

Análise de Clusters I

ACH2036 – Métodos Quantitativos Aplicados à Adm. de Empresas I

Prof. Regis Rossi A. Faria

2º sem. 2020



Créditos: Profa. Ana Amélia Benedito Silva (conteúdo parcial de slides)

Sumário

- Definições e conceitos
- Etapas de aplicação da técnica
- Exemplos

Análise de Conglomerados

- Análise de conglomerados, análise de agrupamentos ou análise de clusters são técnicas de interdependência
- Permitem agrupar casos ou variáveis em um grupo homogêneo, em função de suas similaridades, ou seja, semelhanças
- Os objetos (indivíduos) em cada grupo tendem a ser semelhantes entre si e diferentes dos outros objetos (indivíduos) contidos em outros conglomerados.

Análise de Conglomerados

- é uma técnica exploratória
- permite estudar a estrutura de grupos
- permite identificar *outliers*
- permite levantar hipóteses sobre as associações dos objetos
- é uma técnica não-inferencial, ou seja, não possibilita inferências sobre a população com base na amostra

Análise de Conglomerados

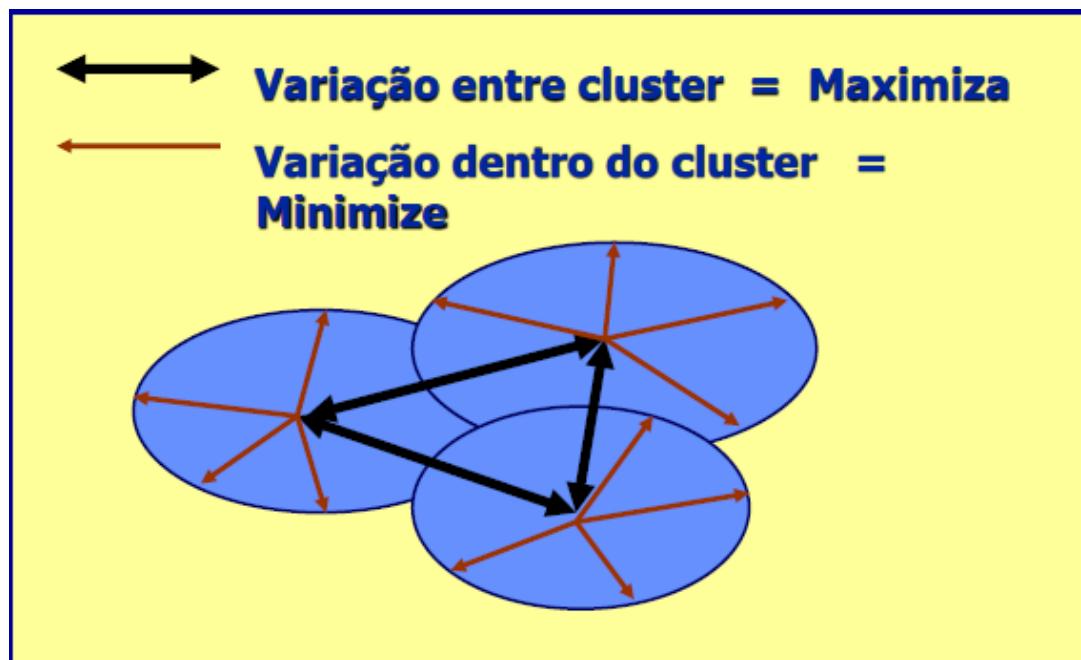
- Técnica utilizada na identificação de padrões de comportamento em banco de dados por meio da formação de grupos homogêneos de observações/instâncias/casos
- Tem aplicabilidade em várias áreas (ex: biologia, psicologia, análise de mercado, produtos, demográfica, etc.)

Peculiaridades da análise de clusters

- Possui forte base matemática, mas não estatística.
- Suposições como normalidade, linearidade e homoscedascidade (importantes em outras técnicas, como Regressão Linear Múltipla) possuem pouco impacto no caso de Análise de clusters

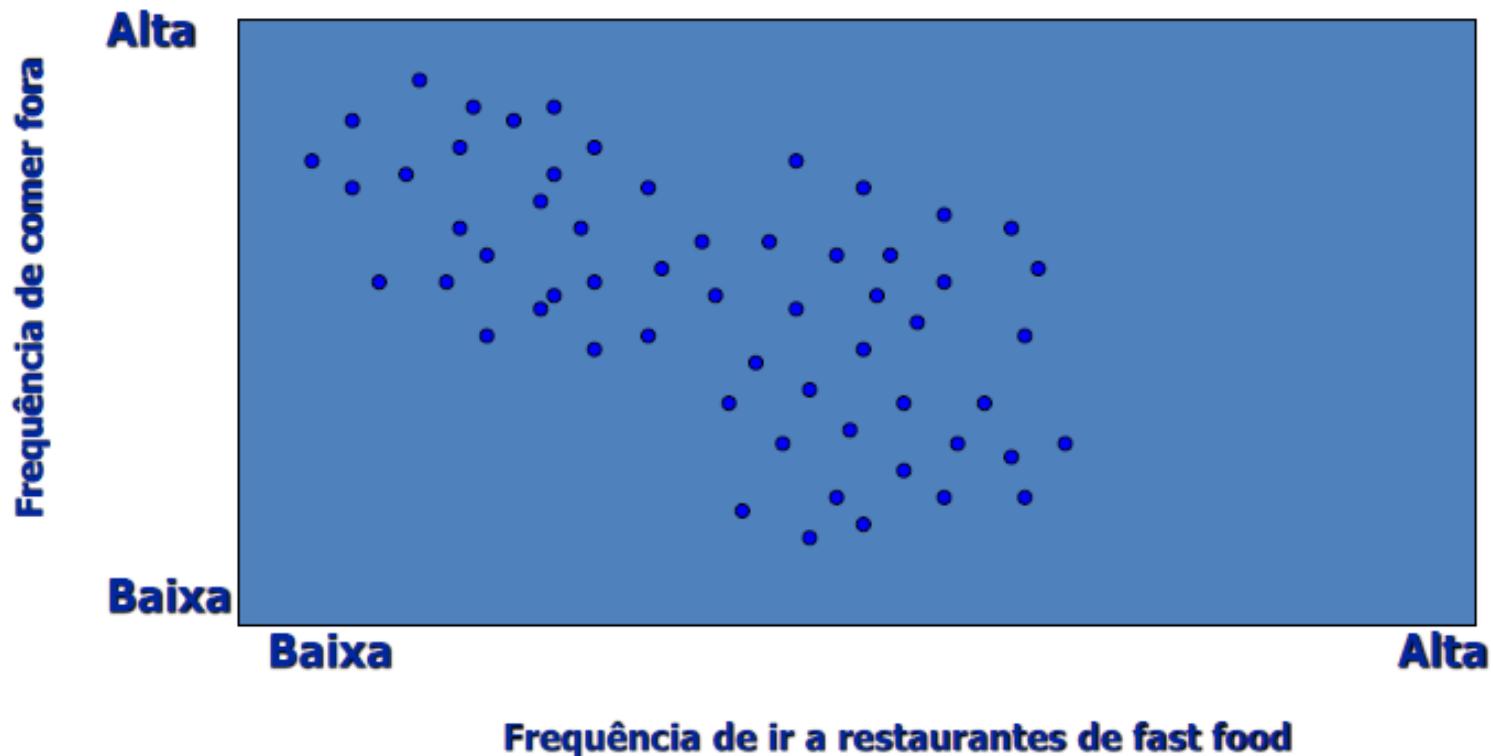
ANÁLISE DE CLUSTERS

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



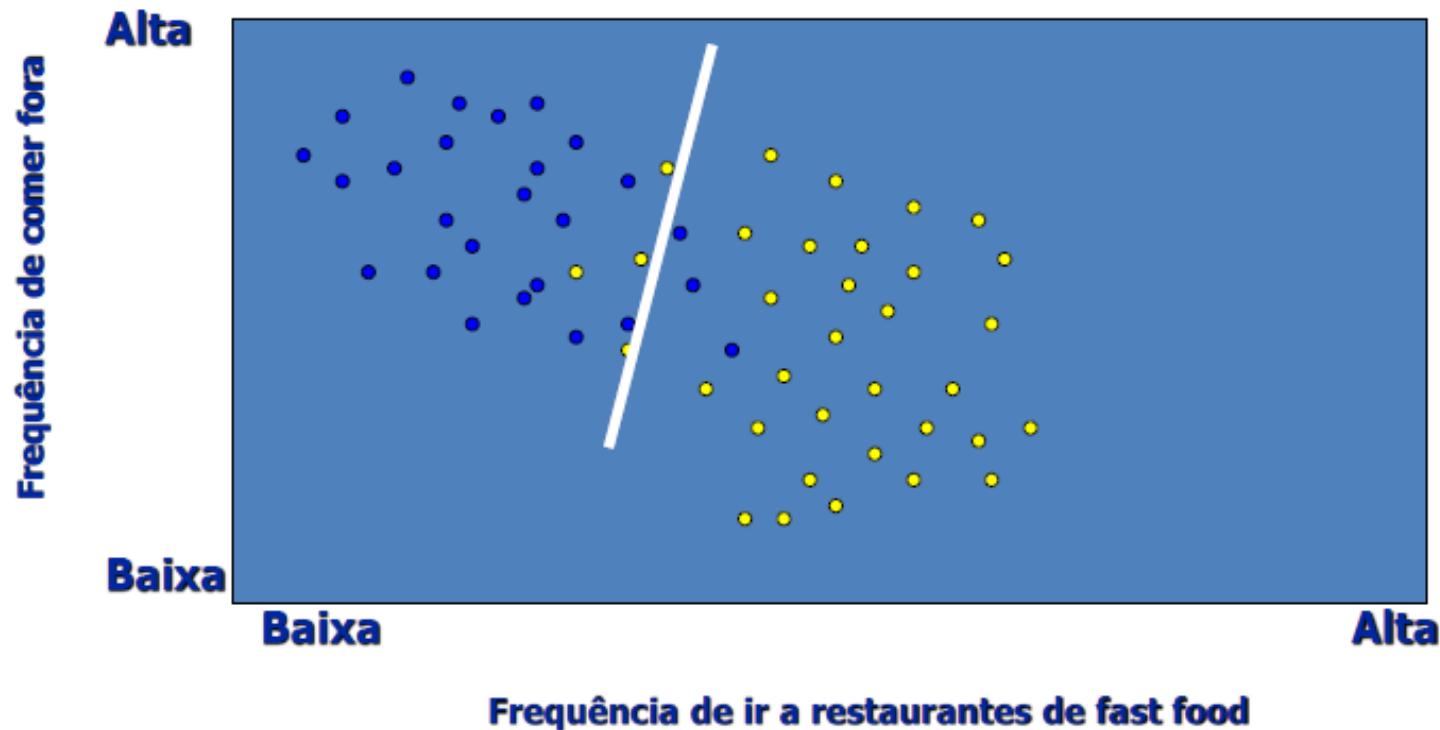
ANÁLISE DE CLUSTERS

Exemplo: agrupar pessoas de acordo com hábitos de alimentação



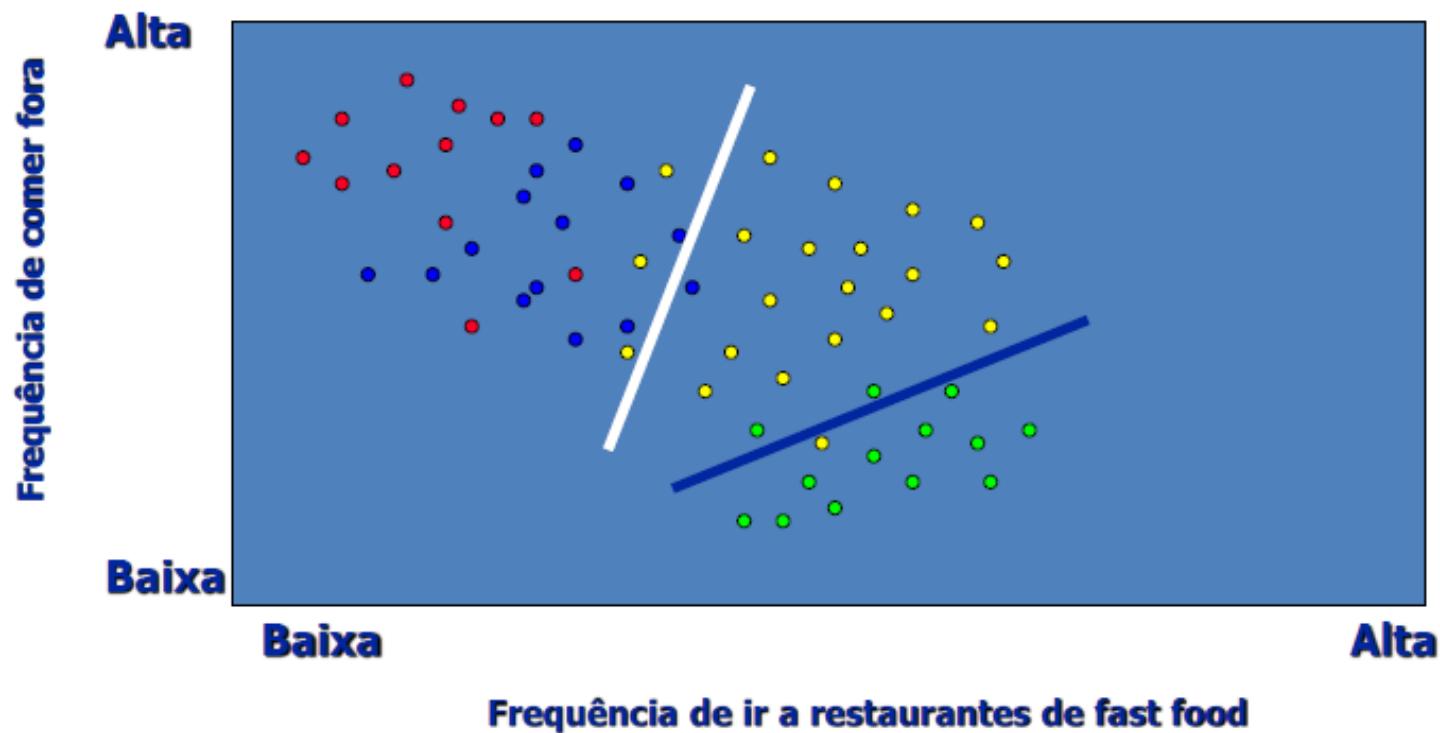
ANÁLISE DE CLUSTERS

Exemplo: agrupar pessoas de acordo com hábitos de alimentação



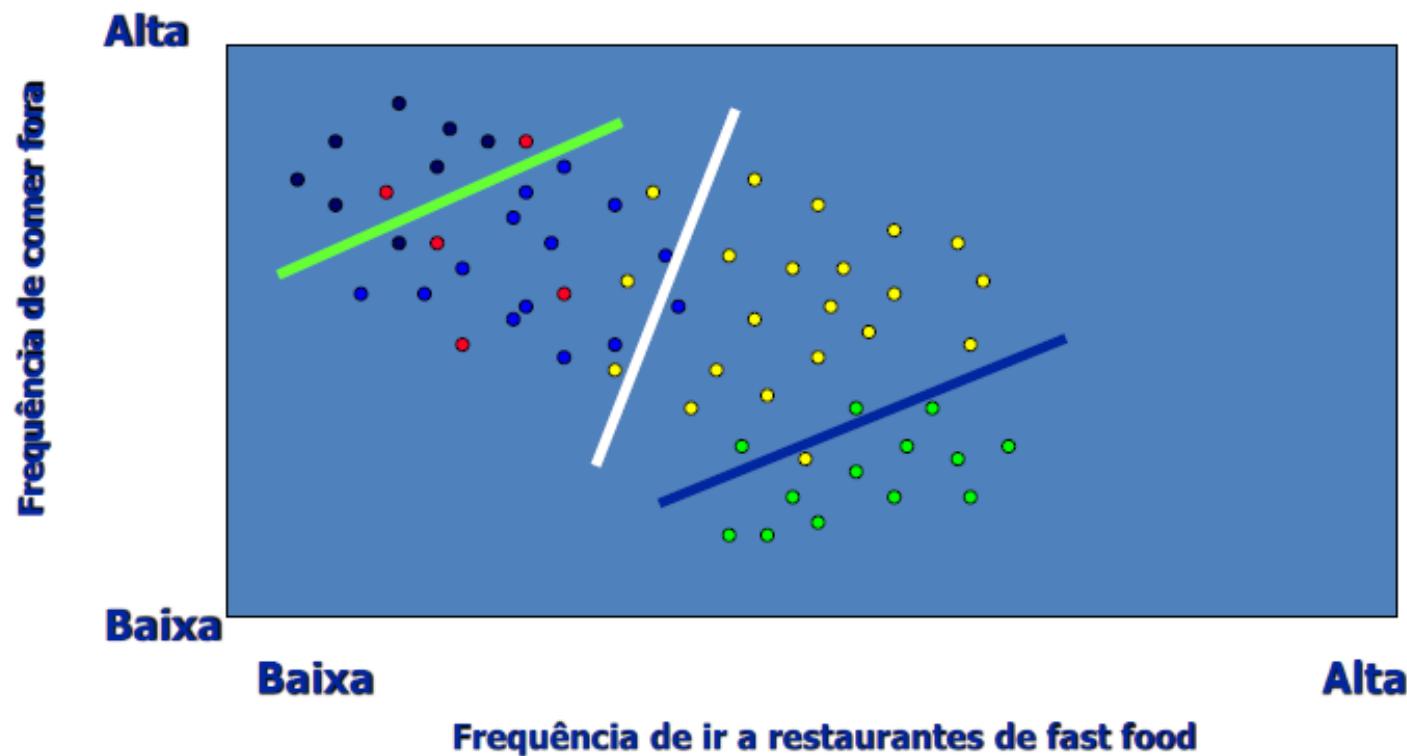
ANÁLISE DE CLUSTERS

Exemplo: agrupar pessoas de acordo com hábitos de alimentação



ANÁLISE DE CLUSTERS

Exemplo: agrupar pessoas de acordo com hábitos de alimentação



ANÁLISE DE CLUSTERS

- Apesar de ser classificada como uma técnica de Análise Multivariada tradicional, a Análise de Clusters é aplicada na maioria das vezes em pesquisas de caráter exploratório.
- É uma técnica para analisar interdependência entre casos/indivíduos segundo determinadas variáveis.
- Não é possível determinar antecipadamente as variáveis dependentes e independentes
- Ao contrário, examina relações de interdependência contidas na estrutura dos dados.

Exemplos

1. Classificar setores censitários de acordo com as diferentes dimensões de justiça/injustiça ambiental
2. Classificar os municípios de SP em função das diferentes dimensões de violência contra a mulher
3. Classificar os bairros do ABC de acordo com a quantidade/perfil dos lançamentos residenciais
4. Classificar os distritos de SP de acordo com as variáveis de infraestrutura e entorno de domicílios

Exemplos

5. Classificar consumidores em relação aos seus hábitos de compra em uma rede de supermercados
6. Um arqueólogo tem dados sobre a localização de restos de cerâmica encontrados em um sitio arqueológico.

Para conhecer como era a organização espacial da tribo que lá habitava, ele necessita ter uma ideia mais precisa da dispersão dessas peças. Há locais com alta concentração de peças? Quantos?

7. Agrupar alunos segundo notas de avaliação

Exemplo 4 - Classificar os distritos de SP de acordo com as variáveis de infraestrutura e entorno de domicílios

distrito	População	IDH	Número de hospitais	Número de escolas
Santana				
Butantã				
Freguesia do Ó				
Pinheiros				
..				
...				
...				
...				
...				
Santo Amaro				
Vila Mariana				

Exemplo 7

- Agrupar alunos segundo notas de avaliação

CONCEITO A – 8,8 a 10

CONCEITO B – 7,0 a 8,7

CONCEITO C – 5,0 a 6,9

CONCEITO D – abaixo de 4,9



Identificação_aluno	nota
1	5,0
2	5,5
3	4,3
4	3,0
5	3,3
6	1,2
7	4,4
8	5,4
9	3,0
10	2,2
11	4,1
12	7,0
13	1,0
14	3,9
15	4,2
16	2,6
17	5,9
18	6,2
19	3,2
20	1,8

1 CONCEITO B

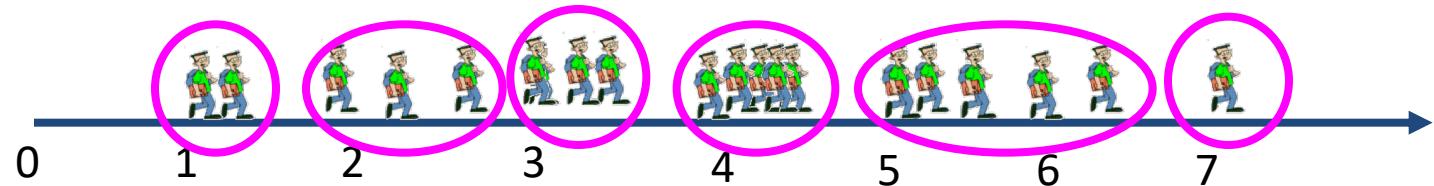
5 CONCEITOS C

14 CONCEITOS D

ANÁLISE DE CLUSTERS

ID_Aluno	NOTA
1	5
2	5,5
3	4,3
4	3
5	3,3
6	1,2
7	4,4
8	5,4
9	3
10	2,2
11	4,1
12	7
13	1
14	3,9
15	4,2
16	2,6
17	5,9
18	6,2
19	3,2
20	1,8

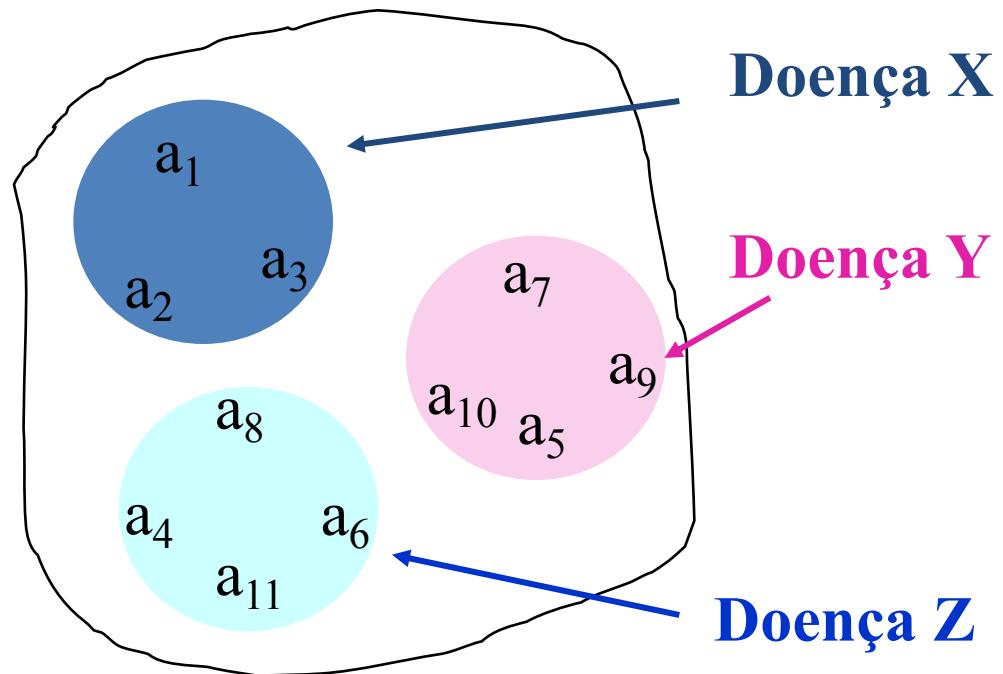
SOLUÇÃO = 6
GRUPOS



Notas avaliação

Agrupamento - Análise de Clusters

a ₁	a	F	1	0	1	1
a ₂	b	M	0	0	1	1
•	c	F	1	1	1	0
•	d	F	1	0	0	0
•	e	M	1	1	0	1
	Nome	Sexo		Sintomas		



Conceito = Doença

Número de Clusters = 3

ANÁLISE DE CLUSTERS

ETAPAS

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. Interpretação e validação dos agrupamentos

Etapa 1 - Análise das variáveis e dos objetos

Seleção de variáveis, identificação de *outliers*

- Cabe ao pesquisador selecionar as variáveis relevantes
- A técnica é muito sensível a *outliers*
 - Deve-se localizar os outliers de cada variável
 - Cabe analisar se devem ou não ser retirados
- É comum que os *outliers* formem grupos isolados

Etapa 1 - Análise das variáveis e dos objetos

Padronização das variáveis

- Medidas/escalas diferentes distorcem a estrutura do agrupamento
- Solução → Padronização resolve problema de diferentes escalas ou magnitudes das variáveis
- Padronização faz com que seja atribuído o mesmo peso para cada variável

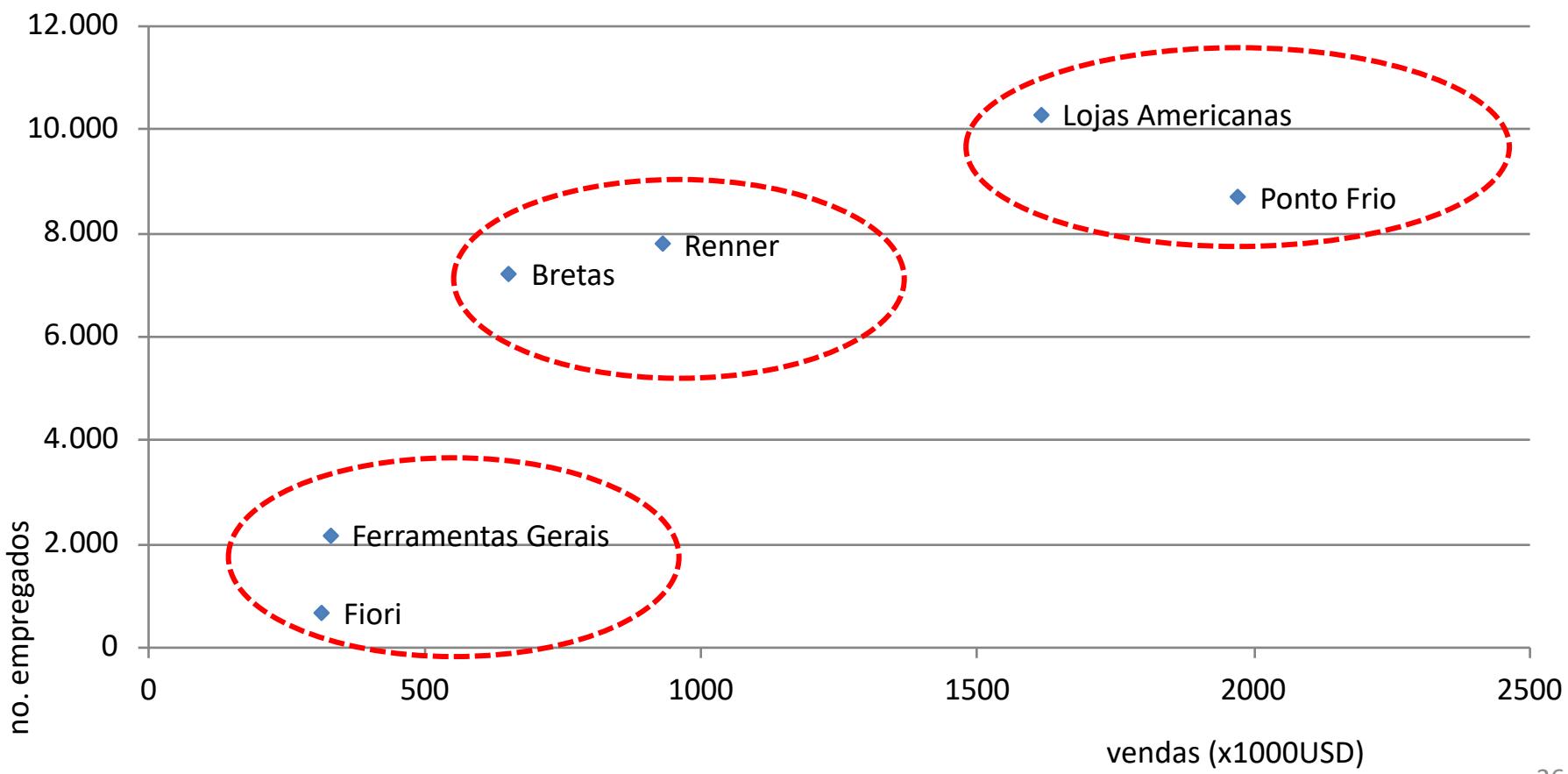
Etapa 1 - Análise das variáveis e dos objetos

Empresas	Vendas (US\$ milhões)	Número empregados
Ferramentas Gerais	327,5	2.150
Fiori	312,2	661
Bretas Supermercados	652,6	7.200
Renner	929	7.764
Lojas Americanas	1.613,5	10.281
Ponto Frio	1.971	8.672

Exemplo: Seria possível separar a amostra de empresas em grupos similares em termos de porte, representado pelas variáveis faturamento e número de empregados?

Etapa 1 - Análise das variáveis e dos objetos

- Exemplo



Etapa 1 - Análise das variáveis e dos objetos

Padronização das variáveis: Tipos de padronização

- z-score
- Método range: -1 a +1
- Método range: 0 a 1
- Método da máxima amplitude
- Método da média =1
- Método do desvio-padrão =1

Etapa 1 - Análise das variáveis e dos objetos

Padronização das variáveis

- Utilização de escalas de medida em grandezas diferentes pode distorcer a análise
- A forma mais utilizada é a padronização (*Z score*), com média zero e desvio padrão 1

$$Z = \frac{x - \text{média}}{\text{desvio padrão}}$$

- A padronização deve ser utilizada com cuidado, pois se existir alguma relação natural refletida nas escalas das variáveis, a padronização pode não ser adequada

Etapa 1 - Análise das variáveis e dos objetos

Padronização das variáveis

- Método range: -1 a +1

$$\frac{x}{\text{amplitude}}$$

- Método range: 0 a +1

$$\frac{x - \text{mínimo}}{\text{amplitude}}$$

- Método de máxima amplitude

$$\frac{x}{\text{máximo}}$$

Etapa 1 - Análise das variáveis e dos objetos

Padronização de Variáveis

- Método da média = $1 \frac{x}{média}$
- Método de desvio padrão = $1 \frac{x}{desvio\ padrão}$

Etapa 1 - Análise das variáveis e dos objetos

Padronização dos Dados com z-score

Empresas	Vendas (US\$ milhões)	Número empregados
Ferramentas Gerais	-0,931	-1,038
Fiori	-0,953	-1,427
Bretas Supermercados	-0,458	0,282
Renner	-0,056	0,429
Lojas Americanas	0,939	1,087
Ponto Frio	1,459	0,666

ANÁLISE DE CLUSTERS

Etapas

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. Interpretação e validação dos agrupamentos

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

- As observações são agrupadas segundo algum tipo de métrica de distância.
- Observações com menor distância entre si são mais semelhantes, logo são aglomerados em um mesmo conglomerado.
- Objetos mais distantes participam de conglomerados distintos.

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

- São classificadas em 3 tipos:
 - Medidas de distância
 - Medidas correlacionais
 - Medidas de associação
- A escolha da medida depende da natureza da variável e da escala de medida

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de distância:

- Distância euclidiana – mais utilizada
- Distância quadrática euclidiana
- Distância de Minkovski
- Distância absoluta, bloco ou Manhattan
- Distância de Mahalanobis
- Distância de Chebychev

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância:

- Distância Euclidiana: a distância entre duas observações (i e j) correspondente à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j) para todas as p variáveis

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

- Em que x_{ik} é o valor da variável k referente à observação i e x_{jk} representa a variável k para a observação j
- Quanto menor a distância, mais similares são as observações

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância:

- Distância Quadrática Euclidiana: a distância entre duas observações (i e j) correspondente à soma dos quadrados das diferenças entre os pares de observações (i e j) para todas as p variáveis

$$d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2$$

- Mais comum
- Quanto menor a distância, mais similares são as observações

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância:

- Distância de Minkowski: a distância euclidiana é um caso particular de uma distância mais geral, chamada de Minkowski

$$d_{ij} = \left(\sum_{k=1}^p (|x_{ik} - x_{jk}|)^n \right)^{1/n}$$

- Se aplicarmos $n = 2$, chegamos a distância euclidiana
- Para $n = 1$ temos a Distância City-Block, ou *Manhattan Distance*

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância:

- Distância Absoluta, Bloco, City-Block ou Manhattan: representa a soma das diferenças absolutas entre os valores das p variáveis para os dois casos

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância:

- Distância Mahalanobis: é a distância estatística entre dois indivíduos i e j , considerando a matriz de covariância para o cálculo das distâncias

$$d_{ij} = \sqrt{(x_i - x_j)'S^{-1}(x_i - x_j)}$$

- Em que S é a estimativa amostral da matriz de variância-covariância Σ dentro dos agrupamentos

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância:

- Distância Chebychev: diferença absoluta máxima entre todas as p variáveis entre duas observações

$$d_{ij} = \max |x_{ik} - x_{jk}|$$

- Em que S é a estimativa amostral da matriz de variâncias-covariâncias Σ dentro dos agrupamentos

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas Correlacionadas:

- Representam similaridade pela correspondência de padrões ao longo das características (X variáveis)
- Correlação de Pearson é a mais utilizada

$$r_{ij} = \frac{\sum_{k=1}^p (x_{1k} - \bar{x}_i)(x_{1j} - \bar{x}_j)}{\sqrt{\sum_{k=1}^p (x_{1k} - \bar{x}_i)^2 \sum_{k=1}^p (x_{1j} - \bar{x}_j)^2}}$$

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Associação:

- Utilizado com variáveis binárias
- Tabela de Contingência

		Indivíduo <i>j</i>		
		1	0	Total
Indivíduo <i>i</i>	1	a	b	a+b
	0	c	d	c+d
Total		a+c	b+d	$p = a+b+c+d$

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Associação:

- Coeficientes de Emparelhamento Simples:

- ✓ Medida de Semelhança (S_{ij})

$$S_{ij} = \frac{a+d}{a+b+c+d}$$

- ✓ Medida de Distância (d_{ij})

$$d_{ij} = \frac{b+c}{a+b+c+d}$$

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Associação:

- Medida de Semelhança (S_{ij})
- Medida de Distância (d_{ij})

$$S_{ij} = \frac{a}{a+b+c}$$

$$d_{ij} = \frac{b+c}{a+b+c}$$

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Matriz de dados padronizados

x_{11}	x_{12}	x_{13}	...	x_{1n}
x_{21}	x_{22}	x_{23}	...	x_{2n}
x_{31}	x_{32}	x_{33}	...	x_{3n}
...
x_{p1}	x_{p2}	x_{p3}	...	x_{pn}

Matriz das distâncias

$d(x_1, x_1) = 0$	0	...	0
$d(x_1, x_2)$	$d(x_2, x_2) = 0$...	0
$d(x_1, x_3)$	$d(x_2, x_3)$...	0
...	0
$d(x_1, x_p)$	$d(x_2, x_p)$...	0

Distância Euclidiana

$$d(x_1, x_2) = \sqrt{(x_{11}-x_{21})^2 + (x_{12}-x_{22})^2 + \dots + (x_{1n}-x_{2n})^2}$$

Exemplo das empresas

Empresas	Vendas (US\$ milhões)	Número de empregados
Ferramentas Gerais (1)	327,5	2150
Fiori (2)	312,2	661
Bretas Supermercados (3)	652,6	7200
Renner (4)	929,0	7764
Lojas Americanas (5)	1613,5	10281
Ponto Frio (6)	1971,0	8672

Exemplo: Seria possível separar a amostra de empresas em grupos similares em termos de porte, representado pelas variáveis faturamento e número de empregados?

Padronização

Empresas	x= Vendas (z-score)	y =Nº de empregados (z-score)
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

Cálculo das distâncias

Empresas	x= Vendas (z-score)	y =Nº de empregados (z-score)
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

$$\text{Distância quadrática euclidiana} = ((x_1-x_2)^2 + (y_1-y_2)^2)^2$$

$$\text{Distância (empresa}_1\text{-empresa}_2\text{)}^2 = (-0,931-(-0,953))^2+(-1,038-(-1,427))^2 = \mathbf{0,152}$$

Matriz de similaridade pela Distância Quadrática Euclideana

	Ferramentas Gerais (1)	Fiori (2)	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
Ferramentas Gerais (1)	0,000					
Fiori (2)	0,152	0,000				
Bretas Supermercados (3)	1,964	3,163	0,000			
Renner (4)	2,916	4,248	0,183	0,000		
Lojas Americanas (5)	8,010	9,898	2,601	1,423	0,000	
Ponto Frio (6)	8,616	10,200	3,824	2,353	0,447	0,000

ANÁLISE DE CLUSTERS

Etapas

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. Interpretação e validação dos agrupamentos

Etapa 3 - Seleção do algoritmo de agrupamento

- Envolve a escolha do **algoritmo de agrupamento** e a decisão quanto ao número de grupos
- **Algoritmo de agrupamento:** qual o procedimento deve ser usado para colocar objetos similares dentro de grupos?
 - Temos os métodos hierárquicos e os não-hierárquicos
- Todo **algoritmo** visa maximizar as diferenças entre os grupos em confronto com a variação dentro dos mesmos (*between-cluster x within-cluster*).

Etapa 3 - Seleção do algoritmo de agrupamento: Método hierárquico e Método não-hierárquico

1. Hierárquicos

- A aglomeração hierárquica se caracteriza pelo estabelecimento de uma hierarquia ou estrutura em forma de árvore, podendo ser aglomerativos ou divisivos

2. Não-hierárquicos (ex: k-means)

- Algoritmo não estabelece uma relação de hierarquia entre os sujeitos e os grupos
- Número inicial de clusters é definido pelo pesquisador

Procedimentos hierárquicos de agrupamento

- Envolvem a construção de uma hierarquia semelhante a uma árvore.
- São de dois tipos: ***aglomerativos*** e ***divisivos***.
- Algoritmos mais populares:
 - (1) *single linkage*
 - (2) *complete linkage*
 - (3) *average linkage*
 - (4) *Ward's method*
 - (5) *centroid method*

Processo de aglomeração

Resumo esquemático:

	Hierárquicos	Não-Hierárquicos
Processo de aglomeração	<i>single linkage</i> <i>complete linkage</i> <i>average linkage</i> <i>Ward</i> <i>centroid method</i>	<i>(K-means clustering)</i> <i>Sequential Threshold</i> <i>Parallel Threshold</i> <i>Optimization</i> <i>Selecting Seed Points</i>
	Divisivos	

Procedimentos hierárquicos de agrupamento

- 1) **Single linkage**: baseado na distância mínima entre dois objetos. É também chamado de *nearest neighbor*
- 2) **Complete linkage**: é baseado na distância máxima, razão pela qual é conhecido como a abordagem do vizinho mais longe (*furthest neighbor*)

Procedimentos hierárquicos de agrupamento

- 3) **Average linkage**: ou ligação média, onde o critério é a distância de todos os indivíduos de um grupo em relação a todos de outro. Tende a produzir grupos com aproximadamente a mesma variância
- 4) **Ward's method**: minimiza a soma dos quadrados entre dois grupos em relação a todas as variáveis. Tende a produzir grupos com mesmo número de observações

Procedimentos hierárquicos de agrupamento

5) Centroid method

- a distância entre os grupos é a distância entre seus centróides, que são os valores médios das observações em relação às variáveis.
- Cada vez que indivíduos são agrupados, um novo centróide é calculado.
- Tanto este método quanto o de *Ward* exigem a distância euclidiana.

Procedimentos não-hierárquicos de agrupamento

- não há relação de hierarquia entre os sujeitos e os grupos
- número inicial de clusters é definido pelo pesquisador
- têm por objetivo uma partição de **n** elementos em **k** grupos de modo que a partição atenda 2 requisitos:
 - coesão interna (ou semelhança interna)
 - isolamento (ou separação)
- não requerem cálculo e armazenamento de uma nova matriz de distância a cada processo
- reduzem tempo computacional
- possibilitam sua aplicação em grandes bases de dados

Procedimentos não-hierárquicos de agrupamento

- K-means
 - O mais popular
 - Produz apenas uma solução ao contrário do hierárquico que fornece uma série de soluções correspondentes a diferentes número de agrupamentos

Procedimentos não-hierárquicos de agrupamento

- K-means
 - Persegue-se o objetivo de minimização da variância interna dos grupos e maximização da variância entre os grupos
 - Pode-se fornecer informações sobre os centróides ou sementes iniciais
 - Se os centróides forem desconhecidos todas as observações são consideradas centróides
 - Verificação é feita pela ANOVA

Procedimentos não-hierárquicos de agrupamento

K-means – Passos:

1. escolhe-se o número **k** de conglomerados
2. assume-se inicialmente um centro de aglomeração (centróide) para cada grupo
3. calcula-se a distância euclidiana de cada sujeito aos centróides e coloca-se cada sujeito no cluster com centro de aglomeração mais próximo
4. quando todos os sujeitos estiverem alocados recalculam-se os novos centróides para cada grupo
5. repetem-se os passos 3 e 4 até os valores dos centróides dos grupos não mais variarem

Procedimentos não-hierárquicos de agrupamento

- **Métodos não-hierárquicos de agrupamento:** ou métodos de partição, atribuem objetos a um grupo uma vez que o número de grupos a ser formado esteja especificado. São referidos como *K-means clustering*
- Seleciona um grupo “semente” (*seed*) como grupo inicial, e todos os objetos próximos são incluídos nesse grupo. Um novo grupo semente é escolhido, e o processo continua até todas as observações serem distribuídas

Procedimentos não-hierárquicos de agrupamento

1) *Sequential threshold* ou princípio sequencial

- Seleciona um grupo semente e inclui todos os objetos dentro de uma distância preestabelecida.
- Após, um novo grupo semente é selecionado, e o processo continua. Quando um objeto é destinado a um grupo semente, ele não é mais considerado nos subsequentes.

Procedimentos não-hierárquicos de agrupamento

2) *Parallel threshold* ou princípio paralelo

- seleciona vários grupos semente e inclui todos os objetos dentro daquele mais próximo. À medida que o processo evolui, as distâncias podem ser ajustadas para incluir menos ou mais objetos

3) *Optimization*

- similar aos anteriores, exceto que ele permite a realocação de objetos em função da maior proximidade com outro grupo.

Vantagens e desvantagens dos métodos hierárquicos

Seleção dos grupos-semente:

- Pode ser aleatório ou escolhidos pelo pesquisador.
- Um dos problemas no primeiro caso é que o resultado final depende da ordem dos dados.

Vantagens e desvantagens dos métodos hierárquicos

(V) são rápidos e exigem menos tempo de processamento;

(V) podem realocar combinações anteriores;

(V) são menos sensíveis a outliers, à medida de distância e a variáveis inapropriadas (quando as ementes são escolhidas pelo pesquisador);

(D) os resultados dependem do processo de escolha dos pontos semente.

(D) não realocam combinações anteriores;

(D) é sensivelmente impactado por *outliers*;

(D) não são apropriados para amostras muito grandes

Combinação dos métodos de agrupamento

- Primeiro, uma técnica hierárquica estabelece o número de grupos, traça o perfil dos núcleos centrais e identifica *outliers*;
- Depois de eliminar eventuais *outliers*, aplica-se um método não-hierárquico, tendo como grupos sementes os núcleos centrais definidos através do método hierárquico.

ANÁLISE DE CLUSTERS

Etapas

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. **Escolha da quantidade de agrupamentos formados**
5. Interpretação e validação dos agrupamentos

Etapa 4 - Escolha da quantidade de agrupamentos formados

- Quantos grupos devem ser formados?
 - Não existe um critério categórico
 - Uma regra de parada (*stopping rule*) simples é examinar a distância entre os grupos a cada passo sucessivo;
 - Outra regra seria adaptar um teste estatístico de significância;
 - Além disso, o pesquisador deve confrontar com o referencial teórico, que pode sugerir um número natural de grupos;
 - Deve-se, ao final, buscar a melhor solução dentre as possíveis.

Etapa 4 - Escolha da quantidade de agrupamentos formados

- A análise de *clusters* deve ser estruturada novamente?
 - Analisar se existe um disparate acentuado entre o tamanho dos grupos, ou se existem grupos com uma ou duas observações (possíveis *outliers*);
 - Comparar a solução final com as expectativas do pesquisador;
 - Bussab refere uma técnica quantitativa para avaliação dos agrupamentos, o Coeficiente de Correlação Cofenética, que relaciona a matriz de distâncias originais com a oriunda da classificação (matriz cofenética); algo em torno de 0,8 já seria bom.

ANÁLISE DE CLUSTERS

Etapas

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. **Interpretação e validação dos agrupamentos**

Etapa 5 - Interpretação e validação dos agrupamentos

- Alguns procedimentos de validação da solução:
 - 1) dividir a amostra em dois grupos;
 - 2) usar outras variáveis conhecidas por discriminar entre os grupos, ou refazer a análise excluindo algumas variáveis;
 - 3) refazer a análise utilizando outros métodos de agrupamento e outras medidas de similaridade.

Etapa 5 - Interpretação e validação dos agrupamentos

- Definindo o perfil da solução: consiste na descrição das características de cada grupo para explicar como elas podem diferir em dimensões relevantes.
 - Utilizam-se dados não previamente incluídos no procedimento de agrupamento (demográficos, psicográficos etc.).
 - O enfoque é na descrição, não do que determinou diretamente os grupos, mas das características dos grupos depois de que eles foram identificados.
 - Pode-se utilizar a análise discriminante: a variável dependente categórica são os grupos.

Exemplos

Exemplo 1: Seria possível separar a amostra de empresas em grupos similares em termos de porte, representado pelas variáveis faturamento e número de empregados?

Empresas	Vendas (US\$ milhões)	Número empregados
Ferramentas Gerais	327,5	2.150
Fiori	312,2	661
Bretas Supermercados	652,6	7.200
Renner	929	7.764
Lojas Americanas	1.613,5	10.281
Ponto Frio	1.971	8.672

Etapa 1 - Análise das variáveis e dos objetos

Empresas	x= Vendas (z-score)	y =Nº de empregados (z-score)
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

Etapa 1 - Análise das variáveis e dos objetos

Padronização dos dados com z-score

Etapa 2 - cálculo das distâncias

Empresas	x= Vendas (z-score)	y =Nº de empregados (z-score)
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

Distância quadrática euclidiana = $d(\text{empresa}_1-\text{empresa}_2)^2 = ((x_1-x_2)^2 + (y_1-y_2)^2)^2$

Distância $(\text{empresa}_1-\text{empresa}_2)^2 = (-0,931-(-0,953))^2+(-1,038-(-1,427))^2 = 0,152$

Etapa 2 - cálculo das distâncias

Matriz de similaridade pela Distância Quadrática Euclidiana

	Ferramentas Gerais (1)	Fiori (2)	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
Ferramentas Gerais (1)	0,000					
Fiori (2)	0,152	0,000				
Bretas Supermercados (3)	1,964	3,163	0,000			
Renner (4)	2,916	4,248	0,183	0,000		
Lojas Americanas (5)	8,010	9,898	2,601	1,423	0,000	
Ponto Frio (6)	8,616	10,200	3,824	2,353	0,447	0,000

Etapa 3 - agrupamento

single linkage – menor distância

Procedimento hierárquico de agrupamento

	Ferramentas Gerais (1)	Fiori (2)	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
Ferramentas Gerais (1)	0,000					
Fiori (2)	0,152	0,000				
Bretas Supermercados (3)	1,964	3,163	0,000			
Renner (4)	2,916	4,248	0,183	0,000		
Lojas Americanas (5)	8,010	9,898	2,601	1,423	0,000	
Ponto Frio (6)	8,616	10,200	3,824	2,353	0,447	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	Ferramentas Gerais (1)	Fiori (2)	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
Ferramentas Gerais (1)	0,000					
Fiori (2)	0,152	0,000				
Bretas Supermercados (3)	1,964	3,163	0,000			
Renner (4)	2,916	4,248	0,183	0,000		
Lojas Americanas (5)	8,010	9,898	2,601	1,423	0,000	
Ponto Frio (6)	8,616	10,200	3,824	2,353	0,447	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
1+2	0,000				
Bretas Super- mercados (3)	1,964	0,000			
Renner (4)	2,916	0,183	0,000		
Lojas Americanas (5)	8,010	2,601	1,423	0,000	
Ponto Frio (6)	8,616	3,824	2,353	0,447	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
1+2	0,000				
Bretas Super- mercados (3)	1,964	0,000			
Renner (4)	2,916	0,183	0,000		
Lojas Americanas (5)	8,010	2,601	1,423	0,000	
Ponto Frio (6)	8,616	3,824	2,353	0,447	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
1+2	0,000				
Bretas Supermercados (3)	1,964	0,000			
Renner (4)	2,916	0,183	0,000		
Lojas Americanas (5)	8,010	2,601	1,423	0,000	
Ponto Frio (6)	8,616	3,824	2,353	0,447	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	3+4	Lojas Americanas (5)	Ponto Frio (6)
1+2	0,000			
3+4	1,964	0,000		
Lojas Americanas (5)	8,010	1,423	0,000	
Ponto Frio (6)	8,616	2,353	0,447	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	3+4	Lojas Americanas (5)	Ponto Frio (6)
1+2	0,000			
3+4	1,964	0,000		
Lojas Americanas (5)	8,010	1,423	0,000	
Ponto Frio (6)	8,616	2,353	0,447	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	3+4	5+6
1+2	0,000		
3+4	1,964	0,000	
5+6	8,010	1,423	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	3+4	5+6
1+2	0,000		
3+4	1,964	0,000	
5+6	8,010	1,423	0,000

Etapa 3 - agrupamento

single linkage – menor distância

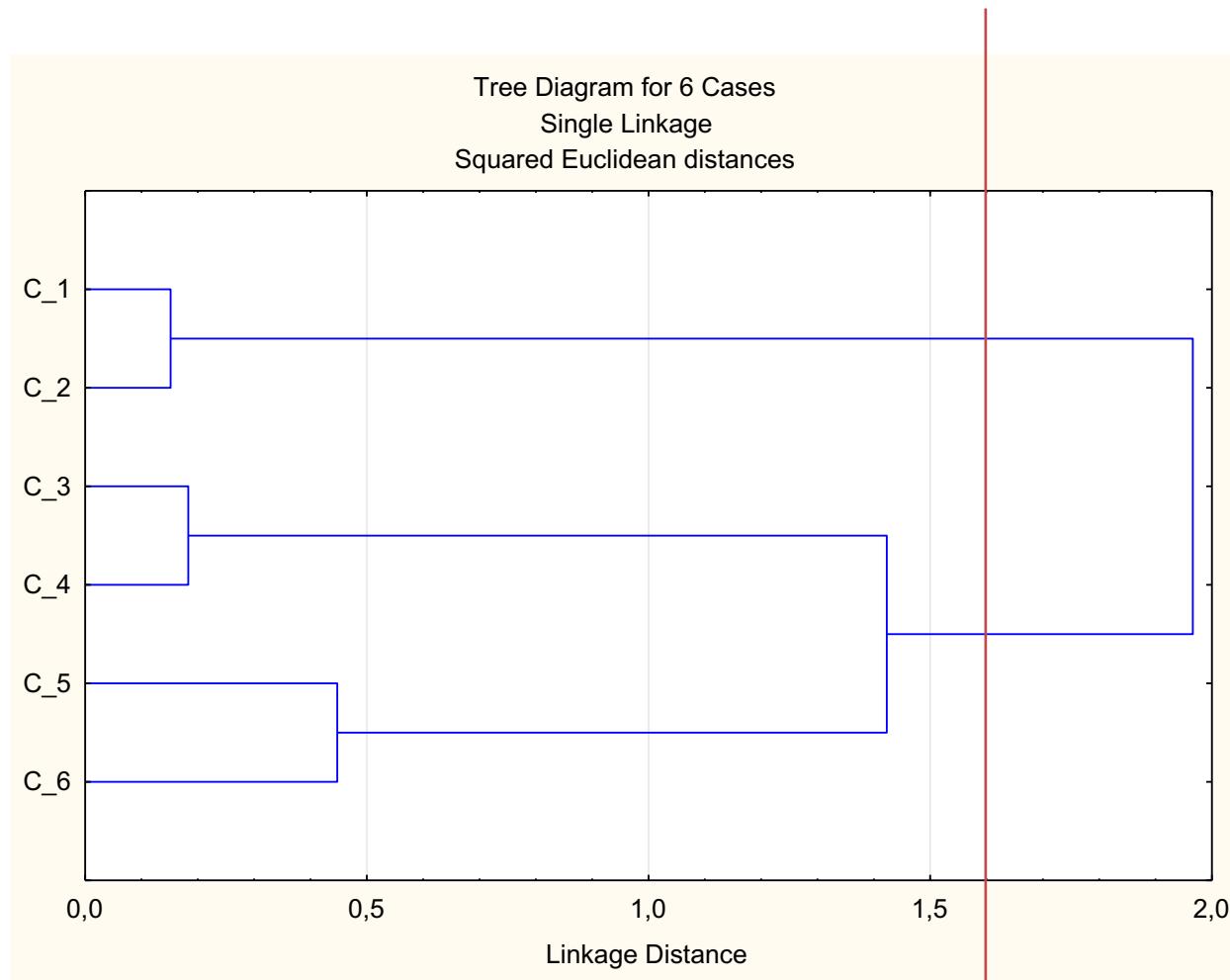
	1+2	3+4	5+6
1+2	0,000		
3+4	1,964	0,000	
5+6	8,010	1,423	0,000

Etapa 3 - agrupamento

single linkage – menor distância

	(1+2)	3+4+5+6
1+2	0,000	
3+4+5+6	1,964	0,000

Dendrogramma



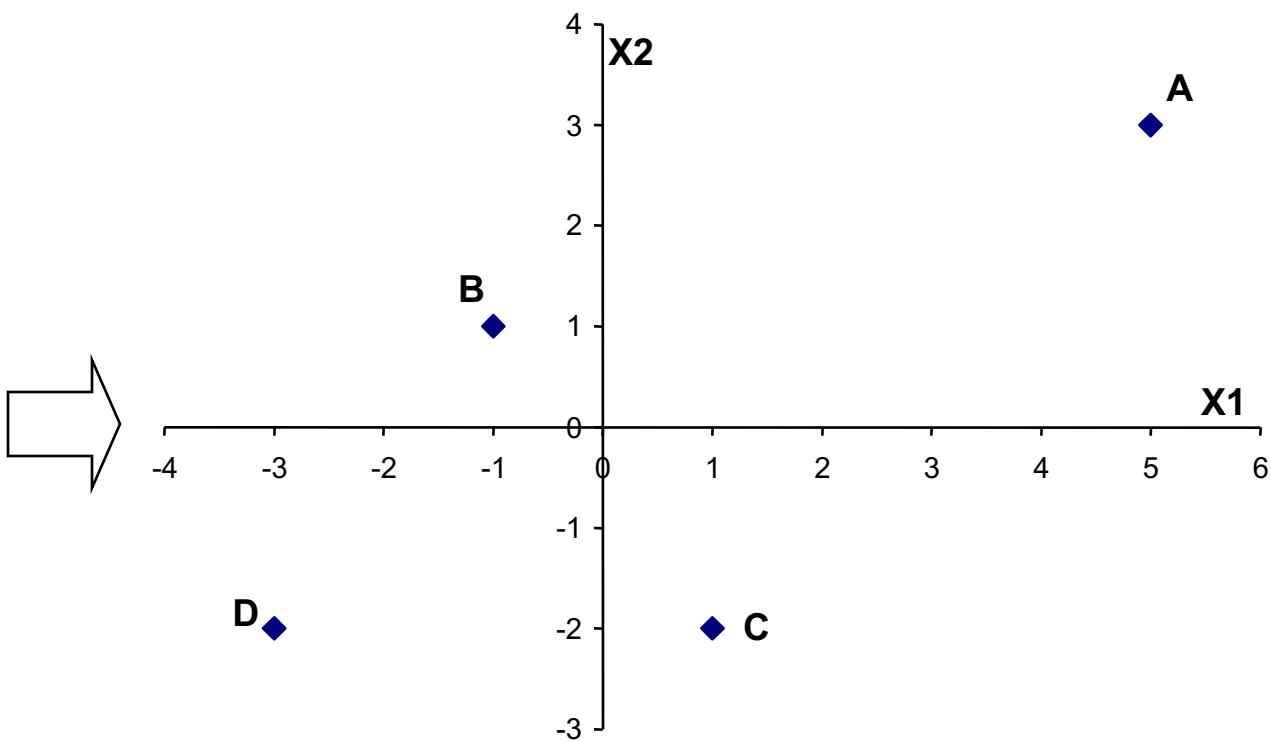
Exemplo 2 – *k-means*

K-Means

Exemplo

Dado o conjunto de 4 objetos ($n=4$), use o algoritmo k-Means para identificar 2 clusters ($k=2$)

Objetos	Coordenadas	
	$X1$	$X2$
A	5	3
B	-1	1
C	1	-2
D	-3	-2



K-Means

Exemplo

Centróides iniciais (seleção aleatória)

Centróide	Coordenadas dos centróides dos clusters	
	X1	X2
C1	2	2
C2	-1	-2

Lista de objetos

Objetos	Coordenadas	
	X1	X2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

K-Means

Exemplo

Centróides iniciais (seleção aleatória)

Centróide	Coordenadas dos centróides dos clusters	
	X1	X2
C1	2	2
C2	-1	-2

Lista de objetos

Objetos	Coordenadas	
	X1	X2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Alocação dos objetos aos clusters

$$\|A - C1\|^2 = (5-2)^2 + (3-2)^2 = 10$$

$$\|A - C2\|^2 = (5+1)^2 + (3+2)^2 = 61$$

A  Cluster com centróide C1

$$\|B - C1\|^2 = (-1-2)^2 + (1-2)^2 = 10$$

$$\|B - C2\|^2 = (-1+1)^2 + (1+2)^2 = 9$$

B  Cluster com centróide C2

$$\|C - C1\|^2 = (1-2)^2 + (-2-2)^2 = 5$$

$$\|C - C2\|^2 = (1+1)^2 + (-2+2)^2 = 4$$

C  Cluster com centróide C2

$$\|D - C1\|^2 = (-3-2)^2 + (-2-2)^2 = 39$$

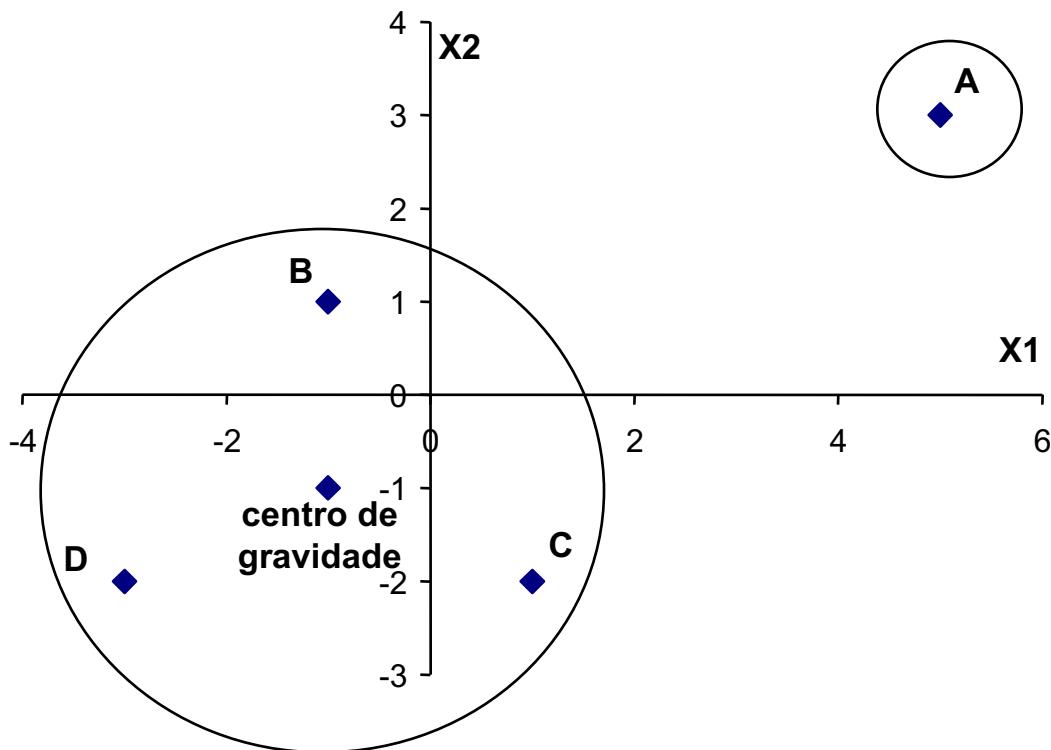
$$\|D - C2\|^2 = (-3+1)^2 + (-2+2)^2 = 4$$

D  Cluster com centróide C2

K-Means

Exemplo

Atualiza os centróides



O centro de gravidade de um cluster é a média dos seus objetos:

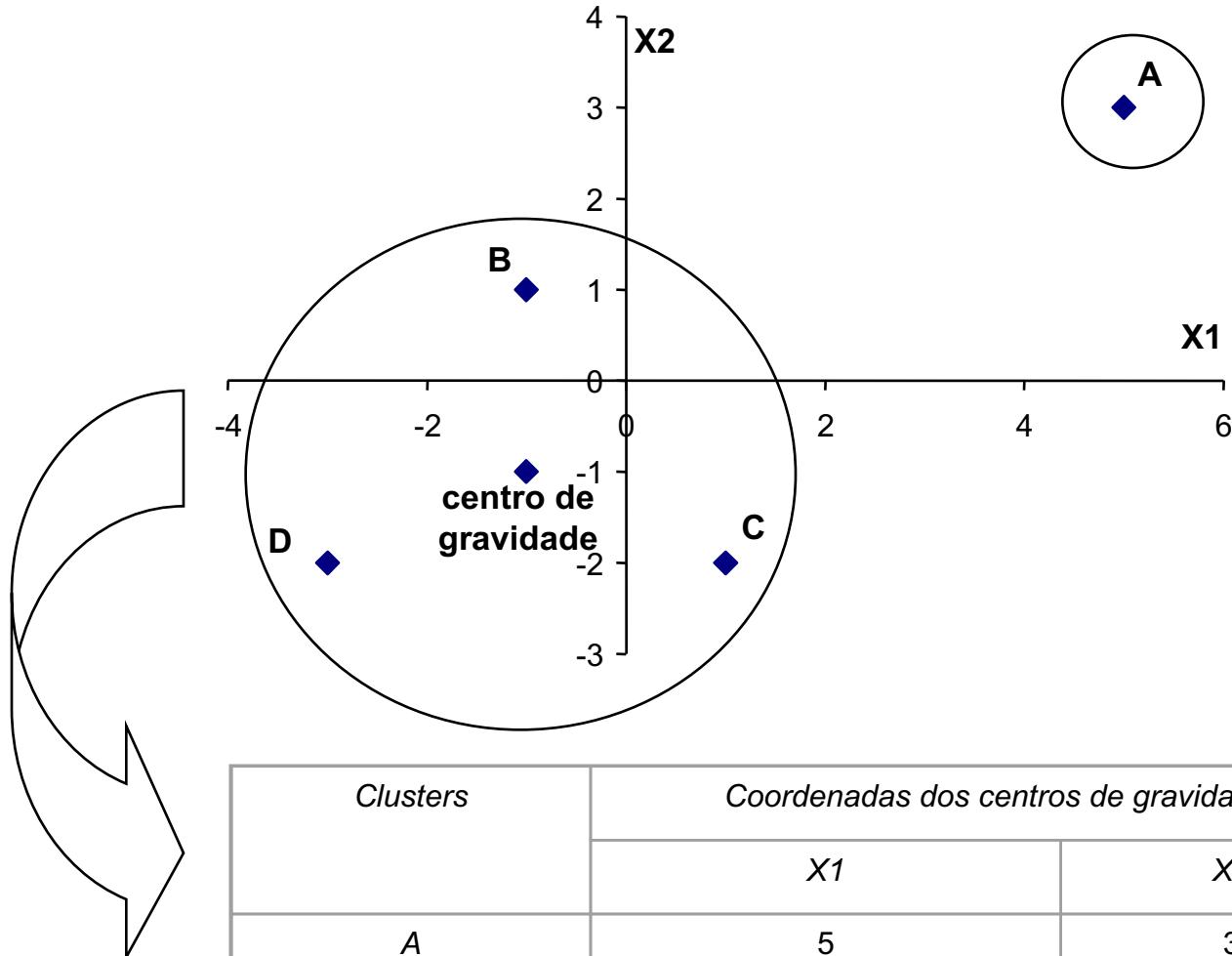
$$\text{Abscissa do centróide BCD} = (-1 + 1 - 3) / 3 = -1$$

$$\text{Ordenada do centróide BCD} = (1 - 2 - 2) / 3 = -1$$

K-Means

Exemplo

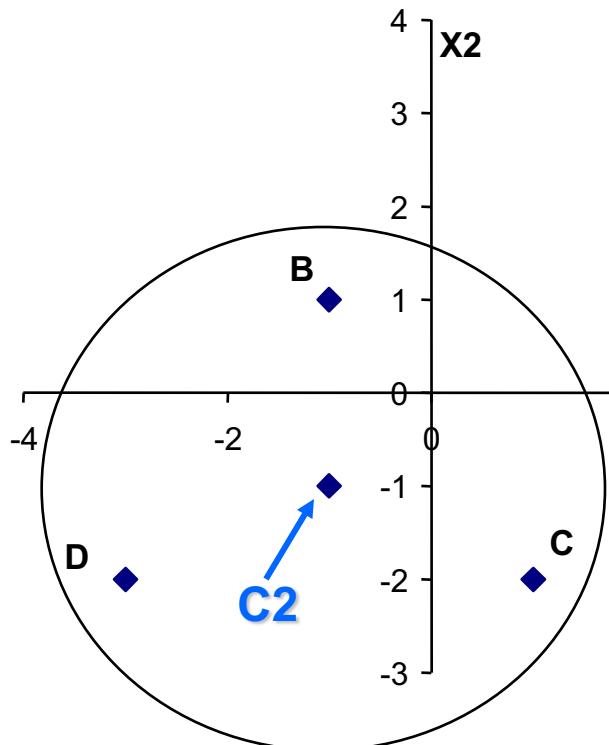
Atualiza os
centróides



Clusters	Coordenadas dos centros de gravidade	
	X_1	X_2
A	5	3
BCD	-1	-1

K-Means

Exemplo



Alocação dos objetos aos clusters

$$\|A - C_1\|^2 = (5-5)^2 + (3-3)^2 = 0$$

$$\|A - C_2\|^2 = (5+1)^2 + (3+1)^2 = 61$$

A \rightarrow Cluster com centróide C1

$$\|B - C_1\|^2 = (-1-5)^2 + (1-3)^2 = 40$$

$$\|B - C_2\|^2 = (-1+1)^2 + (1+1)^2 = 4$$

B \rightarrow Cluster com centróide C2

$$\|C - C_1\|^2 = (1-5)^2 + (-2-3)^2 = 39$$

$$\|C - C_2\|^2 = (1+1)^2 + (-2+1)^2 = 5$$

C \rightarrow Cluster com centróide C2

$$\|D - C_1\|^2 = (-3-2)^2 + (-2-2)^2 = 41$$

$$\|D - C_2\|^2 = (-3+1)^2 + (-2+1)^2 = 5$$

D \rightarrow Cluster com centróide C2

Não houve realocação de objetos, portanto, o algoritmo convergiu e dois clusters foram identificados: A e B,C,D

K-means

<https://www.youtube.com/watch?v=WqMnQuC19Rg>

A partir do minuto 5

Exemplo 3 – Padronização e medida de distância

Etapa 1 – Padronização

Tabela 1.1: Taxas de delitos por 100.000 habitantes por divisão territorial das polícias do Estado de São Paulo (Deinter), em 2002

Deinter	Homicídio doloso	Furto	Roubo	Roubo e furto de veículos
SJRP	10,85	1.500,80	149,35	108,38
RP	14,13	1.496,07	187,99	116,66
Bauru	8,62	1.448,79	130,97	69,98
Campinas	23,04	1.277,33	424,87	435,75
Sorocaba	16,04	1.204,02	214,36	207,06
SP	43,74	1.190,94	1.139,52	909,21
SJC	25,39	1.292,91	358,39	268,24
Santos	42,86	1.590,66	721,90	275,89
Média	23,08	1.375,19	415,92	298,90
DP	13,69	152,05	351,62	273,35

fonte: Secretaria de Segurança Pública do Estado de São Paulo

<http://www.ssp.sp.gov.br/estatisticas/criminais/>, acessada em 11/02/2003.

SJRP: São José do Rio Preto

RP: Ribeirão Preto

SP: São Paulo (capital)

SJC: São José dos Campos

Etapa 1 - Padronização

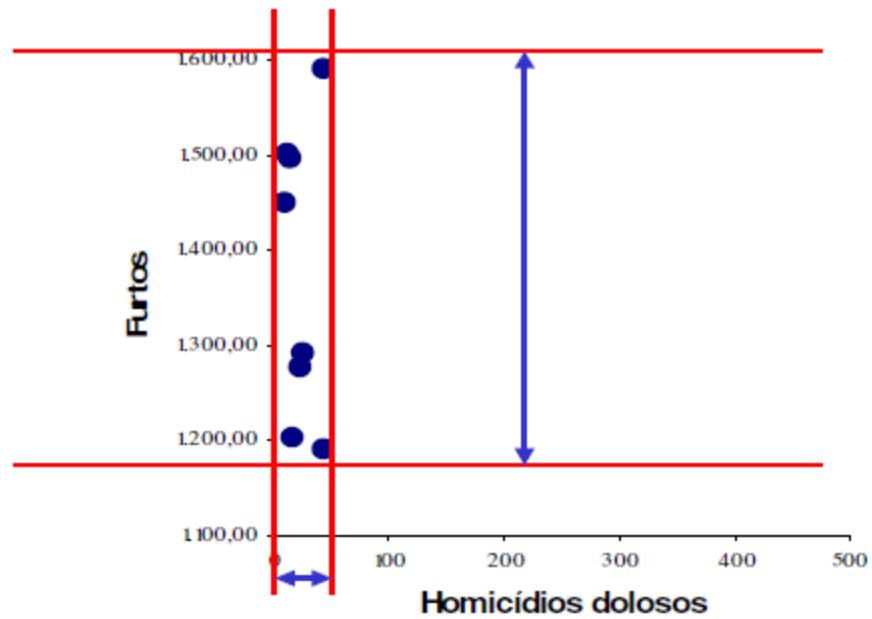


Figura 1.3: Diagrama de dispersão das Deinter's

Etapa 1 - Padronização

Tabela 1.2: Taxas de delitos por 100.000 habitantes padronizadas

Deinter	Homicídio Doloso	Furto
SJRP	-0,89	0,83
RP	-0,65	0,80
Bauru	-1,06	0,48
Campinas	0,00	-0,64
Sorocaba	-0,51	-1,13
SP	1,51	-1,21
SJC	0,17	-0,54
Santos	1,44	1,42
Média	0,00	0,00
DP	1,00	1,00

fonte: Secretaria de Segurança Pública do Estado de São Paulo

$$z\text{-score} = (10,85 - 23,08) / 13,69 = -0,8933$$

Etapa 1 – Dados padronizados

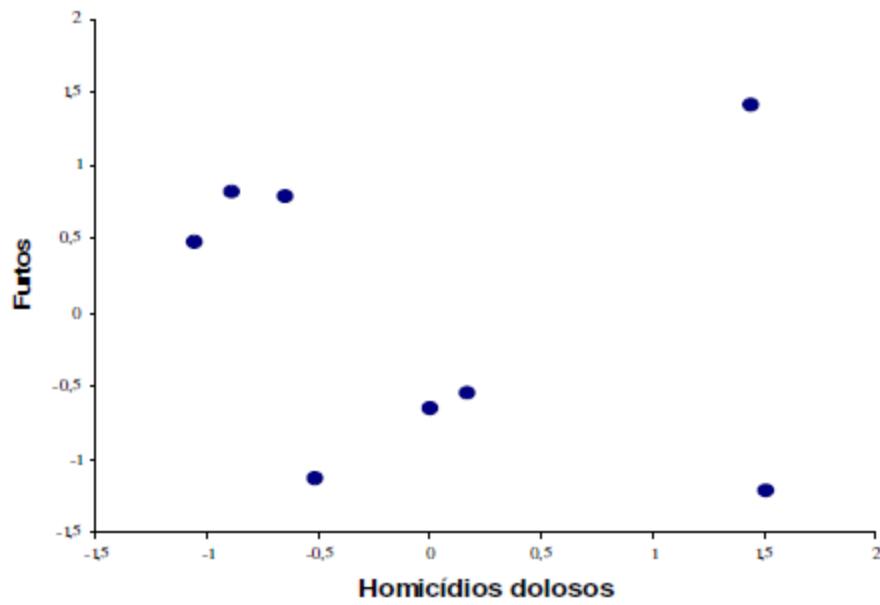


Figura 1.4: Diagrama de dispersão das Deinter's - dados padronizados

Etapa 1 – Ranges efetivos

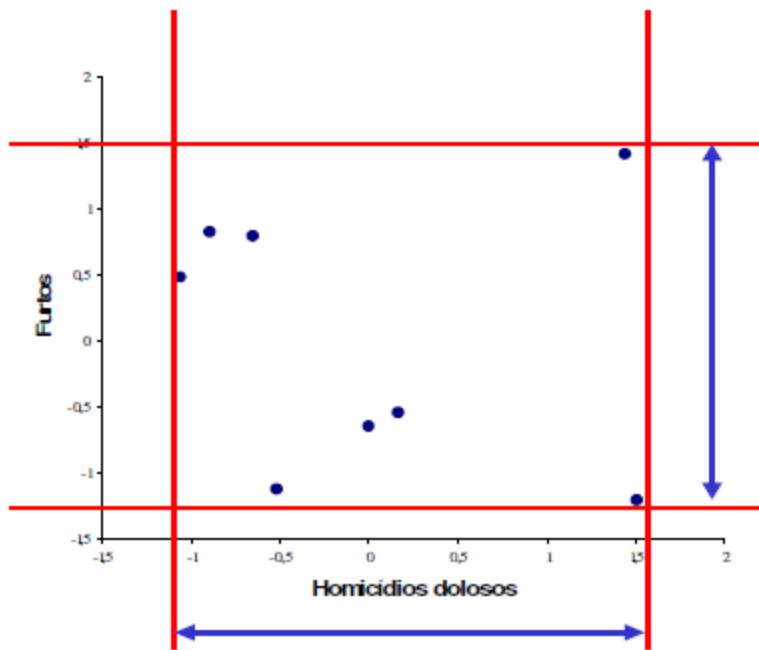


Figura 1.5: Diagrama de dispersão das Deinter's - dados padronizados

Etapa 2 - Medida de distância entre cada par de objetos

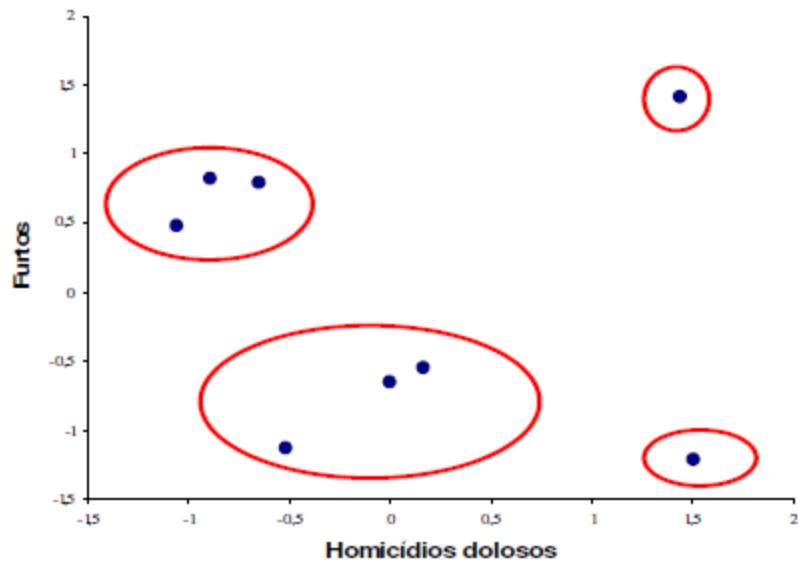


Figura 1.6: Diagrama de dispersão das Deinter's - dados padronizados

Etapa 2 - Seleção da medida de distância

Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Tabela 1.10: Matriz de Distâncias

Deinter	SJRP	RP	Bauru	Campinas	Sorocaba
SJRP	0,00				
RP	0,59	0,00			
Bauru	0,55	1,05	0,00		
Campinas	2,74	2,27	2,89	0,00	
Sorocaba	2,37	2,17	2,24	1,37	0,00

Etapa 2 - Seleção da medida de distância

Tabela 1.11: Matriz de Distâncias (Passo 1)

Deinter	SJRP,Bauru	RP	Campinas	Sorocaba
SJRP, Bauru	0,00			
RP	1,05	0,00		
Campinas	2,89	2,27	0,00	
Sorocaba	2,37	2,17	1,37	0,00

Tabela 1.12: Matriz de Distâncias (Passo 2)

Deinter	SJRP, Bauru, RP	Campinas	Sorocaba
SJRP, Bauru, RP	0,00		
Campinas	2,89	0,00	
Sorocaba	2,37	1,37	0,00

Etapa 2 - Seleção da medida de distância

Tabela 1.13: Resumo do procedimento

Passo	Grupo	Distância
1	SJRP, Bauru	0,55
2	SJRP, Bauru, RP	1,05
3	Campinas, Sorocaba	1,37
4	SJRP, Bauru, RP, Campinas, Sorocaba	2,89

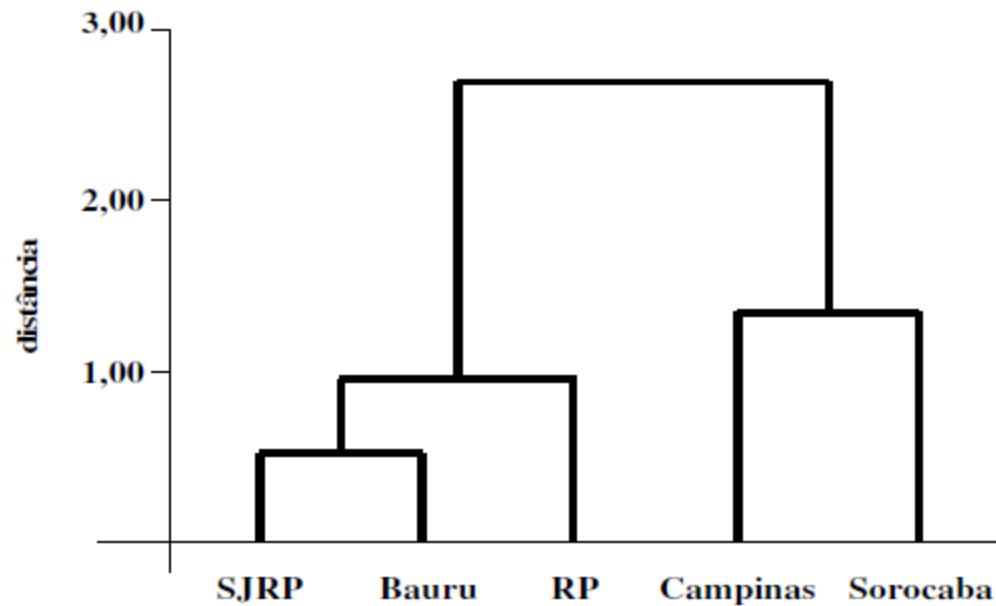


Figura 1.8: Dendrograma - Método do Vizinho mais Longe

FIM