Disciplina: Análise Multivariada I

Prof. Dr. Admir Antonio Betarelli Junior

AULA 5

1 ANÁLISE DE AGRUPAMENTO (AA)

Procedimentos exploratórios são bem úteis no entendimento da natureza complexa de relação multivariada. Encontrar nos dados uma estrutura de agrupamento natural é uma importante técnica exploratória. Isso porque agrupamentos podem fornecer um significado informal para avaliar a dimensionalidade, identificar *outliers* e sugerir hipóteses acerca da estrutura de relações dos dados. Em linhas gerais, a análise de Agrupamentos (*clusters*) consiste em uma técnica de procedimentos exploratórios que busca descobrir agrupamentos naturais de indivíduos (ou variáveis) a partir dos dados observados, agrupando indivíduos com base na similaridade ou distâncias (dissimilaridades).

Nessa técnica, a ideia é maximizar a homogeneidade de indivíduos dentro de grupos, ao mesmo tempo em se maximiza a heterogeneidade entre os grupos. A escolha do método de *cluster* a ser utilizado é balizada por duas questões: *i)* Como medir as similaridades entre indivíduos? *ii)* Como agrupar indivíduos semelhantes?

1.1 Objetivos principais

Tem por objetivo de particionar um conjunto de elementos em dois ou mais grupos (*clusters*) com base na similaridade dos elementos a partir de um conjunto de características especificadas (variáveis aleatórias).

Esse método possui três aplicações mais comuns:

a) classificação de elementos (taxonomia): identificação de grupos naturais dentro dos dados;

b) simplificação de dados: a capacidade de analisar grupos de observações semelhantes em vez de todas as observações individuais;
c) identificação das relações entre os elementos: a estrutura simplificada de análise de *cluster* retrata relações não reveladas.

1.2 Quando usar a técnica

Deve-se usar essa técnica quando a preocupação principal é dividir os elementos da amostra (ou população) em grupos, de forma que os elementos de um mesmo grupo sejam homogêneos com relação às variáveis estudadas e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas variáveis.

2 MEDIDAS DE DISSIMILARIEDADES E SIMILARIDADES

A semelhança entre os indivíduos pode ser medida de acordo com suas similaridades ou, em sentido oposto, por suas dissimilaridades, chamadas de distâncias no espaço das variáveis. Quando os indivíduos são agrupados, a proximidade é usualmente indicada por algum tipo de distância. Vale ressaltar que isto difere do agrupamento de variáveis, que usualmente é feito com base de medidas de associação.

2.1 Medidas de dissimilaridades (distâncias) para variáveis quantitativas

Considere o vetor aleatório, $\mathbf{X}'_j = [X_{j1}, X_{j2}, ..., X_{jp}]$, composto por p variáveis para cada elemento j das n observações da amostra. Assim, utiliza-se uma das medidas de dissimilaridades ou distância, que quanto menor o seu valor, mais similares são os elementos comparados:

a) Distância euclidiana:

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

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

b) Distância generalizada ou ponderada:

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

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.

em que A é a matriz de ponderação. A sua escolha reflete o tipo de informação a ser utilizada na ponderação das diferenças das coordenadas dos vetores comparados. Se ainda $A = diag(S_i^2)^{-1}$, leva em consideração somente a diferença de variabilidade entre as variáveis (S_i^2 é variância amostral da *i-ésima* variável). Já quando $A = S^{-1}$, pondera as possíveis diferenças de variâncias e covariâncias entre as variáveis.

c) Distância de Minkowsky

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

se

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

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

em que $w_i's$ são os pesos de ponderação para as variáveis. A métrica de Minkowsky é menos afetada pela presença de *outliers* do que a distância euclidiana.

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*.

d) Métrica de Canberra

$$d(X_l, X_k) = \sum_{i=1}^{p} \frac{|X_{il} - X_{ik}|}{(X_{il} + X_{ik})} \qquad \because (j \neq l)$$

e) Coeficiente de Czekanowski

$$d(X_l, X_k) = 1 - \frac{2\sum_{i=1}^{p} \min(X_{il} - X_{ik})}{\sum_{i=1}^{p} (X_{il} + X_{ik})} \qquad \because (j \neq l)$$

Para o cálculo da A métrica de Canberra e o coeficiente de Czekanowski, as variáveis devem ser não negativas somente.

2.2 Medidas de similaridades para variáveis qualitativas

Em muitas situações, a pesquisa envolve a análise de variáveis qualitativas. Para tanto, há duas alternativas:

i. transforma-as em quantitativas e, em seguida, usa-se as medidas de distâncias;
ii. trabalha-se com coeficientes de similaridades e, posteriormente, comparam-se os elementos de acordo com a presença ou ausência de certas características.

Para entender o problema de quando se têm variáveis qualitativas, considere 5 variáveis binárias:

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

Existem dois pares (1,1), um par (0,0) e dois pares incompatíveis (0,1;1,0). Logo o quadrado da distância fornece o mesmo número de itens dissimilares:

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

Em geral, deve-se comparar os elementos de acordo com a presença ou ausência de certas características. Elementos "parecidos" devem ter em comum mais itens similares que dissimilares. No exemplo acima, os pares (1,1) e (0,0) são ignorados no computo da distância na sua versão original.

Considere o esquema abaixo, que 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*

Neste esquema, *a* representa a frequência do par (1,1), *b* a do par (1,0), e assim por diante. A partir disso, conforme o quadro abaixo de Johnson e Wichern (2002, p.674), pode-se desenvolver os coeficientes de similaridades para o agrupamento dos itens.

Coefficient	Rationale
1. $\frac{a+d}{p}$	Equal weights for 1–1 matches and 0–0 matches.
2. $\frac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1–1 matches and 0–0 matches.
3. $\frac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\frac{a}{p}$	No 0–0 matches in numerator.
5. $\frac{a}{a+b+c}$	No 0–0 matches in numerator or denominator. (The 0–0 matches are treated as irrelevant.)
6. $\frac{2a}{2a+b+c}$	No 0–0 matches in numerator or denominator. Double weight for 1–1 matches.
7. $\frac{a}{a+2(b+c)}$	No 0–0 matches in numerator or denominator. Double weight for unmatched pairs.
8. $\frac{a}{b+c}$	Ratio of matches to mismatches with 0–0 matches excluded.

*[*p* binary variables; see (12-7).]

Pode-se destacar e exemplificar alguns deles, como segue:

a) <u>concordância simples</u>: considera os pares concordantes, (1,1) e (0,0), em relação ao total de pares.

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

$$\uparrow s(\cdot) \Rightarrow \uparrow \text{similaridade}$$

b) concordância positiva: considera somente o par (1,1), pois (0,0) não necessariamente representa concordância (ideia do caso contrário).

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

$$\uparrow s(\cdot) \Rightarrow \uparrow \text{similaridade}$$

c) concordância de Jaccard: considera a proporção do par (1,1) em relação ao total de pares diferentes de (0,0).

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

$$\uparrow s(\cdot) \Rightarrow \uparrow \text{similaridade}$$

d) distância euclidiana média, um índice de dissimilaridade.

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

$$\uparrow d(\cdot) \Rightarrow \uparrow \text{dissimilaridade ou} \downarrow \text{similaridade}$$

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

2.3 Medidas de similaridades para variáveis quantitativas

Qualquer medida de distância usada para variáveis quantitativas pode ser transformada em um coeficiente 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.

2.4 Alternativas com variáveis quantitativas e qualitativas

Uma situação comum é quando *p* variáveis quantitativas e *q* variáveis qualitativas são observadas nas *n* observações. Assim, pode-se escolher:

a) transformar as variáveis qualitativas em quantitativas ao atribuir valores numéricos às categorias (*ad hoc*). Em seguida, utiliza-se uma das medidas de distância para comparar as $p+q$ variáveis;

b) transformar as variáveis quantitativas em qualitativas categorizando os seus valores por algum critério. Depois disso, utiliza-se uma das medidas de similaridade para comparar as $p+q$ variáveis.

c) construir medidas de semelhança mistas e utilizá-las para a comparação dos elementos amostrais. Elabora-se uma combinação linear entre os dois tipos de variáveis (*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 $\omega_q = \dfrac{q}{p+q}$; e $c_p(\cdot)$ e $c_q(\cdot)$ são coefientes de similaridade.

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

d) Coeficiente de Gower (1971): para cada variável *j*, considera-se um coeficiente de semelhança, s_j, em um intervalo [0,1]. Comparando os elementos *l* e *k*, a similaridade entre os mesmos é dada por:

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

em que $1_j(l,k)$ é uma variável igual a 1 se *l* e *k* podem ser comparados segundo a variável X_j.

Por exemplo, se existir 6 variáveis, porém para *l* somente há os valores de 4 variáveis, então a comparação de *l* e *k* será feita justamente entre as 4 variáveis. Para

usar esse coeficiente, é preciso transformar as medidas de distância das variáveis quantitativas em medidas de similaridades.

2.5 Medidas de similaridades para pares de variáveis

Ao invés dos elementos, as variáveis serão agrupadas. Para tanto, geralmente usa-se os coeficientes de correlação amostral. Entretanto, em algumas aplicações, correlações negativas são substituídas por valores absolutos.

Quando as variáveis são binárias, os dados podem ser agrupados na forma de tabela de contingência. As variáveis, ao contrário 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 aplicada para as variáveis binárias é:

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

Esse número pode ser tomado como uma medida de similaridade entre duas variáveis. Essa correlação é relacionada a uma estatística qui-quadrada ($r^2 = \chi^2 / n$) para testar a independência de duas variáveis categóricas. Para *n* fixado, a grande similaridade é consistente com a ausência de independência.

3 TÉCNICAS HIERARQUICAS DE AGRUPAMENTO

Essas técnicas se dividem entre *Hierárquicas (aglometarivas e divisivas)* e *não hierárquicas*. As técnicas não hierárquicas, número *g* de grupos já deve ser pré-especificado. Já as hierárquicas procuram identificar agrupamentos e o provável número *g* de grupos, por prosseguir por uma série de fusões sucessivas (aglomerativas) ou uma série de sucessivas divisões (divisas).

Nos métodos hierárquicos aglomerativos, os elementos são inicialmente agrupados por suas similaridades, cujos grupos, em etapas posteriores, são combinados de acordo com as suas semelhanças. Eventualmente, como a semelhança diminui, todos os subgrupos são fundidos em um único *cluster*.

Por sua vez, os métodos hierárquicos divisívos trabalham na direção oposta. Um único grupo de elementos é divido em dois subgrupos dissimilares. Esses subgrupos são, em seguida, subdivididos em mais subgrupos dissimilares. O processo continua até que cada subgrupo seja o próprio elemento amostral.

Os resultados de ambos os métodos, aglomerativos e divisivos, podem ser observados em um diagrama de duas dimensões: *dendograma*. O dendograma ilustra as fusões ou divisões que têm sido feitas em níveis sucessivos.

3.1 Técnicas hierárquicas aglomerativas (populares)

A seguir serão apresentadas a técnicas mais populares, como *single linkage* (distância mínima), *complete linkage* (distância máxima), e *avarege linkage* (distância média).

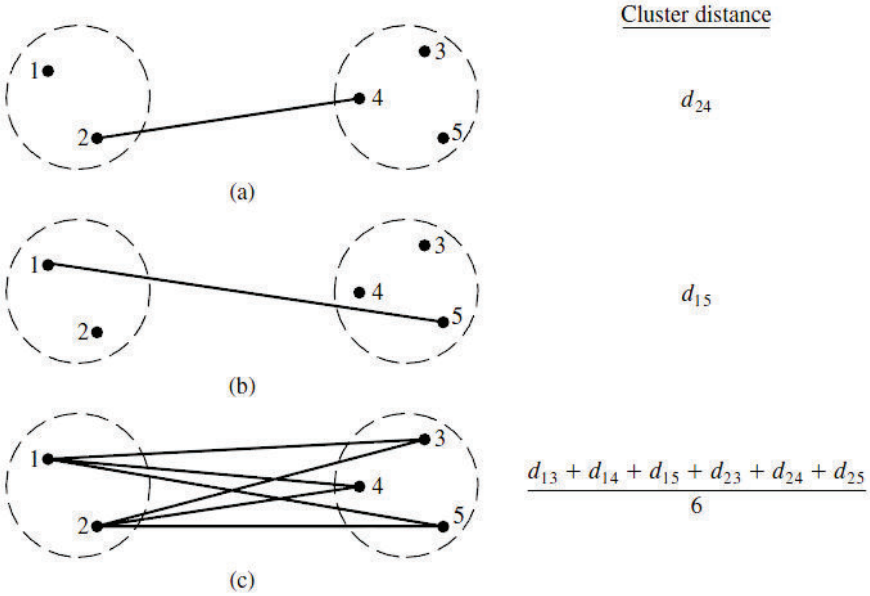


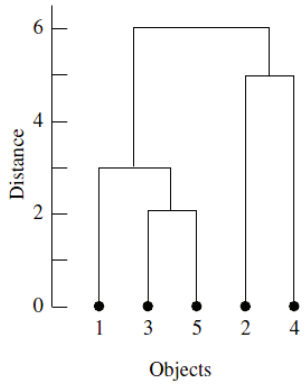
Figure 12.2 Intercluster distance (dissimilarity) for (a) single linkage, (b) complete linkage, and (c) average linkage.

M1. Single linkage: a similaridade entre dois grupos é definida pelos elementos mais próximos, i.e., entre aqueles cuja distância é mínima. Inicialmente o algoritmo agrupa dois elementos em um *cluster* (UV) conforme distância mínima, $Min[D=\{d_{lk}\}]$. As distâncias entre esse *cluster* (UV) e outro qualquer (W) são computadas pelo critério $d_{(UV)W}=\min\{d_{UW},d_{VW}\}$. Em seguida, o passo se repete até que todos os elementos estejam contidos em um único *cluster*. O exemplo 12.4 Johnson e Wichern (2002, p.681-682) apresenta bem os passos dessa técnica, cuja matriz de distância é:

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

Passo 1

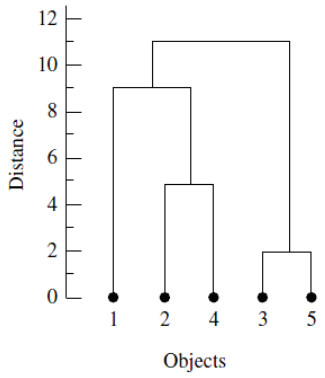
Os elementos 3 e 5 são fundidos em um único grupo, pois $\min(d_{lk})=d_{35}=2$. Esse novo grupo (35) passa a ter as seguintes distâncias mínimas: $d_{(35)1}=\min(d_{31},d_{51})=\min(3,11)=3;\ \ d_{(35)2}=\min(d_{32},d_{52})=7;\ \ d_{(35)4}=\min(d_{34},d_{54})=8$, de acordo com a matriz do passo 2. Em seguida identifica-se a menor distância na matriz do passo 2, $\min(d_{lk})=d_{(35)1}=3$, gerando um novo grupo (135) e preservando as distâncias mínimas entre os elementos (35) e (1). No passo 4, observa-se a formação de dois *clusters:* (135) e (24) de tal modo que os mesmos serão agrupados pela $\min(d_{lk})=d_{(135)(24)}=4$. O dendograma abaixo ilustra a sucessão de construção dos clusters no referido exemplo.



M2. Complete linkage: procede do mesmo modo que a técnica anterior, com uma importante exceção. As distâncias dos novos *clusters* são computadas pelo critério de maior distância entre os elementos combinados, i.e., $d_{(UV)W} = \max\{d_{UW}, d_{VW}\}$. Isto assegura que todos os elementos em um *cluster* estão dentro de uma máxima distância (ou mínima similaridade) de cada outro. Resumidamente, o agrupamento em cada estágio é feito entre os elementos que apresenta o menor valor da distância máxima. Usando a mesma matriz do exemplo anterior:

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

Os elementos 3 e 5 são fundidos em um único grupo, pois $\min(d_{lk}) = d_{35} = 2$. Esse novo grupo (35) passa a ter as seguintes distâncias máximas: $d_{(35)1} = \max(d_{31}, d_{51}) = \max(3,11) = 11$; $d_{(35)2} = \max(d_{32}, d_{52}) = 10$; $d_{(35)4} = 9$, conforme a matriz do passo 2. Posteriormente, observa-se a menor distância na matriz do passo 2, $\min(d_{lk}) = d_{24} = 5$, gerando um novo grupo (24) com distâncias máximas entre os elementos (2) e (4). No passo 4, nota-se a formação de dois *clusters:* (35) e (124) de tal modo que os mesmos serão agrupados conforme $\min(d_{lk}) = d_{(35)(124)} = 10$. Essas sucessivas formações de agrupamentos estão ilustradas no dendograma abaixo.



M3. Average linkage: segue os mesmos passos, porém para computar as distâncias de cada *cluster* formado, utiliza-se a distância média entre os seus membros e os demais elementos amostrais:

$$d_{(UV)W} = \frac{\left(\sum_{l} \sum_{k} d_{lk} \right)}{N_{(UV)} N_W}$$

em que d_{lk} é a distância entre o elemento l no cluster (UV) e o elemento k no cluster W; e $N_{(UV)}$ e N_W são os números de elementos contidos no cluster (UV) e W, respectivamente.

M4. Centroid method: a distância entre dois clusters é definida como sendo a distância entre os vetores de médias (centroide) dos grupos comparados.

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

que é a distância 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.

Contudo, para fazer o agrupamento é necessário voltar aos dados originais e a cada passo para o cálculo da matriz de distâncias.

3.2 Método de agrupamento hierárquico de Ward (1963)

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, a variação (diferenças) entre grupos diminuem e as dissimilaridades (variação) dentre dos grupos aumentam.

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

Ward procurou minimizar as "perdas de informação" diante da junção de dois grupos. Ou melhor, o autor propôs um tratamento para essa "mudança de variação" nos dois casos (inter e intragrupo). Para um cluster *i*, ESS_i é a soma dos desvios de cada elemento no grupo em relação à média do grupo (centroide):

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

sendo n_i o número de elementos no cluster *i*.

No passo k, a soma de quadrados total dentro dos grupos é $SSR = \sum_{i=1}^{g_k} SS_i$. Já a distância entre os *clusters* é definida como:

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

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

Em cada passo do algoritmo de agrupamento, os dois *clusters* que minimizam esta distância são combinados. Esta medida de distância é a diferença entre o valor de *SSR* depois e antes de combinar os dois *clusters* num único conglomerado. Assim, em cada passo o método combina os 2 *clusters* que resultam no menor valor de SSR. Ademais, diferentemente do método de centroide, o de Ward considera a diferença dos tamanhos dos grupos em comparação: $\left(n_l n_i\right)\left(n_l + n_i\right)^{-1}$.

Importante: para o uso do método de Ward, basta que as *p*-variáveis sejam quantitativas e passíveis, portanto, do cálculo de médias.

3.3 Coeficiente de Lance e Williams (1967)

Os autores desenvolveram uma fórmula de recorrência que define, como casos especiais, a maioria dos métodos hierárquicos bem conhecidos, incluindo todos os métodos hierárquicos encontrados no *Stata*. A fórmula de recorrência de Lance-Williams é:

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

em que

d_{ij} é a distância (ou dissimilaridade) entre o cluster *i* e o cluster *j*; $d_{k(ij)}$ é a distância entre o cluster *k* e o novo cluster formado pela combinação do cluster *i* e cluster *j*; e $\alpha_i, \alpha_j, \beta,$ e γ são parâmetros que são conjuntos baseados em um método hierarquico particular.

A fórmula de recorrência 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 diferenças do agrupamento atual. Esta abordagem pode resultar em grandes economias computacionais nos cálculos de cada passo de hierarquia dos elementos amostrais. Esta característica da formula de recorrência permite utilizar uma matriz de similaridade ou dissimilaridade. A tabela abaixo expõe os valores de parâmetros para cada método hierárquico.

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

A fórmula Lance-Williams, usada para o cálculo de agrupamento hierárquico no Stata, é convertida em medidas de dissimilaridade. Antes de realizar agrupamento hierárquico, o software transforma as medidas de similaridade, tanto contínuas e binárias, para dissimilaridades:

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

que é usada transformar a partir da similaridade para uma medida de dissimilaridade, e de volta. Há dois intervalos possíveis para essa fórmula. Medidas de similaridade variam de 0 à 1, resultando em dissimilaridades de 1 à 0. Também, podem as similaridades podem variar de -1 à 1, resultando em dissimilaridades de 2 à 0. Para variáveis contínuas, o

software fornece medidas de dissimilaridades L_2 e L_2^2. A primeira, L_2, é usada para os métodos de ligação tradicionais (simples, completo e média). Já para outros métodos, como o de Ward, utiliza-se L_2^2.

Lance e Williams (1967), Anderberg (1973), Jain e Dubes (1988), Kaufman e Rousseeuw (1990) e Gordon (1999) advertem sobre o uso de muitos tipos de medidas de similaridade e dissimilaridade, apontando que o melhor é usar o padrão, que é distância quadrada euclidiana (ou distância euclidiana), em vez de ter resultados com interpretações difíceis. Na ausência de uma familiaridade maior com os métodos hierárquicos, use a medida de dissimilaridade padrão.

3.4 Comentários gerais

Todos os métodos de agrupamento hierárquico seguem um algoritmo básico, porém com os seus próprios critérios (métrica). Como em muitos métodos de agrupamentos, as fontes de erro e de variações não são formalmente consideradas em nenhum deles, o que torna tais métodos mais sensíveis aos *outliers* (ou "pontos de ruídos"). Inexiste provisão para a realocação dos elementos que podem ter sido agrupados incorretamente em um estágio anterior. Consequentemente, a configuração final de *clusters* deve sempre ser cuidadosamente examinada, verificando se a mesma é sensível.

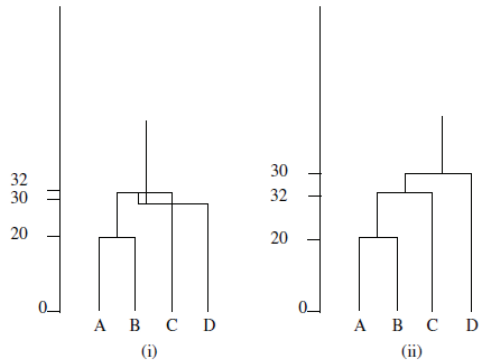
Uma boa ideia é experimentar vários métodos. Se os resultados gerados são aproximadamente consistentes de um método com o outro, então, talvez, um caso de agrupamento natural pode ter alcançado.

A estabilidade de uma solução hierárquica pode, algumas vezes, ser verificada, aplicando pequenas perturbações nas unidades de dados e comparando os resultados entre o antes e o depois do método de agrupamento. Se os grupos forem razoavelmente bem distinguidos, os agrupamentos entre o antes e o depois das perturbações devem se aproximar ou concordar.

Valores comuns na similaridade ou na matriz de distância podem produzir múltiplas soluções para um problema de agrupamento hierárquico. Ou seja, os dendogramas correspondentes aos diferentes tratamentos podem gerar similaridades diferentes, particularmente em níveis menores. Isto não é um problema inerente de qualquer método, em vez disso, múltiplas soluções ocorrem para certos tipos de dados.

Múltiplas soluções não necessariamente são ruins, mas o usuário necessita conhecer as suas existências de maneira que os dendogramas, com os diferentes grupos formados, possam ser corretamente interpretados.

Tais métodos também podem provocar inversões, que ocorrem quando um elemento se junta a um *cluster* existente em uma distância menor que na consolidação anterior. Veja as figuras abaixo. No painel (i), o método de agrupamento junta A e B em uma distância 20. No próximo passo, C é adicionado ao *cluster* (AB) a uma distância 32. Devido à natureza do algoritmo de agrupamento, D é adicionado ao grupo (ABC), a uma distância de 30, inferior à distância a qual se juntou C (AB). Em (i) a inversão é indicado por um dendrograma com cruzamento. Em (ii), a inversão é indicado por um dendrograma sem escala ordenada de distância (não monotônica).



Inversões podem ocorrer quando não existe uma estrutura de c*luster* clara.

3.4.1 Comparações dos métodos

Após a apresentação dos métodos hierárquicos mais tracionais, torna-se oportuno compará-los de forma simples e direta.

a) *single linkage*: pode gerar estruturas geométricas diferentes, mas é incapaz de delinear grupos pouco separados.
b) *complete linkage*: tende a gerar *clusters* de mesmo diâmetro e isolar os *outliers* nos primeiros passos.
c) *avarege linkage*: tende a gerar *clusters* de mesma variância interna, produzindo melhores partições.
d) *Método de Ward*: tende a gerar *clusters* com o mesmo número de elementos, baseado nos princípios de análises de variâncias.

Os três métodos, (a), (b) e (c), são aplicáveis às variáveis quantitativas e qualitativas, enquanto que o método de Ward aplica-se somente para variáveis quantitativas.

Importante: a consistência entre as soluções obtidas pelos métodos pode não ocorrer se há um grande número de dados e variáveis.

4 O NÚMERO DE *CLUSTER* DA PARTIÇÃO FINAL

Existem basicamente 7 procedimentos para alcançar o número de *cluster* numa partição final, a saber:

a) Análise do comportamento do nível de fusão (distância);
b) Análise do comportamento do nível de similaridade;
c) Análise da soma de quadrados entre grupos: coeficiente R^2;
d) Estatística Pseudo F;
e) Correlação semiparcial (Ward);
f) Estatística Pseudo T^2;
g) Estatística CCC (*Cubic Clustering Criterion*);

4.1 Nível de fusão

Quando se avança os estágios de agrupamentos, a similaridade entre os *clusters* existentes diminui (aumenta a distância). Pode-se usar o dendograma ou construir um gráfico entre o número de *cluster* e a distância (fusão) do agrupamento em cada estágio. Se existir pontos de salto relativamente grandes, os mesmos indicam que já se alcançou o número final de *clusters*.

4.2 Nível de similaridade

Computa-se a medida abaixo para detectar pontos em que há decréscimo acentuado na similaridade dos grupos:

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

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

Em geral, faixa de similaridade acima de 90% resulta num número muito grande de grupos.

4.3 Coeficiente R^2

Em cada estágio, pode-se calcular a soma de quadrados intergrupos e intragrupos da partição correspondente. Seja:

$X'_{ij} = (X_{i1j}\ X_{i2j}\ ...\ X_{ipj})$, o vetor aleatório para o j - ésimo elemento do i - ésimo grupo;

$\overline{X}'_{i.} = (\overline{X}_{i1.}\ \overline{X}_{i2.}\ ...\ \overline{X}_{ip.})$, o vetor de médias do i - ésimo grupo;

$\overline{X}' = (\overline{X}_{.1}\ \overline{X}_{.2}\ ...\ \overline{X}_{.p})$, o vetor de média global.

Então, tem - se:

a) a soma de quadrados total:

$$SSTc = \sum_{i=1}^{g^*}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})'(X_{ij} - \overline{X})$$

b) A soma de quadrados total intragrupo (residual):

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

c) A soma de quadrados total intergrupos:

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

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

Assim, quanto maior R^2, maior a soma de quadrado entre grupos e menor a soma de quadrados residual. Procure se há algum "ponto de salto", que deve ser o momento de parada. Observe a queda de R^2 quando o número de g grupos decresce.

4.4 Estatística Pseudo F

Clinski e Harabasz (1974) sugerem o cálculo da estatística *F*, definida em cada estágio do agrupamento:

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

Se *F* apresentar um valor de máximo, logo g^* é a partição ideal dos dados. Na prática, busque o maior valor de *F*, que relaciona com a menor significância, rejeitando a igualdade dos vetores de médias, i.e., leva à partição com maior heterogeneidade entre os grupos. A estatística tem distribuição *F* com *p(g*-1)* e *p(n-g*)* graus de liberdade, quando os *n* elementos são uma amostra aleatória de uma distribuição normal *p*-variada e quando os elementos são alocados aleatoriamente em grupos.

4.5 Correlação semiparcial (Ward)

Se, em um passo do agrupamento, o conglomerado $C_k = C_i \cup C_l$, então o coeficiente de correlação semiparcial (não descrescente) é:

$$SPR^2 = \frac{B_{il}}{SST_c}$$

$$B_{il} = \frac{n_i n_l}{n_i + n_l}\left(\overline{X}_{i.} - \overline{X}_{l.}\right)'\left(\overline{X}_{i.} - \overline{X}_{l.}\right)$$

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

Para cada passo do agrupamento é calculado o referido coeficiente, traçando um gráfico entre o passo e o valor de coeficiente. Neste gráfico, busca-se o ponto da curva em que ocorre um salto maior que os restantes, que deve indicar o número de *clusters* e partição ideal.

4.6 Estatística Pseudo T^2

Proposta por Duda e Hart (1973) e sob um passo do agrupamento, $C_k = C_i \cup C_l$, essa estatística definida como:

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

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

Essa estatística teria distribuição F com p e (n_i+n_l-2) graus de liberdade. Contudo, na prática, alocação aleatória não ocorre, dada partição dos elementos por meio dos métodos de agrupamento com critérios de similaridade. Em cada passo do algoritmo, o valor da estatística é calculado e um gráfico que relaciona o passo *versus* o valor Pseudo T^2. Esse valor máximo relaciona-se com a menor probabilidade de significância do teste, rejeitando a igualdade dos vetores de médias com maior significância.

4.7 Estatística CCC (*Cubic Clustering Criterion*)

Essa estatística compara o R^2 calculado e o seu esperado, $E[R^2]$, supondo que os grupos são gerados com uma distribuição uniforme p-dimensional. Quando os valores da estatística são positivos, os mesmos indicam que $R^2 > E[R^2]$, i.e., a estrutura de *cluster* é diferente da partição uniforme. O número de grupos da partição final relacionado com valores de CCC maiores que 3.

4.8 Resumo

O quadro abaixo expõe um resumo dos 7 procedimentos em relação aos números de passos de agrupamentos realizados. Veja a análise feita no Exemplo 6.8 da Mingoti (2005, p.184 – 187).

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

5 TÉCNICAS NÃO HIERARQUICAS DE AGRUPAMENTO

Tem por objetivo encontrar diretamente uma partição de *n* elementos em *k clusters*, atendendo a dois requisitos: i) semelhança interna e ii) isolamento dos *clusters* formados.

Para encontrar a melhor partição de ordem *k*, é usado algum critério de qualidade da partição, ou seja, são necessários processos de investigação das partições possíveis.

As principais diferenças em relação às técnicas hierárquicas são:

a) definição prévia do número de *clusters*;
b) em cada estágio, novos *clusters* podem ser formados por divisão ou junção de *clusters* inicialmente definidos. Assim, não é mais possível a construção de dendogramas;
c) os algoritmos são iterativos e têm uma maior capacidade de análise do conjunto de dados.

Existem, por exemplo, técnicas como: *k-Médias*, *Fuzzy c-Médias*, *redes neurais artificiais*.

5.1 Técnica de k-médias

MacQueen (1967) sugere essa técnica em que cada elemento amostral é alocado para um cluster que tem um centroide mais próximo (média). Tal técnica é composta pelos seguintes passos:

a) escolher *k* centroides (sementes) para iniciar o processo de partição;
b) comparar cada elemento com o centroide inicial por um tipo de distância (e.g., euclidiana). Os elementos são alocados aos clusters pelo critério de menor distância;
c) após a alocação dos *n* elementos, recalcular os centroides para cada novo cluster formado, repetindo o passo (b) a partir destes novos centroides.
d) repetir os passos (b) e (c) até que todos os elementos estejam bem alocados em seus grupos (i.e., até que nenhuma realocação de elementos seja necessária).

Veja o exemplo 12.12 em Johnson e Wichern (2002, p.695).

Nesse procedimento, a escolha das sementes iniciais influencia no agrupamento final. Assim, seguem algumas sugestões para essa escolha:

Sugestão 1: Use alguma técnica hierárquica para obter os k *clusters* iniciais. Depois disso, calcule o vetor de médias de cada grupo, que serão as sementes iniciais. Isto foi feito no exemplo 12.12 em Johnson e Wichern (2002, p.695).

Sugestão 2: Escolha aletoriamente os k centroides iniciais (e.g. amostragem aleatória simples sem reposição). Em seguida, selecione m amostras aleatórias com k centroides e repetir a amostragem m vezes e, no final, calculam-se os m centroides para cada grupo.

Sugestão 3: Escolha a variável de maior variância dentre as p componentes do vetor X. Posteriormente, divida o domínio da variável em k intervalos. A semente inicial será o centroide de cada intervalo.

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. Em geral, fornece bons resultados quando os elementos são bem discrepantes entre si. Assim, esta sugestão não é recomendável quando os elementos são bem semelhantes.

Caberá o usuário verificar o que está ocorrendo com o seu banco de dados e rearranjá-lo de modo a obter os melhores resultados.

Importante: no exemplo 7.1 da Mingoti (2005, p.194), a autora 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.

5.2 Técnica Fuzzy c-Médias

Também uma técnica iterativa e exige a definição inicial de *k* clusters. Sendo *n* elementos e *p* variáveis aleatórias, busca-se a partição que minimiza:

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

em que V_i é o centroide ponderado do cluster i; $m > 1$ é o parâmetro de Fuzzy; u_{ij} é a probabilidade do elemento X_j de pertencer ao grupo de centróide V_i; $d(X_j, V_i)$ é a distância escolhida.

A função é minimizada quando as probabilidades, u_{ij}, são escolhidas como abaixo:

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

Para encontrar 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 de uma iteração em relação à anterior é menor ou igual a um valor de erro ε pré-estabelecido: $d(V_t, V_{t+1}) < \varepsilon$.

Na técnica Fuzzy, para cada elemento é estimada a probabilidade de pertencer a um dos c *clusters*. A partição final alocará os elementos nos grupos de acordo com a sua maior probabilidade, o que torna possível identificar os elementos que se assemelham a mais de um dos *c* grupos. Em oposição, a técnica de *k*-Médias gera uma partição na qual cada elemento pertence a um único *cluster*. Veja o exemplo 7.2. em Mingoti (2005, p.197).

5.3 Comentários finais

Assim como as técnicas hierárquicas, essas técnicas são sensíveis às dispersões das variáveis de diferentes escalas e por *outliers*. As variáveis de maior dispersão dominam a

definição da distância euclidiana (e.g.). Para evitar isso, pode-se padronizar as variáveis ou usar métodos de agrupamentos com distâncias ponderadas. Aconselha-se que se faça uma análise exploratória prévia para verificar a existência de *outliers*.

Além disso, existem forte argumentos para não fixar o número de *clusters*, como nas técnicas não hierárquicas:

a) se dois ou mais sementes ocasionalmente encontram-se dentro de um *cluster*, os clusters resultantes serão pobremente diferenciados;
b) a existência de *outliers* pode produzir pelo menos um *cluster* com itens muito dispersos;
c) mesmo que os elementos sejam conhecidos para compor os *k clusters*, pode-se deixar de obter informações latentes ou dados de grupos raros que não apareceram na amostra. Assim, os *k* grupos iniciais poderia conduzir aglomerados sem sentido.

Importante: se mesmo em alguns casos, o usuário precise especificar previamente *k clusters*, sempre é uma boa ideia executá-lo por várias escolhas.

Além disso, quando comparados às técnicas não hierárquicas, pode-se afirmar que:

- quando os grupos da partição natural estão bem separados no espaço em relação às *p* variáveis, qualquer método leva a resultados satisfatórios;
- quando a partição natural permite uma grande interseção entre os grupos, o método Fuzzy é mais adequado por estimar a probabilidade de que ele pertença a cada um dos grupos;
- para determinar o número final de grupos, pode ser utilizado o *bootstrap* para definir um intervalo de confiança.