Análise de Agrupamentos

Prof. George von Borries

Departamento de Estatística
Universidade de Brasília

2023

Terminologia

- **Análise de Agrupamentos** (Clustering)

  Encontrar algoritmos ótimos para agrupar objetos que são semelhantes de acordo com algum critério. Não existem grupos pré-definidos. Não sabemos sequer o número de grupos.

  Técnica não supervisionada.

- **Análise de Discriminantes** (Discrimination)

  Encontrar "discriminantes" que separam ao máximo grupos pré-definidos de acordo com uma variável de supervisão.

  Técnica supervisionada.

- **Análise de Classificações** (Classification)

  Encontrar regras de alocação de objetos em grupos pré-definidos de maneira a minimizar o erro de alocação.

Introdução

- Técnica de aprendizado estatístico não supervisionada. Também conhecida como aprendizado de máquinas, mineração de dados, reconhecimento de padrões.

- Objetivos:

  - ✓ Descobrir padrões nos dados que formam grupos de objetos similares;
  - ✓ Exploração/mineração de dados;
  - ✓ Suporte para técnicas visuais;
  - ✓ Identificação de subgrupos em dados complexos;
  - ✓ Identificação de padrões no tempo e espaço;
  - ✓ Redução de dados.

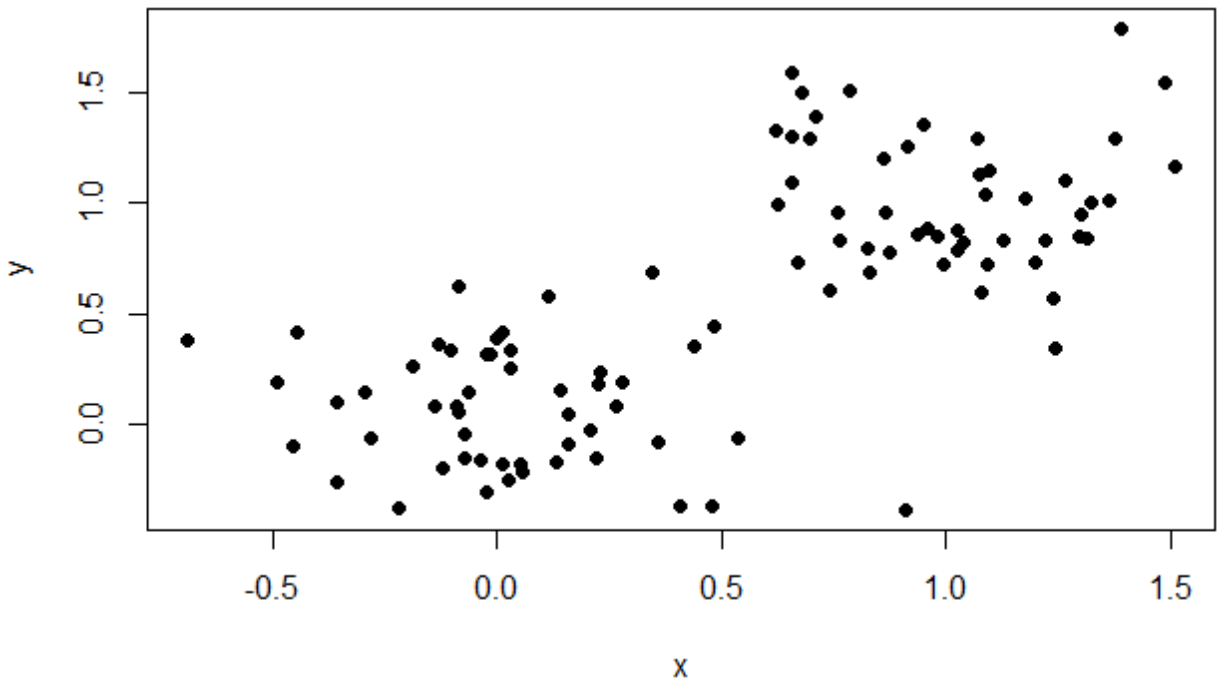
- Áreas de Aplicação: biologia, zoologia, psiquiatria, patologia, sociologia, arqueologia, geografia, geologia, engenharia, medicina, economia, ciência forense,...

- Outras denominações:

  - ✓ Aprendizado não supervisionado;
  - ✓ Reconhecimento de padrões (engenharia);
  - ✓ Taxonomia numérica (biologia e ecologia);
  - ✓ Partição (teoria dos grafos);
  - ✓ Segmentação de dados (marketing);
  - ✓ *Class discovery* (genética).
  - ✓ Mineração de dados (computação).

Exemplo - `ExemplosNA.R: (1) XYdata`

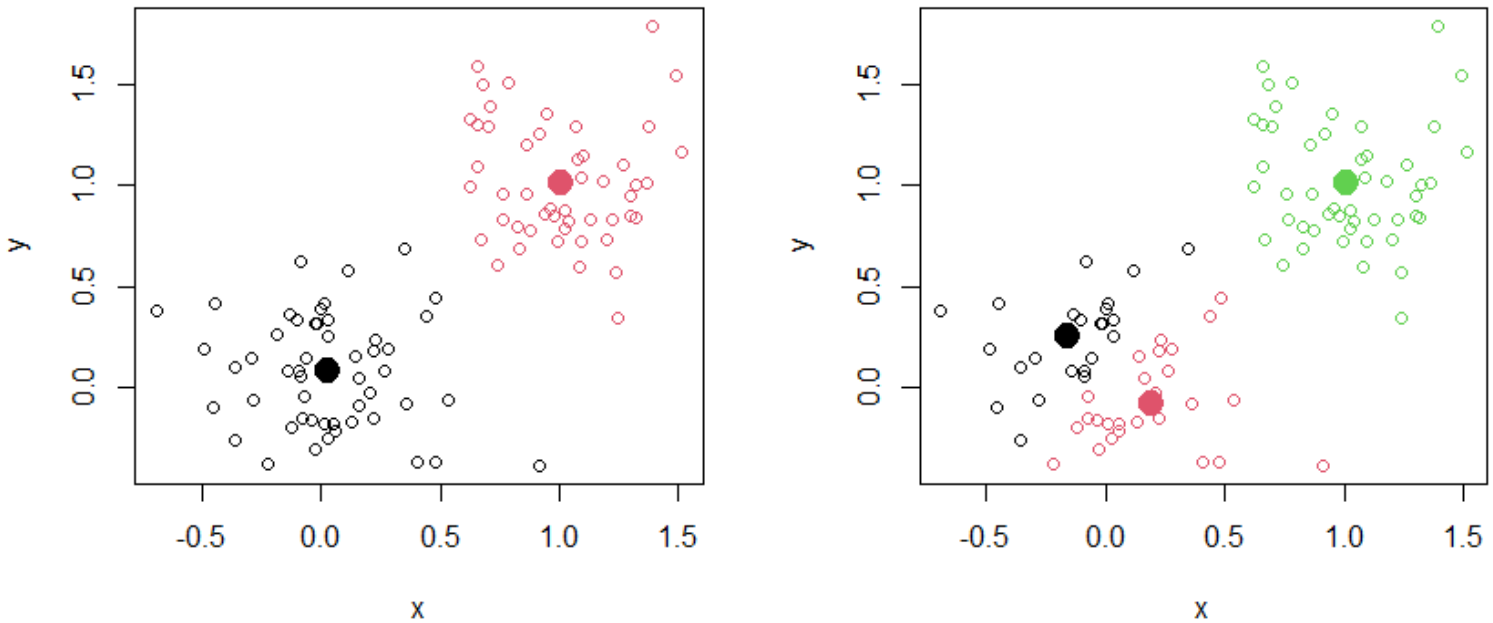
- Sejam $\mathbf{X} \sim \mathrm{N}(0, 0.09)$ e $\mathbf{Y} \sim \mathrm{N}(1, 0.09)$.
  Queremos agrupar pares $(\mathbf{X}, \mathbf{Y})$.



Dados gerados.

Exemplo

- Sejam $\mathbf{X} \sim \mathrm{N}(0, 0.09)$ e $\mathbf{Y} \sim \mathrm{N}(1, 0.09)$.
  Queremos agrupar pares $(\mathbf{X}, \mathbf{Y})$.



Dados agrupados utilizando técnica de K-Médias com g=2 e g=3

Procedimentos

- **Estruturação dos dados**

  Os dados são analisados na forma bruta, visando identificar algum padrão (característica) de agrupamento natural entre eles. As características devem ser escolhidas de modo a agregar o máximo de informação possível sobre o estudo.

- **Modelagem**

  Nesta fase, define-se o conceito de *cluster* e o critério a ser adotado na formação dos grupos.

  - **Medidas de Distância** quantificam o quão homogêneas ou similares duas características são.
  - **Critério de Agrupamento** está intimamente ligado à sensibilidade do pesquisador e ao seu conhecimento a respeito do conjunto de dados.
    Como o pesquisador espera que os dados irão se agrupar? [1]

[1] Certa subjetividade está presente nesta fase e isto é muitas vezes criticado.

- **Otimização** está relacionada à escolha adequada do algoritmo computacional que conseguirá desvendar a estrutura de agrupamento dos dados.
- **Validação** tem objetivo de verificar a compatibilidade dos resultados obtidos e com as expectativas do pesquisador.
- **Interpretação dos Resultados** o especialista na área de aplicação deve integrar os resultados do agrupamento com outras evidências experimentais e análise, a fim de tirar as conclusões corretas.

- Tipos de dados: nominal, ordinal, intervalar e de razão.

  Muitas vezes é útil a mudança de uma escala para outra[2].

[2]Referência: Gan, Ma e Wu. (2007) Data Clustering: Theory, Algorithms and Applications, SIAM.

Técnicas Gráficas

- Primeiro e último passos na análise de agrupamentos.
- Procedimento exploratório ou confirmatório do agrupamento obtido.

  - ✓ Gráfico de Perfis (Profile Plot)
  - ✓ Gráfico de Andrews
  - ✓ Gráfico de Intensidades ou Calor (Heatmap)
  - ✓ Chernoff Faces

- Referências:
  - ✓ Exemplos em `R` em `ExemplosNA.R`.
  - ✓ Moustafa, R.E. e Hadi, A.S. (2009) Grand tour and the Andrews plot, WIRE Computational Statistics, Vol. 1, John Wiley & Sons.
  - ✓ Using `R` to draw a Heatmap from microarray data: `https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/heatmap/`
  - ✓ Ploner, A. (2023) Creating heatmaps using package `Heatplus`. Karolinska Institutet, Stockholm.

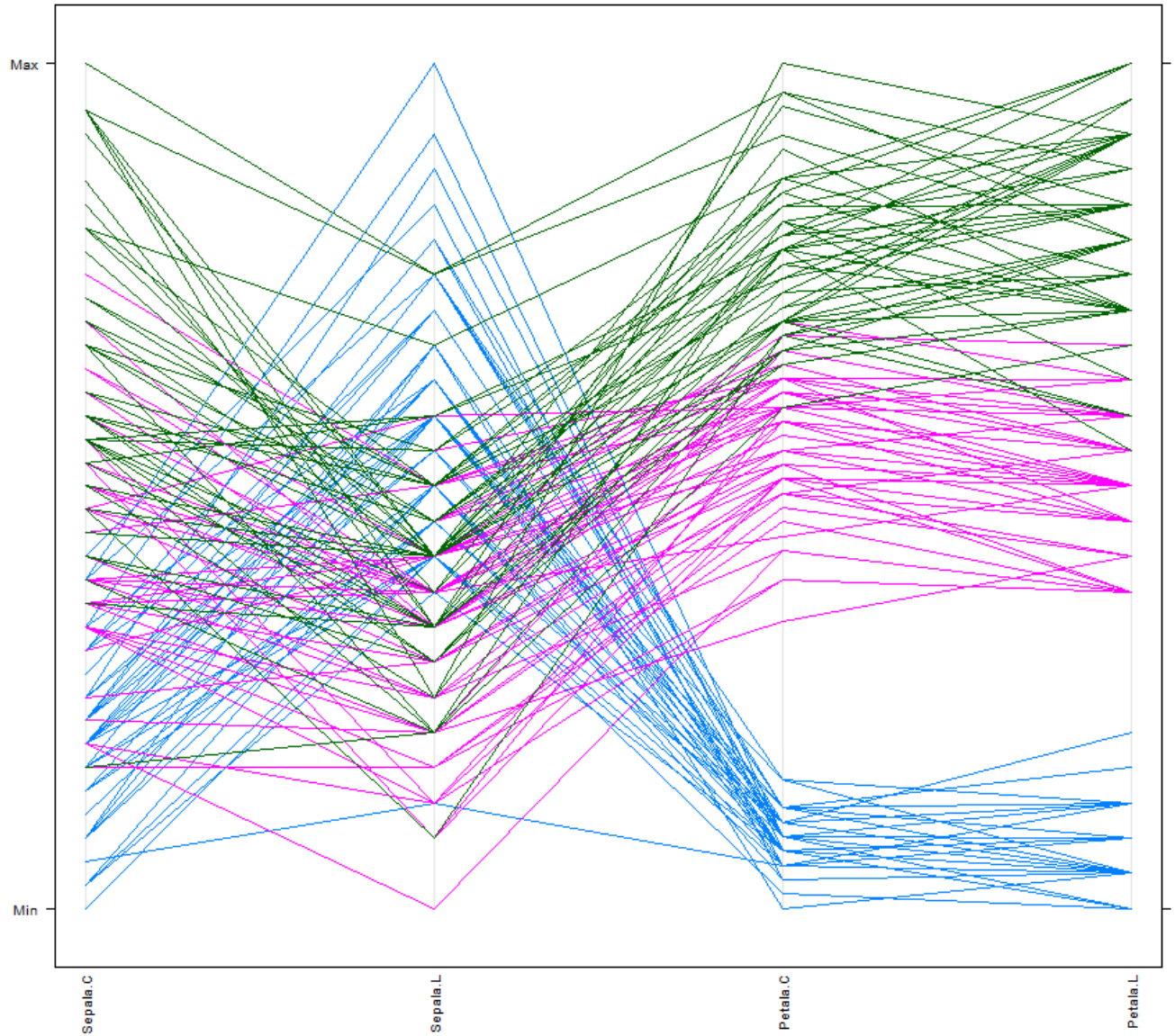
Exemplo: Iris Data `ExemplosNA.R: (2) Analise Grafica`

Iris é um género de plantas com flor, muito apreciado pelas suas diversas espécies, que ostentam flores de cores muito vivas. São, vulgarmente, designadas como lírios (Wikipédia).

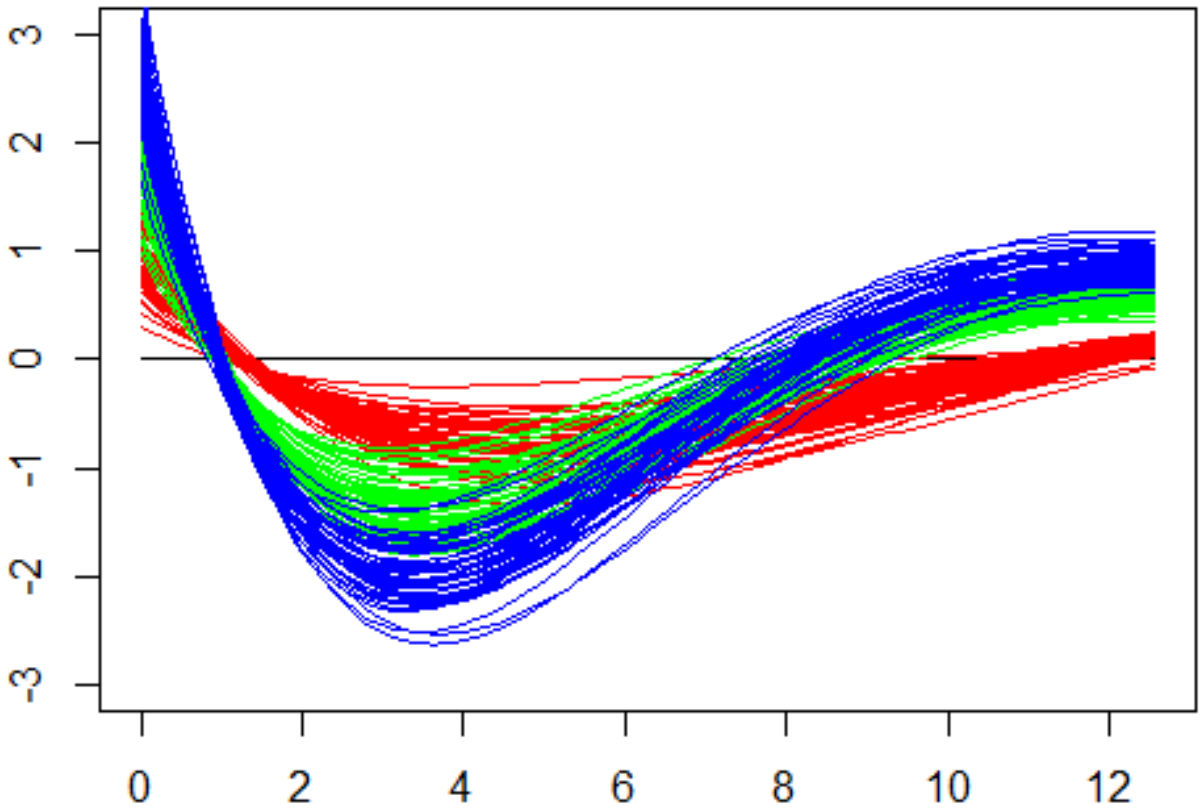
Problema: classificar três tipos diferentes de flores Iris, chamadas setosa, versicolor e virginica. As características disponíveis são: largura e comprimento da sépala (parte da flor que dá sustentação a pétala) e pétala ($p = 4$).

Três tipos de flores Iris: (a) setosa, (b) versicolor, (c) virginica. Fonte: Murphy, K.P. *Machine Learning*, 2012.

Iris Data - Gráfico de Perfis



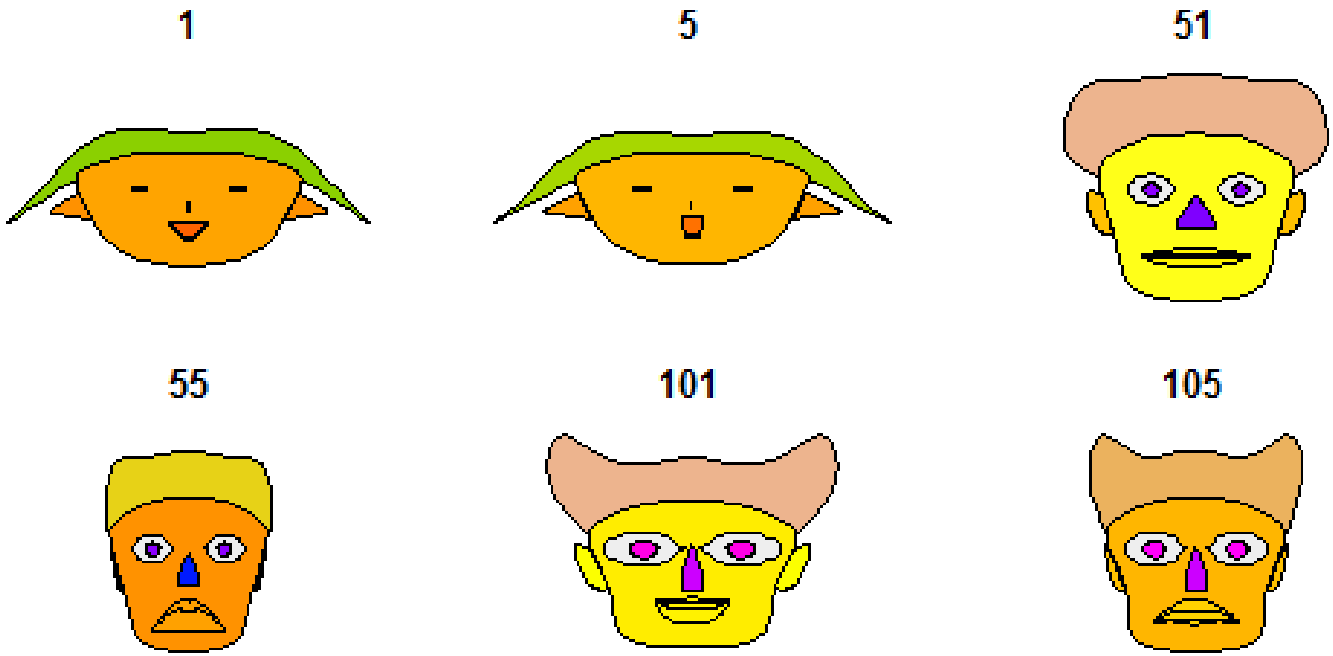
Iris Data - Andrews Plot



Andrews Plot com pacote ANDREWS do R.

$$f(t) = x_1 \times cos(t) + x_2 \times cos(2 \times t)^{0.5} + x_3 \times cos(3 \times t)^{0.5} + \ldots$$

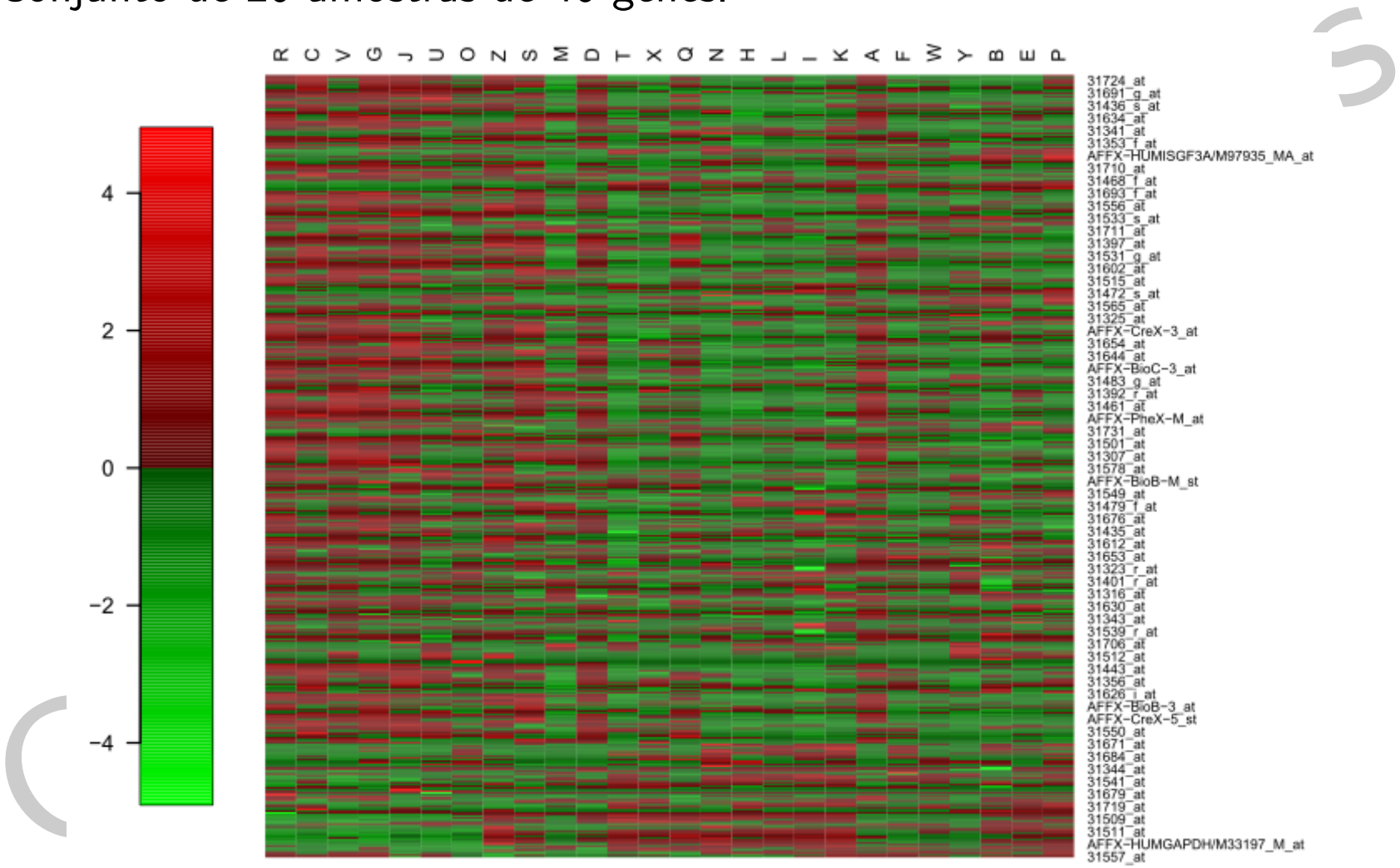
Iris Data - Faces de Chernoff



Faces com pacote APLPACK.
Iris data, sendo 1 e 5 setosa, 51 e 55 versicolor, 101 e 105 virginica.

Expressão Genética - Gráfico de Intensidades

Conjunto de 26 amostras de 46 genes.



Heatmap com pacote HEATPLUS do R.

Ploner, A. (2023) Creating heatmaps using package Heatplus.

Medidas de Proximidade

- Medidas de Dissimilaridade (d):

  Para dois vetores $\mathbf{x}$ e $\mathbf{y}$ no espaço p-dimensional, uma medida d é definida como função de distância se

  - $d(\mathbf{x}, \mathbf{x}) = 0$;
  - $d(\mathbf{x}, \mathbf{y}) \geq 0$;
  - $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$;

  - Estas medidas podem ser extendidas para subgrupos de X e Y.

  - Se $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ chamamos de métrica.
  - Se $d(\mathbf{x}, \mathbf{y}) \leq max(d(\mathbf{x}, \mathbf{z}), d(\mathbf{z}, \mathbf{y}))$ chamamos de ultramétrica.

- Medidas de Similaridade (s):

  Uma medida de similaridade s entre $\mathbf{x}$ e $\mathbf{y}$ é definida se

  - $s(\mathbf{x}, \mathbf{y}) > 0$;
  - $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$;
  - $s(\mathbf{x}, \mathbf{y})$ aumenta se a similaridade de $\mathbf{x}$ e $\mathbf{y}$ aumenta.
  - $s(\mathbf{x}, \mathbf{x}) = 1, \quad 0 \leq s(\mathbf{x}, \mathbf{y}) \leq 1.$

Medidas de Disssimilaridade

- Sejam duas observações p-dimensionais, i.e., $\mathbf{X}^{\mathsf{T}} = [X_1, X_2, \ldots, X_p]$ e $\mathbf{Y}^{\mathsf{T}} = [Y_1, Y_2, \ldots, Y_p]$.

- Distância de Minkowinsky:

$$d(x_i, y_i) = \left(\sum_{i=1}^{p} |x_i - y_i|^r \right)^{1/r}, \quad r \geq 1 \tag{1}$$

- Distância Euclideana: $r = 2$
- Distância de Manhattan: $r = 1$
- Distância máxima: $r \to \infty$.

Problema: geralmente dominadas pelas variáveis de maior escala e só funcionam bem em dados compactos ou com grupos isolados.

Solução: padronização.

- Distância de Karl Pearson:

$$d(x_i, y_i) = \sum_{i=1}^{p} \frac{(x_i - y_i)^2}{s_i^2} \tag{2}$$

onde s_i^2 é a variância (covariância) amostral de (x_i, y_i).

Problema: Sensível a outliers.

Solução: Pesos diferentes para diferentes observações.

$$d(x_i, y_i) = \left(\sum_{i=1}^{p} w_i |x_i - y_i|^r \right)^{1/r} \qquad w_i \geq 0 \tag{3}$$

- Distância de Manhattan (**X** e **Y** com valores perdidos)

$$d(x, y) = \sum_{i=1}^{p} \frac{w_i |x_i - y_i|}{\sum_{i=1}^{p} w_i} \tag{4}$$

sendo $w_i = 1$ se (x, y) observados e $w_i = 0$ c.c.

- Distância de Mahalanobis (generalização):

$$d(x, y) = \sqrt{(x - y)' \Sigma^{-1} (x - y)}. \quad (5)$$

Vantagens: aplica um esquema de pesos aos dados e é invariante a transformações não singulares, i.e., se $z = cx$ e $r = cy$, então $d(x, y) = d(z, r)$. Provar.

Problema: Σ^{-1} deve ser estimado. Em pequenas amostras não possui inversa em muitas situações. Em dados superdimensionados o cálculo pode ser inviável. Computacionalmente intensiva.

- Podemos aplicar um esquema de pesos a esta distância.

$$d(x, y) = \sqrt{(x - y)' \Delta \Sigma^{-1} \Delta (x - y)}. \quad (6)$$

onde Δ é uma matriz diagonal de pesos.

Os números $d(a, b)$ são dispostos numa matriz quadrada $(n \times n)$ denominada matriz de dissimilaridade.

$$\mathrm{M}_{\mathrm{diss}}(\mathbf{D}) = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

sendo $d_{ab} = d(x_a, x_b)$ com relação a alguma medida de dissimilaridade $d(\cdot, \cdot)$ e $D = \{x_1, x_2, ..., x_n\}$.

Medidas de Similaridade

- Produto Interno:

$$s(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y} = \sum_{i=1}^{p} x_i y_i, \tag{7}$$

Utilizado em medidas normalizadas. x e y tem o mesmo comprimento. Assim, $-a^2 \leq s(x, y) \leq a^2$ e depende apenas do ângulo entre os vetores x e y.

- Medida relacionada:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}. \tag{8}$$

onde $\|\mathbf{x}\|$ e $\|\mathbf{y}\|$ são os comprimentos de $\mathbf{x}$ e $\mathbf{y}$.

Invariante a rotações, mas não é invariante a transformações lineares.

Problema: uso de dados normalizados para agrupamento.

- O coeficiente de correlação de Pearson ($s_p(\mathbf{x}, \mathbf{y})$) tem $\mathbf{x}$ e $\mathbf{y}$ centralizados.
- $-1 \leq s_p \leq 1$. Pode ser transformado em $0 \leq s_p \leq 1$.

  s_p não depende de x e y, mas dos vetores de diferenças $x_i - \bar{x}$ e $y_i - \bar{y}$, $i = 1, \ldots, p$.

- $d(\mathbf{x}, \mathbf{y})$ pode ser obtida de $s(\mathbf{x}, \mathbf{y})$ por transformação:

$$d(\mathbf{x}, \mathbf{y}) = \frac{1 - s(\mathbf{x}, \mathbf{y})}{2} \quad \in [0, 1] \tag{9}$$

Os números $s(a, b)$ são dispostos numa matriz quadrada $(n \times n)$ denominada matriz de similaridade.

$$\mathrm{M}_{\mathrm{sim}}(\mathbf{D}) = \begin{pmatrix} 1 & s_{12} & \cdots & s_{1n} \\ s_{21} & 1 & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & 1 \end{pmatrix}$$

sendo $s_{ab} = s(x_a, x_b)$ com relação a alguma medida de similaridade $s(\cdot, \cdot)$. e $D = \{x_1, x_2, ..., x_n\}$.

Medidas de Concordância (Dados Binários)

- Variáveis binárias são variáveis que podem receber exatamente dois valores: "sim/não", "0/1", "masculino/feminino".
- Seguindo a abordagem proposta por Gower, as variáveis binárias podem ser divididas em dois tipos:
  - **Simétricas:** os dois valores assumidos por cada variável são igualmente importantes. Exemplos: "casado/solteiro", "canhoto/destro".
  - **Assimétricas:** um dos valores carrega mais importância do que o outro. Exemplo: Para a cor de uma flor, considera-se: "é vermelha" $= 1$, "não é vermelha" $= 0$, sendo que, se $x_{if} = 1$ e $x_{jf} = 1$, significa dizer que as flores i e j possuem a mesma cor e são vermelhas, enquanto que, $x_{if} = 0$ e $x_{jf} = 0$ implica que as flores não são vermelhas (de mesma cor ou não).

- Sejam dois vetores assumindo valores 0 e 1:

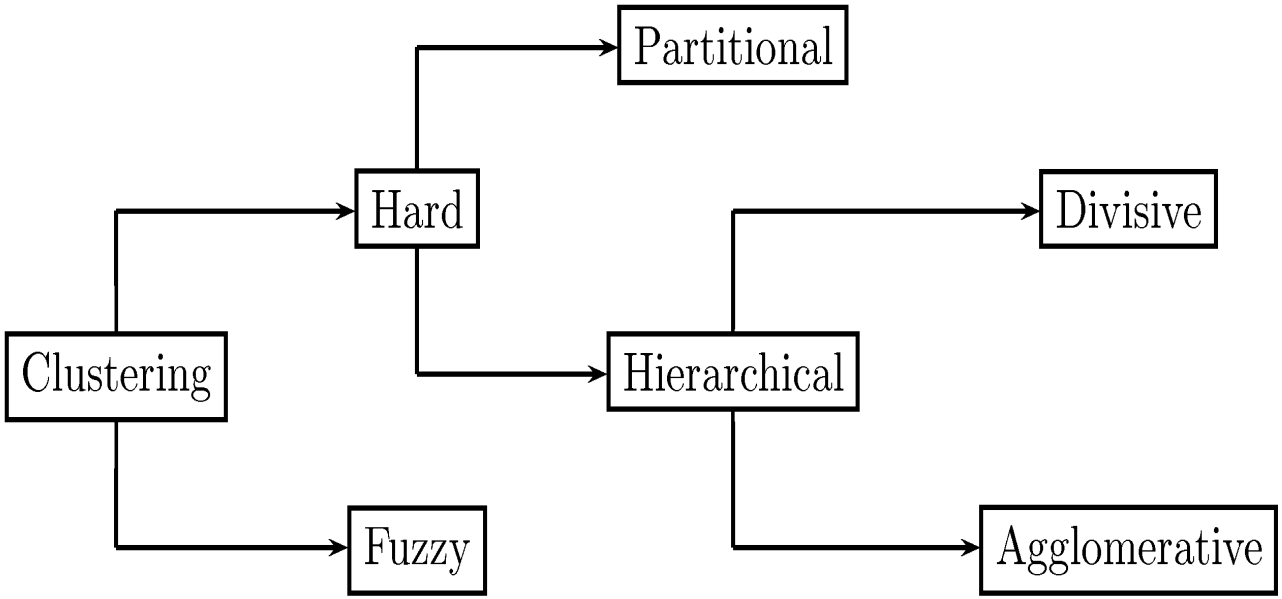
	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	n

Medida	Similaridade	Intervalo
Jaccard	a/(a+b+c)	[0,1]
Sokal-Sneath	2(a+d)/(2(a+d)+b+c)	[0,1]
Rogers-Tanimoto	(a+d)/(a+2(b+c)+d)	[0,1]
Russell-Rao	a/n	[0,1]
Sokal-Michener	(a+d)/n	[0,1]
Dice	a/(2a+b+c)	[0,1/2]
Kulzinsky	a/(b+c)	$[0,\infty]$

Exemplos: `ExemplosNA.R: (3) Matriz de Distancias`

Formação de Grupos

- Algoritmos



Fonte: Gan et al., 2007.

Algoritmos

- **"Hard Clustering"**: partição de X em m grupos $G_1, G_2, \ldots, G_m$, de maneira que
  - $G_i \neq \phi, i = 1, \ldots, m$
  - $\cup_{i=1}^{m} G_i = \mathrm{X}$
  - $G_i \cap G_j = \phi, i \neq j, i, j = 1, \ldots, m.$

- **"Fuzzy Clustering"**: X é dividido em m grupos através de m funções u_j tais que
  - $u_j : \mathrm{X} \rightarrow [0, 1], j = 1, \ldots, m.$
  - $\sum_{j=1}^{m} u_j(x_i) = 1, i = 1, \ldots, N.$
  - $0 < \sum_{i=1}^{N} u_j(x_i) < N, j = 1, \ldots, m.$

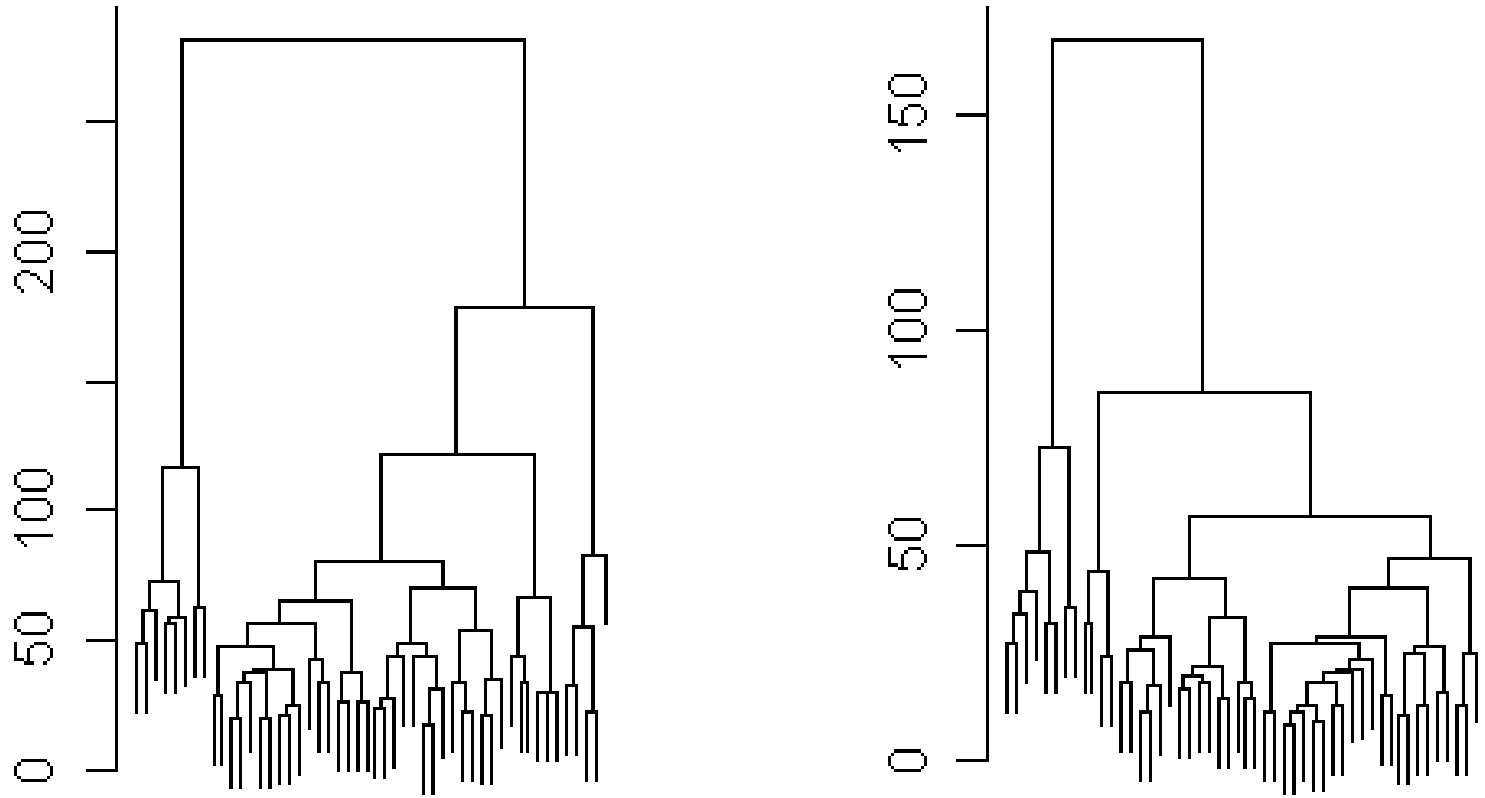
Algoritmos Hierárquicos

- **Aglomerativo:** inicia com cada observação em um grupo. A cada passo, as observações vão sendo agrupadas ao seu par mais similar de acordo com algum critério de similaridade. Termina com todas as observações num único grupo (*cluster*).

  AGNES: *Agglomerative Nesting*

- **Divisivo:** inicia com todas as observações num único *cluster*. A cada passo, as observações vão se separando de acordo com algum critério de similaridade (ou dissimilaridade). Termina com as observações separadas em diferentes grupos (*clusters*).

  DIANA: *Divisive Analysis*



Exemplos de Dendogramas Aglomerativo e Divisivo
para um mesmo conjunto de dados.

Algoritmos Hierárquicos Aglomerativos

- **Ligação Simples**: une os elementos mais próximos.

$$\mathrm{D} = \begin{bmatrix} 0_{11} & & & & \\ 9_{21} & 0_{22} & & & \\ 3_{31} & 7_{32} & 0_{33} & & \\ 6_{41} & 5_{42} & 9_{43} & 0_{44} & \\ 11_{51} & 10_{52} & 2_{53} & 8_{54} & 0_{55} \end{bmatrix}$$

(9_{21} indica que a distância do elemento 2 ao elemento 1 é igual a 9)

- $\min(\mathrm{d}_{ij}) = 2 = \mathrm{d}_{53} = \mathrm{d}_{35}$
- A união é feita pelas distâncias mínimas (*nearest neighbor*)
  - $\mathrm{d}_{(35)1} = \min\{\mathrm{d}_{31}, \mathrm{d}_{51}\} = \min\{3, 11\} = 3$
  - $\mathrm{d}_{(35)2} = \min\{\mathrm{d}_{32}, \mathrm{d}_{52}\} = \min\{7, 10\} = 7$
  - $\mathrm{d}_{(35)4} = \min\{\mathrm{d}_{34}, \mathrm{d}_{54}\} = \min\{9, 8\} = 8$

$$\mathrm{D} = \begin{bmatrix} 0_{(35)(35)} & & & \\ 3_{1(35)} & 0_{11} & & \\ 7_{2(35)} & 9_{21} & 0_{22} & \\ 8_{4(35)} & 6_{41} & 5_{42} & 0_{44} \end{bmatrix}$$

- **Ligação Simples**

$$D = \begin{bmatrix} 0 & & & \\ 3_{1(35)} & 0 & & \\ 7_{2(35)} & 9_{21} & 0 & \\ 8_{4(35)} & 6_{41} & 5_{42} & 0 \end{bmatrix}$$

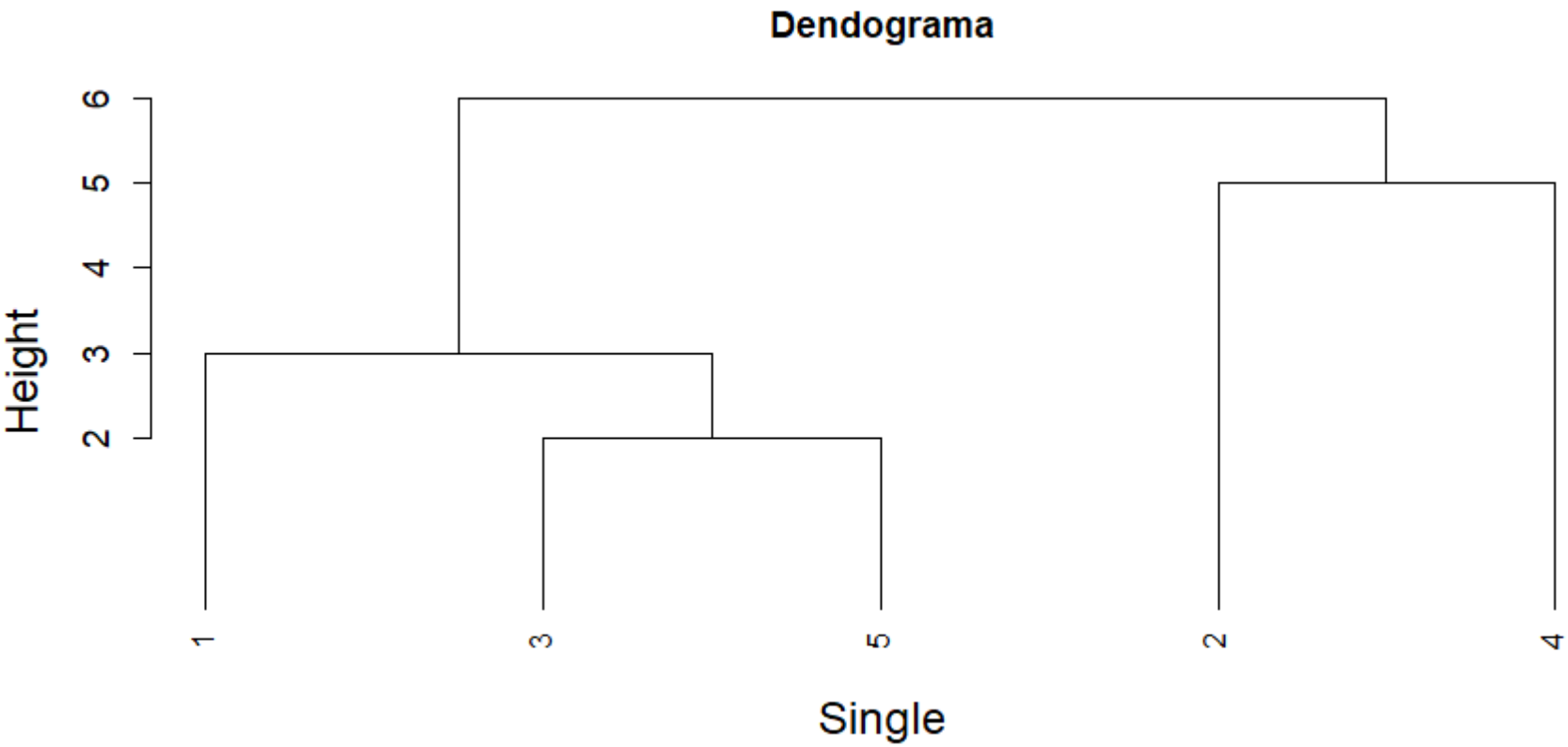
- $\min(d_{ij}) = 3 = d_{1(35)}$
  - $d_{(1;35)2} = \min\{d_{12}, d_{(35)2}\} = \min\{9, 7\} = 7$
  - $d_{(1;35)4} = \min\{d_{14}, d_{(35)4}\} = \min\{6, 8\} = 6$

$$D = \begin{bmatrix} 0 & & \\ 7_{(1;35)2} & 0 & \\ 6_{(1;35)4} & 5_{42} & 0 \end{bmatrix}$$

- **Ligação Simples**

$$D = \begin{bmatrix} 0 & & \\ 7_{(1;35)2} & 0 & \\ 6_{(1;35)4} & 5_{42} & 0 \end{bmatrix}$$

- $\min(d_{ij}) = 5 = d_{24}$
  - $d_{(1;35)(24)} = \min\{d_{(1;35)2}, d_{(1;35)4}\} = \min\{7, 6\} = 6$



Algoritmos Hierárquicos Aglomerativos

- **Ligação Completa**: une os elementos mais distantes.

$$\mathrm{D} = \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

- $\min(\mathrm{d}_{ij}) = 2 = \mathrm{d}_{53} = \mathrm{d}_{35} \leftarrow$ seleção pelo mínimo!
  - $\mathrm{d}_{(35)1} = \max\{\mathrm{d}_{31}, \mathrm{d}_{51}\} = \max\{3, 11\} = 11$
  - $\mathrm{d}_{(35)2} = \max\{\mathrm{d}_{32}, \mathrm{d}_{52}\} = \max\{7, 10\} = 10$
  - $\mathrm{d}_{(35)4} = \max\{\mathrm{d}_{34}, \mathrm{d}_{54}\} = \max\{9, 8\} = 9 \leftarrow$ uniões pelo máximo!

$$\mathrm{D} = \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & 5 & 0 \end{bmatrix}$$

...continua...

Algoritmos Hierárquicos Aglomerativos

- **Ligação Média**: une os pontos médios.

$$\mathrm{D} = \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

- $\min(\mathrm{d}_{ij}) = 2 = \mathrm{d}_{53} = \mathrm{d}_{35} \leftarrow$ seleção pelo mínimo!
  - $\mathrm{d}_{(35)1} = \text{média}\{\mathrm{d}_{31}, \mathrm{d}_{51}\} = \text{média}\{3, 11\} = 7$
  - $\mathrm{d}_{(35)2} = \text{média}\{\mathrm{d}_{32}, \mathrm{d}_{52}\} = \text{média}\{7, 10\} = 8.5$
  - $\mathrm{d}_{(35)4} = \text{média}\{\mathrm{d}_{34}, \mathrm{d}_{54}\} = \text{média}\{9, 8\} = 8.5 \leftarrow$ uniões pelas médias!

$$\mathrm{D} = \begin{bmatrix} 0 & & & \\ 7 & 0 & & \\ 8.5 & 9 & 0 & \\ 8.5 & 6 & 5 & 0 \end{bmatrix}$$

...continua...

- **Ligação Simples, Completa e Média no R**

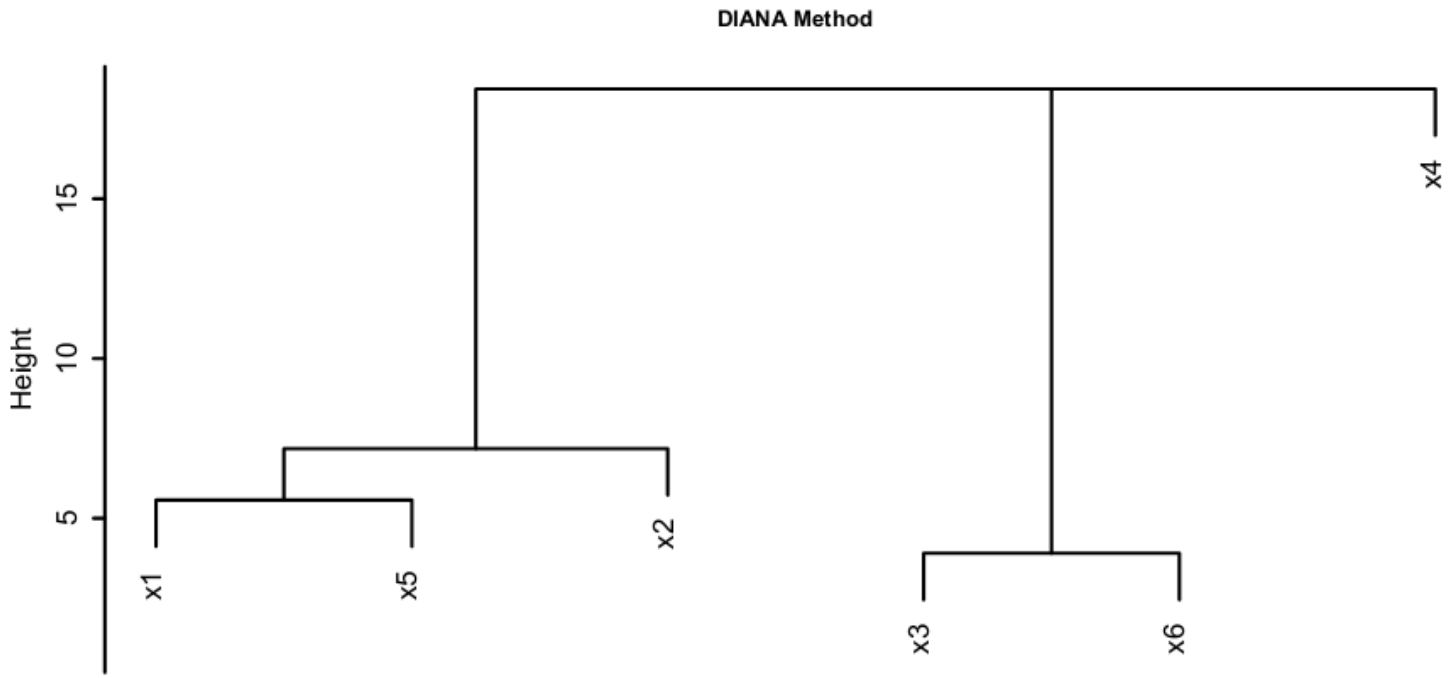
```
require(graphics)

m <- 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)
m <- matrix(m,5,5)

d <- as.dist(m)

hcs <- hclust(d, "single")
plot(hcs)
plot(hcs, hang = -1, cex.axis=1.2, cex.lab = 1.5,
     xlab="Single", main="Dendograma")

hcc <- hclust(d, "complete")
plot(hcc)
plot(hcc, hang = -1, cex.axis=1.2, cex.lab = 1.5,
     xlab="Complete", main="Dendograma")

hcc <- hclust(d, "average")
plot(hcc)
plot(hcc, hang = -1, cex.axis=1.2, cex.lab = 1.5,
     xlab="Average", main="Dendograma")
```

Algoritmos Hierárquicos Aglomerativos

- **Algoritmo da Média** ponderada de grupos;
- **Algoritmo de Ward** seleciona novos grupos com base na menor soma de quadrados do erro (SSE) de todas as possíveis combinações de pares;
- **ENERGY** ϵ-clustering (Székly e Rizzo - 2005).
  - Une objetos baseado na definição de energia;
  - Equivale a métodos tradicionais para certos paramêtros iniciais.
- **AGNES** (Agglomerative Nesting);
- e muitos outros...
- Referência: Kaufman, L. e Rousseeuw, P.J. (2005) Findig Groups in Data: An Introduction to Cluster Analysis. Wiley.

Algoritmos Hierárquicos Divisivos

- **DIANA** (DIvisive ANAlysis)



- Problema com algoritmos hierárquicos:

  Para classificar 1000 objetos temos

  499500 possíveis fusões e

  $5.357.543 \times 10^{300}$ possíveis divisões.

Algoritmos Não-Hierárquicos (Partitional)

- Procuram encontrar, diretamente, uma partição de n elementos em C grupos.
- Algoritmos buscam k observações que representam a estrutura de dados em cada grupo.
- Possuem funções objetivo próprias para avaliar a qualidade das partições.
- Os grupos são representados por centróides (possuem forma convexa).
- Eficientes em bancos com grande volume de informação.

Algoritmos Não-Hierárquicos

- K-means (MacQueen - 1967)
  - Mais popular dos algoritmos;
  - Número partições pré-estabelecido;
  - Calcula um centro para cada partição (μ_c, $c = 1, ..., k$);
  - Minimiza a soma de quadrados de cada observação em relação a μ_c;
  - Atualiza o grupos até não existirem objetos a designar.

- PAM - Partitioning around medoids (Kaufman e Rousseeuw - 1990);
  - Similar a K-means utilizando MEDOIDS em vez de médias.
  - MEDOIDS: objeto do grupo para o qual a dissimilaridade média para os demais objetos do grupo é mínima.

- CLARA - Clustering large applications;
  - Reduz o número de comparações no algoritmo PAM.

- SOM - Self organizing maps (Kohonen - 1989)
  - Idéia de redes neurais;
  - Permite uma visualização dos agrupamentos em 2 ou 3 dimensões.

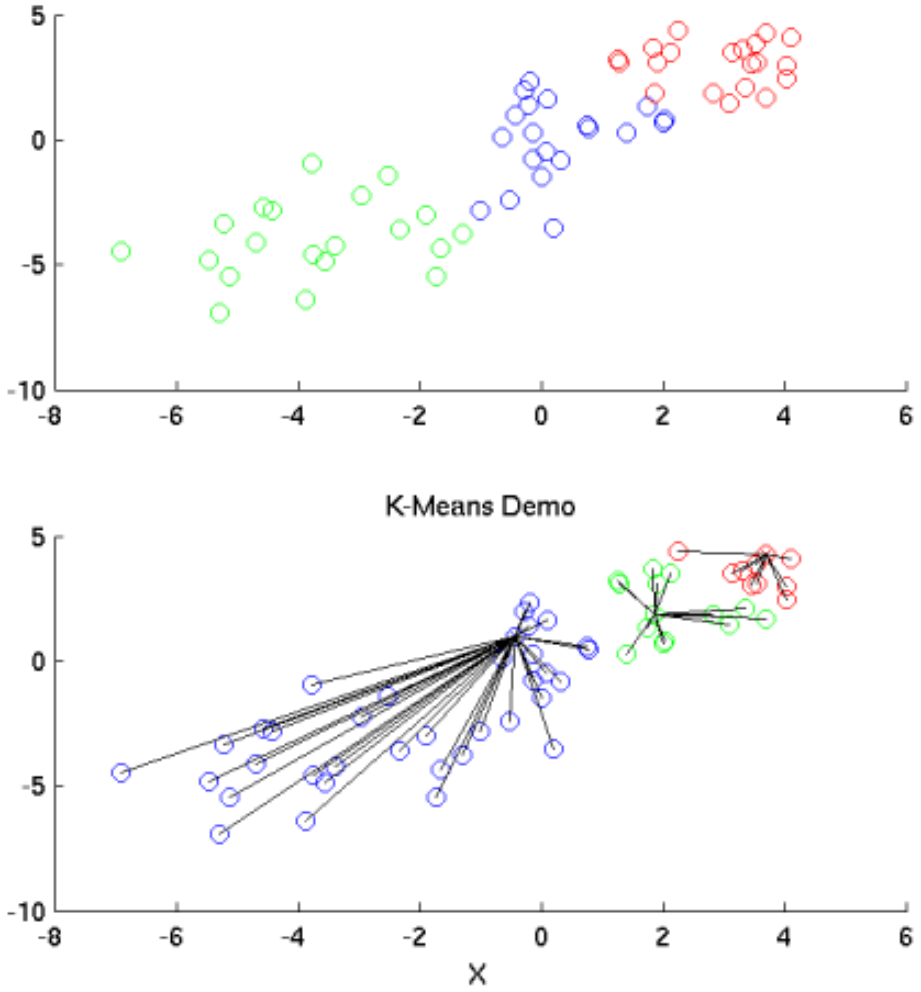
Algoritmo K-means

1. Considere os dados e um número k de grupos (cluster).
2. Faça um dos seguintes passos:
   - Alocar os elementos aleatoriamente aos k grupos e calcular $\bar{\boldsymbol{x}}_k$.
   - Especificar os centróides de cada grupo, $\bar{\boldsymbol{x}}_k$.
3. Calcule a distância Euclideana de cada item ao respectivo centróide:

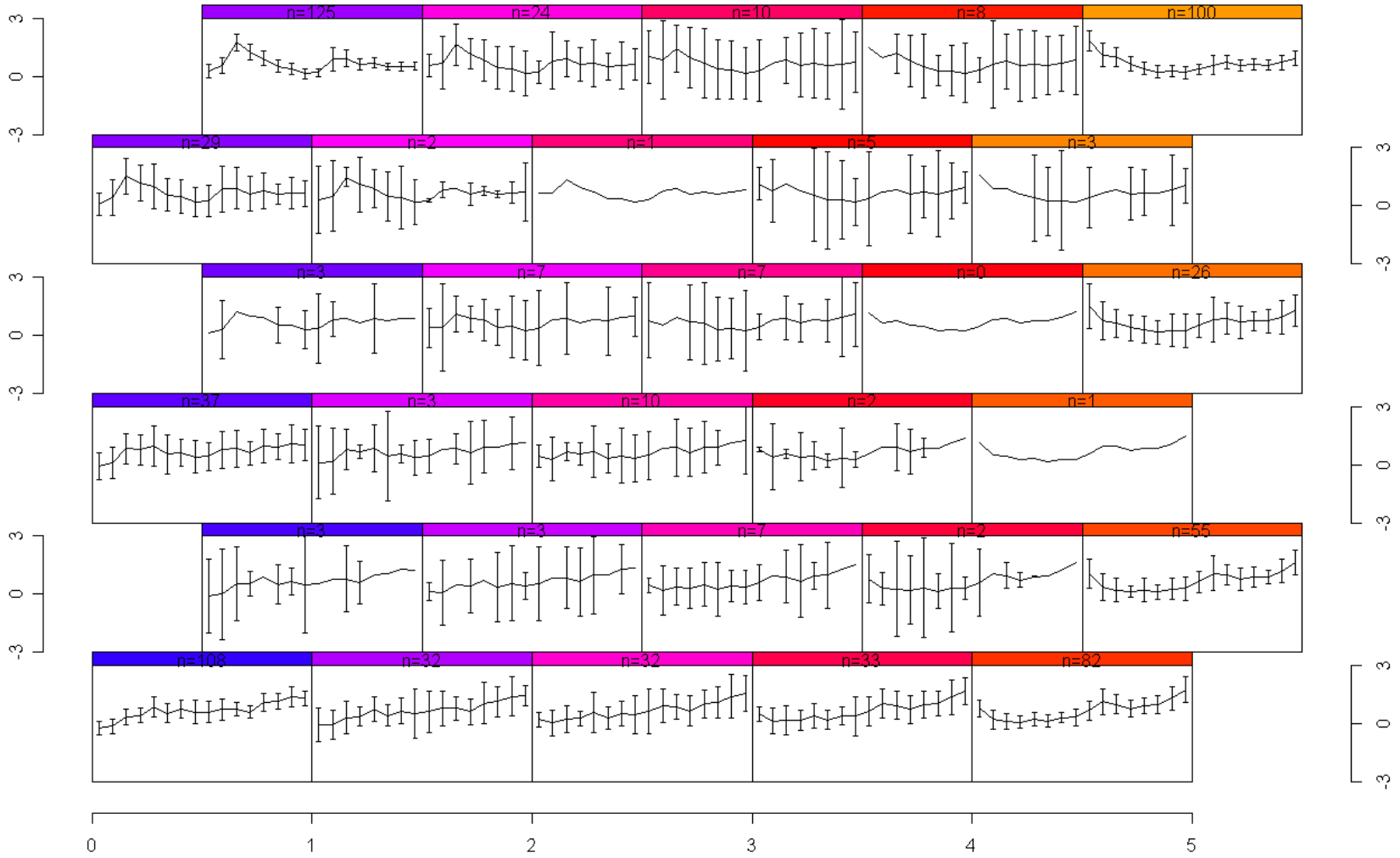
$$\text{ESD} = \sum_{k=1}^{k} \sum_{c(i)=k} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)'(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k),$$

   tal que $\bar{\boldsymbol{x}}_k$ é o centróide do k-ésimo grupo e $c(i)$ é o cluster contendo $\boldsymbol{x}_i$.
4. Realocar cada item ao grupo com centróide mais próximo de maneira a reduzir ESD. Atualizar os centróides dos grupos após cada realocação.
5. Repetir os passos (3) e (4) até que não exista itens para realocar.

Exemplo K-means



Exemplo SOM

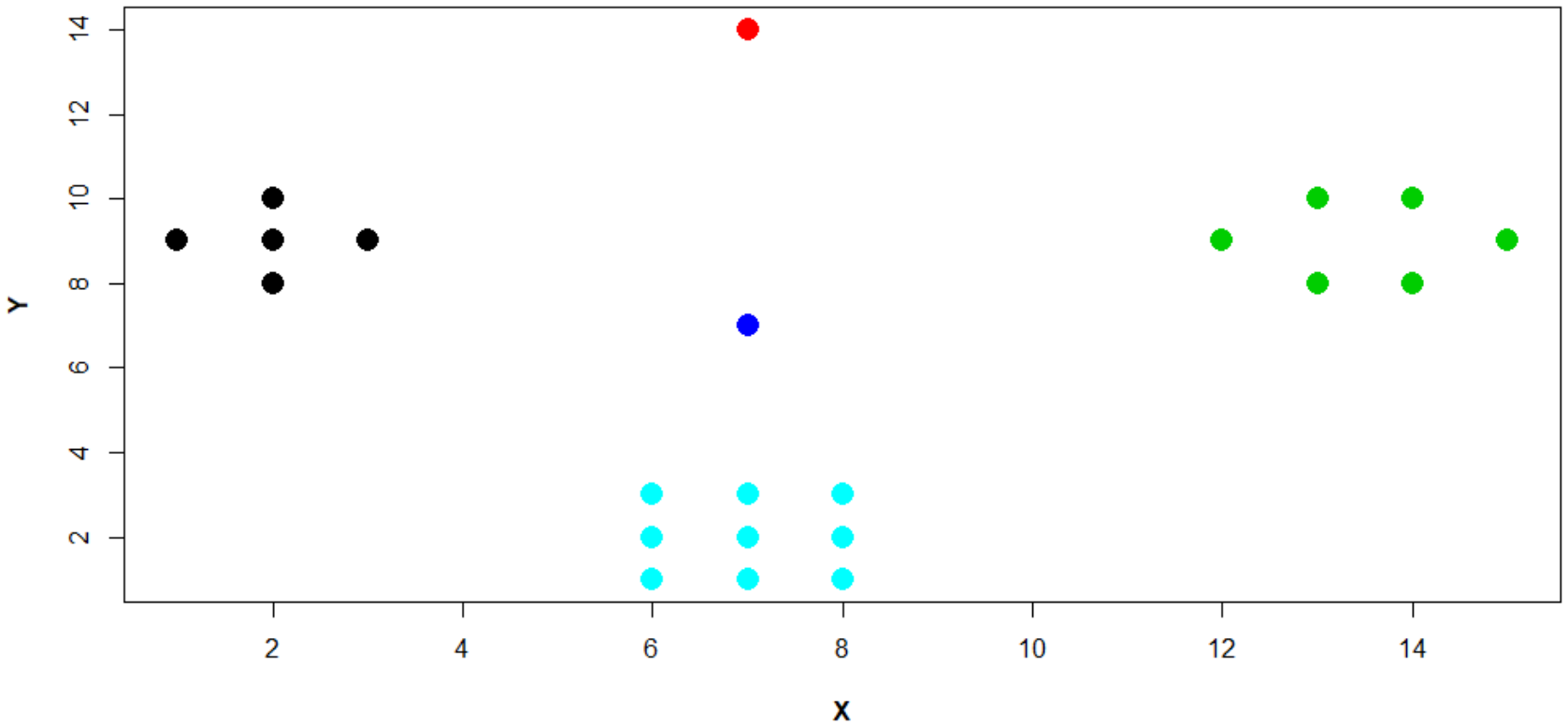


Exemplo de grid SOM bidimensional produzido com o pacote R.

Algoritmos Fuzzy

Por que utilizar algoritmos Fuzzy?

FANNY: Fuzzy Analysis



Mistura de Distribuições

- Algoritmos Baseados em Mistura de Distribuições:

  - Algoritmos hierárquicos aglomerativos que permitem alocação probabilística (algoritmo fuzzy);

  - Muito explorados em estudos envolvendo muitas variáveis e poucas observações.

  - Modelo de Mistura Finita[3]

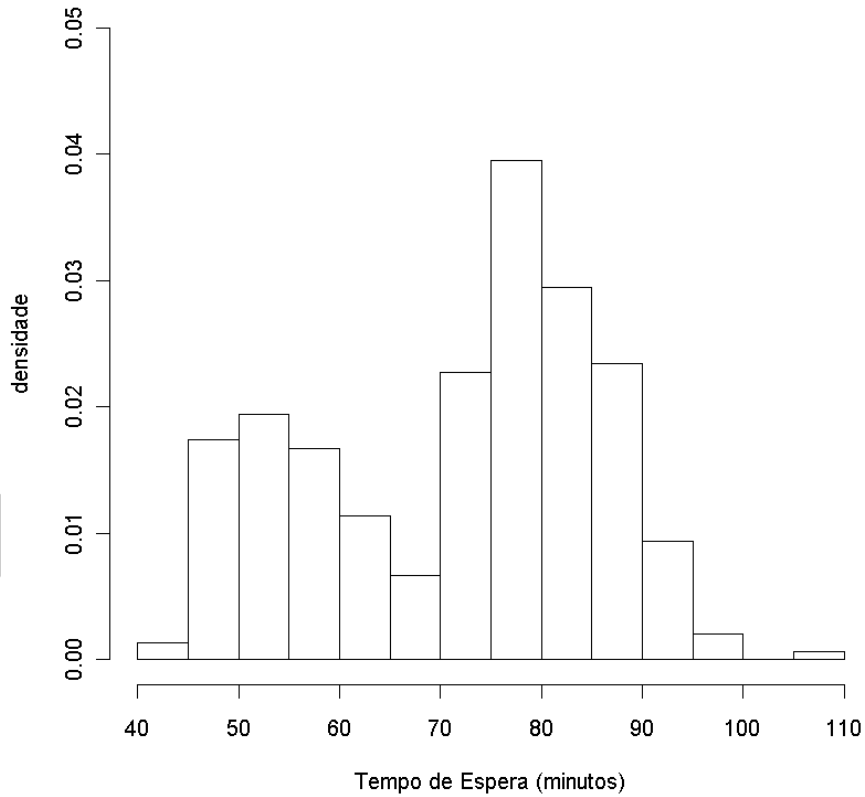
$$f(y_i) = p(y_i) = \sum_{g=1}^{G} \pi_g f_g(y_i|\theta_g) \tag{10}$$

onde, pi_g é a probabilidade de uma observação ter sido gerada pela g-ésima componente, e $f_g(y_i|\theta_g)$ é a densidade da g-ésima componente com parâmetros θ_g.

[3] Bouveyron et al. (2019) Model-based Clustering and Classification for Data Science (with Applications in R). Cambridge.

Modelo baseado em Misturas

Qual modelo utilizar para estes dados?



Modelos de Misturas de Distribuições (Mixture Models)

O que é um modelo baseado em misturas?

- Modelos utilizados quando uma única distribuição não é capaz de representar a densidade em estudo.
- $p_\theta(y) = \pi_1 f_1(y_1|\theta_1) + \pi_2 f_2(y_2|\theta_2) + ... + \pi_G f_G(y_G|\theta_G)$
- $\pi_g, g = 1, \ldots, G$ é a probabilidade de um componente da desconhecido da mistura;
- $\sum_i \pi_g = 1$;
- $f_g(y_i|\theta_g)$ é um modelo probabilístico desconhecido e
- θ_g são parâmetros desconhecidos do g-ésimo componente da mistura.
- θ é o conjunto de $(\pi_1, \ldots, \pi_G, \theta_1, \ldots, \theta_G)$.

- Quando os dados são multivariados, f_g é geralmente normal com densidade ϕ_g parametrizado pela vetor de médias μ_g e pela matriz de variância-covariância Σ_g, i.e,

$$\phi_g(y_i|\mu_g, \Sigma_g) = |2\pi\Sigma_g|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_i - \mu_g)^\top \Sigma_g^{-1}(y_i - \mu_g)\right\}.$$

- A estimação é feita pela Máxima Verossilhança, utilizando o Algoritmo EM (Expectation-Maximization).

Estimação por Máxima Verossimilhança (MLE)

- Sejam n observações multivariadas (y_i, z_i) em que y_i é observado e z_i não é observado. Se (y_i, z_i) são iid's, a verossimilhança completa dos dados é

$$\mathcal{L}_c(y, z|\theta) = \prod_{i=1}^{n} f(y_i, z_i|\theta),$$

em que $y = (y1, \ldots, y_n)$ e $z = (z_1, \ldots, z_n)$.

- A verossimilhança observada pode ser obtida integrando z da verossimilhança completa,

$$\mathcal{L}_o(y|\theta) = \int \mathcal{L}_c(y, z|\theta) dz.$$

- A verossimilhança observada ou verossimilhança da mistura pode ser escrita como,

$$\mathcal{L}_o(y|\theta) = \prod_{i=1}^{n} \sum_{i=1}^{G} \pi_g f_g(y_i|\theta_g) = \prod_{i=1}^{n} \sum_{i=1}^{G} \pi_g \phi_g(y_i|\mu_g, \Sigma_g).$$

O Algoritmo EM

- Alterna entre os passos Expectation e Maximization.
- Expectation (E): a esperança condicional da função $\log \mathcal{L}_c = \ell_c$ dada a informação observada e as estimativas atuais dos parâmetros são calculados.
- Maximization (M): obtemos os parâmetros que maximizam ℓ_c do passo E.
- No MLE para modelo de misturas, a parte não observada dos dados, z, envolve quantidades que são introduzidas de forma a reformular o problema para EM.
- No EM para modelos de misturas, os dados completos são considerados (y_i, z_i) em que $z_i = (z_{i,1}, \ldots, z_{i,G})$ é a porção não observada com

$$z_{i,g} = 1 \quad \text{se} \quad y_i \in \text{grupo } g \quad e \quad z_{i,g} = 0 \quad \text{c.c..}$$

- Assumimos que z_i são iid, conforme uma distribuição multinomial de G categorias com probabilidades $\pi_1, \ldots, \pi_G$.
- Assumimos que

$$f(y_i|z_i) = \prod_{g=1}^{G} f_g(y_i|\theta_g)^{z_{i,g}}$$

e assim,

$$\ell_c(\theta_g, \pi_g, z_{i,g}|y) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{i,g} \log[\pi_g f_g(y_i|\theta_g)]$$

é a função de log-verossimilhança dos dados completos.

- No passo E, a interação s do algoritmo EM é dada por

$$\hat{z}_{i,g}^{(s)} = \frac{\hat{\pi}_g^{(s-1)} f_g(y_i|\hat{\theta}_g^{(s-1)})}{\sum_{h=1}^{G} \hat{\pi}_h^{(s-1)} f_h(y_i|\hat{\theta}_h^{(s-1)})}$$

- A quantidade $\hat{z}_{i,g}^{(s)} = \mathrm{E}[z_{i,g}|y_i, \theta_1, \ldots, \theta_G]$ para o modelo de mistura $p_\theta(y)$.
- Para mistura de normais multivariadas,

  - o passo E é dado por ϕ_g no lugar de f_g.

  - o passo M as estimativas dos parâmetros tem forma fechada e são

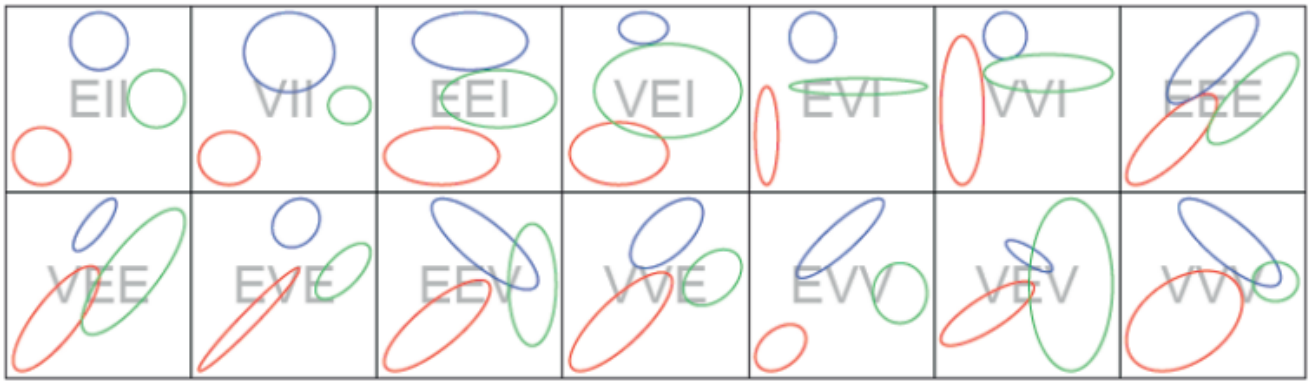
$$\hat{\pi}_g^{(s)} = \frac{\hat{n}_g^{(s-1)}}{n}; \quad \hat{\mu}_g^{(s)} = \frac{\sum_{i=1}^n \hat{z}_{i,g}^{(s-1)} y_i}{\hat{n}_g^{(s-1)}}; \quad \hat{n}_g^{(s-1)} = \sum_{i=1}^n \hat{z}_{i,g}^{(s-1)}.$$

- A estimativa $\hat{\Sigma}_g^{(s)}$ depende de sua parametrização. Vários modelos são propostos.

Identifier	Model	Distribution	Volume	Shape	Orientation
E		Univariate	Equal		
V		Univariate	Variable		
EII	λI	Spherical	Equal	Equal	NA
VII	$\lambda_g I$	Spherical	Variable	Equal	NA
EEI	λA	Diagonal	Equal	Equal	Axis-aligned
VEI	$\lambda_g A$	Diagonal	Variable	Equal	Axis-aligned
EVI	λA_g	Diagonal	Equal	Variable	Axis-aligned
VVI	$\lambda_g A_g$	Diagonal	Variable	Variable	Axis-aligned
EEE	Σ	Ellipsoidal	Equal	Equal	Equal
VEE	$\lambda_g D A D^T$	Ellipsoidal	Variable	Equal	Equal
EVE	$\lambda D A_g D^T$	Ellipsoidal	Equal	Variable	Equal
EEV	$\lambda D_g A D_g^T$	Ellipsoidal	Equal	Equal	Variable
VVE	$\lambda_g D A_g D^T$	Ellipsoidal	Variable	Variable	Equal
EVV	$\lambda D_g A_g D_g^T$	Ellipsoidal	Equal	Variable	Variable
VEV	$\lambda_g D_g A D_g^T$	Ellipsoidal	Variable	Equal	Variable
VVV	Σ_g	Ellipsoidal	Variable	Variable	Variable

Parametrizações de Σ_g no agrupamento por modelo baseado em misturas.

A é uma matriz diagonal.



Exemplos de contornos para os diferentes modelos.

Exemplo de Aplicação do Algoritmo EM

- Considere um modelo GMM (Gaussian Mixture Model) com a mistura de duas normais, i.e.,

$$p_\theta(y) = \pi_1 \mathrm{N}(\mu_1, \sigma_1^2) + \pi_2 \mathrm{N}(\mu_2, \sigma_2^2)$$

- Note que $\pi_2 = 1 - \pi_1$ e seja $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$.

- O **Log da função de Verossimilhança** é

$$\ell(\theta|y) = \sum_i \log \{\pi_1 \phi(y_i, \mu_1, \sigma_1^2) + 1 - \pi_1 \phi(y_i, \mu_2, \sigma_2^2)\}$$

e

$$\ell(\theta|y_i) = \log p_{z_i}(y_i|\theta_{z_i}) + \log \pi_{z_i}$$

$$= \sum_{j=1}^{k} \{I(z_i = j) \log p_j(y_i|\theta_j) + I(z_i = j) \log \pi_j\}$$

($\phi(y, \mu, \sigma^2)$ é a densidade da $\mathrm{N}(\mu, \sigma^2)$)

Exemplo de Aplicação do Algoritmo EM

- **Passo E**

$$\hat{p}_{ij} = E\{I(Z_i = j)|y_i, \theta^0\} = P(Z_i = j|y_i, \theta^0)$$

$$= \frac{\pi_j^0 p_j(y_i|\theta_j^0)}{p_{\theta^0}(y_i)} = \frac{\pi_j^0 p_j(y_i|\theta_j^0)}{\sum_j \pi_j^0 p_j(y_i|\theta_j^0)}$$

Que no caso normal será:

$$\hat{p}_{ij} = \frac{\pi_j \mathrm{N}(y_i|\mu_j^0, \sigma_j^0)}{\sum_j \pi_j \mathrm{N}(y_i|\mu_j^0, \sigma_j^0)}$$

- Significado: o "peso" que a j-ésima distribuição tem em explicar a observação, corresponde ao peso desta função ao observarmos y_i considerando $\theta = \theta^0$.

Exemplo de Aplicação do Algoritmo EM

- **Passo M**: atualização de cada θ_j é baseada na maximização (em separado) de

$$\sum_i \hat{p}_{ij} \log p_j(y_i|\theta_j)$$

e

$$\pi_j^{(1)} = \frac{\sum_i \hat{p}_{ij}}{n}$$

Que no caso normal tem forma explicita.

- j-ésimo parâmetro: $\theta_j = (\mu_j, \sigma_j)$.
- Verossimilhança ponderada:

$$-\frac{1}{2} \sum_i \hat{p}_{ij} \left\{ \log \sigma_j^2 + \frac{(y_i - \mu_j)^2}{\sigma_j^2} \right\}$$

- Atualizações ponderadas:

$$\mu_j^{(1)} = \frac{\sum_i \hat{p}_{ij} y_i}{\sum_i \hat{p}_{ij}} \quad \text{e} \quad \sigma_j^{2(1)} = \frac{\sum_i \hat{p}_{ij} (y_i - \mu_j^1)^2}{\sum_i \hat{p}_{ij}}$$

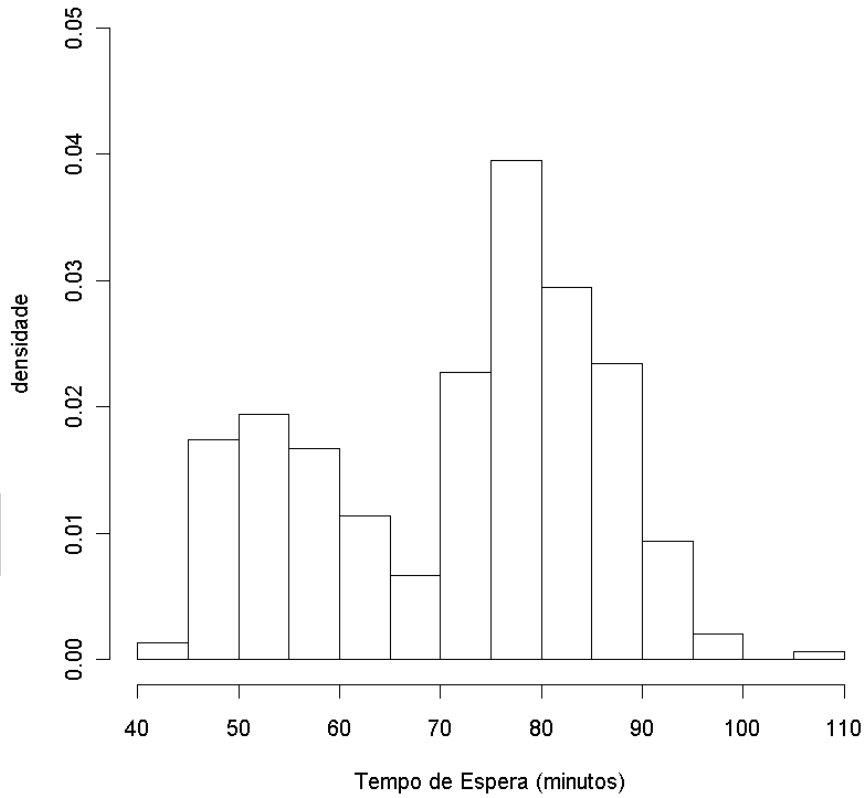
Exemplo de Aplicação do Algoritmo EM

- Programa R

```
y <- c(rnorm(300,54,5),rnorm(700,80,7.5))
n <- 1000

par(mfrow=c(1, 1))

hist(y, freq=F)

p <- 0.1; mu1 <- 50; sig1 <- 4 ; mu2 <- 70; sig2 <- 7

for (i in 1:400){
    deny <- p*dnorm(y,mu1,sig1) + (1-p)*dnorm(y,mu2,sig2)
    p1 <- p*dnorm(y,mu1,sig1)/deny
    p2 <- 1 - p1

# updates:

p <- sum(p1)/n
mu1 <- sum(p1*y)/sum(p1)
mu2 <- sum(p2*y)/sum(p2)
sig11 <- sum(p1 *(y-mu1)^2)/sum(p1); sig1<- sqrt(sig11)
sig22 <- sum(p2 *(y-mu2)^2)/sum(p2); sig2<- sqrt(sig22)
}

resp <- c(p,mu1,sig1 ,mu2,sig2)
resp

px <- c(rnorm(p*1000,mu1,sig1),rnorm((1-p)*1000,mu2,sig2))

lines(density(px))
```

Exemplo de Aplicação do Algoritmo EM

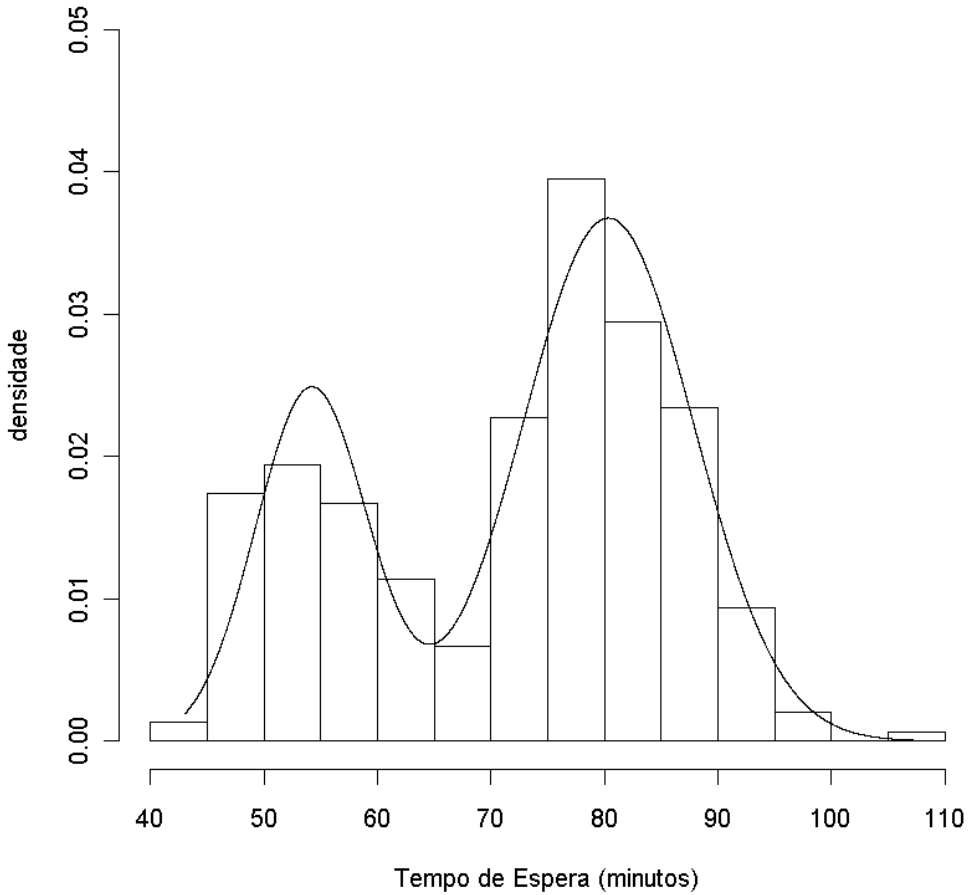
Convergência do Algoritmo EM (Visualização de 19 iterações)
Valores iniciais: $p = 0.1$, $\mu_1 = 50$, $\sigma_1 = 4$, $\mu_2 = 70$, $\sigma_2 = 7$.



Exemplo de Aplicação do Algoritmo EM

Convergência do Algoritmo EM

Valores finais: $p = 0.307$, $\mu_1 = 54.186$, $\sigma_1 = 4.939$, $\mu_2 = 80.347$, $\sigma_2 = 7.52$.



Dificuldades com os algoritmos tradicionais

- Precisam de especificação prévia do número de grupos;
- Sensibilidade a *outliers*;
- Falta de robustez;
- Apresentam resultados diferentes em transformações monótonas;
- Problemas de inversão (matriz de covariância);
- Suposições sobre distribuição dos dados em cada grupo;
- Desenvolvidos para análises com $p < n$.
- Tempo de processamento e alocação de memória.

Técnicas Alternativas de Agrupamento

- Gene-shaving (Hastie et al. - 2001);
- Agrupamentos baseados em suavização (Jiang et al. - 2003);
- CLIFF - filtragem iterativa (Xing e Karp - 2001);
- ...

Problemas

- Pouca flexibilidade computacional;
- Péssima documentação;
- Incompatibilidade com diferentes sistemas operacionais;
- Nem sempre funcionam como descritos...

Determinação do Número de Grupos

1. RMSSTD (Root Mean Squared Standard Deviation)

   Grupo formado, G_q, contém variâncias $s^2_{q1}, \ldots, s^2_{qp}$ para as p variáveis (agrupando observações).

$$\text{RMSSTD}_q = \sqrt{\frac{1}{p}\sum_{j=1}^{p} s^2_{qj}}$$

   Para grupos compactos, a medida deve ser pequena.

Determinação do Número de Grupos

2. R^2

$$\mathrm{R}^2 = \frac{\mathrm{SSB}}{\mathrm{SSTo}}$$

em que

$$\mathrm{SSB} = \sum_{i=1}^{g} n_i(\bar{X}_{i.} - \bar{X})'(\bar{X}_{i.} - \bar{X})$$

e

$$\mathrm{SSTo} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})'(X_{ij} - \bar{X})$$

O valor decresce sempre que novos grupos são incluídos. Devemos procurar um ponto de salto de R^2 para saber aonde parar.

Determinação do Número de Grupos

3. Pseudo F

$$F = \frac{\mathrm{SSB}/(g-1)}{\mathrm{SSR}/(n-g)} = \frac{(n-g)}{(g-1)} \times \frac{\mathrm{R}^2}{1-\mathrm{R}^2}$$

em que

$$\mathrm{SSR} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})'(X_{ij} - \bar{X}_{i.})$$

Busca-se o maior valor de Pseudo F .

Validação do Agrupamento

- **Índice Ajustado de Rand**: é um procedimento que permite comparar o agrupamento com algum critério externo, i.e., um agrupamento prévio ou resultado padrão.

  Esta é uma medida de concordância entre dois critérios. Agrupamentos com elementos nos mesmos grupos tem valor esperado $\mathrm{ARI} \approx 1$ e agrupamentos em que todos os elementos mudam de grupos tem valor esperado $\mathrm{ARI} \approx 0$.

  O procedimento está disponível no pacote `MCLUST` do `R`.

Referências:

- Hubert, L. (1985) Comparing Partitions. *Journal of Classification*, Springer-Verlag.

- Yeung, K.Y. e Ruzzo, W.L. (2001) *The Adjusted Rand Index*. Texto técnico.

Validação do Agrupamento

- **Outras Medidas de validação**: medidas internas, externas, de estabilidade e biológicas.

Referência:

- Brock, G.; Pihur, V.; Datta, S. e Datta, S. (2008) clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, Vol. 25, Issue 4.

Pacotes e funções no R

- `CCLUST`, `CLASS` (`som`), `CLUSTER` (`agnes`, `clara`, `diana`, `fanny`, `mona`, `pam`), `E1071` (`bclust`, `cmeans`), `FLEXMIX`, `FPC`, `HOPACH`, `MCLUST`, `STATS` (`hclust`, `heatmap`, `kmeans`), `PROXY`.
- https://cran.r-project.org/web/views/Cluster.html para visão mais completa de recursos no R.

Exemplos

1. `ExemplosCL.R`

`(1) XYdata`: agrupamento via k-means de dados gerados de duas normais.

`(2) Analise Grafica`

`(3) Matriz de Distancias`: medidas de similaridade e concordância.

`(4) Algoritmos`: Algoritmos hierárquicos e não hierárquicos.

`(5) k-Means`: analise k-Means do arquivo geyser.rda (ver descrição dos dados no arquivo Izenman-Geyser.pdf).

Algumas medidas para determinação de grupos e validação do agrupamento (em comparação com agrupamento hierárquico) são apresentadas.

`(6) MCLUST`: contornos de densidade para modelo de mistura normal bidimensional para dados geyser.rda.

Exemplos

2. von Borries, G.; Wang, H. (2009) Partition clustering of high dimensional low sample size data based on p-values. *Computational Statistics and Data Analysis*, vol. 53.
3. Lins, R. (2019) Implementação computacional de algoritmos para agrupamento de dados HDLSS e HDLLSS. Apresentação na 64a RBRAS, Cuibá-MT.

   https://github.com/rafaelslins/Rclust
4. dos Santos, L.B. (2021) Agrupamento por misturas finitas normais com aplicação a dados de expressão gênica. Trabalho de conclusão do Bacharelado em Estatística, UnB.