Análise Multivariada de Dados - Aula 02

Análise de Conglomerados I: Medidas de Distância e Similaridade

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

Conteúdo

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

- vimos alguns objetivos da Análise Multivariada;
- tivemos uma visão geral dos métodos multivariados;
- lidamos com o problema da representação de dados multivariados;
- começamos a exposição da Análise de Conglomerados.

O que é a Análise de Conglomerados?

O que é a Análise de Conglomerados?

Definição 2.1

A **análise de agrupamentos/conglomerados** diz respeito à identificação de grupos de objetos similares.

O que é Análise de Conglomerados?

A **análise de agrupamentos/conglomerados** representa um conjunto de técnicas exploratórias que podem ser aplicadas quando há a intenção de se verificar a existência de comportamentos semelhantes entre observações (indivíduos, empresas, municípios, países, etc) em relação a determinadas variáveis e o objetivo de se criarem grupos, ou clusters, em que prevaleça a homogeneidade interna.

O que é Análise de Conglomerados?

O pesquisador pode optar por elaborar uma análise de agrupamentos quando tiver o objetivo de ordenar e alocar as observações em grupos e, a partir de então, estudar qual a quantidade interessante de clusters formados.

O que é a Análise de Conglomerados?

Muitos são os procedimentos para que seja elaborada uma análise de agrupamentos, visto que existem diferentes medidas de distância ou de semelhança para, respectivamente, variáveis métricas ou binárias.

O que é a Análise de Conglomerados?

Além disso, definida a medida de distância ou de semelhança, o pesquisador ainda precisa determinar, entre diversas possibilidades, o método de aglomeração das observações, a partir de determinados critérios hierárquicos ou não hierárquicos.

Medidas de Distância

Medidas de Distância

Conforme discutimos, a primeira etapa para a elaboração de uma análise de agrupamentos consiste em definir a medida de distância (dissimilaridade) ou de semelhança (similaridade) que servirá de base para que cada observação seja alocada em determinado grupo.

Medidas de Distância

As medidas de distância são frequentemente utilizadas quando as variáveis do banco de dados forem essencialmente métricas, visto que, quanto maiores as diferenças entre os valores das variáveis de duas determinadas observações, menor a similaridade entre elas (ou, em outras palavras, maior a dissimilaridade).

Medidas de Distância

Retomando conceitos da Geometria, aprendemos que a distância entre dois pontos $P = (x_1, y_1)$ e $Q = (x_2, y_2)$ no plano cartesiano é dada pela fórmula

$$d_{PQ} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2},$$

e representa o comprimento do segmento $\overline{PQ}$.

Medidas de Distância

O número d_{PQ} também é chamado **distância Euclidiana**, e representa a menor distância entre os pontos P e Q.

Medidas de Distância

Mas nem sempre considerar a distância Euclidiana é o mais adequado.

Medidas de Distância

Medidas de Distância

Assim, vamos ver alguns Exemplos de distância que podem usados para realizar uma análise de conglomerados. Vamos exemplificar todas elas por meio da planilha "aula-02-exemplo-nutella".

Definição 3.1 (Distância quadrática euclidiana)

Alternativamente à distância euclidiana, pode ser utilizada quando as variáveis apresentarem pequena dispersão de seus valores, fazendo com que o uso da distância euclidiana ao quadrado facilite a interpretação dos outputs da análise e a alocação das observações nos grupos. Sua expressão é dada por:

$$d_{pq} = \sum_{j=1}^{k} (X_{jp} - X_{jq})^2$$
$$= (X_{1p} - X_{1q})^2 + (X_{2p} - X_{2q})^2 + ... + (X_{kp} - X_{kq})^2.$$

Definição 3.2 (Distância de Minkowski)

É a expressão de medida de dissimilaridade mais geral a partir da qual outras derivam. É dada por:

$$d_{pq} = \left[\sum_{j=1}^{k} (|X_{jp} - X_{jq}|)^m \right]^{\frac{1}{m}}$$

em que m assume valores inteiros e positivos ($m = 1, 2, ...$). Podemos verificar que a distância euclidiana é um caso particular da distância de Minkowski, quando $m = 2$.

Definição 3.3 (Distância de Manhattan)

Também conhecida por distância absoluta ou bloco, não leva em consideração a geometria triangular inerente à expressão inicial de Pitágoras e considera apenas as diferenças entre os valores de cada variável. Sua expressão, também um caso particular da distância de Minkowski quando $m = 1$, é dada por:

$$d_{pq} = \sum_{j=1}^{k} |X_{jp} - X_{jq}|.$$

Definição 3.4 (Distância de Chebychev)

Também conhecida por distância infinita ou máxima, é um caso particular da distância de Manhattan por considerar, para duas determinadas observações, apenas a máxima diferença entre todas as j variáveis em estudo. Sua expressão é dada por:

$$d_{pq} = \max |X_{jp} - X_{jq}|,$$

também um caso particular da distância de Minkowski quando $m \to \infty$.

Definição 3.5 (Distância de Canberra)

Utilizada para os casos em que as variáveis apresentam apenas valores positivos, assume valores entre 0 e j (número de variáveis). Sua expressão é dada por:

$$d_{pq} = \sum_{j=1}^{k} \frac{|X_{jp} - X_{jq}|}{X_{jp} + X_{jq}}$$

Medidas de Distância

Na presença de variáveis métricas, o pesquisador ainda pode fazer uso da correlação de Pearson, que pode propiciar informações importantes quando o intuito for agrupar linhas do banco de dados.

Medidas de Distância

Além da decisão sobre a escolha da medida de distância, o pesquisador também deve verificar se os dados precisam ser preliminarmente tratados.

Medidas de Distância

Nos exemplos abordados até o presente momento, tomamos o cuidado de considerar variáveis métricas sempre com valores na mesma unidade de medida.

Medidas de Distância

Entretanto, caso as variáveis sejam medidas em unidades distintas (por exemplo, renda em $R\$$, quantidade de filhos, etc), a intensidade das distâncias entre as observações poderá ser influenciada arbitrariamente pelas variáveis que eventualmente apresentarem maior magnitude de seus valores, em detrimento das demais.

Medidas de Distância

Nessas situações, o pesquisador deve padronizar os dados, a fim de que a arbitrariedade das unidades de medida seja eliminada, fazendo cada variável ter a mesma contribuição sobre a medida de distância considerada.

Medidas de Distância

O método mais comumente utilizado para padronização das variáveis é conhecido por **procedimento zscore**.

Definição 3.6

Para uma variável X, o **zscore** do valor X_j é dado por

$$ZX_j = \frac{X_j - \mu}{\sigma},$$

onde μ é a média e σ o desvio-padrão da variável X respectivamente.

Medidas de Distância

Dessa forma, independentemente da magnitude dos valores e da natureza das unidades de medida das variáveis originais de um banco de dados, todas as respectivas variáveis padronizadas pelo procedimento zscore terão média igual a 0 e desvio-padrão igual a 1, o que garante a eliminação de eventuais arbitrariedades das unidades de medida sobre a distância entre cada par de observações.

Medidas de Distância

Além disso, o procedimento Zscores tem a vantagem de não alterar a distribuição da variável original.

Medidas de Distância

Portanto, caso as variáveis originais apresentem unidades de medida distintas, as expressões das medidas de distância devem ter os termos X_{jp} e X_{jq} substituídos, respectivamente, por ZX_{jp} e ZX_{jq}.

Medidas de Distância

Embora a correlação de Pearson não seja uma medida de dissimilaridade é relevante comentar que seu uso também requer que as variáveis sejam padronizadas por meio do procedimento zscore caso não apresentem as mesmas unidades de medida.

Medidas de Distância

Caso o intuito fosse agrupar variáveis, que é o objetivo da análise fatorial, a padronização de variáveis por meio do procedimento zscores seria, de fato, irrelevante, dado que a análise consistiria em avaliar a correlação entre colunas do banco de dados.

Medidas de Distância

Como o nosso objetivo, por outro lado, é agrupar linhas do banco de dados que representam as observações, a padronização das variáveis faz-se necessária para a elaboração de uma correta análise de agrupamentos.

Medidas de Similaridade

Medidas de Similaridade

Imagine agora que tenhamos a intenção de calcular a distância entre observações provenientes de um banco de dados com todas referentes à presença ou ausência de alguma características de interesse (vide planilha "aula-02-similaridade").

Medidas de Similaridade

Nessa situação, é comum que a presença ou ausência de determinada característica seja representada por uma variável binária, ou **dummy**, que assume valor 1, caso a característica ocorra, e 0, caso contrário.

Medidas de Similaridade

É importante ressaltar que o artificio das variáveis binárias não gera problemas de ponderação arbitrária, oriunda das categorias das variáveis, ao contrário do que ocorreria caso fossem atribuídos valores discretos $(1, 2, 3, ...)$ para cada categoria de cada variável qualitativa.

Medidas de Similaridade

Nesse sentido, caso determinada variável qualitativa apresente k categorias, serão necessárias $(k-1)$ variáveis binárias que representarão a presença ou a ausência de cada uma das categorias, ficando todas as variáveis binárias iguais a 0 para o caso de ocorrer a categoria de referência.

Medidas de Similaridade

Nosso intuito é intuito criar medidas de semelhança entre observações utilizando coeficientes que levam em consideração a similaridade de respostas 1-1 e 0-0, sem que necessariamente esses pares tenham a mesma importância relativa.

Medidas de Similaridade

Para que possamos apresentar essas medidas, é necessário construir uma tabela de frequências absolutas de respostas 0 e 1 para cada par de observações quaisquer p e q.

Medidas de Similaridade

Observação q \ Observação p	1	0	Total
1	a	b	$a + b$
0	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Medidas de Similaridade

Com base nessa tabela, apresentamos, a seguir, as principais medidas de semelhança existentes, lembrando que a adoção de cada uma depende dos pressupostos e dos objetivos do pesquisador.

Definição 4.1 (Medida de emparelhamento simples)

Essa medida, que iguala os pesos das respostas convergentes 1-1 e 0-0, tem sua expressão dada por:

$$s_{pq} = \frac{a+d}{a+b+c+d}.$$

Definição 4.2 (Medida de Jaccard)

Essa medida não leva em conta a frequência do par de respostas 0-0, considerada irrelevante. Sua expressão geral é dada por:

$$s_{pq} = \frac{a}{a+b+c}.$$

Definição 4.3 (Medida de Dice)

É similar ao coeficiente de Jaccard, porém dobra o peso sobre a frequência de pares de respostas em convergência do tipo 1-1. Sua expressão é dada por:

$$s_{pq} = \frac{2a}{2a + b + c}.$$

Definição 4.4 (Medida antiDice)

Assim como as medidas de Jaccard e de Dice, a medida antiDice também ignora a frequência de pares de respostas 0-0. Sua expressão é dada por:

$$s_{pq} = \frac{a}{a + 2(b + c)}.$$

Definição 4.5 (Medida de Russell e Rao)

Essa medida privilegia no cálculo de seu coeficiente apenas as similaridades das respostas 1-1. Foi proposta por Russell e Rao (1940), tendo sua expressão dada por:

$$s_{pq} = \frac{a}{a+b+c+d}.$$

Definição 4.6 (Medida de Ochiai)

Esse coeficiente é indefinido quando uma ou ambas as observações estudadas apresentarem os valores de todas as variáveis iguais a 0. Se esse fato ocorrer para apenas um dos dois vetores, a medida de Ochiai é considerada igual a 0. Sua expressão é dada por:

$$s_{pq} = \frac{a}{\sqrt{(a+b)(a+c)}}.$$

Definição 4.7 (Medida de Yule)

Essa medida de semelhança para variáveis binárias oferece como resposta um coeficiente que varia de -1 a 1. Conforme podemos verificar, por meio de sua expressão apresentada a seguir, o coeficiente gerado é indefinido se um ou ambos os vetores comparados apresentarem todos os valores iguais a 0 ou 1:

$$s_{pq} = \frac{ad - bc}{ad + bc}.$$

Definição 4.8 (Medida de Rogers e Tanimoto)

Essa medida, que dobra o peso das respostas discrepantes 0-1 e 1-0 em relação ao peso das combinações de respostas convergentes do tipo 1-1 e 0-0, foi inicialmente proposta por Rogers e Tanimoto (1960). Sua expressão, que passa a ser igual à da medida antiDice quando a frequência de respostas 0-0 for igual a 0 ($d = 0$), é dada por:

$$s_{pq} = \frac{a + d}{a + d + 2(b + c)}.$$

Definição 4.9 (Medida de Sneath e Sokal)

Ao contrário da medida de Rogers e Tanimoto, essa medida, proposta por Sneath e Sokal (1962), dobra o peso das respostas convergentes do tipo 1-1 e 0-0 em relação ao das demais combinações de respostas (1-0 e 0-1). Sua expressão, que passa a ser igual à da medida Dice quando a frequência de respostas do tipo 0-0 for igual a 0 ($d = 0$), é dada por:

$$s_{pq} = \frac{2(a+d)}{2(a+d)+b+c}.$$

Definição 4.10 (Medida de Harnann)

Hamann (1961) propôs essa medida de semelhança para variáveis binárias com o intuito de que fossem subtraídas as frequências de respostas discrepantes (1-0 e 0-1) do total de respostas convergentes (1-1 e 0-0). Esse coeficiente, que varia de -1 (divergência total de repostas) a 1 (convergência total de respostas), é igual a duas vezes a medida de emparelhamento simples menos 1. Sua expressão é dada por:

$$s_{pq} = \frac{(a+d)-(b+c)}{a+b+c+d}.$$

Medidas de Similaridade

Analogamente ao discutido quando do cálculo das medidas de dissimilaridade, é visível que medidas de similaridade diferentes geram resultados distintos, o que pode fazer, quando da elaboração do método de aglomeração, que as observações sejam alocadas em diferentes agrupamentos homogêneos, dependendo da escolha da medida para análise.

Medidas de Similaridade

Lembramos que não faz sentido algum aplicar o procedimento de padronização zscore para o cálculo das medidas de semelhança discutidas nesta seção, visto que as variáveis utilizadas para a análise de agrupamentos são binárias.

Medidas de Similaridade

Definida a medida a ser utilizada, com base nos objetivos de pesquisa, na teoria subjacente e em sua experiência e intuição, o pesquisador deve partir para a definição do esquema de aglomeração.

Medidas de Similaridade

Este será o foco das próximas Aulas.

Comentários Finais

Comentários Finais

Em resumo, na aula de hoje nós lidamos com:

- medidas de distância;
- medidas de similaridade.

Comentários Finais

Nas próximas aulas focaremos nos Esquemas de Aglomeração.

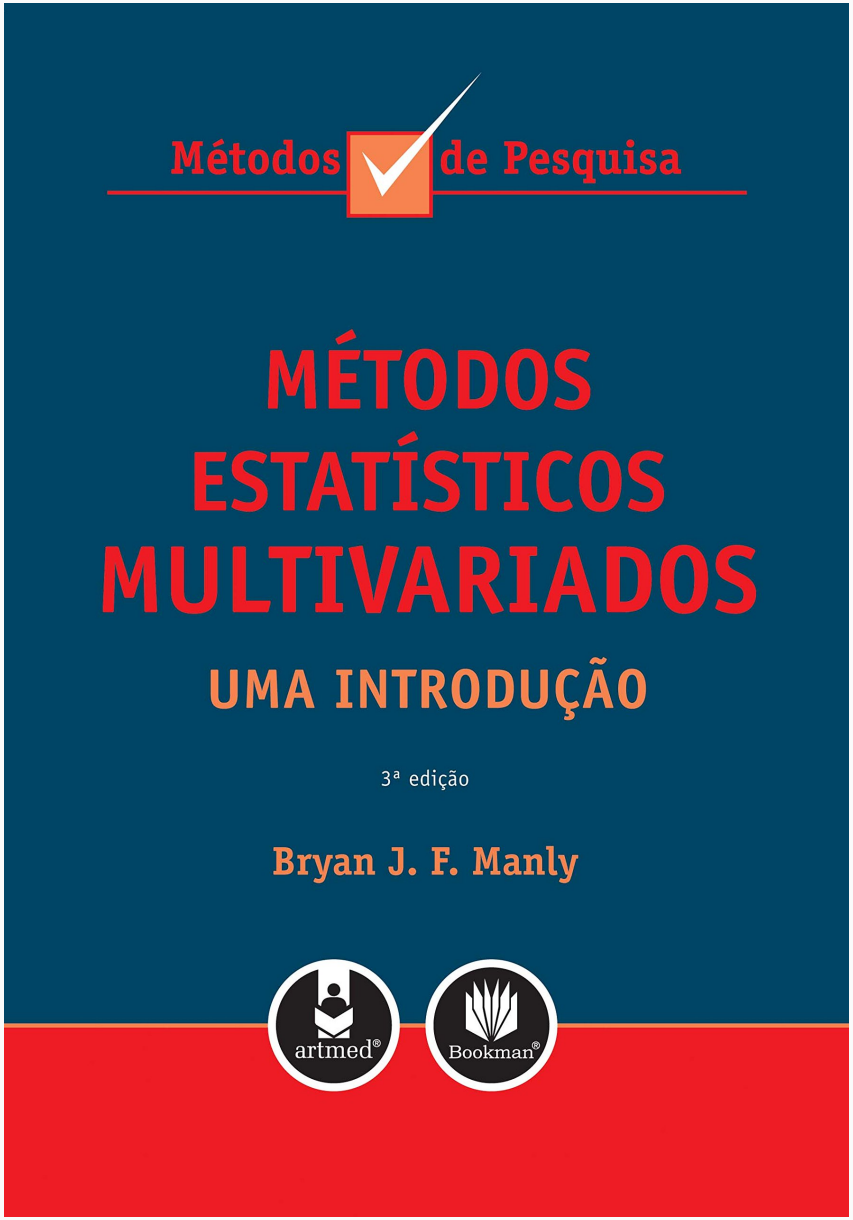
Comentários Finais

ATIVIDADE PARA ENTREGAR (E COMPOR A NOTA N1)

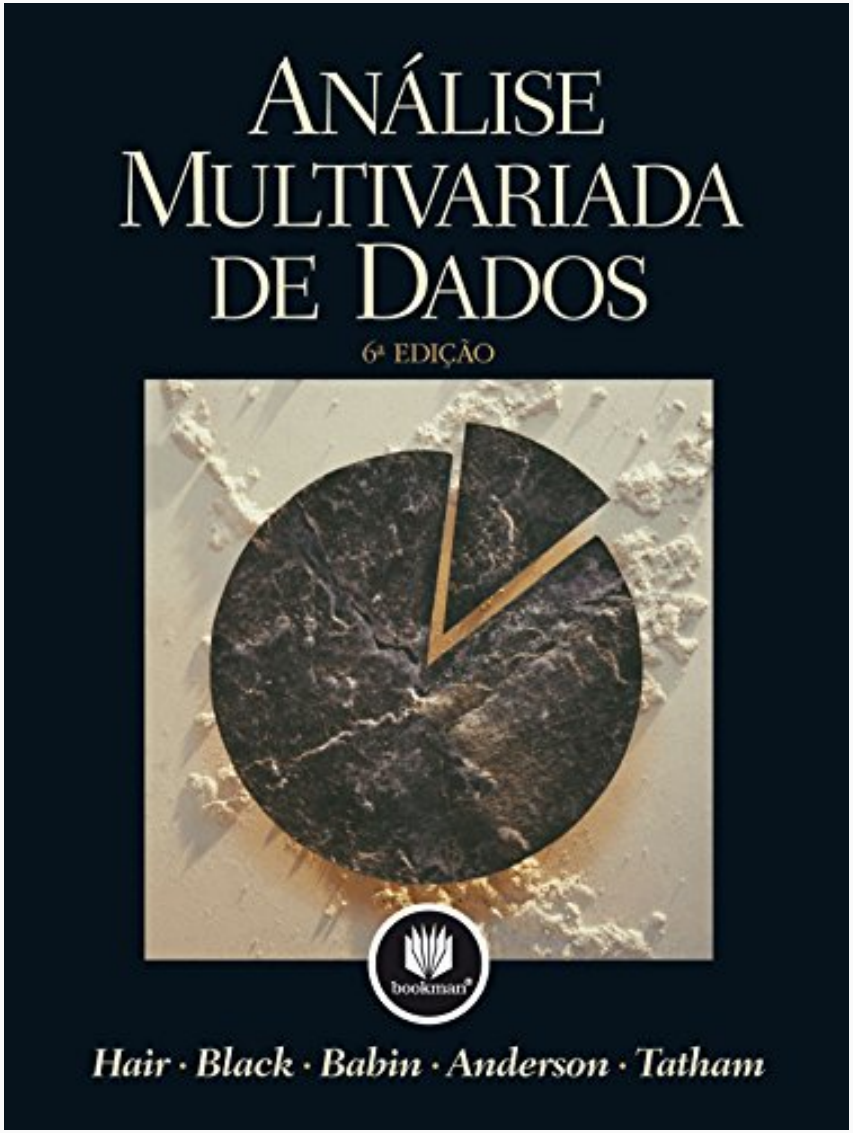
Resolva em grupos de até 4 integrantes os Exercícios 2.1, 2.2, 2.3.

Referências

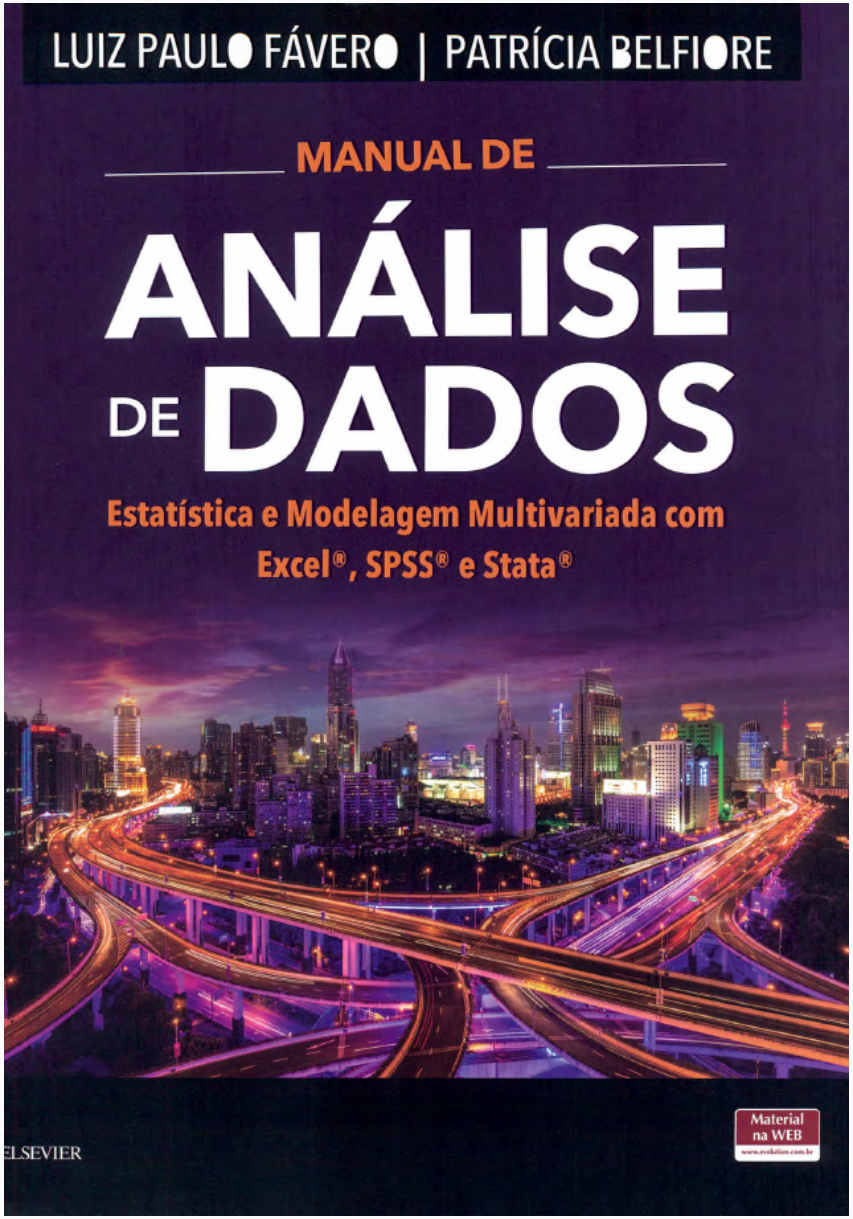
Referências



Referências



Referências



Bons Estudos!