Análise Multivariada

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

1. Introdução
2. Vetores Aleatórios
3. Normal Multivariada
4. Componentes Principais
5. Análise Fatorial
6. Análise de Agrupamentos
7. Referências

Análise de Agrupamentos

Análise de Agrupamentos

- Procurar por uma estrutura de grupos "naturais" dos dados
  - √ É uma importante técnica exploratória
- Objetivo básico:
  - √ Descobrir agrupamentos naturais dos itens (ou variáveis)
- Mesmo sem noção precisa de um agrupamento natural, em geral, somos capazes de agrupar visualmente objetos em gráficos

- São necessários:
  - √ Medidas de similaridade (ou distância)
  - √ Desenvolvimento de escala quantitativa para medir associação (similaridade) entre os dados
  - √ Algoritmos para ordenar objetos em grupos

Medidas de Similaridade

- Há muita subjetividade na escolha de uma medida de similaridade
- Considerações importantes:
  - √ Natureza das variáveis
    - (discreta, contínua, binária)
  - √ Escala das medidas
    - (nominal, ordinal, intervalar, razão)

- Agrupamentos de itens (unidades ou casos)
  - √ Proximidade é usualmente indicada por algum tipo de distância
- Agrupamento de variáveis:
  - √ Usualmente são agrupadas com base em coeficientes de correlação ou medidas de associação

Distâncias para Pares de Itens

- Sejam as observações:
  - √ **x**' = [x_1, x_2, ..., x_p] e **y**' = [y_1, y_2, ..., y_p]
- Distância Euclidiana:

$$d(\mathrm{x},\mathrm{y}) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2\ldots(x_p-y_p)^2} = \sqrt{(\mathrm{x}-\mathrm{y})'(\mathrm{x}-\mathrm{y})}$$

- Distância generalizada ou ponderada:

$$d(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{x}-\mathrm{y})'\mathrm{A}(\mathrm{x}-\mathrm{y})}$$

  - √ A é matriz de ponderação positiva definida
  - √ **A** = $\mathbf{S}^{-1}$ (distância de Mahalanobis)
    - Não podem ser calculadas sem conhecimento prévio dos grupos
  - √ Se A = I (distância Euclidiana)
  - √ Se A = diagonal(1/*p*) (distância Euclidiana média)

- Métrica de Minkowski:

$$d(\mathrm{x},\mathrm{y}) = \left[\sum_{i=1}^{p} w_i |x_i - y_i|^m\right]^{1/m}$$

  - √ w_i: peso de ponderação para as variáveis
  - √ m = 1, d(**x**, **y**) mede distância "city block" ou Manhatan
  - √ m = 2, d(**x**, **y**) é a distância Euclidiana
  - √ variar m muda a ponderação dada a diferenças maiores ou menores
  - √ A métrica de Minkowski é menos afetada pela presença de valores discrepantes na amostra do que a distância Euclidiana.

Métricas para Variáveis Não-Negativas

- Métrica de Camberra:

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

- Métrica de Czekanowski:

$$d(\mathrm{x},\mathrm{y}) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} x_i + y_i}$$

Distância

- Qualquer medida de distância d(P, Q) entre dois pontos P e Q é válida, desde que satisfaça as seguintes propriedades. R é um ponto intermediário:
  - √ d(P, Q) = d(Q, P)
  - √ d(P, Q) > 0 se P ≠ Q
  - √ d(P, Q) = 0 se P = Q
  - √ d(P, Q) ≤ d(P, R) + d(R, Q) – desigualdade triangular

- Itens representados por medidas qualitativas
  - √ os pares de itens são frequentemente comparados com base na presença ou ausência de certas características
  - √ Itens similares têm mais características comuns que os itens dissimilares
  - √ Presença ou ausência de característica é descrita por variável indicadora (binária):

	X_1	X_2	X_3	X_4	X_5
Item i	1	0	0	1	1
Item j	1	1	0	1	0

- Para $j = 1, 2, ..., p$, sejam:

  x_{ij}: escore da j-ésima variável do i-ésimo item

  x_{kj}: escore da j-ésima variável do k-ésimo item

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{se } x_{ij} = x_{kj} = 1 \text{ ou } x_{ij} = x_{kj} = 0 \\ 1 & \text{se } x_{ij} \neq x_{kj} \end{cases}$$

  - √ A distância Euclidiana $\sum_{i=1}^{p}(x_{ij} - x_{kj})^2$ é a contagem das discordâncias
  - √ Grandes distâncias correspondem a muitas discordâncias
  - √ Essa medida de similaridade pondera igualmente concordâncias e discordâncias

- No exemplo:

	X_1	X_2	X_3	X_4	X_5
Item i	1	0	0	1	1
Item j	1	1	0	1	0

$$\sum_{i=1}^{p}(x_{ij} - x_{kj})^2 = (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 = 2$$

- Muitas vezes uma concordância 1–1 é uma indicação mais forte de similaridade que uma concordância 0–0

Coeficientes de Similaridade

- Há vários esquemas para definir coeficientes de similaridade:
- Seja a tabela de contingência abaixo:

		Item k		
		1	0	Total
Item i	1	a	b	a + b
	0	c	d	c + d
	Total	a + c	b + d	p = a + b + c + d

Exemplo 12.2

- O significado das palavras muda ao longo da história
  - √ O significado dos números constitui uma exceção
- Uma primeira comparação de línguas poderia ser baseada nos numerais

- Numerais em 11 línguas

English (E)	Norwegian (N)	Danish (D)	Dutch (Du)	German (G)	French (F)	Spanish (S)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (F)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
tem	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

√ Comparação das línguas pela 1ª. letra dos números

- Números concordantes: tem a mesma 1ª. letra
- Números discordantes: caso contrário

	E	N	D	Du	G	F	S	I	P	H	F
E	10										
N	8	10									
D	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
F	4	4	4	1	3	10					
S	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
F	1	1	1	1	1	1	1	1	1	2	10

√ Inglês e Norueguês – 1ª.s letras: 8 em 10

√ Inglês, norueguês, dinamarquês, holandês e alemão

– aparentam formar um grupo

√ Francês, espanhol, italiano e polonês

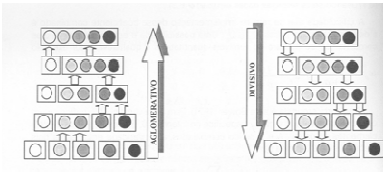
– podem ser agrupados

√ Húngaro e filandês parecem estar sozinhos

Métodos de Agrupamentos Hierárquicos

- Raramente podemos examinar todas as possibilidades de agrupamentos
  - √ Há algoritmos de agrupamento que não têm de verificar todas as configurações
- Técnicas de Agrupamento Hierárquicas
  - √ Procedimentos que realizam uma série de sucessivas fusões (ou uma série de sucessivas divisões)

- Técnicas Hierárquicas:
  - √ Aglomerativas
  - √ Divisivas



√ Em geral, são usadas em análises exploratórias dos dados com o objetivo de:

– identificar possíveis agrupamentos

– estimar o valor provável do número de grupos g

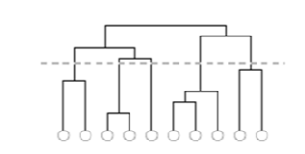
- Técnicas Não-Hierárquicas:
  - √ É necessário que o valor do número de grupos já esteja pré-especificado pelo pesquisador

Métodos Hierárquicos Aglomerativos

1. Cada elemento constitui um cluster de tamanho 1
   - √ Há *n* clusters
2. Em cada estágio do algoritmo os pares de conglomerados mais similares são combinados (novo conglomerado)
   - √ Em cada estágio do processo, o número de conglomerados vai sendo diminuído

3. Propriedade de Hierarquia:
   - √ Em cada estágio do algoritmo, cada novo conglomerado formado é um agrupamento de conglomerados formados nos estágios anteriores
     - Se 2 elementos aparecem juntos em algum estágio do processo, eles aparecerão juntos em todos os estágios subsequentes
   - √ Uma vez unidos, estes elementos não poderão ser separados

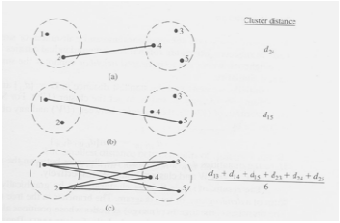
4. Dendograma (ou Dendrograma):



√ Representa a árvore (ou história) do agrupamento
- Escala Vertical: nível de similaridade (ou dissimilaridade)
- Eixo Horizontal: elementos amostrais na ordem relacionada à história do agrupamento

Métodos de Agrupamentos

- Medida de similaridade (ou distância) entre 2 conglomerados



- Método de Ligação Simples (Single Linkage):

√ Similaridade entre dois conglomerados é definida pelos dois elementos mais parecidos entre si
- distância mínima ou vizinho mais próximo

$C_1 = \{X_1, X_2\}$ e $C_2 = \{X_3, X_4, X_5\}$

$$d(C_1, C_2) = \min\{d(X_j, X_k)\}, \quad j \neq k, \quad j = 1,2 \text{ e } k = 3,4,5$$

√ Em cada estágio do processo de agrupamento os dois conglomerados que são mais similares (mais próximos) são combinados em um único *cluster*.

Exemplo 12.4

- Matriz de Distâncias (**D**):

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

√ $\min\{d_{ik}\} = d(5,3) = 2$

√ cluster (35)

$d(35, 1) = \min\{d(3, 1), d(5, 1)\} = \min\{3, 11\} = 3$

$d(35, 2) = \min\{d(3, 2), d(5, 2)\} = \min\{7, 10\} = 7$

$d(35, 4) = \min\{d(3, 4), d(5, 4)\} = \min\{9, 8\} = 8$

√ cluster (135)

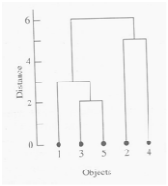
$d(135, 2) = \min\{d(35, 2), d(1, 2)\} = \min\{7, 9\} = 7$

$d(135, 4) = \min\{d(35, 4), d(1, 4)\} = \min\{8, 6\} = 6$

√ cluster (1354)

$d(1354,2) = d(135, 2) = 7$

- Dendograma:



√ Os resultados intermediários são o principal interesse

Exemplo 12.5

- Numerais em 11 línguas (continuação 12.2)

	E	N	D	Du	G	F	S	I	P	H	F
E	0										
N	2	0									
D	2	1	0								
Du	7	5	6	0							
G	6	4	5	5	0						
F	6	6	6	9	7	0					
S	6	6	5	9	7	2	0				
I	6	6	5	9	7	1	1	0			
P	7	7	6	10	8	3	3	4	0		
H	9	8	8	8	9	10	10	10	10	0	
F	9	9	9	9	9	9	9	9	9	8	0

√ Menores distâncias entre pares de línguas:

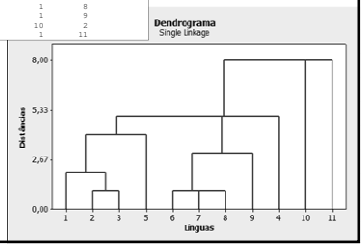
- d(D,N)=1; d(I,F) = 1; d(I,S)=1
- Como d(F,S) = 2, podemos fundir apenas IF ou IS

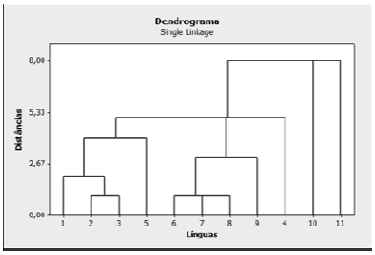
- Análise de Aglomerado – Ligação Simples

Cluster Analysis of Observations: M1

Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined	New cluster	Number of obs. in new cluster
1	10	90	1	7 8	7	2
2	9	90	1	6 7	6	3
3	8	90	1	2 3	2	2
4	7	80	2	1 2	1	3
5	6	70	3	6 9	6	4
6	5	60	4	1 5	1	4
7	4	50	5	1 6	1	8
8	3	50	5	1 4	1	9
9	2	20	8	10 11	10	2
10	1	20	8	1 10	1	11





√ Norueguês(2) + dinamarquês (3); francês(6) + espanhol(7) + Italiano(8) aglomeram-se na mínima distância

√ No próximo passo o inglês (1) é adicionado ao grupo ND

√ Húngaro (10) e finlandês (11) são mais similares um com o outro que com outros clusters

- Método de Ligação Completa (Complete Linkage):
  - √ Similaridade entre dois conglomerados é definida pelos dois elementos menos parecidos entre si
    - distância máxima ou vizinho mais distante

$C_1 = \{X_1, X_2\}$ e $C_2 = \{X_3, X_4, X_5\}$

$$d(C_1, C_2) = \max\{d(X_j, X_k)\}, \quad j \neq k, \quad j = 1, 2 \text{ e } k = 3, 4, 5$$

√ Em cada estágio, a distância (similaridade) entre os *clusters* é determinada pela distância (similaridade) entre os dois elementos, em cada *cluster*, que são mais distantes.

- Garante que todos os itens em cada cluster estão com a máxima distância (mínima similaridade) entre eles.

- Método da Média das Distâncias (AverageLinkage):
  - √ Similaridade entre dois conglomerados é definida pela distância média de todos os pares de itens
    - cada membro do par pertence a grupos diferentes

$C_1 = \{X_1, X_2\}$ e $C_2 = \{X_3, X_4, X_5\}$

$$d(C_1, C_2) = \sum_{j \in C_1} \sum_{k \in C_2} \frac{n_1 n_2}{n_1 + n_2} d(X_j, X_k)$$

- n_1: quantidade de elementos do *cluster* 1
- n_2: quantidade de elementos do *cluster* 2.

- Podem ser usadas distâncias ou similaridades
- Pode ser usado para agrupar variáveis e itens
- Mudanças na atribuição de distâncias (similaridade) podem afetar o arranjo da configuração final de *clusters*, mesmo que as alterações preservem as ordenações relativas.

- Método do Centróide:
  - √ Distância entre dois grupos é definida como sendo a distância entre os vetores de médias (centróides)
    - cada membro do par pertence a grupos diferentes

$C_1 = \{X_1, X_2\}$ e $C_2 = \{X_3, X_4, X_5\}$

$$d(C_1, C_2) = \left[(\bar{\mathbf{C}}_1 - \bar{\mathbf{C}}_2)'(\bar{\mathbf{C}}_1 - \bar{\mathbf{C}}_2)\right]^{1/2}$$

Distância Euclidiana entre os dois grupos

$$\bar{\mathbf{C}}_1 = \frac{1}{2}(X_1 + X_2)$$

$$\bar{\mathbf{C}}_2 = \frac{1}{3}(X_3 + X_4 + X_5)$$

- É método direto e simples, mas em cada passo é necessário retornar aos dados originais para o cálculo da matriz de distâncias
  - √ exige mais tempo computacional
- Não pode ser usado em situações em que se dispões apenas da matriz de distâncias entre os *n* elementos amostrais
  - √ Ao contrário dos métodos simple, complete e average linkage
- Quanto maior a quantidade de elementos amostrais (*n*) e de variáveis (*p*), menor a chance de empates entre valores da matriz de distâncias

Exemplo

- Dados 6 indivíduos de uma comunidade:
  - √ Renda (em salários mínimos)
  - √ Idade
  - √ Dados: (Fonte: Mingoti, 2005)

Indivíduo	Renda	Idade
A	9,60	28
B	8,40	31
C	2,40	42
D	18,20	38
E	3,90	25
F	6,40	41

  - √ Agrupamento pelo método do centróide

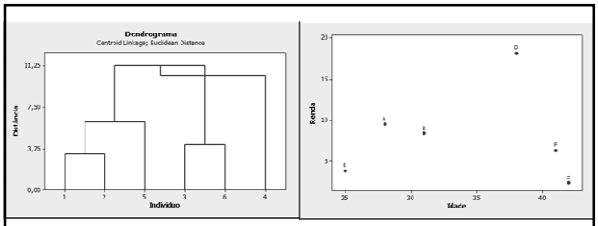
- Matriz de distâncias Euclidianas:

```
Matrix M3

  0,0000   3,2311  15,7429  13,1894   6,4413  13,3881
  3,2311   0,0000  12,5300  12,0433   7,5000  10,1980
 15,7429  12,5300   0,0000  16,2985  17,0660   4,1231
 13,1894  12,0433  16,2985   0,0000  19,3259  12,1754
  6,4413   7,5000  17,0660  19,3259   0,0000  16,1941
 13,3881  10,1980   4,1231  12,1754  16,1941   0,0000
```

- Histórico do agrupamento:

Passo	g	Fusão	Distância (nível)
1	5	{A} e {B}	3,23
2	4	{C} e {F}	4,12
3	3	{A,B} e {E}	6,80
4	2	{A,B,E} e {C,F}	13,81
5	1	{A,B,E,C,F} e {D}	12,91



√ o nível de fusão do passo 5 foi menor que do passo 4

√ Isso poderá ocorrer no método do centróide quando, em algum passo do algoritmo, houver <u>empates</u> entre valores da matriz de distâncias *D*

√ Quanto maior for o tamanho amostral e de variáveis, menor será a chance de ocorrência desta situação

Método de *Ward*

- Objetivo do procedimento:
  - √ Minimizar a perda de informação ao juntar 2 grupos
- Partição desejada:
  - √ A que produz os grupos mais heterogêneos entre si, com elementos homogêneos dentro de cada grupo
- Fundamento do método:
  - √ Em cada passo do agrupamento há mudança de variação entre os grupos e dentro dos grupos
  - √ Procedimento também denominado de mínima variância

- Métodos anteriores:
  - √ quando se passa de ($n - k$) para ($n - k - 1$) grupos o nível de fusão aumenta (nível de similaridade decresce) e a qualidade da partição decresce.
  - √ Variação entre grupos diminui e a variação dentro dos grupos a

Procedimento

1. Cada elemento é considerado um único *cluster;*
2. Em cada passo calcula-se a soma da distância Euclidiana dentro dos grupos:

$$\mathrm{SSR} = \sum_{i=1}^{g_k} \mathrm{SS}_i$$

- √ SSR: soma dos quadrados total (dentro) dos grupos
- √ g_k: número de grupos no passo k
- √ $\mathrm{SS_i}$: soma dos quadrados do cluster i

$$\mathrm{SS}_i = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)'(X_{ij} - \bar{X}_i)$$

- √ $\mathrm{SS_i}$: soma dos quadrados do cluster i
- √ n_i: quantidade de elementos do *cluster Ci* (passo k)
- √ X_{ij}: vetor de observações do j-ésimo elemento amostral que pertence ao i-ésimo conglomerado
- √ $\bar{X}_i$: centróide do *cluster i*

3. Em cada passo do algoritmo, combinam-se os dois conglomerados que minimizam a distância entre os conglomerados C_i e C_j, definida como:

$$d(C_i, C_j) = \left[\frac{n_i n_j}{n_i + n_j}\right](\bar{X}_i - \bar{X}_j)'(\bar{X}_i - \bar{X}_j)$$

√ d(C_i, C_j) é a soma de quadrados entre os clusters C_i e C_j

- Comentários:
  - √ Em cada passo, o método combina os dois conglomerados que resultam no menor valor de SSR
  - √ Prova-se que d(C_i, C_j) é a diferença entre o valor de SSR depois e antes de se combinar os clusters C_i e C_j.
  - √ Os métodos de Ward e do centróide usam o vetor de médias amostrais como representantes da informação global dos conglomerados em cada passo do processo de agrupamento
  - √ A distância considera a diferença dos tamanhos dos conglomerados na comparação
    - $n_i n_j/(n_i + n_j)$ penalizam as comparações (maiores grupos ➔ maiores distâncias

- O método do centróide não tem qualquer ponderação em relação ao tamanho dos *clusters*
- Para usar o método de Ward basta que as variáveis sejam quantitativas
  - √ Para o cálculo do vetor de médias
  - √ Não depende de se conhecer a distribuição da populacão
- Sob certas condições, há uma relação entre o método de Ward e o método de máxima verossimilhança
  - √ Se a distribuição das variáveis for normal p-variada

- O método de Ward baseia-se na noção de que espera-se que os *clusters* de observações multivariadas tenham forma aproximadamente elíptica
- É um precursor de métodos de aglomeração não-hierárquicos
  - √ Otimizam algum critério para dividir os dados em um número determinado de grupos elípticos

Exemplo

- Dados 6 indivíduos de uma comunidade:
  - √ Renda (em salários mínimos)
  - √ Idade
  - √ Dados: (Fonte: Mingoti, 2005)

Indivíduo	Renda	Idade
A	9,60	28
B	8,40	31
C	2,40	42
D	18,20	38
E	3,90	25
F	6,40	41

  - √ Agrupamento pelo método de Ward

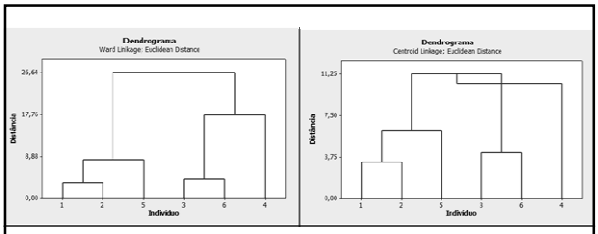
- Matriz de distâncias Euclidianas:

Matrix M3

0,0000	3,2311	15,7429	13,1894	6,4413	13,3881
3,2311	0,0000	12,5300	12,0433	7,5000	10,1980
15,7429	12,5300	0,0000	16,2985	17,0660	4,1231
13,1894	12,0433	16,2985	0,0000	19,3259	12,1754
6,4413	7,5000	17,0660	19,3259	0,0000	16,1941
13,3881	10,1980	4,1231	12,1754	16,1941	0,0000

- Histórico do agrupamento:

Passo	g	Fusão	Distância (nível)
1	5	{A} e {B}	3,23
2	4	{C} e {F}	4,12
3	3	{A,B} e {E}	8,21
4	2	{C,F} e {D}	17,61
5	1	{A,B,E} e {C,F,D}	26,64

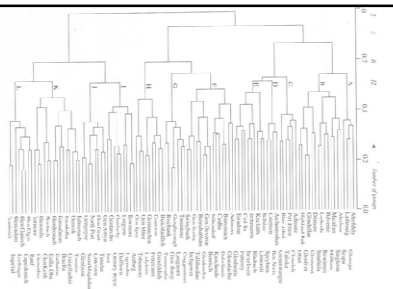


√ Os grupos finais foram os mesmos obtidos com o método do centróide

√ Não houve inversão

Exemplo 12.11 – *Pure Malt*

- Agrupamento de 109 marcas de *scotch* de diferentes destilarias
- 68 variáveis binárias para medir as características do whiskey
  - √ relacionadas com: cor, corpo,aroma, etc.
- Objetivos:
  - √ Determinar os principais tipos de whiskies
  - √ Determinas suas principais características
  - √ Saber se os grupos correspondem a diferentes regiões
    - – são afetados por solo, temperatura, condições da água
- Variáveis binárias são escaladas



√ Identificados 12 grupos de scotchs (A a L)

√ Os grupos têm uma importante componente geográfica

– Apóia a hipótese de que os whiskies são afetados por fatores geográficos (água, solo, microclima, temperatura, etc)não apenas pela extertise das destilarias.

Métodos Hierárquicos – Comentários Finais

- Fontes de erros e de variação não são formalmente considerados nos procedimentos hierárquicos
  - √ Significa que esses métodos são sensíveis a *outliers* ou pontos de perturbação
- Deve-se sempre verificar a sensibilidade da configuração dos grupos
  - √ Os métodos não permitem a realocação de objetos que possam ter sido agrupados incorretamente nos estágios iniciais

- É recomendado tentar vários métodos de agrupamento e de atribuição de distâncias (similaridades)
- Empates na matriz de distâncias podem produzir múltiplas soluções ao problema de agrupamento hierárquico
- A maioria dos métodos produz clusters esféricos ou elípticos

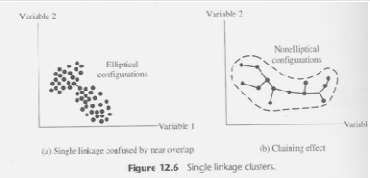


Figure 12.6 Single linkage clusters.

- O método de ligação simples é um dos poucos métodos que pode delinear cluster não-elípticos
  - √ Tem a capacidade de gerar estruturas geométricas diferentes
  - √ Tem a tendência de formar strings longas (chaining)
  - √ Entretanto, ele é incapaz de perceber grupos pouco separados

- Os clusters formados pelo método de ligação simples não serão modificados por qualquer atribuição de distância (similaridade) que dá as mesmas ordenações relativas
  - √ Em particular, qualquer um dos coeficientes de similaridade monotônicos (Tabela 12.2)
- O método de ligação completa tende a produzir conglomerados de aproximadamente mesmo diâmetro
  - √ Tem a tendência de isolar os valores discrepantes nos estágios iniciais do agrupamento

- O método da média das distâncias tende a produzir conglomerados de aproximadamente mesma variância interna
  - √ Em geral, produz melhores partições que os métodos de ligação simples e completa
- Os métodos de ligação simples, completa e da média podem ser utilizados tanto para variáveis quantitativas quanto para variáveis qualitativas
- Os métodos do centróide e de Ward são apropriados apenas para variáveis quantitativas

- O método de Ward tende a produzir grupos com aproximadamente o mesmo número de elementos e tem como base principal os princípios de análise de variância
- Com um número maior de dados amostrais (n) ou de variáveis (p), necessariamento não irá ocorrer a igualdade das soluções apresentadas pelos vários métodos
  - √ Espera-se sempre que haja uma certa consistência entre as soluções obtidas por métodos diferentes

Métodos para Encontrar o Número *g* de Clusters da Partição Final

- Problema de agrupamento:
  - √ Como escolher o número final (g) de grupos que define a partição do conjunto de dados?
  - √ Qual o passo k em que o algoritmo de agrupamento deve ser interrompido?

- Critério 1 – Análise do comportamento do nível de fusão (distância)
  - √ À medida que o algoritmo avança, a similaridade dos grupos diminui (distância aumenta)
  - √ Gráfico do passo (ou número de grupos) *vs.* nível de distância (nível de fusão)
    - – Verifica-se a existência de "saltos" relativamente grandes
      - ▪ Pontos indicadores do momento ideal de parada (número final de conglomerados)
      - ▪ Se observados vários pontos de "saltos" pode-se delimitar uma região de prováveis valores do número de grupos g (deve ser investigado por outro procedimento)
  - √ Pode-se usar o dendograma quando n não for muito grande

- Critério 2 – Análise do comportamento do nível de similaridade
  - √ Similar ao critério 1
    - – Observa-se o nível de similaridade (ao invés da distância)
  - √ Nível de similaridade:

$$S_{ij} = \left(1 - \frac{d_{ij}}{\max(d_{rs}),\ r, s = 1, \ldots, n)}\right) \times 100\%$$

  - – $\max(d_{rs})$: maior distância entre os n elementos amostrais na matriz de distâncias D_{nxn} do início do processamento

√ Procura-se detectar pontos em que haja um decrescimento acentuado na similaridade dos conglomerados unidos
  - indicam a interrupção do algoritmo de agrupamento
  - número final de *clusters* (g) está relacionado com o estágio em que o algoritmo foi interrompido

√ Em geral, a escolha de valores de similaridade acima de 90% leva a um número de grupos muito elevado

- Critério 3 – Análise da soma dos quadrados entre grupos: R^2

  √ É possível calcular a soma de quadrados **entre** *clusters* e **dentro** dos grupos, em cada passo do procedimento

  √ Em partição com g* grupos, sejam:
  - $\mathbf{X}'_{ij} = (X_{i1,j}, X_{i2,j}, ..., X_{ip,j})$
  - vetor de medidas observadas para o j-ésimo elemento amostral do i-ésimo grupo
  - $\overline{\mathbf{X}}_{i\cdot}' = (\overline{X}_{i1\cdot}, \overline{X}_{i2\cdot}, ..., \overline{X}_{ip\cdot})$
  - vetor de médias do i-ésimo grupo (sem considerar partição)
  - $\overline{\mathbf{X}}' = (\overline{X}_{\cdot 1}, \overline{X}_{\cdot 2}, ..., \overline{X}_{\cdot p})$

$$\bar{X}_{\cdot r} = \frac{1}{n}\sum_{i=1}^{g^*}\sum_{j=1}^{n_i} X_{i_r j},\ r = 1, 2 \ldots, p$$

√ Soma dos quadrados total corrigida para a média global em cada variável

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

√ Soma dos quadrados total dentro dos grupos da partição

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

√ Soma dos quadrados total entre os *g** grupos

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

√ Coeficiente R^2 da partição: $R^2 = \frac{\mathrm{SSB}}{\mathrm{SST}_C}$

√ Quanto maior o valor de R^2, maior será a soma de quadrados entre grupos e menor será a soma de quadrados residual SSR

√ Procedimento para escolha de *g*

- Gráfico do passo do agrupamento *vs.* R^2
- Procurar identificar algum ponto de 'salto' relativamente grande em relação aos demais
  - indica momento ideal da parada
- Gráfico é sempre decrescente
- maior valor de g*, menor a variabilidade interna dos grupos e maior será o valor de R^2
  - máximo $R^2 = 1$ (para g* = n)

- Estratégia:
  - √ Definir uma região de valores plausíveis para o número de grupos *g*
  - √ Utilizar o critério 3 dentro da região estabelecida

- <u>Critério 4</u> – Estatística Pseudo F
  - √ (Calniski e Harabasz, 1974)
  - √ Cálcular estatística *F* em cada passo do agrupamento

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

g*: número de grupos da partição em análise

  - √ Idéia do teste:
    - Em cada passo do agrupamento estaria sendo feito um teste F de análise de variância
  - √ Importante:
    - Na prática, não ocorre alocação aleatória
    - A maioria dos métodos usa métodos de agrupamento com critérios determinísticos para partição dos dados

√ Se os elementos amostrais são provenientes de uma distribuição normal p-variada e quando os elementos são alocados aleatoriamente nos grupos

√ $F \sim F_{p(g^*-1),\, p(n-g^*)}$

√ Se *F* é monotonicamente crescente com g*, os dados sugerem que não há qualquer estrutura 'natural' de partição dos dados

√ Se *F* apresentar um valor máximo, o número de conglomerados corresponderá à partição 'ideal'

√ Busca-se o maior valor de *F*

- Busca-se partição com maior heterogeneidade dos grupos
- valor relacionado com a menor probabilidade de significância do teste
- Estaria rejeitando a igualdade de vetores de médias populacionais com maior significância

- <u>Critério 5</u> – Correlação Semiparcial (Método de Ward)

√ Em determinado passo, $C_k = C_i \cup C_j$

$$\mathrm{SPR}^2 = \frac{B_{ij}}{\mathrm{SST}_C}$$

Coeficiente de correlação parcial da partição

$$B_{ij} = \frac{n_i n_j}{n_i + n_j}(\bar{\mathbf{X}}_{i\cdot} - \bar{\mathbf{X}}_{j\cdot})'(\bar{\mathbf{X}}_{i\cdot} - \bar{\mathbf{X}}_{j\cdot})$$

Distância entre grupos – Método de Ward

1. Calcula-se SPR^2 em cada passo
2. Gráfico passo *vs.* SPR^2
3. Busca-se no gráfico salto consideravelmente maior que os restantes
4. Ponto indica partição ideal
   (parada do algoritmo de agrupamento)

√ A função SPR2 é não decrescente

√ Se o agrupamento dos dados foi feito pelo método de Ward, o critério do coeficiente de correlação semiparcial equivalerá à aplicação do critério 1.

- Critério 6 – Estatística Pseudo T^2

√ Em determinado passo, $C_k = C_i \cup C_j$

$$T^2 = \frac{B_{ij}}{\left[\sum_{r\in C_i} \|\mathbf{X}_{ir} - \bar{\mathbf{X}}_{i\cdot}\|^2 + \sum_{r\in C_j} \|\mathbf{X}_{jr} - \bar{\mathbf{X}}_{j\cdot}\|^2\right](n_i + n_j)^{-1}}$$

$$\|\mathbf{X}_{sr} - \bar{\mathbf{X}}_{s\cdot}\| = (\mathbf{X}_{sr} - \bar{\mathbf{X}}_{s\cdot})'(\mathbf{X}_{sr} - \bar{\mathbf{X}}_{s\cdot}),\ s = i, j$$

√ Sob as suposições de normalidade p-variada e alocação aleatória dos grupos

$T^2 \sim F_{p,\,(ni+nj-2)}$

√ Na prática, não se tem alocação aleatória dos grupos

√ Ideia do teste:

– Teste de comparação de média de dois grupos, unidos para formar novo grupo

√ Gráfico passo *vs.* valor da Pseudo T^2

– Busca-se no gráfico o valor máximo

√ O valor de *g* correspondente ao máximo (ou aquele imediatamente anterior) é escolhido como o número provável de grupos da partição final

√ Busca-se o maior valor de T^2

– aquele relacionado com a menor probabilidade de significância

(Rejeita a igualdade dos vetores de média com maior significância)

– Se a igualdade entre os vetores de médias é rejeitada, os dois clusters deveriam ser unidos para formar um único agrupamento

- Critério 6 – Estatística CCC (*Cubic Clustering Criterium*)
  - √ Sarle (1983)
  - √ Obtida comparando-se o valor esperado do coeficiente R2 com a aproximação do valor esperado de r2 sob a suposição de que os grupos são gerados de acordo com uma uniforme p-dimensional
  - √ *CCC* indicaria a presença de estrutura de agrupamento diferente da partição uniforme
  - √ A quantidade de grupos da partição final estaria relacionada com valores de $CCC > 3$
  - √ Está implementada no software estatístico SAS

Exemplo 6.8

Mingoti, 2005

- Dados relativos a 21 países (ONU, 2002)
- Variáveis:
  - √ Expectativa de vida
  - √ Educação
  - √ Renda (PIB)
  - √ Estabilidade política e de sergurança
- Método de agrupamento: Ward
- Conjunto de dados: *BD_multivariada.xls/paises*

- Minitab

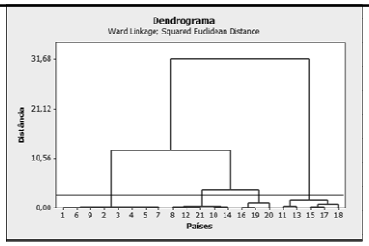
Cluster Analysis of Observations: Índice de Ex; Índice de Ed; Índice PIB; ...

Squared Euclidean Distance, Ward Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	20	99,995	0,0006	2	3	2	2
2	19	99,966	0,0042	1	6	1	2
3	18	99,934	0,0081	4	5	4	2
4	17	99,923	0,0095	2	4	2	4
5	16	99,821	0,0219	12	21	12	2
6	15	99,613	0,0473	1	9	1	3
7	14	99,510	0,0598	16	19	16	2
8	13	99,462	0,0657	8	12	8	3
9	12	99,377	0,0761	2	7	2	5
10	11	98,999	0,1221	10	14	10	2
11	10	98,963	0,1266	15	17	15	2
12	9	98,622	0,1681	1	2	1	8
13	8	97,997	0,2445	11	13	11	2
14	7	97,535	0,3008	8	10	8	5
15	6	94,602	0,6587	15	18	15	3
16	5	92,489	0,9166	16	20	16	3
17	4	88,114	1,4505	11	15	11	5
18	3	71,202	3,5145	8	16	8	8
19	2	1,220	12,0549	1	8	1	16
20	1	-159,594	31,6803	1	11	1	21

Final Partition
Number of clusters: 1

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	21	25,7654	0,987565	2,2069



- Visualmente, é razoável definir 4 grupos de países
  - √ Analisar queda de similaridade entre os passos 16 e 18

- Resultados da análise de agrupamento:

Passo	g*	Similaridade	Distância	R^2	Pseudo F	SP R^2	Pseudo T^2	CCC
1	20	99,99	0,001	1,000	4520,0	0,0000		
2	19	99,97	0,004	1,000	1193,0	0,0001		
3	18	99,93	0,008	1,000	705,0	0,0002		
4	17	99,92	0,009	1,000	576,0	0,0002	2,2	
5	16	99,82	0,022	0,999	388,0	0,0004		
6	15	99,61	0,047	0,998	241,0	0,0009	11,3	
7	14	99,51	0,060	0,997	183,0	0,0012		
8	13	99,46	0,066	0,996	158,0	0,0013	3,0	
9	12	99,38	0,760	0,994	143,0	0,0015	12,6	
10	11	99,00	0,122	0,992	123,0	0,0024		
11	10	98,96	0,127	0,989	115,0	0,0025		
12	9	98,62	0,168	0,986	107,0	0,0033	6,9	
13	8	98,00	0,245	0,981	98,4	0,0047		
14	7	97,54	0,301	0,976	93,5	0,0058	4,3	
15	6	94,60	0,659	0,963	77,8	0,0128	5,2	
16	5	92,49	0,917	0,945	68,8	0,0178	15,3	
17	4	88,11	1,450	0,917	62,5	0,0281	4,2	-0,12
18	3	71,20	3,514	0,849	50,5	0,0682	14,2	-0,65
19	2	1,22	12,055	0,615	30,3	0,2339	31,8	-1,80
20	1	-159,59	31,680	0,000		0,6148	30,3	

  - √ Do passo 17 para 18:
    - Perda mais acentuada de similaridade
    - O valor de R^2 passa de 0,917 para 0,849
    - Valores da Pseudo F e do CCC decrescem substancialmente
    - Pseudo T^2 e SPR^2 crescem acentuadamente

- Medidas descritivas dos grupos formados:

Grupos (SQ)	Países	Média			
		Expectativa de vida	Educação	PIB	Estabilidade política
1 (0,157) $n_1 = 8$	Austrália, Canadá, Cingapura, Estados Unidos, França, Japão Reino Unido, Uruguai	0,8838	0,9538	0,9075	1,1850
2 (0,255) $n_2 = 5$	Argentina, Brasil, China, Cuba, Egito	0,7660	0,8140	0,6740	0,3380
3 (1,240) $n_3 = 5$	Angola, Colômbia, Nigéria, Paraguai, Serra Leoa	0,5060	0,5900	0,4940	-1,3660
4 (0,488) $n_4 = 3$	Etiópia, Mocambique, Senegal	0,3400	0,3633	0,3767	-0,3433
Global n = 21	Todos	0,6881	0,7495	0,6776	0,1580

  - √ Grupo 1 – 'Primeiro Mundo'
    - países com maiores índice de desenvolvimento
  - √ Grupo 4 – alguns países africanos
    - menores índices em todas as variáveis

√ Variável estabilidade política e segurança:
- Grande diferença de comportamento dos grupos 1 e 2 em relação aos grupos 3 e 4
- Grupo 1 é o de maior estabilidade e o grupo 3 de menor

√ Dispersão interna é menor no grupo 1 e maior no grupo 3

Técnicas Hierárquica e Seleção de Variáveis

- Os métodos hierárquicos podem ser úteis na seleção das variáveis mais importantes na caracterização de determinada situação
- Métodos de ligação simples, completa e da média
  - √ É necessária apenas matriz inicial que represente proximidade (ou similaridade) entre os elementos amostrais
  - √ É necessário escolher uma matriz inicial que represente o relacionamento dessas variáveis
    - Interesse: agrupar as variáveis mais similares entre si (separar aquelas com informações diferenciadas)

Variáveis quantitativas:

- Pode-se usar coeficiente de correlação de Pearson
  - √ Expressa similaridade com relação à associação linear
  - √ Quanto maior seu valor absoluto, maior a aproximação entre as variáveis
- Matriz de correlação amostral não é uma matriz de distâncias (ou proximidades)
  - √ Transformação mais simples

    $D_{pxp} = 1 - Abs(R_{pxp})$
- Podem ser usados coeficientes de correlação não paramétricos
  - √ Spearman, Kendall, etc.

Exemplo

- Matriz de correlação amostral (R):

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1					
X_2	0,57	1				
X_3	0,51	0,60	1			
X_4	0,39	0,38	0,43	1		
X_5	0,46	0,32	0,40	0,50	1	
X_6	0,35	0,72	0,45	0,58	0,30	1

√ X_2 e X_6 são mais similares ($r_{26} = 0,72$)

- $D_{6x6} = 1 - Abs(R_{6x6})$

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0					
X_2	0,43	0				
X_3	0,49	0,40	0			
X_4	0,61	0,62	0,57	0		
X_5	0,54	0,68	0,60	0,50	0	
X_6	0,65	0,28	0,55	0,42	0,70	0

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0					
X_2	0,43	0				
X_3	0,49	0,40	0			
X_4	0,61	0,62	0,57	0		
X_5	0,54	0,68	0,60	0,50	0	
X_6	0,65	0,28	0,55	0,42	0,70	0

- Método de Ligação Simples

Passo	g	Fusão	Nível Fusão
1	5	X_2 e X_6	0,28
2	4	X_2, X_6 e X_3	0,40
3	3	X_2, X_6, X_3 e X_4	0,42
4	2	X_2, X_6, X_3, X_4 e X_1	0,43
5	1	X_2, X_6, X_3, X_4, X_1 e X_5	0,50

- No passo 3
  - √ $C_1 = \{X_2, X_6, X_3, X_4\}$
  - √ $C_2 = \{X_1\}$
  - √ $C_3 = \{X_5\}$

Suponha escolher 3 dentre as 6 variáveis:
X_1
X_5
Uma das variáveis de C_1

- Medidas de similaridade para variáveis categóricas:
  - √ Coeficiente qui-quadrado
  - √ Coeficiente de contingência de Pearson
  - √ Coeficiente de concordância de Kappa
- Outros Coeficientes:
  - √ Podem-se desenvolver medidas de associação (similaridade) análogos aos coeficientes estabelecidos anteriormente (Tabela 12.2) – Troca-se *p* por *n*.

- Variáveis Binárias:
  - √ Os dados podem ser agrupados na forma de tabela de contingência
  - √ Para cada par de variáveis, há *n* itens categorizados na tabela

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

- Correlação Momento-Produto

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

  - √ Pode ser tomado como medida de similaridade entre as duas variáveis
  - √ r está relacionado com a estatística χ^2 para teste de independência entre duas variáveis categóricas

$$r^2 = \frac{\chi^2}{n}$$

  - √ Para *n* fixo, uma correlação (similaridade) grande é consistente com a ausência de independência

Comentários

- Há várias maneiras de medir similaridade entre pares de objetos:
  - √ distâncias (12-1 a 12-5)
  - √ Coeficientes (Tabela 12-2) – para agrupar itens
  - √ Correlações – para agrupar variáveis
- Podem ser usadas frequências

Exemplo 12.8

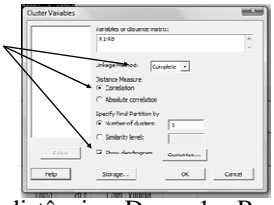
- Agrupamento de variáveis (Ligação Completa)
  - √ Dados de 22 concessionárias públicas (USA)
  - √ Variáveis:
    - X_1: renda/dívidas
    - X_2: taxa de retorno de capitais
    - X_3: custo por capacidade instalada (kW)
    - X_4: fator de carga anual
    - X_5: pico de demanda (crescimento último ano)
    - X_6: Vendas (kWh por ano)
    - X_7: participação nucleares (%)
    - X_8: custo total de combustível ($ por kWh)
  - √ Dados: *BD_multivariada.xls/public_utilities*

- Coeficiente de correlação para medir similaridade
  - √ variáveis com grandes correlações negativas são consideradas muito dissimilares
  - √ variáveis com grandes correlações positivas são consideradas muito similares
  - √ distância entre *clusters* é medida como menor similaridade entre grupos
- Matriz de correlações:

Correlations: X1; X2; X3; X4; X5; X6; X7; X8

	X1	X2	X3	X4	X5	X6	X7
X2	0,643						
X3	-0,103	-0,348					
X4	-0,082	-0,086	0,100				
X5	-0,259	-0,260	0,435	0,033			
X6	-0,152	-0,010	0,028	-0,288	0,176		
X7	0,045	0,211	0,115	-0,164	-0,019	-0,374	
X8	-0,013	-0,328	0,005	0,486	-0,007	-0,561	-0,185

- Minitab **Stat > Multivariate > Cluster Variables ➔**

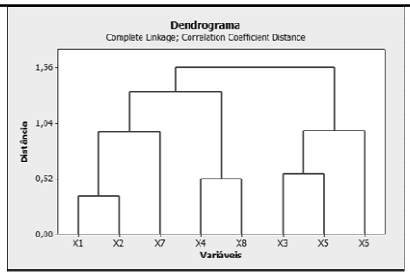


- Matriz de distâncias: $D_{8x8} = 1 - R_{8x8}$

Data Display

Matriz de Distâncias

0,00000	0,35726	1,10279	1,08203	1,25911	1,15167	0,95520	1,01337
0,35726	0,00000	1,34755	1,08634	1,26011	1,00962	0,78856	1,32766
1,10279	1,34755	0,00000	0,89969	0,56463	0,97201	0,88534	0,99478
1,08203	1,08634	0,89969	0,00000	0,96652	1,28794	1,16416	0,51450
1,25911	1,26011	0,56463	0,96652	0,00000	0,82358	1,01913	1,00713
1,15167	1,00962	0,97201	1,28794	0,82358	0,00000	1,37369	1,56053
0,95520	0,78856	0,88534	1,16416	1,01913	1,37369	0,00000	1,18509
1,01337	1,32766	0,99478	0,51450	1,00713	1,56053	1,18509	0,00000



√ Variáveis: (1,2), (4,8), (3,5) aglomeram-se a um nível intermediário

√ Variáveis 7 e 6 permanecem isoladas até os estágios finais

√ Agrupamento final:
(12478) e (356)

Técnicas de Agrupamento Não Hierárquicas

- Objetivo:
  - √ Encontrar diretamente uma partição de *n* elementos em *k* grupos
  - √ Requisitos:
    - coesão interna (semelhança interna)
    - isolamento (separação) dos clusters formados
- Busca da "melhor" partição de ordem *k*
  - √ Satisfaz algum critério de qualidade
  - √ Procedimentos computacionais para investigar partições 'quase' ótima
    (inviável a busca exaustiva)

- Métodos Não Hierárquicos *vs.* Hierárquicos :
  - √ Especificação prévia do número de cluster
  - √ (ao contrário das técnicas aglomerativas)
  - √ Novos grupos podem ser formados pela divisão (ou junção) de grupos já combinados:
    - Se em um passo do algoritmo, dois elementos tiverem sido colocados em um mesmo grupo, não significa que estarão juntos na partição final
    - Não é mais possível a construção de dendogramas
  - √ Em geral, são do tipo iterativo

√ Tem maior capacidade de analisar grande número de dados

√ A matriz de distância não tem de ser calculada e os dados básicos não precisam ser armazenados durante a execução do procedimento

√ Métodos hierárquicos são mais adequados para agrupar itens que variáveis

Métodos Não Hierárquicos – Estrutura

- Iniciam-se:
  1. partição inicial de itens em grupos
  2. conjunto inicial de sementes que formarão o núcleo dos clusters
- Escolha das configurações iniciais pode afetar partição final
  - √ Viés na escolha das sementes iniciais
  - √ Alternativas:
    - Seleção aleatória de sementes
    - Partição aleatória de itens em grupos iniciais

Métodos Não Hierárquicos – Procedimentos

Alguns procedimentos:

- Método das k-Médias (*k-Means*)
  - √ mais conhecido e popular
- Método Fuzzy c-Médias
- Redes Neurais Artificais

Método das *k*-Médias

- Provavelmente, um dos mais conhecidos e mais utilizados
- Idèia Básica:
  - √ Cada elemento amostral é alocado àquele *cluster* cujo centróide é o mais próximo do elemento

Passos do Procedimento

1. Escolhem-se *k* centróides para inicializar o processo de partição
   - √ Sementes ou protótipos
2. Cada elemento do conjunto de dados é comparado com cada centróide inicial
   - √ Alocação ao centróide menos distante
   - √ Em geral, utiliza-se distância Euclidiana
   - √ Aplicação a todos os *n* elementos amostrais

3. Cálculo dos novos centróides para cada grupo formado no passo (2)
   - √ Repetição do passo (2), considerando os novos valores dos centróides
4. Os passos (2) e (3) são repetidos até que todos os elementos amostrais estejam "bem alocados" em seus grupos
   - √ "Bem alocados" = não é necessária realocação de elementos

Exemplo 12.12

- Agrupamento pelo Método das *k*-Médias:
  - √ Medidas das variáveis X_1 e X_2:

	Observações	
Item	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

  - √ Dividir em $k = 2$ grupos de maneira que os itens de um *cluster* sejam os mais próximos um dos outros e que estejam distantes em *clusters* diferentes

- Implementação:
  - √ Partição arbitrária em 2 *clusters*: (AB) e (CD)
  - √ Cálculo das coordenadas $(\bar{x}_1, \bar{x}_2)$ dos centróides:

Cluster	$\bar{x}_1$	$\bar{x}_2$
AB	2	2
CD	–1	–2

  - √ Distância euclidiana de cada item

	AB	CD
A	$d(A,AB) = (5-2)^2 + (3-2)^2 =$ 10	$d(A,CD) = (5+1)^2 + (3+2)^2 =$ 61
B	$d(B,AB) = (-1-2)^2 + (1-2)^2 =$ 10	$d(B,CD) = (-1+1)^2 + (1+2)^2 =$ 9
C	$d(C,AB) = (1-2)^2 + (-2-2)^2 =$ 17	$d(C,CD) = (1+1)^2 + (-2+2)^2 =$ 4
D	$d(D,AB) = (-3-2)^2 + (-2-2)^2 =$ 41	$d(D,CD) = (-3+1)^2 + (-2+2)^2 =$ 4

B é agrupado ao *cluster* (CD)

  - √ Cálculo das coordenadas $(\bar{x}_1, \bar{x}_2)$ dos centróides:

Cluster	$\bar{x}_1$	$\bar{x}_2$
A	5	3
BCD	–1	–1

  - √ Distância euclidiana de cada item

	A	BCD
A	$d(A,A) = (5-5)^2 + (3-3)^2 =$ 0	$d(A,BCD) = (5+1)^2 + (3+1)^2 =$ 52
B	$d(B,A) = (-1-5)^2 + (1-3)^2 =$ 40	$d(B,BCD) = (-1+1)^2 + (1+1)^2 =$ 4
C	$d(C,A) = (1-5)^2 + (-2-3)^2 =$ 41	$d(C,BCD) = (1+1)^2 + (-2+1)^2 =$ 5
D	$d(D,A) = (-3-5)^2 + (-2-3)^2 =$ 89	$d(D,BCD) = (-3+1)^2 + (-2+1)^2 =$ 5

  - √ O agrupamento se mantém e o processo pára

- Agrupamento Final ($k = 2$)
  - √ A e (BCD)
- Comentários:
  - √ Para verificar a estabilidade da solução é recomendável reiniciar o algoritmo com uma nova partição inicial
  - √ Uma tabela de centróides e das variâncias dentro dos grupos auxilia a delinear as diferenças entre os grupos

Sugestões para Escolha Cuidadosa das Sementes

- Sugestão 1: Uso de técnicas hierárquicas aglomerativas:
  - √ Utiliza-se algum método de agrupamento hierárquico para se obter os k grupos iniciais
  - √ Calcula-se o vetor de médias de cada grupo
  - √ Esses vetores são usados como sementes iniciais

- Sugestão 2: Escolha aleatória:
  - √ As k sementes iniciais são escolhidas aleatoriamente dentro do conjunto de dados
  - √ Sugestão amostragem aleatória simples sem reposição
    (estratégica simples, mas sem eficiência)
  - √ Melhoria de eficiência na escolha:
    - Selecionar m amostras aleatórias, constituídas de k sementes
    - Cálculo do vetor de médias das k sementes selecionadas para cada grupo
    - Esses vetores constituem os centróides de inicialização do processo de agrupamento das k-médias

- Sugestão 3: Escolha por meio de uma variável aleatória:
  - √ Escolhe-se uma variável aleatória dentre as *p* componentes em consideração
    - a variável por si só já induz uma certa "partição natural" dos dados
  - √ Divide-se o domínio da variável em *k* intervalos
  - √ A semente inicial será o centróide de cada intervalo

- Sugestão 4: Observação dos valores discrepantes do conjunto de dados
  - √ Análise estatística para buscar *k* elementos discrepantes no conjunto de dados
    - Discrepância em relação às *p* variáveis observadas
  - √ Cada um desses elementos será a semente

- Sugestão 5: Escolha prefixada
  - √ Método não muito recomendável, pois, tem um alto grau de subjetividade
  - √ Sementes escolhidas arbitrariamente
  - √ Pode ser usadas em casos em há grande conhecimento do problema
    - buca-se validar solução já existente

- Sugestão 6: Os k primeiros valores do banco de dados
  - √ Usado como *default* pela maioria dos softwares
  - √ Pode trazer bons resultados quando os k primeiros elementos amostrais são discrepantes entre si
    (Não é recomendável quando são semelhantes)

Exemplo 7.1

Mingoti, 2005 – Continuação Ex. 6.8

- Dados relativos a 21 países (ONU, 2002)
- Variáveis:
  - √ Expectativa de vida
  - √ Educação
  - √ Renda (PIB)
  - √ Estabilidade política e de sergurança
- Método de agrupamento: k-Médias
- Conjunto de dados: *BD_multivariada.xls/paises*

- Utiliza-se da Análise pelo Método de Ward:
  - √ $k = g = 4$ grupos para partição dos países
  - √ Sementes iniciais = centróides *clusters* finais
- Partição final:
  - √ a mesma obtida anteriormente

Grupos (SQ)	Países	Média			
		Expectativa de vida	Educação	PIB	Estabilidade política
1 (0,157) $n_1 = 8$	Austrália, Canadá, Cingapura, Estados Unidos, França, Japão Reino Unido, Uruguai	0,8838	0,9538	0,9075	1,1850
2 (0,255) $n_2 = 5$	Argentina, Brasil, China, Cuba, Egito	0,7660	0,8140	0,6740	0,3380
3 (1,240) $n_3 = 5$	Angola, Colômbia, Nigéria, Paraguai, Serra Leoa	0,5060	0,5900	0,4940	-1,3660
4 (0,488) $n_4 = 3$	Etiópia, Mocambique, Senegal	0,3400	0,3633	0,3767	-0,3433
Global $n = 21$	Todos	0,6881	0,7495	0,6776	0,1580

- Sementes iniciais: Reino Unido, Brasil, Serra Leoa e Moçambique
  - √ Obtém-se mesma partição final

- Sementes iniciais: 4 primeiros países do banco

Grupos (SQ)	Países	Média			
		Expectativa de vida	Educação	PIB	Estabilidade política
1 (0,091) $n_1 = 7$	Austrália, Canadá, Estados Unidos, França, Japão, Reino Unido, Uruguai	0,8843	0,9657	0,9071	1,1529
2 (0,748) $n_2 = 6$	Argentina, Brasil, China, Cuba, Egito, Moçambique	0,6783	0,7400	0,6271	0,3150
3 (2,188) $n_3 = 7$	Angola, Colômbia, Nigéria, Paraguai, Serra Leoa,Etiópia, Senegal	0,4729	0,5243	0,4626	-1,1514
4 (0,488) $n_4 = 1$	Cingapura	0,8800	0,8700	0,9100	1,4100
Global $n = 21$	Todos	0,6881	0,7495	0,6776	0,1580

  - √ Cingapura foi separada do *cluster* 1
  - √ Moçambique deslocado para grupo do Brasil
  - √ Grupo da Colômbia acrescido de Etiópia e Senegal

- √ Soma de quadrados dentro dos grupos:
  - Nova solução aumentou variabilidade dentro dos grupos 2 e 3

- Minitab:

K-means Cluster Analysis: Índice de Ex; Índice de Ed; Índice PIB; Estabilidade

Final Partition

Number of clusters: 4

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	7	2,188	0,528	0,856
Cluster2	1	0,000	0,000	0,000
Cluster3	7	0,091	0,105	0,204
Cluster4	6	0,748	0,308	0,641

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Grand centroid
Índice de Expectativa de vida	0,4729	0,8800	0,8843	0,6783	0,6881
Índice de Educação	0,5243	0,8700	0,9657	0,7400	0,7495
Índice PIB	0,4629	0,9100	0,9071	0,6217	0,6776
Estabilidade_politica_violência	-1,1514	1,4100	1,1529	0,3150	0,1576

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	0,0000	2,6545	2,4231	1,5048
Cluster2	2,6545	0,0000	0,2744	1,1575
Cluster3	2,4231	0,2744	0,0000	0,9364
Cluster4	1,5048	1,1575	0,9364	0,0000

- Soma dos Quadrados:

  √ ANOVA para cada uma das *p* variáveis

One-way ANOVA: Índice de Expectativa de vida versus C35

Source	DF	SS	MS	F	P
C35	3	0,6311	0,2104	6,83	0,003
Error	17	0,5234	0,0308		
Total	20	1,1545			

One-way ANOVA: Índice de Educação versus C35

Source	DF	SS	MS	F	P
C35	3	0,6974	0,2325	7,22	0,002
Error	17	0,5475	0,0322		
Total	20	1,2449			

One-way ANOVA: Índice PIB versus C35

Source	DF	SS	MS	F	P
C35	3	0,7644	0,2548	14,94	0,000
Error	17	0,2900	0,0171		
Total	20	1,0544			

One-way ANOVA: Estabilidade_política_violência versus C35

Source	DF	SS	MS	F	P
C35	3	20,6458	6,8819	70,23	0,000
Error	17	1,6658	0,0980		
Total	20	22,3116			

Soma Quadrados Total Corrigida

SSTc = 25,7654

Soma dos Quadrados Total dentro dos grupos

SSR = 3,0267

- Cálculo R^2:

$$\text{SSB} = \text{SST}_\text{C} - \text{SSR} = 25,7654 - 3,0267 = 22,7387$$

$$R^2 = \frac{\text{SSB}}{\text{SST}_\text{C}} = \frac{22,7387}{25,7654} = 0,883$$

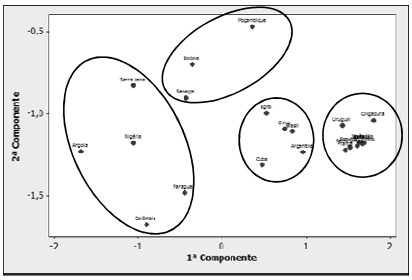
- Comparação da qualidade das partições

Partição	Soma Quadrados Residual	Variação Residual Média	R^2
k-Médias (Ward)	2,1406	0,5351	0,917
k-Médias (*)	3,0267	0,7557	0,883

(*) Os 4 primeiros elementos do banco de dados

√ A solução do método das k-Médias, utilizando como sementes iniciais as sementes de Ward, é melhor do que a solução encontrado quando os 4 primeiros elementos do banco de dados são usados como semente

- Visualização espacial dos grupos:
  - √ 2 primeiras componentes principais com base na matriz de covariâncias amostral



  - √ É possível visualizar claramente os 4 grupos
    - *k*-médias com sementes de Ward

Comentários Finais

- A escolha das sementes iniciais de agrupamento podem influenciar o agrupamento final
  - √ Se duas ou mais sementes situarem-se em um único *cluster*, os grupos resultantes serão pouco diferenciados
  - √ A existência de *outlier* pode produzir pelo menos um grupo com muitos itens dispersos

- Há fortes argumentos para não se fixar o número de *clusters k*
  - √ Mesmo sabendo-se que a população consiste de *k* grupos, dependendo do método de amostragem, pode não aparecer na amostra os dados provenientes de um grupo mais raro
    - Forçar *k* grupos levaria a *clusters* sem sentido
  - √ Em casos em que o algoritmo requer o uso de um valor especificado de *k*, é sempre uma boa idéia executar novamente o algoritmo para diversas escolhas de *k*

Referências

Bibliografia Recomendada

- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1998
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- LATTIN, J.; CARROLL, J. D.; GREEN, P. E. *Análise de Dados Multivariados*. Cengage, 2011.