,

$$\mathbf{D} = \begin{bmatrix} 0 & 200,0225 & 500,0160 \\ 200,0225 & 0 & 300,0817 \\ 500,0160 & 300,0817 & 0 \end{bmatrix}.$$

Podemos pensar que no primeiro exemplo as unidades da primeira componente estão em metros e no segundo em centímetros. **Como muda a relação de proximidade das observação?**

```
> y1 <- c(2, 5); y2 <- c(4, 2); y3 <- c(7, 9)
> A <- rbind(y1, y2, y3); A
  [,1] [,2]
y1    2    5
y2    4    2
y3    7    9

> dist(A, method = "euclidean")
      y1      y2
y2 3.605551
y3 6.403124 7.615773
```

1.2 Agrupamento hierárquico

O número de maneiras de particionar n indivíduos em g grupos ou clusters é dado por

$$N = \frac{1}{g!} \sum_{k=1}^g \binom{g}{k} (-1)^{g-k} k^n.$$

Alguns métodos de agrupamento hierárquicos são:

1. **Vizinho mais próximo:** a distância entre dois clusters A e B é definida como a distância mínima entre os pontos em A e os pontos em B , isto é,

$$D(A, B) = \min \{d(\mathbf{y}_i, \mathbf{y}_j) \text{ para } \mathbf{y}_i \in A \text{ e } \mathbf{y}_j \in B\}$$

em que $d(\mathbf{y}_i, \mathbf{y}_j)$ é a distância Euclidiana (ou outra distância ou métrica) entre os vetores \mathbf{y}_i e \mathbf{y}_j .

Em cada etapa a distância $D(A, B)$ é encontrada para cada par de clusters e mesclamos os dois clusters com a menor distância em um único cluster. Depois que os dois clusters são mesclados, o procedimento é repetido para na próxima etapa: as distâncias entre todos os pares de clusters são calculadas novamente e o par com distância mínima é mesclado em um único cluster.

Exemplo 1.2. Considere o seguinte conjunto de dados para exemplificar o método do vizinho mais próximo.

Cidade	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Atlanta	16,5	24,8	106	147	1112	905	494
Boston	4,2	13,3	122	90	982	669	954
Chicago	11,6	24,7	340	242	808	609	645
Dallas	18,1	34,2	184	293	1668	901	602
Denver	6,9	41,5	173	191	1534	1368	780
Detroit	13,0	35,7	477	220	1566	1183	788

Passo 1: calcular as distâncias Euclidianas com todas as cidades.

	Atlanta	Boston	Chicago	Dallas	Denver
Boston	536.6419				
Chicago	516.3700	447.4033			
Dallas	590.1753	833.0708	924.0035		
Denver	693.5741	914.9784	1073.3948	527.6673	
Detroit	716.1962	881.0858	971.5271	464.4677	358.6654

Neste caso, as cidades mais próximas são Detroit e Denver (358,6654). Portanto, $G_1 = \{\text{Detroit, Denver}\}$.

Passo 2: calcular as distâncias Euclidianas entre Atlanta, Boston, Chicago, Dallas e G_1 .

	Atlanta	Boston	Chicago	Dallas
Atlanta	0			
Boston	536.6419	0		
Chicago	516.3700	447.4033	0	
Dallas	590.1753	833.0708	924.0035	0
G1	693.5741	881.0858	971.5271	464.4677

Note que todos os elementos desta matriz são obtidos da matriz de proximidade do passo 1. Neste caso as cidades mais similares são Boston e Chicago (447,4033). Portanto, $G_2 = \{\text{Boston, Chicago}\}$.

Passo 3: calcular as distâncias Euclidianas entre Atlanta, G_2 , Dallas e G_1 .

	Atlanta	G2	Dallas
Atlanta	0		
G2	516.3700	0	
Dallas	590.1753	833.0708	0
G1	693.5741	881.0858	464.4677

Neste caso os grupos mais similares são Dallas e G_1 (464.4677). Portanto, $G_3 = \{\text{Dallas}, G_1\}$.

Passo 4: calcular as distâncias Euclidianas entre Atlanta, G_2 , G_3 .

	Atlanta	G2
Atlanta	0	
G2	516.3700	0
G3	590.1753	833.0708

Neste caso os grupos mais similares são Atlanta e G_2 (516.3700). Portanto, $G_4 = \{\text{Atlanta}, G_2\}$.

Passo 5: calcular as distâncias Euclidianas entre G_3 e G_4 .

	G4
G3	590.1753

Portanto, temos o grupo $G_5 = \{G_3, G_4\}$.

Finalmente, o dendrograma é uma representação gráfica do método de agrupamento. Neste podemos identificar a ordem de construção dos grupos.

```
> require(fpc)

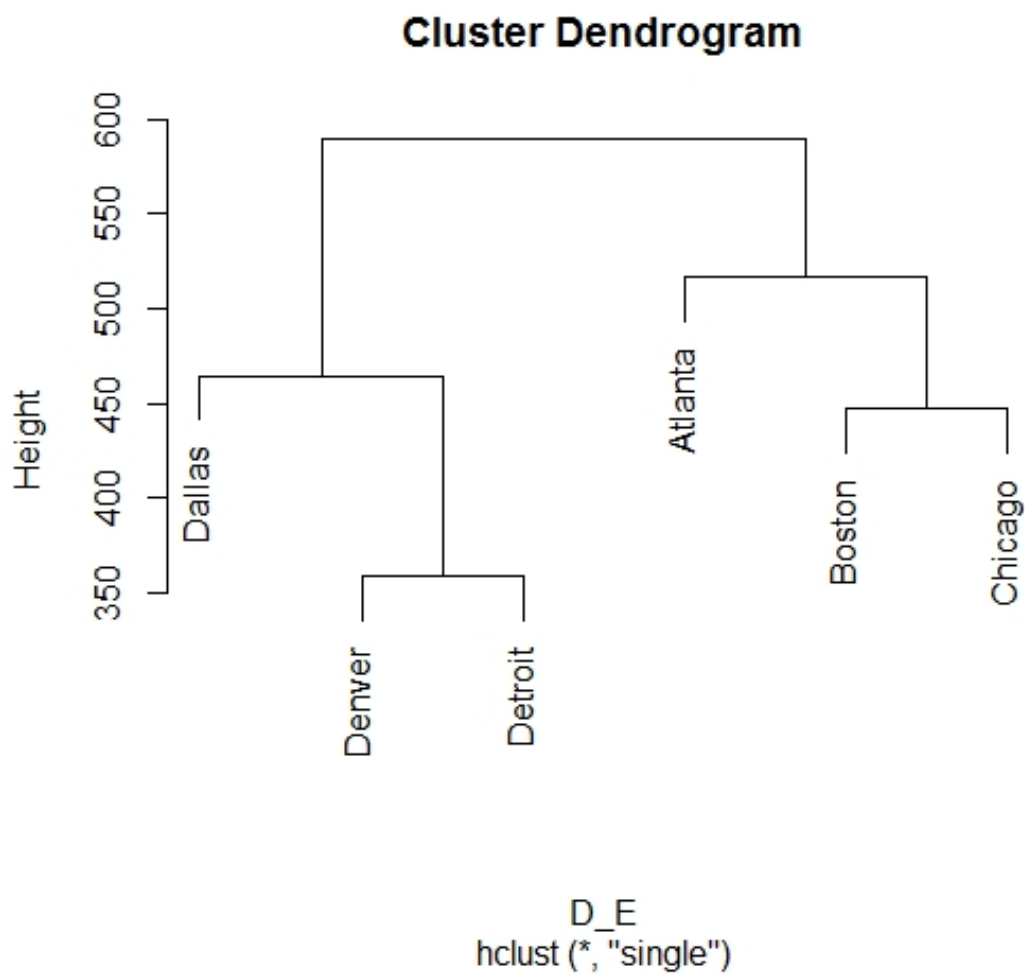
> # Distâncias
> D_E <- dist(Dados, method = "euclidean")

> # O vizinho mais próximo
> m <- hclust(D_E, method="single")
> m$height
[1] 358.6654 447.4033 464.4677 516.3700 590.1753
```

```
> # Fazer dois grupos
> grupos <- cutree(m, k=2); grupos
Atlanta  Boston Chicago  Dallas  Denver Detroit
          1          1      1        2          2          2

> # Fazer três grupos
> grupos <- cutree(m, k=3); grupos
Atlanta  Boston Chicago  Dallas  Denver Detroit
          1          2      2        3          3          3

> plot(m)
```



2. **Vizinho mais distante:** a distância entre dois clusters A e B é definida como a distância máxima entre os pontos em A e os pontos em B , isto é,

$$D(A, B) = \max \{d(\mathbf{y}_i, \mathbf{y}_j) \text{ para } \mathbf{y}_i \in A \text{ e } \mathbf{y}_j \in B\}$$

em que $d(\mathbf{y}_i, \mathbf{y}_j)$ é a distância Euclidiana (ou outra distância ou métrica) entre os vetores \mathbf{y}_i e \mathbf{y}_j .

Exemplo 1.3. Usar conjunto de dados do Exemplo 1.2 para exemplificar o método do vizinho mais distante.

Passo 1: calcular as distâncias Euclidianas com todas as cidades.

	Atlanta	Boston	Chicago	Dallas	Denver
Boston	536.6419				
Chicago	516.3700	447.4033			
Dallas	590.1753	833.0708	924.0035		
Denver	693.5741	914.9784	1073.3948	527.6673	
Detroit	716.1962	881.0858	971.5271	464.4677	358.6654

Neste caso, as cidades mais próximas são Detroit e Denver (358,6654). Portanto, $G_1 = \{\text{Detroit}, \text{Denver}\}$.

Passo 2: calcular as distâncias Euclidianas entre Atlanta, Boston, Chicago, Dallas e G_1 .

	Atlanta	Boston	Chicago	Dallas
Atlanta	0			
Boston	536.6419	0		
Chicago	516.3700	447.4033	0	
Dallas	590.1753	833.0708	924.0035	0
G_1	716.1962	914.9784	1073.3948	527.6673

Neste caso as cidades mais similares são Boston e Chicago (447.4033). Portanto, $G_2 = \{\text{Boston, Chicago}\}$.

Passo 3: calcular as distâncias Euclidianas entre Atlanta, G_2 , Dallas e G_1 .

	Atlanta	G2	Dallas
Atlanta	0		
G2	536.6419	0	
Dallas	590.1753	924.0035	0
G1	716.1962	1073.3948	527.6673

Neste caso os grupos mais similares são Dallas e G_1 (527.6673). Portanto, $G_3 = \{\text{Dallas, } G_1\}$.

Passo 4: calcular as distâncias Euclidianas entre Atlanta, G_2 , G_3 .

	Atlanta	G2
Atlanta	0	
G2	536.6419	0
G3	716.1962	1073.3948

Neste caso os grupos mais similares são Atlanta e G_2 (536.6419). Portanto, $G_4 = \{\text{Atlanta, } G_2\}$.

Passo 5: calcular as distâncias Euclidianas entre G_3 e G_4 .

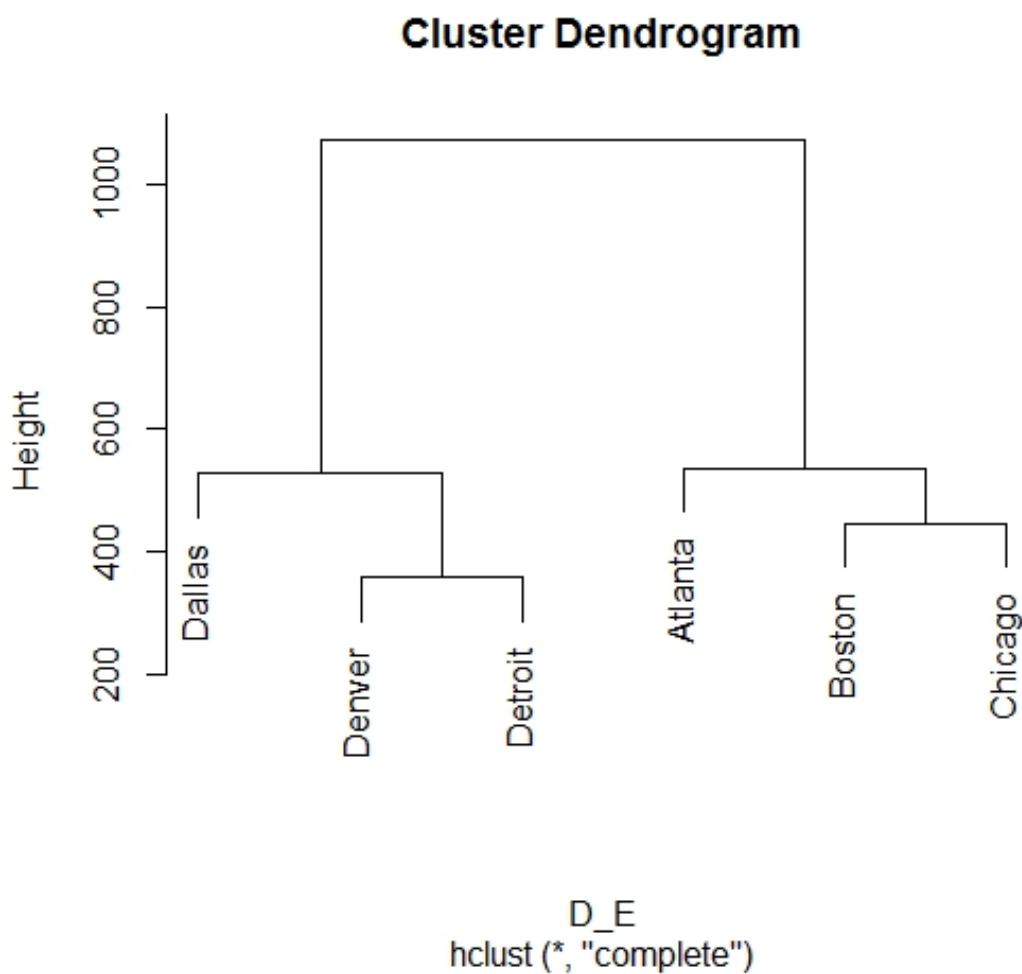
	G4
G3	1073.3948

Portanto, temos o grupo $G_5 = \{G_3, G_4\}$.

```
> # O vizinho mais distante
> m <- hclust(D_E, method="complete")
```

```
> m$height
[1] 358.6654 447.4033 527.6673 536.6419 1073.3948

> plot(m)
```



3. **Centroide:** a distância entre os conglomerados A e B é definida como a distância Euclidiana entre a média dos vetores (geralmente chamados de centroides) dos dois clusters

$$D(A, B) = d(\bar{y}_A, \bar{y}_B),$$

em que \bar{y}_A e \bar{y}_B são as médias dos vetores em A e B , respectivamente.

4. **Ligação média:** a distância entre os conglomerados A e B é definida como

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\mathbf{y}_i, \mathbf{y}_j),$$

em que n_A e n_B são o número de pontos em A e B , respectivamente.

Exercício: usar o conjunto de dados do exemplo associado ao vizinho mais próximo para exemplificar o método do centroide e da ligação média. Fazer no R (mediante algum pacote ou programando) e manualmente (pode usar a ajuda do R).