Análise Multivariada

Aula 5: Análise de Agrupamentos (*Clusters*)

Prof. Admir Antonio Betarelli Junior

Juiz de Fora

Estrutura

- Parte I. Introdução.
- Parte II. Medidas de dissimilariedades e similaridades.
- Parte III. Técnicas hierárquicas de agrupamento.
- Parte IV. Técnicas para a partição final.
- Parte V. Técnicas não hierárquicas de agrupamento.

Parte I. Introdução

Análise de Cluster

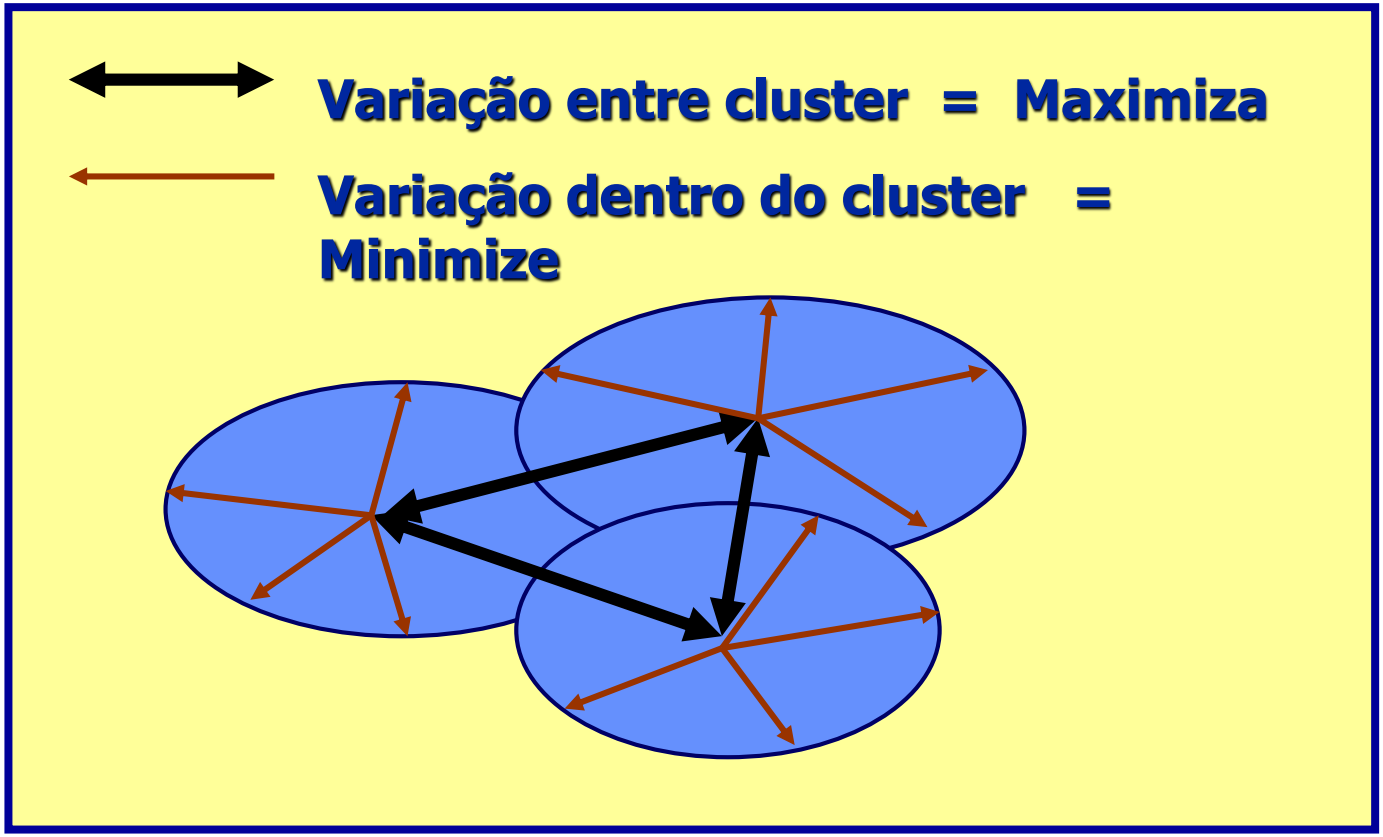
- Encontrar nos dados uma estrutura de agrupamento natural é uma importante técnica exploratória.

- Permite avaliar a dimensionalidade, identificar *outliers* e sugerir hipóteses acerca da estrutura de relações.

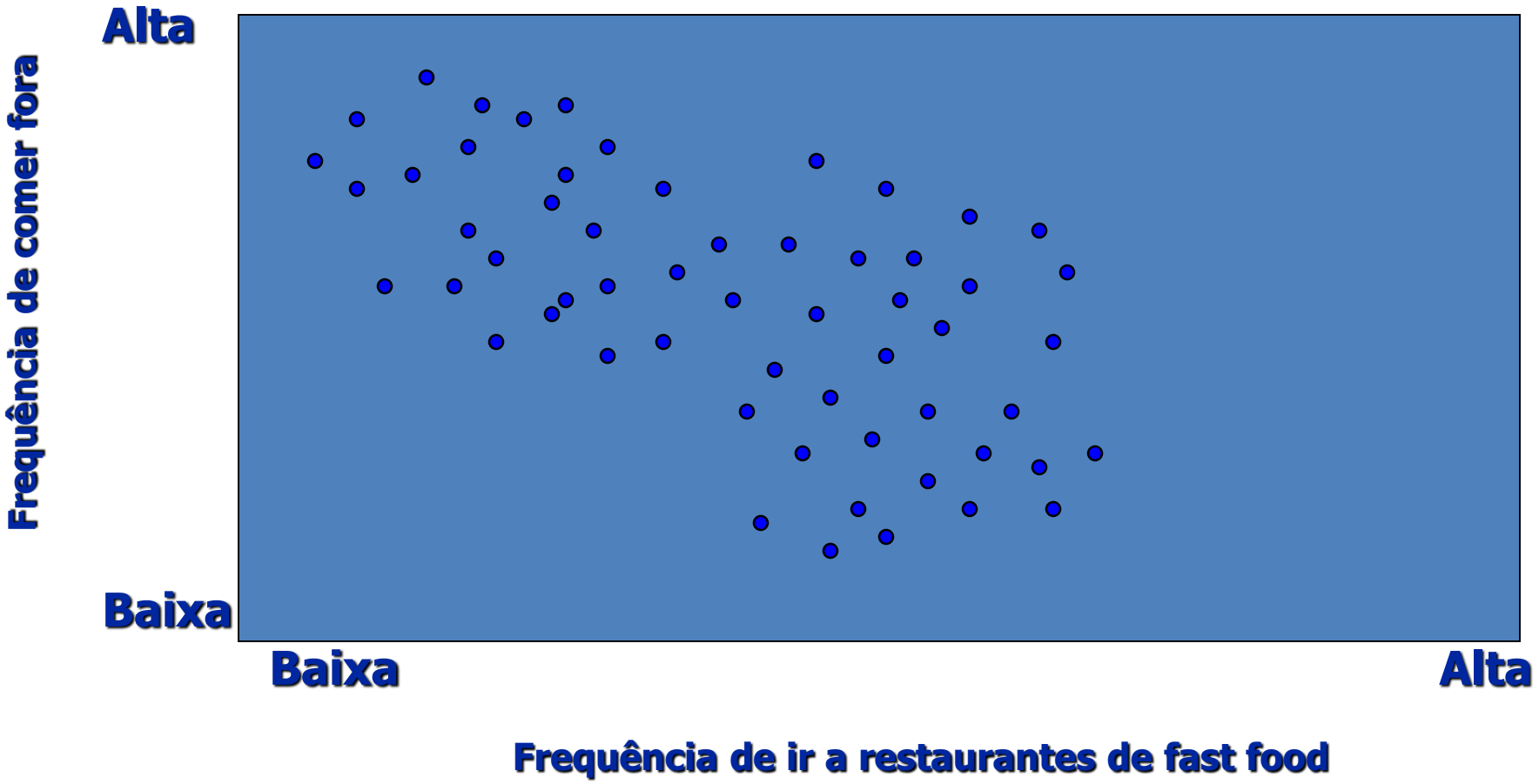
- Busca descobrir agrupamentos naturais de indivíduos (ou variáveis) a partir dos dados observados, agrupando indivíduos com base na similaridade ou distâncias (dissimilaridades).

Análise de Cluster

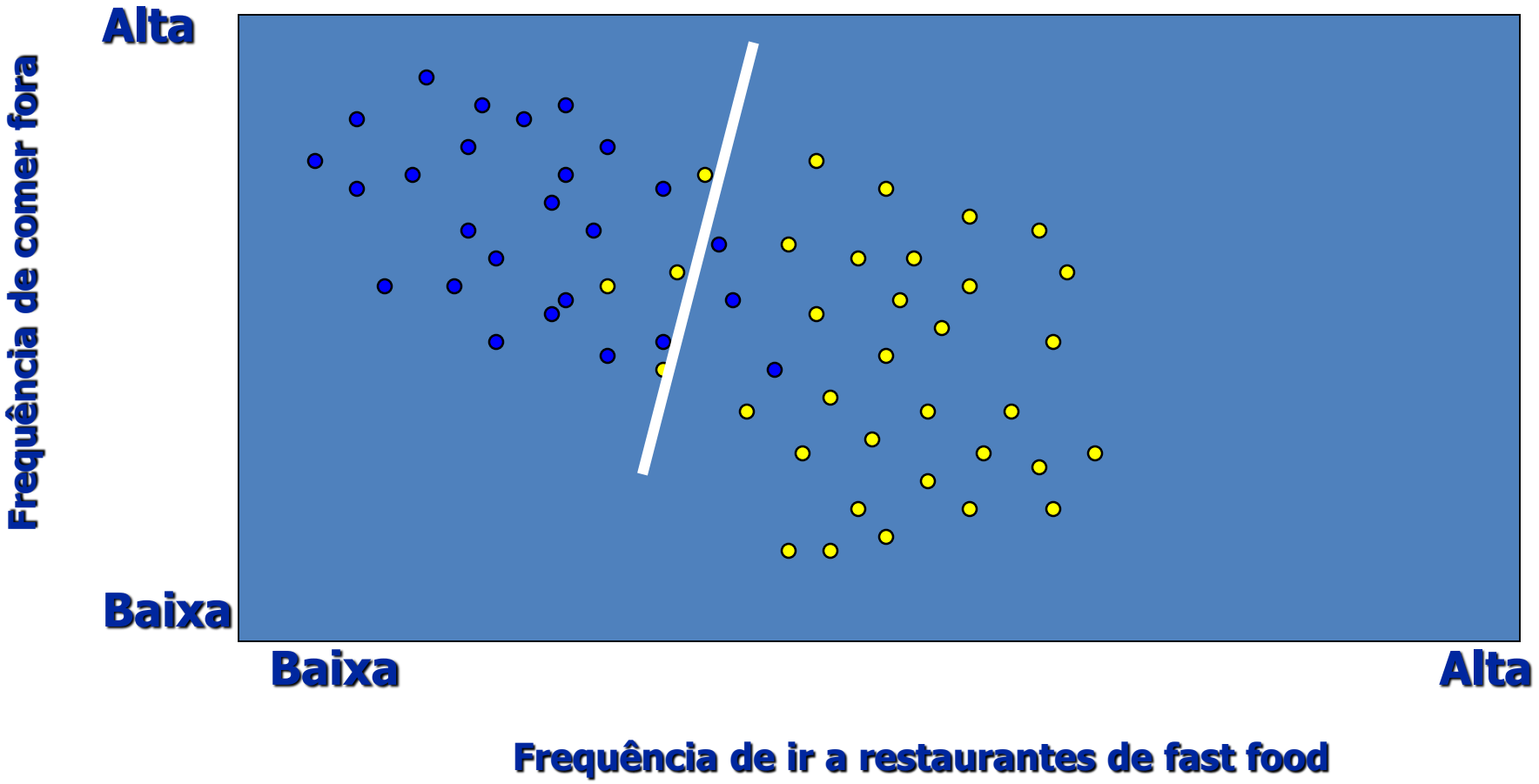
- Maximiza a homogeneidade de indivíduos dentro de grupos, e maximiza a heterogeneidade entre os grupos.



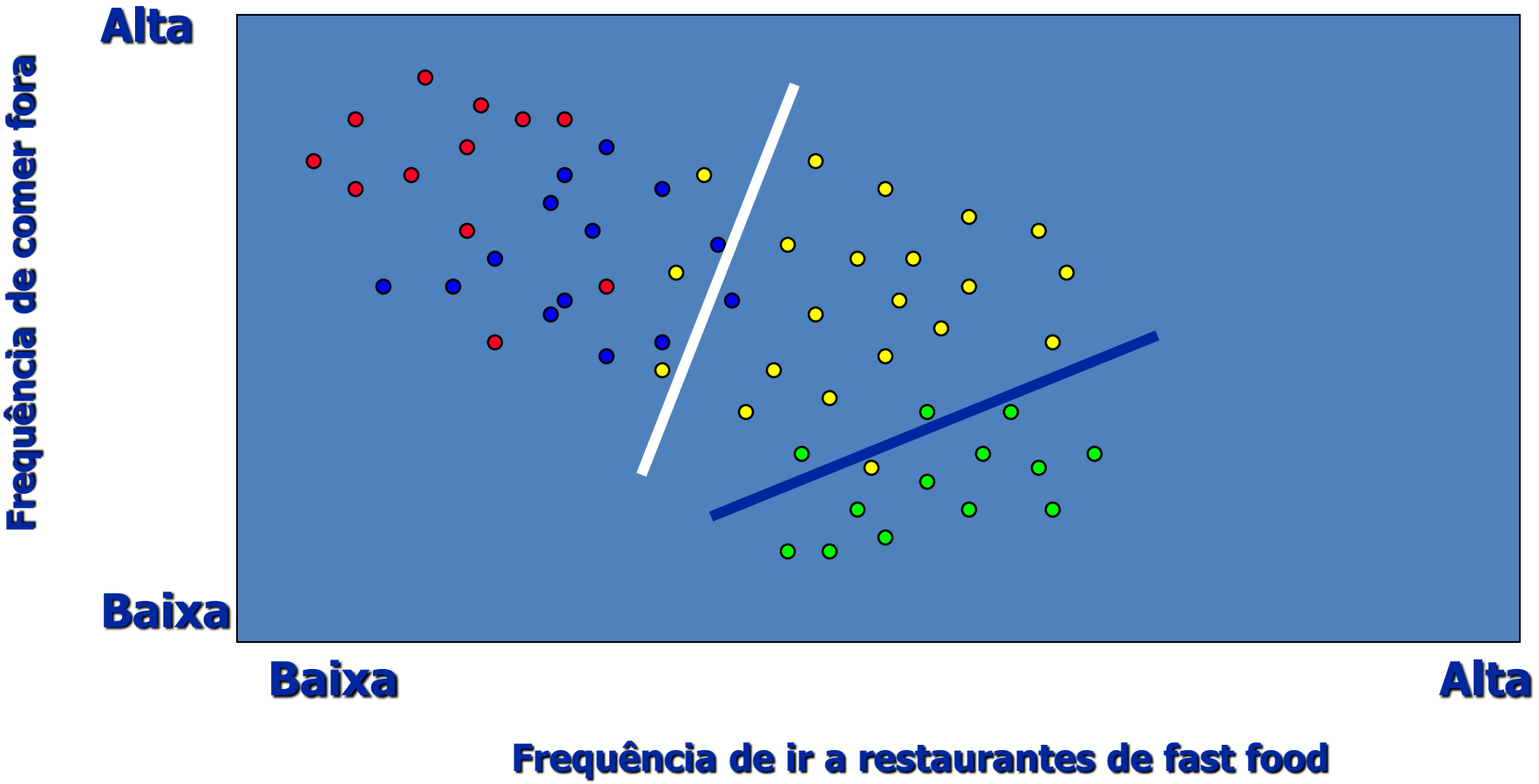
Análise de Cluster



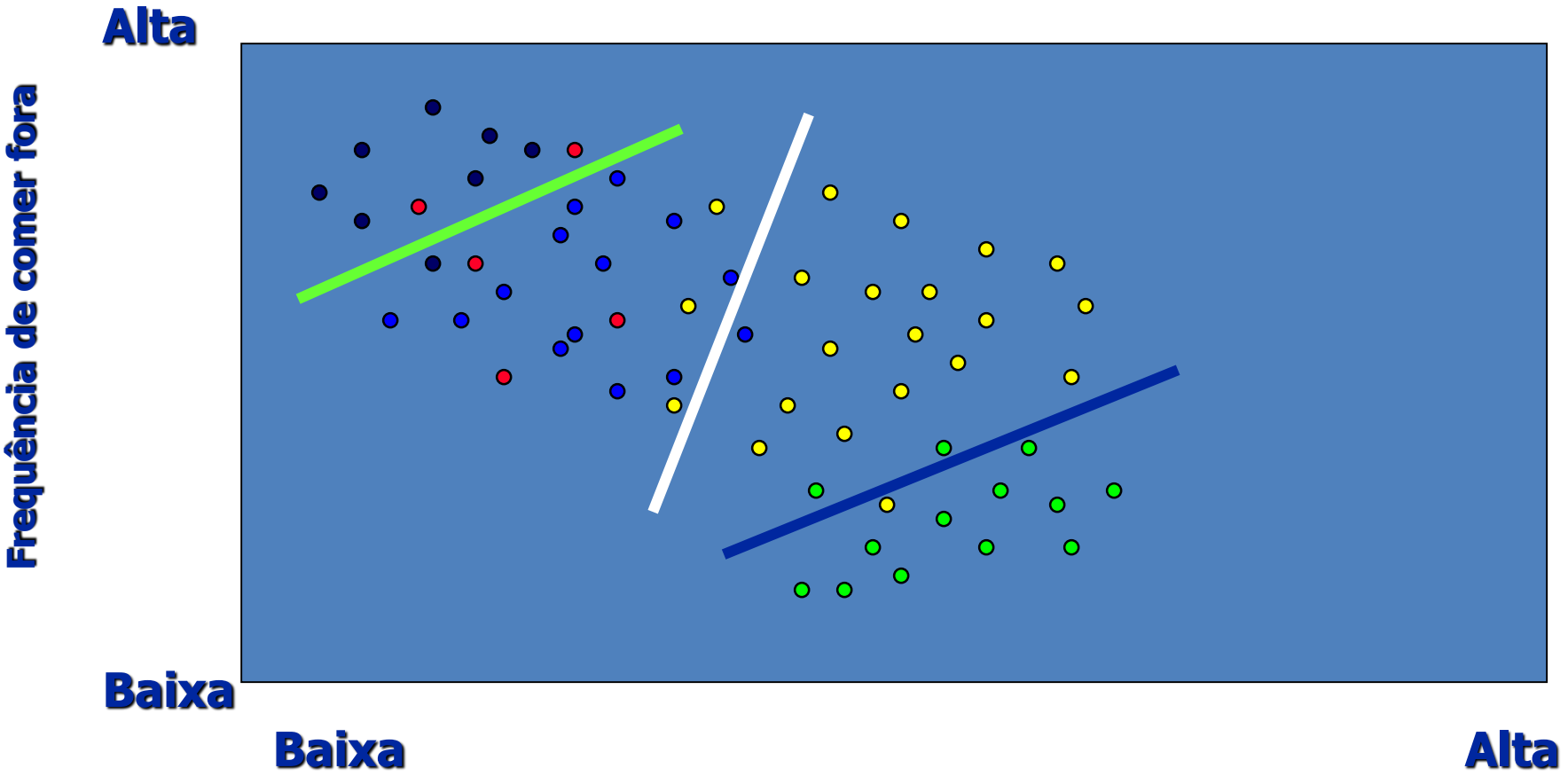
Análise de Cluster



Análise de Cluster



Análise de Cluster



Objetivos gerais

- Particionar os elementos em 2 ou mais clusters com base na similaridade deles a partir de um conj. de variáveis.

- Possui 3 aplicações comuns:

  - **classificação de elementos (taxonomia);**

  - **simplificação de dados;**

  - **identificação das relações entre os elementos.**

Críticas à Análise de Cluster

- A análise de agrupamento é descritiva, a-teórica, e não inferencial.
- . . . vai sempre criar *clusters*, independentemente da existência real de qualquer estrutura nos dados.
- A solução de *cluster* não é generalizável porque é totalmente dependente das variáveis utilizadas como a base para a medida de similaridade.

Quando usar?

- Quando a preocupação principal é dividir os elementos em grupos, de forma que os elementos de um mesmo grupo sejam homogêneos e os elementos em grupos diferentes sejam heterogêneos.

- Considerações teóricas, conceituais e práticas devem ser observadas ao selecionar variáveis para a AA.

- **Como medir similaridades entre indivíduos?**

- **Como agrupar indivíduos semelhantes?**

Parte II. Medidas de dissimilariedades e similaridades

Similaridade

- Similaridade entre os elementos é uma medida empírica de correspondência, ou semelhança, entre os elementos a serem agrupados.

- **Três técnicas** dominam as aplicações na AA:
  - **Tipos de distância** (proximidade):variáveis quantitativas.
  - **Medida de similaridade** => variáveis qualitativas
  - **Associação** => agrupamentos das variáveis.

Dissimilaridade – Var. quantitativas

- Seja o vetor aleatório, $\mathbf{X}'_j = [X_{j1}, X_{j2}, \ldots, X_{jp}]$, com p variáveis para cada elemento j dos n elementos.

- Utilizam-se medidas de distância (dissimilaridades): ↓ seu valor → ↑ similares são os elementos comparados.

a) Distância euclidiana:

$$d(X_l, X_k) = [(X_l - X_k)'(X_l - X_k)]^{1/2} = \left[\sum_{i=1}^{p}(X_{il} - X_{ik})^2\right]^{1/2} \qquad \because (j \neq l)$$

i.e., 2 elementos são comparados em cada variável i.

Dissimilaridade – Var. quantitativas

b) Distância generalizada ou ponderada:

$$d(X_l, X_k) = \left[(X_l - X_k)' A (X_l - X_k)\right]^{1/2} \qquad \because (j \neq l)$$

se

$A = I \Rightarrow d(\cdot)$ é uma euclidiana .

$A = S^{-1} \Rightarrow d(\cdot)$ é uma Mahalanobis.

$A = diag(1/p) \Rightarrow d(\cdot)$ é uma euclidiana média.

A reflete a ponderação. Se $A = diag(S_i^2)^{-1}$ => considera somente a ≠ de variabilidade entre as variáveis. Já quando $A = S^{-1}$ => pondera as possíveis ≠s de variâncias e covariâncias entre as variáveis.

Dissimilaridade – Var. quantitativas

c) Distância de Minkowsky:

$$d(X_l, X_k) = \left[\sum_{i=1}^{p} w_i \left| X_{il} - X_{ik} \right|^{\lambda} \right]^{1/\lambda} \qquad \because (j \neq l)$$

se

$\lambda = 1 \Rightarrow d(\cdot)$ é uma city - block ou Manhattan.

$\lambda = 2 \Rightarrow d(\cdot)$ é uma euclidiana .

$w_i's$ são os pesos de ponderação para as variáveis.

A métrica de Minkowsky é menos afetada pela presença de *outliers* do que a distância euclidiana.

Dissimilaridade – Var. quantitativas

- As distâncias entre os elementos são armazenadas em uma matriz de distâncias:

$$\underset{(nxn)}{D} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} \\ & 0 & d_{23} & d_{24} \\ & & 0 & d_{34} \\ & & & 0 \end{bmatrix}$$

em que d_{lk} representa a distância do elemento l ao elemento k.

Similaridade – Var. qualitativas

- Há 2 alternativas:
  - Transforma em quantitativas e usa-se as medidas de distâncias.
  - Trabalha-se com coeficientes de similaridades, comparando os elementos de acordo com a presença ou ausência de certas características.

Similaridade – Var. qualitativas

- Para entender o problema com variáveis qualitativas:

	Variáveis				
	1	**2**	**3**	**4**	**5**
Item l	1	0	0	1	1
Item k	1	1	0	1	0

- Há 2 pares (1,1), 1 par (0,0) e 2 pares incompatíveis (0,1;1,0).

$$\sum_{i=1}^{5}\left(X_{il}-X_{ik}\right)^2=(1-1)^2+(0-0)^2+(0-1)^2+(1-0)^2=2$$

- Deve-se comparar os itens diante da presença ou ausência de características. Os pares (1,1) e (0,0) são ignorados na distância.

Similaridade – Var. qualitativas

- O esquema organiza a frequência de similaridades e dissimilaridades para os elementos l e k.

		Elemento k		
		1	**0**	**Total**
Elemento l	**1**	a	b	**a+b**
	0	c	d	**c+d**
	Total	**a+c**	**b+d**	**p = a+b+c+d**

- a é a frequência do par (1,1), b a do par (1,0), e assim por diante.

Similaridade – Var. qualitativas

- Desenvolve-se os coef. de similaridades para os itens:

a) **<u>concordância simples</u>**:

$$s(l,k) = \frac{a+d}{p} \Rightarrow \text{exemplo anterior} : \frac{3}{5} = 0.6 \qquad \because \uparrow s(\cdot) \Rightarrow \uparrow \text{similaridade}$$

b) **<u>concordância positiva</u>:** (0,0) não necessariamente representa concordância (ideia do caso contrário).

$$s(l,k) = \frac{a}{p} \Rightarrow \text{exemplo anterior} : \frac{2}{5} = 0.4 \qquad \because \uparrow s(\cdot) \Rightarrow \uparrow \text{similaridade}$$

Similaridade – Var. qualitativas

c) **<u>concordância de Jaccard</u>:** proporção do par (1,1) em relação ao total [-(0,0)].

$$s(l,k) = \frac{a}{a+b+c} \Rightarrow \text{exemplo anterior} : \frac{2}{4} = 0.5 \qquad \because \uparrow s(\cdot) \Rightarrow \uparrow \text{similaridade}$$

d) **<u>distância euclidiana média</u>:** índice de dissimilaridade.

$$d(l,k) = \left(\frac{c+b}{p} \right)^{1/2} \Rightarrow \text{exemplo anterior} : \sqrt{\frac{2}{5}} = 0.63 \qquad \because \uparrow d(\cdot) \Rightarrow \downarrow \text{similaridade}$$

$$\text{em que } s(l,k) = 1 - d(\cdot)^2 \Rightarrow \text{similaridade simples}$$

Similaridade – Var. quantitativas

- Qualquer distância usada para var. quantitativas pode ser transformada em um coef. de similaridade:

$$s(l,k) = 1 - d^*(l,k)$$

$$d^*(l,k) = \frac{d(l,k) - \min(D)}{\max(D) - \min(D)}$$

em que :

$\min(D)$ é o menor e $\max(D)$ é o maior valor dos elementos fora da diagonal de D.

Variáveis quantitativas e qualitativas

- Uma situação comum é quando p var. quantitativas e q var. qualitativas são observadas nos n itens. Pode-se:

a) **Var. qualitativas => quantitativas** ao atribuir valores às categorias (*ad hoc*). Depois, usa-se uma medida de distância para comparar as $p+q$ var.;

b) **Var. quantitativas => qualitativas** categorizando os seus valores. Depois, usa-se uma medida de similaridade para comparar as $p+q$ var.

Variáveis quantitativas e qualitativas

c) **<u>Construir medidas de semelhança mistas</u>** e usá-las para a comparação dos elementos. Tem-se uma combinação linear entre as var. (p e q).

$$c(l,k) = \omega_p c_p(l,k) + \omega_q c_q(l,k)$$

em que $\omega_p = \dfrac{p}{p+q}$ e $\omega_q = \dfrac{q}{p+q}$; $c_p(\cdot)$ e $c_q(\cdot)$ são coef. de similaridade.

- A definição dos pesos de ponderação, ω , permite que os coef. tenham o intervalo de variação. Para manter $c_p(\cdot)$ e $c_q(\cdot)$ na mesma direção e o mesmo padrão, usa-se $s(l,k) = 1 - d^*(l,k)$ no caso das quantitativas.

Variáveis quantitativas e qualitativas

d) **Coeficiente de Gower (1971):** para cada var. j, considera-se um coef. , s_j, em um intervalo [0,1]. Comparando os elementos, l e k, as suas similaridades:

$$d(l,k) = \left(\frac{\sum_{j=1}^{p+q} 1_j(l,k) s_j(l,k)}{\sum_{j=1}^{p+q} 1_j(l,k)} \right)$$

$1_j(l,k)$ é uma variável igual a 1 se l e k podem ser comparados pela var. X_j.

- E.g., se existir 6 var., porém para l há valores de 4 var., então compara-se l e k para 4 var..
- Usa-se $s(l,k) = 1 - d^*(l,k)$ no caso das quantitativas.

Similaridades para pares de variáveis

- Ao invés dos elementos, as variáveis serão agrupadas.
- Usa-se a matriz de correlação (R). Pode-se obter a matriz de distância a partir de R (valores absolutos):

$$\underset{(pxp)}{D} = \underset{(pxp)}{1} - ABS(\underset{(pxp)}{R})$$

$$\uparrow r_{\substack{ik \\ i \neq k}} = \frac{s_{ik}}{\sqrt{s_{ii}s_{kk}}} \Rightarrow \downarrow d_{\substack{ik \\ i \neq k}} \quad \forall i,k = 1,2,\ldots,p$$

Similaridades para pares de variáveis

- **Para variáveis são binárias**, os dados são agrupados por tabela de contingência. As variáveis, ao invés dos itens, delineiam as categoriais.

		Variável k		
		1	**0**	**Total**
Variável l	**1**	a	b	**a+b**
	0	c	d	**c+d**
	Total	**a+c**	**b+d**	**n = a+b+c+d**

- A correlação é:

$$r(l,k) = \frac{ad - bc}{[(a+b)(c+d)(a+c)(b+d)]^{1/2}}$$

Parte III. Técnicas hierárquicas de agrupamento

Métodos para construção de Clusters

- ***Não hierárquicos:*** o n° g de grupos é pré-especificado.

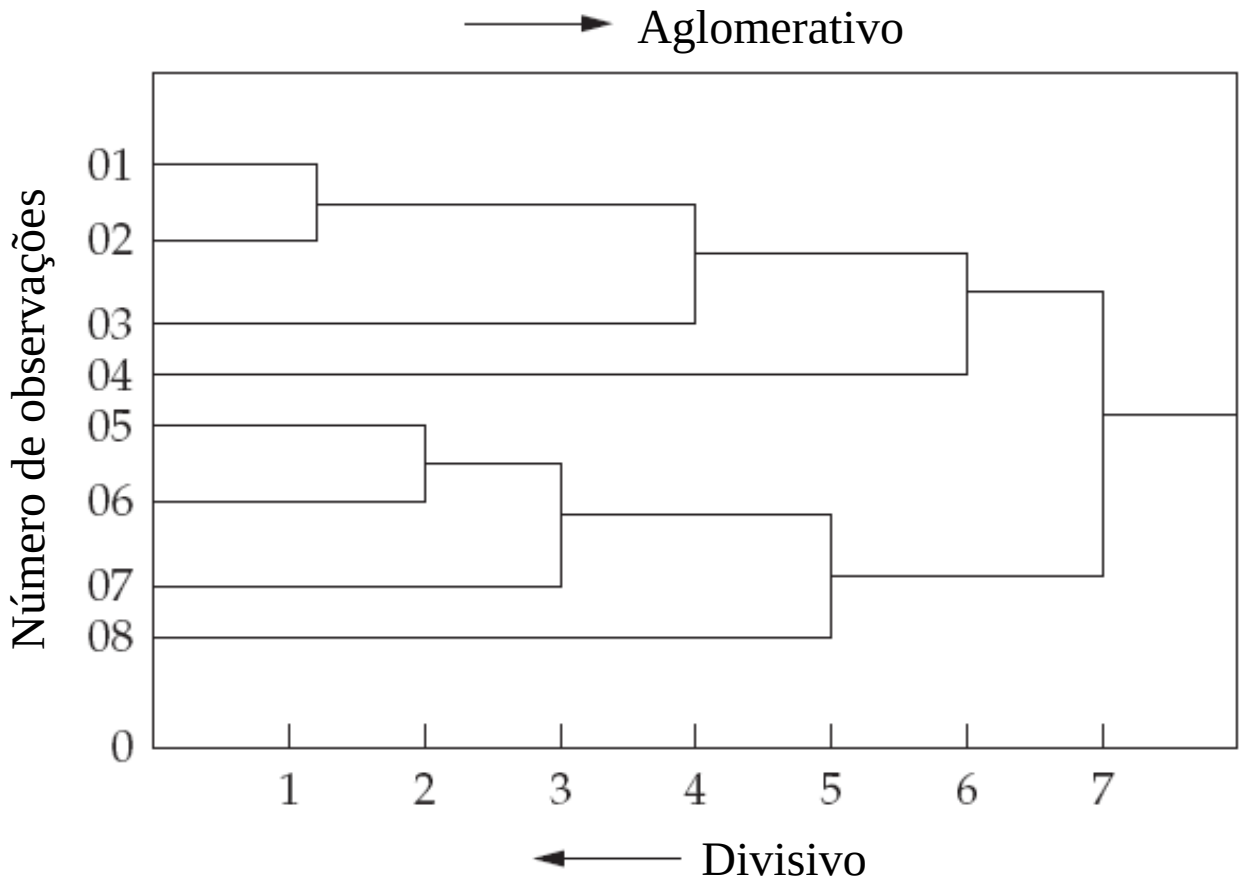
- ***Hierárquicos:*** identificam agrupamentos e o provável o n° g de grupos, por:

  a) Uma série de fusões sucessivas (técnicas *aglomerativas*);

  b) Ou uma série de sucessivas divisões (técnicas *divisas*).

  - Os resultados de ambos, aglomerativos e divisivos, são observados no *dendograma*, que ilustra as fusões ou divisões feitas em níveis sucessivos.

Métodos hierárquicos



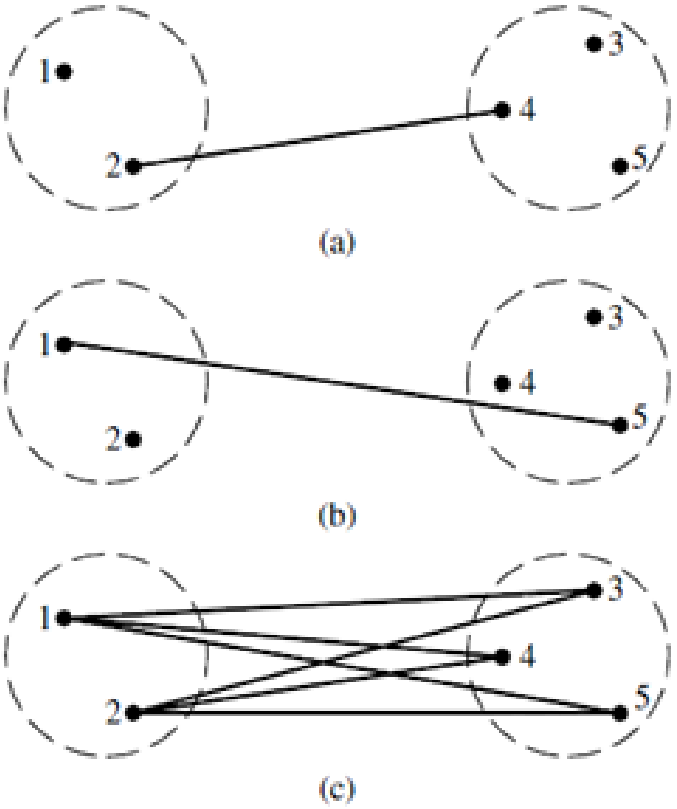
Dendograma ilustrando o agrupamento hierárquico

Técnicas hierárquicas aglomerativas

- Inicia com todos os elementos sendo o próprio *cluster*.
- Usando a medida de similaridade, combina 2 elementos mais semelhantes em um novo *cluster*, agora contendo 2 itens.
- Repete o procedimento de agrupamento usando a medida de similaridade para combinar os dois itens mais semelhantes ou combinações de itens de outro *cluster*.
- Continua o processo até que todos os itens estejam em um único *cluster*.

Técnicas hierárquicas aglomerativas

- Single Linkage (a)
- Complete Linkage (b)
- Average Linkage (c)



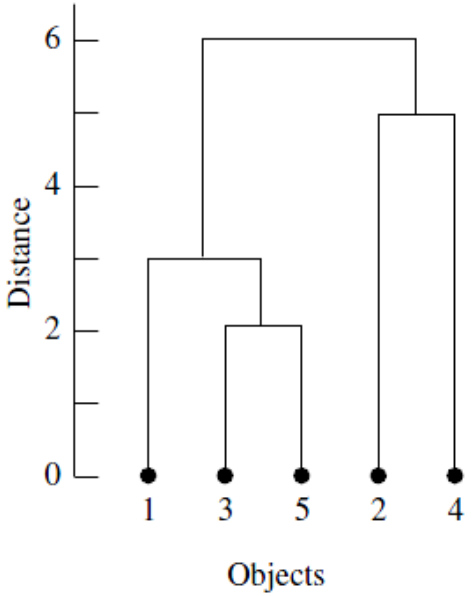
- Centroid Method.
- Ward's Method.

Técnicas hierárquicas aglomerativas

- **Single Linkage:**

$$D_{(nxn)} = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix}\begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & (2) & 8 & 0 \end{bmatrix} \Rightarrow \underbrace{\begin{matrix} (35) \\ 1 \\ 2 \\ 4 \end{matrix}\begin{bmatrix} 0 & & & \\ (3) & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix}}_{\text{Passo 2}} \Rightarrow \underbrace{\begin{matrix} (135) \\ 2 \\ 4 \end{matrix}\begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & (5) & 0 \end{bmatrix}}_{\text{Passo 3}} \Rightarrow \underbrace{\begin{matrix} (135) \\ (24) \end{matrix}\begin{bmatrix} 0 & \\ (6) & 0 \end{bmatrix}}_{\text{Passo 4}}$$

$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{Passo 1}}$



- **Passo 1:** item 3 e 5 serão agrupados: $Min[D = \{d_{lk}\}]$
- **Passo 2:** as distâncias do grupo (35) serão: $d_{(35)k} = \min\{d_{3k}, d_{5k}\}$.

$$d_{(35)1} = \min(d_{31}, d_{51}) = \min(3,11) = 3; \quad d_{(35)2} = \min(d_{32}, d_{52}) = 7; \quad d_{(35)4} = \min(d_{34}, d_{54}) = 8$$

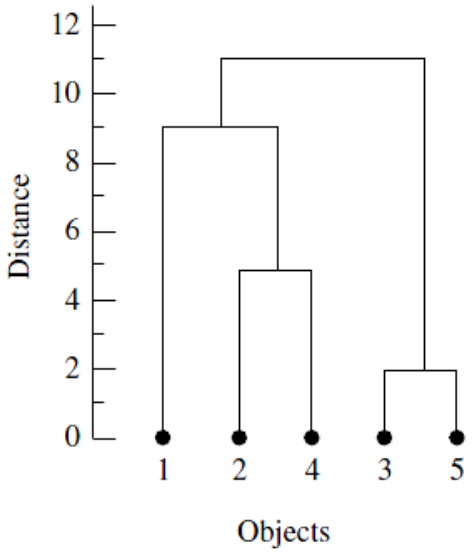
  - Depois roda novamente: $Min[D = \{d_{lk}\}]$; e continua os estágios de agrupamento.

Técnicas hierárquicas aglomerativas

- **Complete Linkage:**

$$D_{(nxn)} = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix}\begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & (2) & 8 & 0 \end{bmatrix} \Rightarrow \underbrace{\begin{matrix} (35) \\ 1 \\ 2 \\ 4 \end{matrix}\begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & (5) & 0 \end{bmatrix}}_{\text{Passo 2}} \Rightarrow \underbrace{\begin{matrix} (35) \\ 1 \\ (24) \end{matrix}\begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & (9) & 0 \end{bmatrix}}_{\text{Passo 3}} \Rightarrow \underbrace{\begin{matrix} (35) \\ (124) \end{matrix}\begin{bmatrix} 0 & \\ (11) & 0 \end{bmatrix}}_{\text{Passo 4}}$$

$\underbrace{\qquad\qquad}_{\text{Passo 1}}$



- **Passo 1:** item 3 e 5 serão agrupados: $Min[D = \{d_{lk}\}]$
- **Passo 2:** as distâncias do grupo (35) serão: $d_{(35)k} = \max\{d_{3k}, d_{5k}\}$.

$$d_{(35)1} = \max(d_{31}, d_{51}) = \max(3,11) = 11; \quad d_{(35)2} = \max(d_{32}, d_{52}) = 10; \quad d_{(35)4} = 9$$

  - Depois roda novamente: $Min[D = \{d_{lk}\}]$ e continua os estágios de agrupamento.

Técnicas hierárquicas aglomerativas

- **Average linkage**: segue os mesmos passos, porém para computar as distâncias de cada *cluster* formado, utiliza-se a distância média:

$$d_{(UV)W} = \frac{\left(\sum_l \sum_k d_{lk}\right)}{N_{(UV)} N_W}, d_{lk} \text{ é a distância entre } l \text{ no cluster (UV) e } k \text{ no cluster W;}$$

- **Centroid method:** a distância entre dois *clusters* é aquela entre as médias (centroide) dos *clusters* formados**:**

$$d_{(UV)W} = (\overline{X}_{UV} - \overline{X}_W)'(\overline{X}_{UV} - \overline{X}_W)$$

é a distância euclidiana ao quadrado entre os vetores de médias $\overline{X}_{UV}$ e $\overline{X}_W$.

O agrupamento em cada passo se dá pelo menor valor da distância.

Técnicas hierárquicas aglomerativas

- **Ward method**: a partição "desejada" é aquela que produz os grupos mais heterogêneos possíveis entre si e o mais possível homogêneo internamente.

- Quando se passa de (n-k) para (n-k-1) *clusters*, a qualidade de partição decresce, pois o nível de fusão aumenta e o nível de similaridade decresce. Ou seja:

$$C_1 \cup C_2 = C \Rightarrow \begin{cases} \downarrow \neq \text{entre os grupos } (C_1, C_2) \\ \uparrow \neq \text{dentro do grupo } (C) \end{cases}$$

- Ward buscou minimizar as "perdas de informação",i.e., tratar essa "mudança de variação" nos 2 casos (inter e intragrupo).

Técnicas hierárquicas aglomerativas

- **Ward method**:
  - Inicia tratando cada item como um *cluster*. Agrupa-os por $\text{Min } d_{jk}$
  - Depois, para um *cluster i*, há ESS_i, que é a soma dos desvios de cada item em relação à média no *cluster* :

$$ESS_i = \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{i.}\right)\left(X_{ij} - \overline{X}_{i.}\right)$$

sendo n_i o número de elementos no cluster i.

  - No passo k, a soma de quadrados dentro dos *clusters* é:

$$SSR = \sum_{i=1}^{g_k} SS_i$$

Técnicas hierárquicas aglomerativas

- **Ward method**: var. quantitativas para o cálculo de médias.
  - A distância entre os *clusters* é definida como:

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i} \right] (\bar{X}_l - \bar{X}_i)'(\bar{X}_l - \bar{X}_i)$$

que é a soma dos quadrados entre os cluster C_l e C_i.

  - Em cada passo, 2 *clusters* são combinados pela *Min d(·)*. A *d(·)* é a ≠ entre o valor de SSR depois e antes de combiná-los.
  - Esta combinação resulta no menor valor de SSR.
  - Centroide ≠ Ward, que trata a ≠ dos tamanhos dos *clusters* comparação. $\left[\frac{n_l n_i}{n_l + n_i} \right]$

Técnicas hierárquicas aglomerativas

- **Coeficiente de Lance e Williams (1967)** : fórmula de recorrência que define a maioria dos métodos hierárquicos bem conhecidos (Stata):

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right|$$

d_{ij} é a distância entre o cluster i e o cluster j; $d_{k(ij)}$ é a distância entre o cluster k e o novo cluster formado pela junção do i e j; e $\alpha_i, \alpha_j, \beta,$ e γ são parâmetros de um método.

- Permite, a cada novo nível do agrupamento hierárquico, a dissimilaridade entre o grupo recém-formado e o resto dos grupos a ser calculado a partir das ≠s do agrupamento atual.
- economias computacionais .

Técnicas hierárquicas aglomerativas

- **Coeficiente de Lance e Williams (1967)** :

Clustering linkage method	α_i	α_j	β	γ
Single	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Average	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Weighted average	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Centroid	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$-\alpha_i\alpha_j$	0
Median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward's	$\frac{n_i+n_k}{n_i+n_j+n_k}$	$\frac{n_j+n_k}{n_i+n_j+n_k}$	$\frac{-n_k}{n_i+n_j+n_k}$	0

Técnicas hierárquicas aglomerativas

- **Coeficiente de Lance e Williams (1967)** : é convertida em medidas de dissimilaridade.

$$d(l,k) = 1 - s(l,k)$$

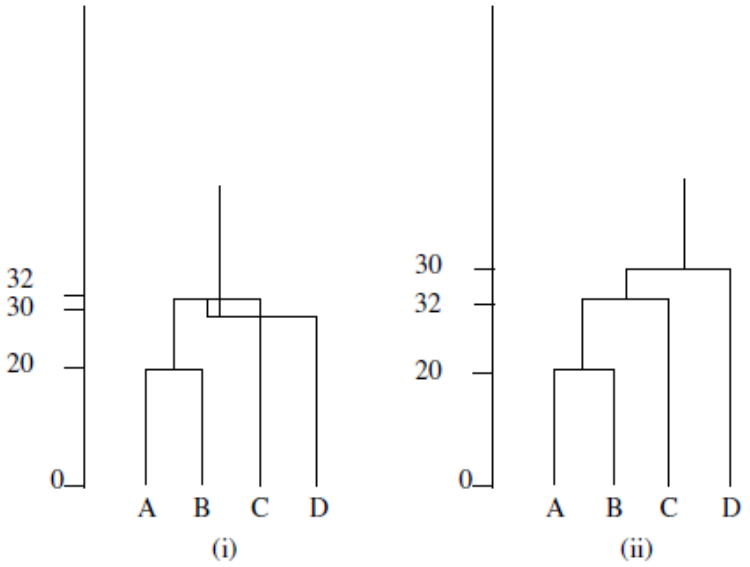
- Há 2 intervalos possíveis: i) similaridade [0,1]=> dissimilaridade [1,0]; ii) similaridade [-1,1] => dissimilaridade [2,0].
- O *software* fornece medidas de dissimilaridades:
  - L_2 : simples, completo e média.
  - L_2^2 : outros, como Ward.

Técnicas hierárquicas aglomerativas

- **Considerações gerais:**
  - Todas as técnicas seguem um algoritmo básico, porém com seus critérios (métrica).Na maioria delas, as variações não são tratadas, => sensíveis aos *outliers.*
  - Não aponta os itens agrupados incorretamente em um estágio anterior. Análise cuidadosa.
  - Aplique várias técnicas. Se a configuração for ≈ consistente => agrupamento natural.
  - Pode-se testar a estabilidade da solução por perturbações nos itens e comparar os resultados (antes/depois). Se os *clusters* forem distinguidos, os resultados (antes/depois) se aproximam.

Técnicas hierárquicas aglomerativas

- **Considerações gerais:**
  - Valores comuns na distância => múltiplas soluções em níveis menores. O usuário necessita conhecê-las (não são ruins).
  - Podem provocar inversões. Ocorrem quando inexiste uma estrutura de cluster clara. Use o m. centroide para solucioná-las. D é adicionado ao grupo (ABC), a uma distância de 30, inferior à distância a qual se juntou C (AB).



Técnicas hierárquicas aglomerativas

- **Comparações dos métodos:**

  a) **single linkage:** estruturas geométricas diferentes, mas é incapaz de delinear grupos pouco separados**.**

  b) **complete linkage:** *clusters* de mesmo diâmetro e isolam os *outliers* nos primeiros passos**.**

  c) **avarege linkage:** *clusters* de mesma variância interna, produzindo melhores partições.

  d) **Ward:** *cluster* com o mesmo n° de itens, baseado nos princípios de análises de variâncias.

(a), (b) e (c) : var. quantitativas e qualitativas; (d): var. quantitativas

Parte IV. Técnicas para a partição final

Técnicas para a partição final

1. Nível de fusão (distância);
2. Nível de similaridade;
3. Coeficiente R^2;
4. Estatística Pseudo F;
5. Correlação semiparcial (Ward);
6. Estatística Pseudo T^2;
7. Estatística CCC (*Cubic Clustering Criterion*);

TÉCNICAS PARA A PARTIÇÃO FINAL

1. **Nível de fusão:** avanço dos passos => ↓ similaridade (↑ d) entre os *clusters*. No dendograma, se existir um salto grande, já se alcançou o n° de *cluster* final.

2. **Nível de similaridade:** detecta pontos em que há decréscimo acentuado na similaridade dos grupos. N° de *cluster* final com acima 90%.

$$S_{il} = \left(1 - \frac{d_{il}}{\max\{d_{jk}\}}\right).100 \qquad \because\ j,k = 1,2,\ldots,n$$

em que $\max\{d_{jk}\}$ é a maior distância entre os n elementos de D no primeiro estágio.

TÉCNICAS PARA A PARTIÇÃO FINAL

3. **<u>Coeficiente R^2</u>:** calcula-se a soma de quadrados intergrupos e intragrupos de uma partição. Seja *j* item e *i* grupo, então:

$X'_{ij} = \left(X_{i1\,j}\ X_{i2\,j}\ \ldots\ X_{ipj}\right)$; $\overline{X}'_{i.} = \left(\overline{X}_{i1.}\ \overline{X}_{i2.}\ \ldots\ \overline{X}_{ip.}\right)$, médias i grupo; $\overline{X}' = \left(\overline{X}_{.1}\ \overline{X}_{.2}\ \ldots\ \overline{X}_{.p}\right)$.

a) Soma de quadrados total : $SSTc = \sum_{i=1}^{g^*}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}\right)'\left(X_{ij} - \overline{X}\right)$

b) Soma de quadrados total intragrupo : $SSR = \sum_{i=1}^{g^*} SS_i = \sum_{i=1}^{g^*}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i.}\right)'\left(X_{ij} - \overline{X}_{i.}\right)$

c) Soma de quadrados total intergrupos : $SSB = \sum_{i=1}^{g^*} n_i\left(\overline{X}_{i.} - \overline{X}\right)'\left(\overline{X}_{i.} - \overline{X}\right)$

Logo : $R^2 = \dfrac{SSB}{SSTc}$

↑R^2 => ↑ SSB e ↓ SSR. Procure se há algum "ponto de salto". Observe a ↓ R^2 quando ↓g grupos.

TÉCNICAS PARA A PARTIÇÃO FINAL

4. **Estatística Pseudo F:** se *F* apresentar um valor de máximo, logo g^* é a partição ideal dos dados:

$$F = \frac{SSB(g^* - 1)^{-1}}{SSTc(n - g^*)^{-1}} = \left(\frac{n - g^*}{g^* - 1}\right)\left(\frac{R^2}{1 - R^2}\right)$$

5. **Correlação semiparcial (Ward):** em um passo, $C_k = C_i \cup C_l$, SPR² será:

$$SPR^2 = \frac{B_{il}}{SST_c}, \qquad B_{il} = \frac{n_i n_l}{n_i + n_l}\left(\bar{X}_{i.} - \bar{X}_{l.}\right)'\left(\bar{X}_{i.} - \bar{X}_{l.}\right)$$

em que B_{il} é a distância intergrupos (Ward).

busca-se o um salto maior que os restantes, o que deve indicar o número de *clusters* e partição ideal.

TÉCNICAS PARA A PARTIÇÃO FINAL

6. **Estatística Pseudo T^2:** em um passo, $C_k = C_i \cup C_l$:

$$P.T^2 = \frac{B_{il}}{\left[\sum_{j \in C_i} \left\| \overline{X}_{ij} - \overline{X}_{i.} \right\|^2 + \sum_{j \in C_l} \left\| \overline{X}_{lj} - \overline{X}_{l.} \right\|^2 \right] \left(n_i + n_l - 2\right)^{-1}};$$

$$\left\| \overline{X}_{kj} - \overline{X}_{k.} \right\| = \left[\left(X_{kj} - \overline{X}_{k.}\right)' \left(X_{kj} - \overline{X}_{k.}\right) \right]^{\frac{1}{2}}$$

   - busca-se ponto de máximo para um número g de grupos.

7. **Estatística CCC (*Cubic Clustering Criterion*)**: compara o R^2 calculado e o seu esperado, $E[R^2]$, supondo que os clusters são gerados por distribuição uniforme *p*-dimensional. Se CCC>3 (bom), R2 > E[R2], i.e., a estrutura de cluster é ≠ da partição uniforme.

TÉCNICAS PARA A PARTIÇÃO FINAL

Indicador	Observação
Nível de fusão (distância)	Salto do ↑D: parar no passo anterior
Nível de similaridade	Salto da ↓S: parar no passo anterior (≈ 90%)
Coeficiente R^2	Salto da ↓ R^2: parar no passo anterior (≥ 90%)
Estatística Pseudo F	Salto da ↓ F: parar no passo anterior
Correlação Semiparcial (SPR^2)	Salto do ↑ SPR^2: parar no passo anterior
Pseudo T^2 ($P.T^2$)	Salto do ↓ $P.T^2$: parar no anterior ou vigente.
Estatística CCC	Salto do ↓ CCC: parar no passo anterior

Parte V. Técnicas não hierárquicas de agrupamento

Técnicas não hierárquicas

- Encontrar diretamente uma partição de n itens em k clusters, por 2 requisitos: semelhança interna e isolamento dos *clusters* formados.
- **Não hierárquicas ≠ hierárquicas**:
  - definição prévia do número de clusters;
  - em cada estágio, novos *clusters* podem ser formados por divisão ou junção de clusters inicialmente definidos. Sem dendogramas;
  - os algoritmos são iterativos e têm uma maior capacidade de análise do conjunto de dados.

Técnicas não hierárquicas: k-médias

- Cada item é alocado para um cluster que tem um centroide mais próximo (média). Passos:

a) escolher k centroides (sementes) para iniciar o processo de partição;

b) comparar cada item com o centroide inicial por uma distância (e.g., euclidiana). Os itens são alocados aos clusters pelo $min\ d(\cdot)$;

c) Após a alocação dos *n* itens, recalcular os centroides para cada novo cluster formado, repetindo o passo (b) com estes novos centroides.

d) repetir os passos (b) e (c) até que todos os elementos estejam bem alocados em seus grupos.

Técnicas não hierárquicas: k-médias

- Em k-médias, a escolha das sementes iniciais influencia na partição final. Assim, seguem algumas sugestões para essa escolha:
- **Sugestão 1**: Use alguma técnica hierárquica para obter os k clusters iniciais. Calcule o vetor de médias de cada grupo, as sementes iniciais.
- **Sugestão 2:** escolha aletoriamente os k centroides iniciais. Selecione m amostras aleatórias com k centroides e repetir a amostragem *m* vezes e, no final, calcula-se os *m* centroides para cada grupo.
- **Sugestão 3:** Escolha a variável de maior variância. Em seguida, divida o domínio da variável em k intervalos. A semente inicial será o centroide de cada intervalo.

Técnicas não hierárquicas: k-médias

- **Sugestão 4:** Escolha os k *outliers* identificados, que serão as sementes iniciais.
- **Sugestão 5:** Escolha prefixada (*ad hoc*) – não muito recomendável.
- **Sugestão 6:** selecione os k primeiros valores do banco de dados. Grande parte dos *softwares* usa como padrão esta sugestão para atribuir as sementes iniciais. Fornece bons resultados quando os itens são bem discrepantes entre si. Logo, não é recomendável quando os elementos são bem semelhantes.

- Mingoti (2005, p.194) aponta que a solução da k—médias, utilizando como sementes iniciais a técnica de Ward, gera melhores resultados que a solução e k-médias, usando os quatro primeiros valores

Técnicas não hierárquicas: Fuzzy c-Médias

- Técnica iterativa e exige a definição inicial de k clusters. Sendo n itens e p variáveis aleatórias, busca-se a partição que minimiza:

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n}\left(u_{ij}\right)^{m} d\left(X_j, V_i\right)$$

V_i é o centroide ponderado do cluster i; $m > 1$ é o parâmetro de Fuzzy;

u_{ij} é a probabilidade do item X_j de pertencer ao grupo de centróide V_i;

$d\left(X_j, V_i\right)$ é a distância escolhida.

- A função é minimizada quando as probabilidades:

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d(X_j, V_i)}{d(X_j, V_k)}\right)^{2/(m-1)}\right]^{-1} \qquad \text{em que } V_i = \frac{\sum_{j=1}^{n}\left(u_{ij}\right)^{m} X_j}{\sum_{j=1}^{n}\left(u_{ij}\right)^{m}}$$

Técnicas não hierárquicas: Fuzzy c-Médias

- Para ter solução final, deve-se ter os centroides e probabilidades iniciais, u_{ij} , geradas de uma distribuição uniforme [0,1].
- Os centroides se modificam a cada iteração e o processo cessa quando a distância entre os centroides dos 2 últimos passos é: $d(V_t, V_{t+1}) < \varepsilon$.
- Nessa técnica, a partição final alocará os itens nos *clusters* conforme a sua maior probabilidade, o que torna possível identificar os itens que se assemelham a mais de um *cluster*.
- Em oposição, a técnica de k-Médias gera uma partição na qual cada elemento pertence a um único cluster.

Técnicas não hierárquicas: comentários

- Tais técnicas são também sensíveis às escalas e aos *outliers*. As variáveis de maior dispersão dominam na distância euclidiana.
- Pode-se padronizar as variáveis ou usar distâncias ponderadas.
- **Há razões para não fixar o n° de *clusters***, como nessas técnicas:
  - se 2 ou mais sementes estão dentro de um *cluster*, os clusters resultantes serão pobremente diferenciados;
  - Outliers => pelo menos 1 cluster com itens muito dispersos.
  - Mesmo que saiba os itens nos k clusters, perde-se grupos raros e latentes na amostra. Os k grupos iniciais => partição sem sentido.

Técnicas não hierárquicas: comentários

- Comparando às técnicas, pode-se afirmar:
  - quando os grupos estão bem separados, qualquer técnica leva a resultados satisfatórios;
  - quando há interseção inicial entre os grupos, Fuzzy é melhor por gerar a probabilidade dos itens;
  - para definir o n° final de grupos, pode aplicar bootstrap a fim de delinear um intervalo de confiança.