UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
DEPARTAMENTO DE ADMINISTRAÇÃO

RAD5053 – Análise de Dados Multivariados I

Prof. Dr. Evandro Marcos Saidel Ribeiro **Lista 05 – Análise de Agrupamento**

PARTE 1: Exercícios do CAPÍTULO 8 do livro "Análise de Dados Multivariados; Lattin, Carroll e Green".

Vários arquivos estão disponíveis nos exercícios da Lista 04 no STOA.

8.3 Arquivo 02_SIMILARITY.XLSX

Considere a matriz de julgamentos de similaridade descrita no problema 7.2 (e disponível no arquivo *SIMILARITY*). Faça uma análise de agrupamento desses dados e descreva seus resultados. Como as suas conclusões se comparam àquelas da análise de MDS dos mesmos dados?

8.4 Arquivo 09_STORE_SHARE.XLSX

Considere os dados sobre detergentes de lavar roupa que as famílias adquirem em lojas descritas no problema 7.9 (e disponíveis no arquivo *STORE_SHARE*). Faça uma análise de agrupamento para identificar grupos de consumidores com padrões similares de compra de detergente para lavar roupa. Descreva seus resultados.

8.7 Arquivo 06_BASKET.XLSX

Considere a coincidência dos dados da cesta de mercado descrita no problema 7.7 (e disponível no arquivo *BASKET*).

a. Examine os dados você mesmo usando uma análise de agrupamentos. Qual método você considera ser o mais apropriado? Por quê? Como os dados precisam ser escalonados (se é que precisam)? Faça uma apresentação sucinta de suas descobertas.

8.9 Arquivo 13_INTL_FOODS.XLSX

Considere os dados sobre a proporção de famílias em diferentes países com diferentes tipos de alimentos (descritos no problema 7.13 e disponíveis no arquivo *INTL_FOODS*).

a. Proponha e calcule uma medida apropriada de similaridade entre os países (baseada em padrões similares de consumo de alimento). Usando a medida proposta, faça uma análise de agrupamentos dos países e descreva seus resultados.
b. Proponha e calcule uma medida apropriada de similaridade entre alimentos (baseada em padrões similares de demanda dos países). Utilizando a medida proposta, realize uma análise de agrupamentos dos países e descreva seus resultados.
c. Como os *insights* em sua análises dos itens "a" e "b" anteriores comparam-se àqueles obtidos na análise de revelação dos dados no problema 7.13?

8.10 Arquivo 10_COFFEE.XLSX

Grover e Srinivasan (1987) examinaram o comportamento de mudança de preferência de mais de 4.500 consumidores por 11 marcas e tipos diferentes de cafés instantâneos. A Tabela 8.13 (disponível no arquivo *COFFEE*) mostra a mudança de comportamento relatado por um subconjunto de 1.553 consumidores.

Note, por exemplo, que o número de vezes em que observamos a mudança do descafeinado normal High Point e do descafeinado normal Sanka é 43, mas o número de vezes em que observamos uma mudança entre o descafeinado normal High Point e o descafeinado liofilizado Sanka é somente 1. Assim, podemos concluir que há um grau de competição muito maior entre as diferentes marcas do mesmo formato do que entre formatos diferentes.

a. Qual medida você usaria para refletir o nível de competição entre as diferentes alternativas de café instantâneo?
b. Usando a medida proposta no item "a", realize uma análise de agrupamento dos dados de Grover e Srinivasan. Como você descreveria a estrutura competitiva do mercado de café instantâneo?

Tabela 8.13 Mudando a preferência por marcas e tipos de café instantâneo

	1	2	3	4	5	6	7	8	9	10	11
High Point Decaffeinated Regular	93	7	17	19	18	43	1	4	6	7	10
Tasters Choice Caffeinated Freeze Dried	9	80	12	11	24	7	4	2	6	3	3
Tasters Choice Decaffeinated Freeze Dried	9	14	46	3	7	7	4	2	2	0	9
Folgers Caffeinated Regular	19	18	4	82	29	14	0	4	9	2	6
Maxwell House Caffeinated Regular	26	11	6	35	184	24	3	11	18	6	6
Sanka Decaffeinated Regular	15	13	8	13	28	127	4	3	3	8	8
Sanka Decaffeinated Freeze Dried	2	0	3	2	1	7	17	3	0	1	4
Maxim Caffeinated Freeze Dried	4	3	4	3	6	5	2	27	1	0	4
Nescafe Caffeinated Regular	5	3	2	4	16	4	0	1	46	9	2
Nescafe Decaffeinated Regular	6	1	4	1	5	9	0	0	11	15	2
Brim Decaffeinated Freeze Dried	10	4	4	4	2	10	2	2	5	2	27

PARTE 2: Exercícios do CAPÍTULO 12 do livro "Applied Multivariate Statistical Analysis; Johnson & Wichern" (5ªed)

Exemplo 12.3 Similarities of 11 languages - p677, ou exemplo 12.2 na

Example 12.2 (Measuring the similarities of 11 languages) The meanings of words change with the course of history. However, the meaning of the numbers 1, 2, 3, ... represents one conspicuous exception. Thus, a first comparison of languages might be based on the numerals alone. Table 12.2 gives the first 10 numbers in English, Polish, Hungarian, and eight other modern European languages. (Only languages that use the Roman alphabet are considered, and accent marks, cedillas, diereses, etc., are omitted.) A cursory examination of the spelling of the numerals in the table suggests that the first five languages (English, Norwegian, Danish, Dutch, and German) are very much alike. French, Spanish, and Italian are in even closer agreement. Hungarian and Finnish seem to stand by themselves, and Polish has some of the characteristics of the languages in each of the larger subgroups.

Table 12.2 Numerals in 11 Languages

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
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	neljä
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
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

The words for 1 in French, Spanish, and Italian all begin with *u*. For illustrative purposes, we might compare languages by looking at the *first letters* of the numbers. We call the words for the same number in two different languages *concordant* if they have the same first letter and *discordant* if they do not. From Table 12.2, the table of concordances (frequencies of matching first initials) for the numbers 1–10 is given in Table 12.3. We see that English and Norwegian have the same first letter for 8 of the 10 word pairs. The remaining frequencies were calculated in the same manner.

The results in Table 12.3 confirm our initial visual impression of Table 12.2. That is, English, Norwegian, Danish, Dutch, and German seem to form a group. French, Spanish, Italian, and Polish might be grouped together, whereas Hungarian and Finnish appear to stand alone. ■

Table 12.3 Concordant First Letters for Numbers in 11 Languages

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
Fr	4	4	4	1	3	10					
Sp	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	
Fi	1	1	1	1	1	1	1	1	1	2	10

In our examples so far, we have used our visual impression of similarity or distance measures to form groups. We now discuss less subjective schemes for creating clusters.

12.3: Considere o exemplo 12.3 acima e utilize métodos hierárquicos de agrupamento para criar grupos de idiomas.

12.1. Certain characteristics associated with a few recent U.S. presidents are listed in Table 12.11.

Table 12.11

President	Birthplace (region of United States)	Elected first term?	Party	Prior U.S. congressional experience?	Served as vice president?
1. R. Reagan	Midwest	Yes	Republican	No	No
2. J. Carter	South	Yes	Democrat	No	No
3. G. Ford	Midwest	No	Republican	Yes	Yes
4. R. Nixon	West	Yes	Republican	Yes	Yes
5. L. Johnson	South	No	Democrat	Yes	Yes
6. J. Kennedy	East	Yes	Democrat	Yes	No

(a) Introducing appropriate binary variables, calculate similarity coefficient 1 in Table 12.1 for pairs of presidents.

Hint: You may use birthplace as South, non-South.

12.6. The distances between pairs of five items are as follows:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 4 & 0 & & & \\ 6 & 9 & 0 & & \\ 1 & 7 & 10 & 0 & \\ 6 & 3 & 5 & 8 & 0 \end{bmatrix} \end{array}$$

Cluster the five items using the single linkage, complete linkage, and average linkage hierarchical methods. Draw the dendrograms and compare the results.

12.7. Sample correlations for five stocks were given in Example 8.5. These correlations, rounded to two decimal places, are reproduced as follows:

$$\begin{array}{l} \\ \text{JP Morgan} \\ \text{Citibank} \\ \text{Wells Fargo} \\ \text{Royal DutchShell} \\ \text{ExxonMobil} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{JP Morgan} & \text{Citibank} & \text{Wells Fargo} & \text{Royal DutchShell} & \text{Exxon Mobil} \end{array} \\ \begin{bmatrix} 1 & & & & \\ .63 & 1 & & & \\ .51 & .57 & 1 & & \\ .12 & .32 & .18 & 1 & \\ .16 & .21 & .15 & .68 & 1 \end{bmatrix} \end{array}$$

Treating the sample correlations as similarity measures, cluster the stocks using the single linkage and complete linkage hierarchical procedures. Draw the dendrograms and compare the results.

12.9. The vocabulary "richness" of a text can be quantitatively described by counting the words used once, the words used twice, and so forth. Based on these counts, a linguist proposed the following distances between chapters of the Old Testament book Lamentations (data courtesy of Y. T. Radday and M. A. Pollatschek):

Lamentations chapter	1	2	3	4	5
1	0				
2	.76	0			
3	2.97	.80	0		
4	4.88	4.17	.21	0	
5	3.86	1.92	1.51	.51	0

Cluster the chapters of Lamentations using the three linkage hierarchical methods we have discussed. Draw the dendrograms and compare the results.

12.11. Suppose we measure two variables X_1 and X_2 for four items A, B, C, and D. The data are as follows:

	Observations	
Item	x_1	x_2
A	5	4
B	1	−2
C	−1	1
D	3	1

Use the K-means clustering technique to divide the items into $K = 2$ clusters. Start with the initial groups (AB) and (CD).

12.12. Repeat Example 12.11, starting with the initial groups (AC) and (BD). Compare your solution with the solution in the example. Are they the same? Graph the items in terms of their (x_1, x_2) coordinates, and comment on the solutions.

12.13. Repeat Example 12.11, but start at the bottom of the list of items, and proceed up in the order D, C, B, A. Begin with the initial groups (AB) and (CD). [The first potential reassignment will be based on the distances $d^2(D, (AB))$ and $d^2(D, (CD))$.] Compare your solution with the solution in the example. Are they the same? Should they be the same?