Análise de dados multivariados I

Tópico:

Análise de Conglomerados ou Agrupamentos (ou ainda, Cluster Analysis)

Bibliografia:

- R.A. **Johnson**, Applied Multivariate Statistical Analysis, Prentice Hall, 1992
- L.J. **Corrar**; E. **Paulo**; J.M. **Dias Filho**, Analise multivariada ..., Atlas, 2007
- L.P. **Favero**; et al, Análise de dados - modelagem multivariada ..., Campus, 2009.

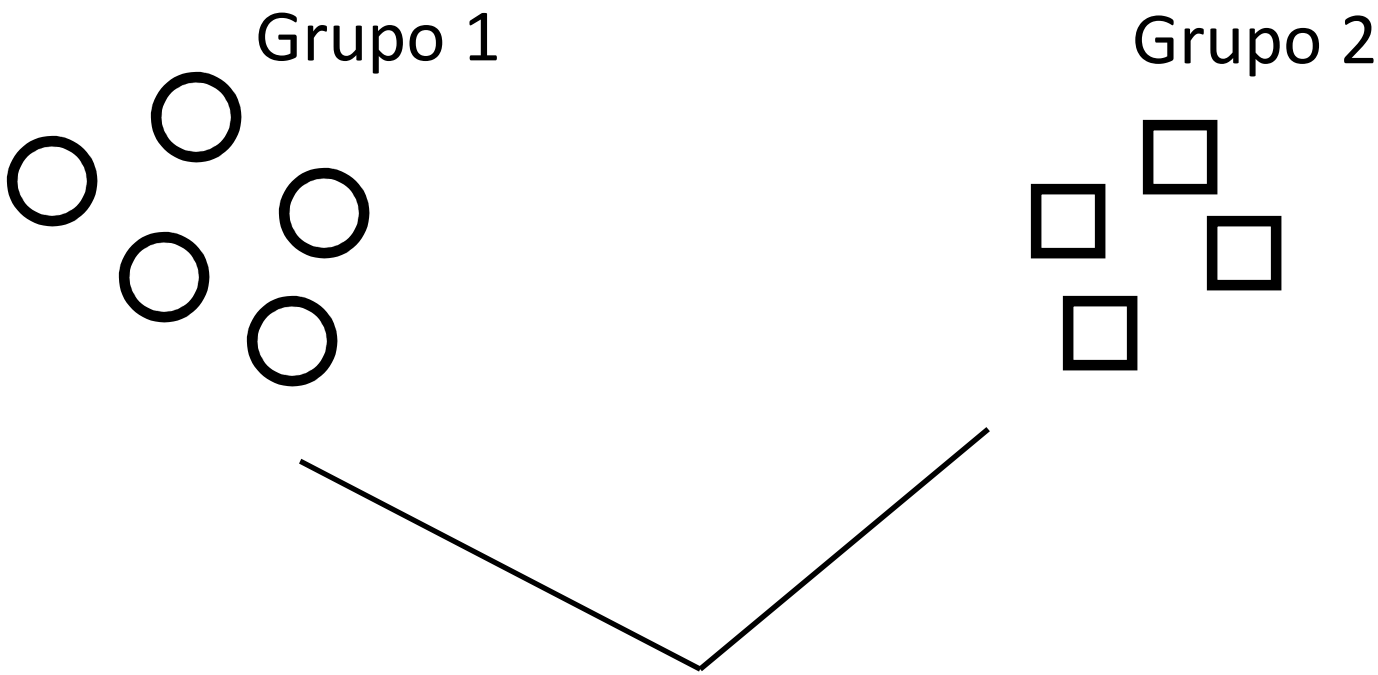
Análise de Agrupamentos

Análise de agrupamentos (cluster): Objetivo dessa técnica é agrupar objetos (itens) semelhantes segundo suas características (variáveis).

Embora vamos considerar **agrupamentos de objetos**, não existem impedimentos para realizar o agrupamento de variáveis semelhantes segundo as realizações obtidas pelos objetos amostrados. Análise fatorial se mostra mais robusta para o agrupamento de variáveis.

Análise de Agrupamentos

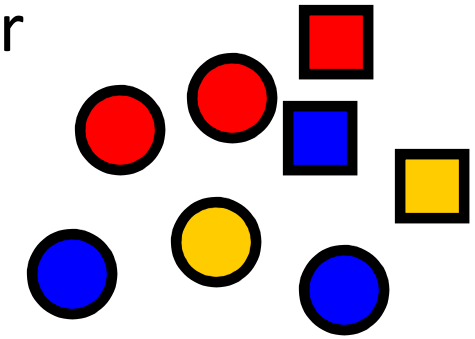
Finalidade: Encontrar grupos de modo a
- maximizar a homogeneidade dentro dos grupos
- maximizar a heterogeneidade entre os grupos



Dois grupos classificados pela forma

Análise de Agrupamentos

Formar os agrupamentos não é tão simples, pois os objetos são, em geral, definidos por várias características



Possibilidades:

2 grupos:



3 grupos:



A formação dos agrupamentos vai depender da definição de **"similaridade"**.

Análise de Agrupamentos - Aplicações

- Identificar grupos de investimento de acordo com perfis de risco;

- Identificar segmentos homogêneos de consumidores, a fim de estabelecer programas específicos para cada público;

- Agrupar empresas com base em indicadores financeiros;

- …
- Exemplo: Häagen-Dazs (Malhotra 4ª Ed, Cap. 20, p. 571)

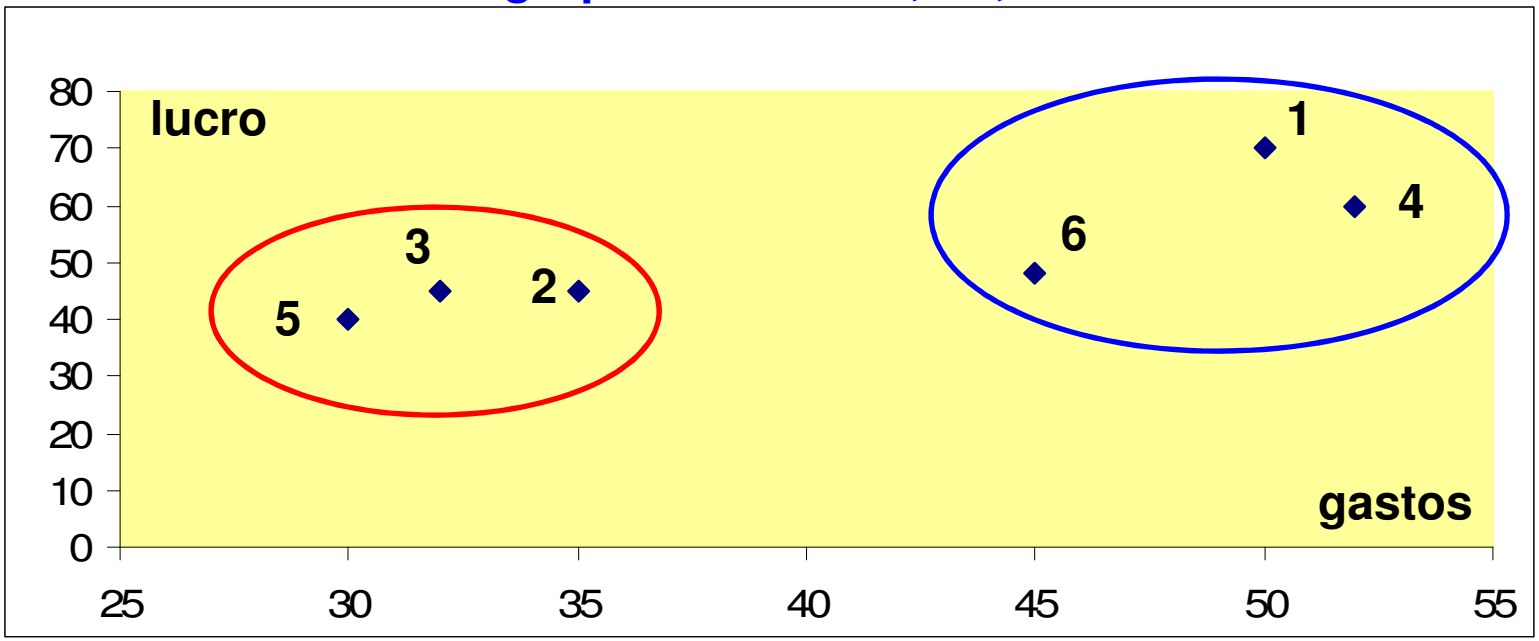
Análise de Agrupamentos – Qual critério?

Poderíamos procurar grupos graficamente.

	gastos	lucro
empresa1	50	70
empresa2	35	45
empresa3	32	45
empresa4	52	60
empresa5	30	40
empresa6	45	48

Com duas variáveis poderíamos fazer um diagrama de dispersão e agrupar baseado na distância entre pontos.

Agrupamento 1: E2, E3, E5
Agrupamento 2: E1, E4, E6

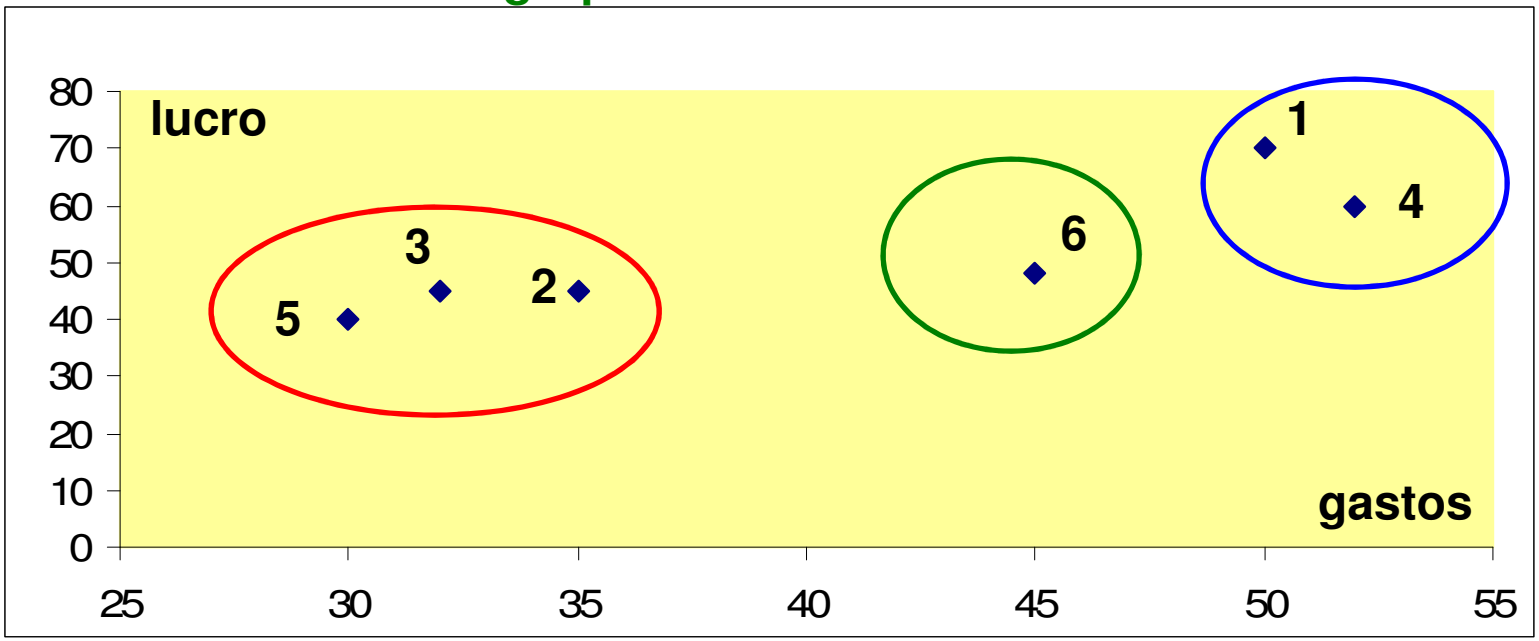


Análise de Agrupamentos – Quantos grupos?

	gastos	lucro
empresa1	50	70
empresa2	35	45
empresa3	32	45
empresa4	52	60
empresa5	30	40
empresa6	45	48

Poderíamos formar três grupos:

Agrupamento 1: E2, E3, E5
Agrupamento 2: E1, E4
Agrupamento 3: E6



Etapas de uma análise de conglomerados

- Análise das variáveis e dos objetos a serem agrupados (seleção das variáveis, identificação de outliers e padronização de variáveis).

- Seleção da medida de **distância** ou **semelhança** entre cada par de objetos.

- Seleção do algoritmo de agrupamento: método **hierárquico** ou **não-hierárquico**.

- Escolha da quantidade de agrupamentos formados interpretação e validação dos agrupamentos.

Variáveis e objetos a serem agrupados

A estrutura que os grupos refletem dependem da seleção das variáveis.

A técnica de análise de agrupamentos não distingue se as variáveis são ou não relevantes (**Essa tarefa fica a cargo do pesquisador)**. Sugestões:

- Pesquisas passadas
- Teoria
- Hipóteses que estão sendo testadas
- Presença do julgamento e intuição do pesquisador

A técnica é sensível a observações atípicas (outliers?). Observações atípicas podem formar grupos isolados, que podem ser interesse do pesquisador.

Padronização das variáveis

- Elimina viés introduzido por diferenças de escala
- Z_scores: $$Z = \frac{x - \bar{x}}{s}$$
- Recomenda-se quando há diferentes escalas entre as variáveis
- Tem impacto substancial na solução final: padronização minimiza diferenças de grupo, pois se grupos são bem separados por uma dada variável Xi, a variância dessa variável será grande.

Medida de similaridade e de dissimilaridade

Precisamos medir correspondências entre objetos. Ou seja medidas de similaridade e de dissimilaridade.

Os métodos mais usados são:

- Medidas de distância (**dissimilaridade**)
- Medidas de correlação (**similaridade**)
- Medidas de associação (**dissimilaridade** ou **similaridade** com variáveis nominais)

Dissimilaridade

Note que estamos interessados em agrupar objetos. As medidas de distância são entre dois objetos (itens).

objetos

	X_1	X_2	...	X_p
Item 1	x_{11}	x_{12}	...	x_{1p}
Item 2	x_{21}	x_{22}	...	x_{2p}
Item 3	x_{31}	x_{32}	...	x_{3p}
...	...	...	...	...
Item n	x_{n1}	x_{n2}	...	x_{np}

Medidas de distância

Medidas de distância são consideradas medidas de dissimilaridade pois quanto maior a distância, menor a semelhança entre objetos, e vice-versa.

Medidas de distância $\mathbf{d_{ij}}$ entre o item **i** e o item **j**

	X_1	X_2	...	X_p
Item 1	x_{11}	x_{12}	...	x_{1p}
...	x_{21}	x_{22}	...	x_{2p}
Item i	x_{i1}	x_{i2}	...	x_{ip}
...	...	...	...	...
Item j	x_{j1}	x_{j2}	...	x_{jp}
...	...	...	...	...
Item n	x_{n1}	x_{n2}	...	x_{np}

Podemos obter $\mathbf{d_{ij}}$ a partir das variáveis.

Como?

Algumas medidas de distância

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

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

Distância quadrática euclidiana : $d_{ij}^{\,2} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$

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

Distância euclidiana padronizada : $d_{ij} = \sqrt{\sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{\sqrt{s_{kk}}} \right)^2}$

Algumas medidas de distância

$$\text{Distância de Minkowski: } d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{n} \right)^{\frac{1}{n}}$$

$n = 1$: é conhecida por "métrica do quarteirão" (métrica city-block)

$n = 2$: representa a distância euclidiana

variações de n causam trocas nos pesos dados a pequenas e a grandes diferenças.

$$\text{Distância } \textit{de Mahalanobis}: \quad d_{ij} = \sqrt{(x_i - x_j)' \, S^{-1} (x_i - x_j)}$$

$$\textit{onde}: \quad x'_i = [\, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip} \,]$$

$$x'_j = [\, x_{j1}, x_{j2}, x_{j3}, \ldots, x_{jp} \,]$$

$$S^{-1} = \textit{inversa da matriz de covariância amostral}$$

Similaridade: Medidas correlacionais

$$\text{Correlação de Pearson : } r_{ij} = \frac{\sum_{k=1}^{p}\left(x_{ik} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_j\right)}{\sqrt{\sum_{k=1}^{p}\left(x_{ik} - \bar{x}_i\right)^2 \sum_{k=1}^{p}\left(x_{jk} - \bar{x}_j\right)^2}}$$

onde : x_{ik} = valor da variável k para o objeto i

x_{jk} = valor da variável k para o objeto j

$\bar{x}_i$ = média das variáveis para o objeto i

$\bar{x}_j$ = média das variáveis para o objeto j

p = número de variáveis

O valor de r varia entre -1 e 1, sendo que o zero significa que não há associação. Quanto maiores as correlações, mais próximas as observações.

Exemplo

Um exemplo: vamos encontrar a **matriz de distâncias** para o seguinte exemplo:

	gasto	lucro
Empresa 1	**50**	**70**
Empresa 2	**35**	**45**
Empresa 3	**32**	**45**

Vamos utilizar a distância quadrática:

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

Temos:

$$\text{distância entre empresa 1 e empresa 2}: d_{12}^{\,2} = (50-35)^2 + (70-45)^2 = 850$$

$$\text{distância entre empresa 1 e empresa 3}: d_{13}^{\,2} = (50-32)^2 + (70-45)^2 = 949$$

$$\text{distância entre empresa 2 e empresa 3}: d_{23}^{\,2} = (35-32)^2 + (45-45)^2 = 9$$

Matriz de distância para o exemplo

$$D = \begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{pmatrix} = \begin{pmatrix} 0 & 850 & 949 \\ 850 & 0 & 9 \\ 949 & 9 & 0 \end{pmatrix}$$

Propriedades observadas no exemplo:

-elementos nulos na diagonal (distância entre um mesmo objeto é nula)

-Matriz simétrica (distância de i até j = distância de j até i)

Métodos hierárquico e não-hierárquico

Os agrupamentos hierárquicos: são realizados por sucessivas fusões ou por sucessivas divisões.

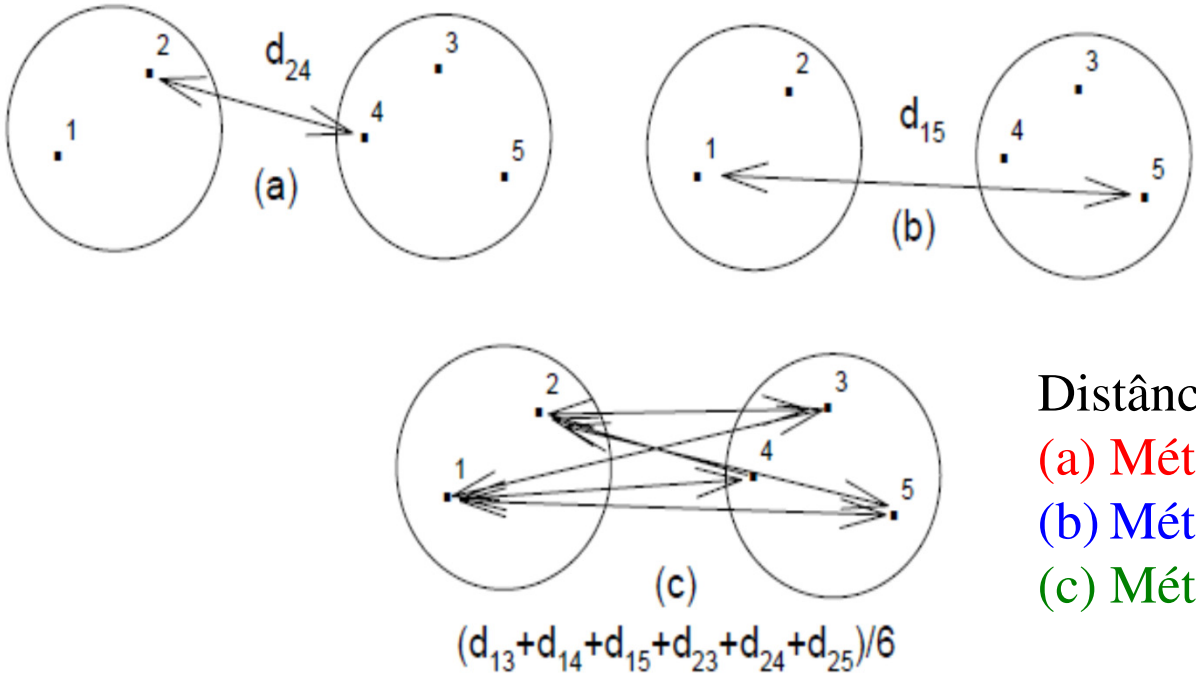
métodos hierárquicos aglomerativos: iniciam com tantos grupos quanto aos objetos, ou seja, cada objeto forma um agrupamento. Inicialmente, os objetos mais similares são agrupados e fundidos formando um único grupo. Eventualmente o processo é repetido, e com o decréscimo da similaridade, todos os subgrupos são fundidos, formando um único grupo com todos os objetos.

Os métodos hierárquicos divisivos: Um único subgrupo inicial existe com todos os objetos e estes são subdivididos em dois subgrupos de tal forma que exista o máximo de semelhança entre os objetos dos mesmos subgrupos e a máxima dissimilaridade entre elementos de subgrupos distintos. Estes subgrupos são posteriormente subdivididos em outros subgrupos dissimilares. O processo é repetido até que haja tantos subgrupos quantos objetos.

Consideraremos agrupamentos **hierárquicos aglomerativos** ("Linkage Methods")

Vamos considerar:

- Método de ligação simples (mínima distância ou vizinho mais próximo),
- Método ligação completa (máxima distância ou vizinho mais distante) e
- Método da ligação média (distância média).



Distâncias entre os grupos para:
(a) Método da ligação simples,
(b) Método da ligação completa e
(c) Método da ligação média.

Algoritmo para grupamentos hierárquicos aglomerativos com n objetos (itens ou variáveis).

1. Iniciar com n grupos, cada um com um único elemento e com uma matriz simétrica n x n de dissimilaridades (distâncias) $D=\{d_{U,V}\}$.

2. Buscar na matriz D o par de grupos mais similar (menor distância) e fazer a distância entre os grupos mais similares U e V igual a $d_{u\,v}$.

3. Fundir os grupos U e V e nomeá-lo por (UV). Recalcular e rearranjar as distâncias na matriz D (a) eliminando as linhas e colunas correspondentes a U e V e (b) acrescentando uma linha e coluna com as distâncias entre o grupo (UV) e os demais grupos.

4. Repetir os passos 2 e 3 num total de (n-1) vezes (todos os objetos estarão em único grupo). Anotar a identidade dos grupos que vão sendo fundidos e os respectivos níveis (distâncias) nas quais isto ocorre.

Método: Ligação simples (Single Linkage, Nearest Neighbor)

No passo 3 será usado: $d_{(UV),W} = \min\{ d_{U,W}, d_{V,W}\}$.
Vamos utilizar um exemplo onde temos 4 objetos (A, B, C, D) com a seguinte matriz de distância:

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

Os objetos menos distantes devem, inicialmente, ser agrupados.

Mímina distância entre i e j = $\min(d_{i,j}) = d_{A,B} = 3$.

Próximo passo: juntar A com B formando o grupo (AB) e em seguida calcular as distâncias deste grupo e os objetos remanescentes.

Método: Ligação simples (Single Linkage, Nearest Neighbor)

As distâncias dos vizinhos mais próximos são:

$d_{(AB),C} = \min\{ d_{A,C}, d_{B,C} \} = \min\{7, 9\} = 7$

$d_{(AB),D} = \min\{ d_{A,D}, d_{B,D} \} = \min\{8, 6\} = 6$

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

E obtemos a nova matriz D para o próximo passo:

$$D = \begin{array}{c} \\ AB \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & D \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

A menor distância é entre D e C, $d_{D,C} = 5$. Então DC formam um grupo no nível (distância) 5.

Método: Ligação simples (Single Linkage, Nearest Neighbor)

Recalculando a distância:

$d_{(DC),(AB)} = \min\{ d_{D,(AB)}, d_{C,(AB)} \}$

$= \min\{6, 7\} = 6$

$$D = \begin{array}{c} \\ AB \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & D \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

E a nova matriz D será:

$$D = \begin{array}{c} \\ DC \\ AB \end{array} \begin{array}{c} \begin{array}{cc} DC & AB \end{array} \\ \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array}$$

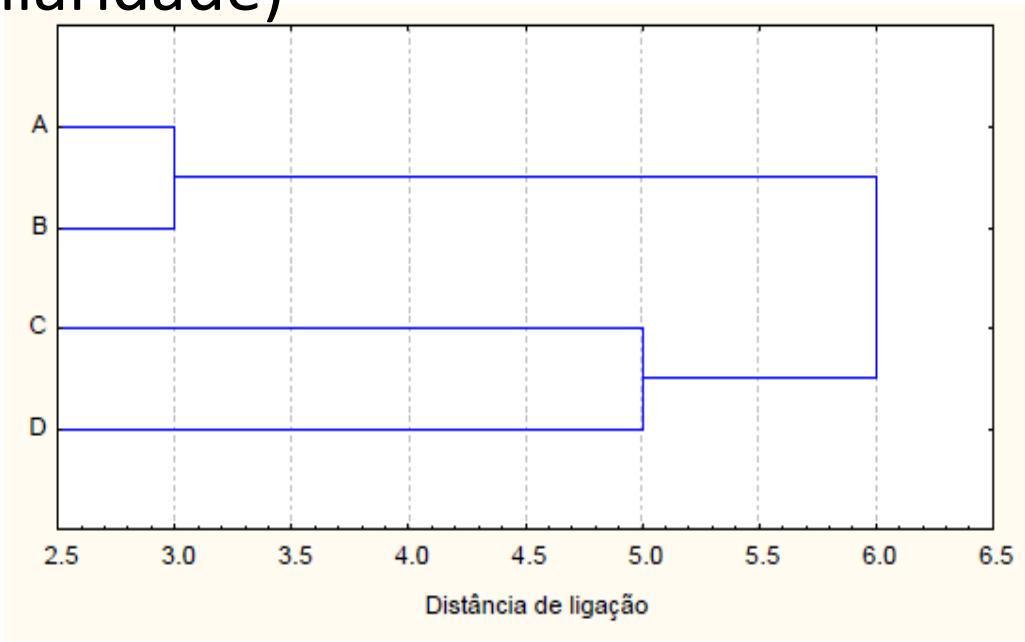
O grupo DC é agrupado com o grupo AB na distância 6

Método: Ligação simples (Single Linkage, Nearest Neighbor)

Dendrograma (ou diagrama de árvore):

Os agrupamentos são mostrados graficamente. Os ramos representam os agrupamentos e os nodos representam os pontos onde os agrupamentos se fundem ao longo do eixo de distâncias (ou similaridade)

No exemplo visto:



Dendrograma para agrupar 4 objetos (A, B, C e D) pelo método da ligação simples (vizinho mais próximo).

Ligação Completa (Complete Linkage, Furthest Neighbor)

Parecido com o anterior, mas a distância entre grupos é tomada como a "máxima" distância entre dois elementos de cada grupo.

No passo 3 é usado: $d_{(UV),W} = \max\{ d_{U,W}, d_{V,W}\}$

Vamos utilizar o mesmo exemplo com 4 objetos (A, B, C, D) e matriz de distância:

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

Os objetos menos distantes devem, inicialmente, ser agrupados.

Mímina distância entre i e j = $\min(d_{i,j}) = d_{A,B} = 3$.

Próximo passo: juntar A com B formando o grupo (AB) e em seguida calcular as distâncias deste grupo e os objetos remanescentes.

Ligação Completa (Complete Linkage, Furthest Neighbor)

As distâncias dos vizinhos mais distantes são:

$d_{(AB),C} = \max\{ d_{A,C}, d_{B,C} \} = \max\{7, 9\} = 9$

$d_{(AB),D} = \max\{ d_{A,D}, d_{B,D} \} = \max\{8, 6\} = 8$

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

E obtemos a nova matriz D para o próximo passo:

A menor distância é entre D e C, $d_{D,C} = 5$.
Então DC formam um grupo no nível 5.
Recalculando a distância:

$$D = \begin{array}{c} \\ AB \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & D \end{array} \\ \begin{bmatrix} 0 & & \\ 9 & 0 & \\ 8 & 5 & 0 \end{bmatrix} \end{array}$$

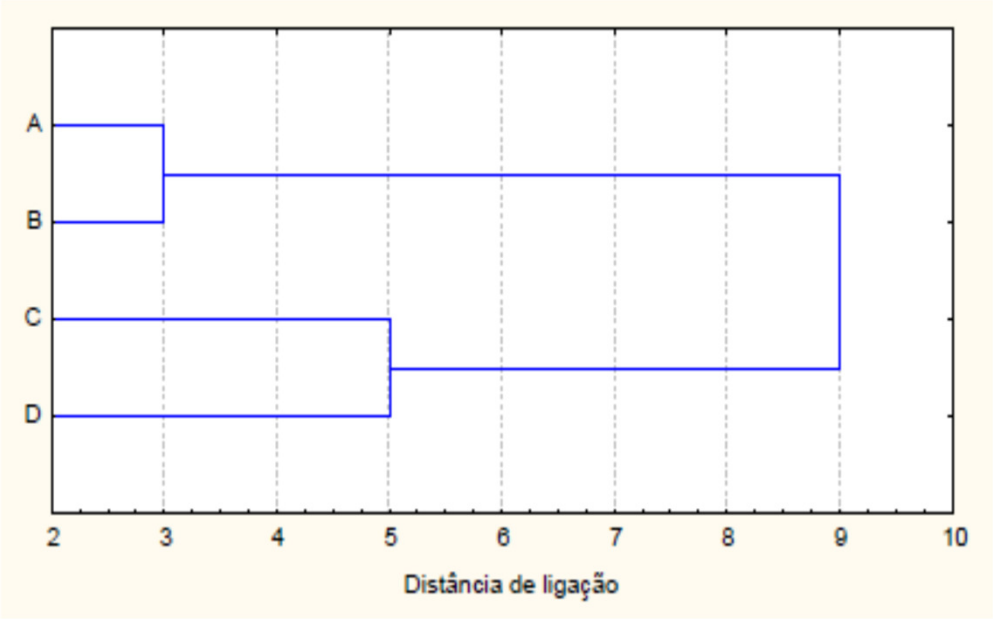
$d_{(DC),(AB)} = \max\{ d_{D,(AB)}, d_{C,(AB)} \}$
$= \min\{8, 9\} = 9$

Ligação Completa (Complete Linkage, Furthest Neighbor)

E a nova matriz D será:

$$D = \begin{array}{c} \\ DC \\ AB \end{array} \begin{array}{c} \begin{array}{cc} DC & AB \end{array} \\ \begin{bmatrix} 0 & \\ 9 & 0 \end{bmatrix} \end{array}$$

O grupo DC é agrupado com o grupo AB a distância 9



Dendrograma para agrupar 4 objetos (A, B, C e D) pelo método da ligação completa (vizinho mais distânte).

Ligação Média (Average Linkage, Between Groups)

Análogo aos anteriores, com exceção de que a distância entre grupos é tomada como a média da distância entre dois elementos de cada grupo.

No passo 3 é usado: $d_{(UV),W}$ = média{ $d_{U,W}$, $d_{V,W}$}

Vamos utilizar o mesmo exemplo com 4 objetos (A, B, C, D) e matriz de distância:

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

Os objetos menos distantes são agrupados.

Mímina distância entre i e j

= min($d_{i,j}$) = $d_{A,B}$ = 3 .

Próximo passo: juntar A com B formando o grupo (AB) e em seguida calcular as distâncias deste grupo e os objetos remanescentes.

Ligação Média (Average Linkage, Between Groups)

As distâncias entre os grupos são baseadas na média das distâncias entre todos os elementos de um grupo com relação ao elementos de outro grupo

$d_{(AB),C} = (d_{A,C} + d_{B,C})/2\} = (7 + 9)/2 = 8$

$d_{(AB),D} = (d_{A,D} + d_{B,D}) = (8+6)/2 = 7$

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

E obtemos a nova matriz D para o próximo passo:

A menor distância é entre D e C, $d_{D,C} = 5$.
Então DC formam um grupo no nível 5.

$$D = \begin{array}{c} \\ AB \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccc} AB & C & D \end{array} \\ \begin{bmatrix} 0 & & \\ 8 & 0 & \\ 7 & 5 & 0 \end{bmatrix} \end{array}$$

Recalculando a distância:

$d_{(DC),(AB)} = (d_{D,(AB)} + d_{C,(AB)})/2 = (7+8)/2 = 7,5$

Ligação Média (Average Linkage, Between Groups)

E a nova matriz D será:

$$D = \begin{array}{c} \\ DC \\ AB \end{array} \begin{array}{c} \begin{array}{cc} DC & AB \end{array} \\ \begin{bmatrix} 0 & \\ 7{,}5 & 0 \end{bmatrix} \end{array}$$

O grupo DC é agrupado com o grupo AB a distância 7,5



Dendrograma para agrupar 4 objetos (A, B, C e D) pelo método da ligação média.

Diferentes Métodos

Note que os métodos podem produzir:

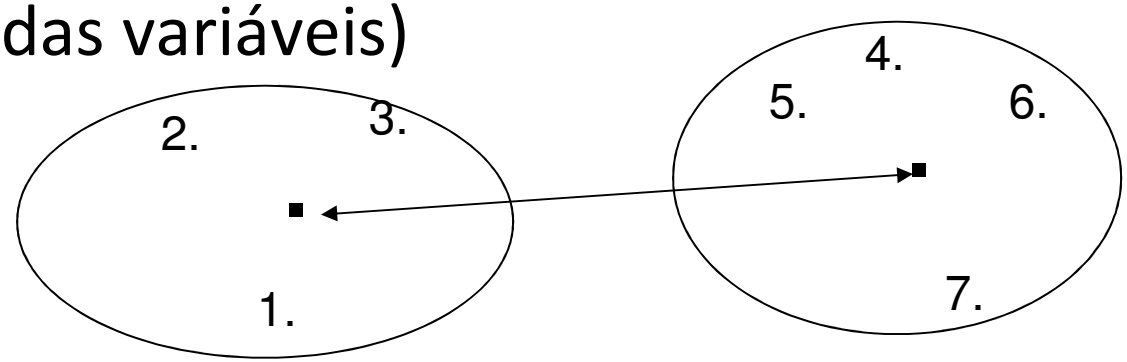
- Agrupamentos com objetos alocados de maneiras diferentes.

- Distâncias de fusão entre grupos diferentes.

Para o exemplo anteriormente considerado, os três métodos não produziram diferenças na alocação dos objetos mas apenas nas distâncias de fusão dos grupos.

Diferentes Métodos

Método do centróide: baseia-se na distância entre centróides (média das variáveis)



Método de Ward: Tende a produzir agrupamentos com aproximadamente mesmo número de itens.

- São calculadas médias (centróides) para cada grupo.
- É calculado o quadrado das distâncias entre as médias e os valores das variáveis para cada item.
- Somam-se os quadrados: variância
-Minimiza-se a variância dentro dos grupos: Error Sum of Squares (ESS).

Métodos não hierárquicos

- Formar agrupamentos cujo número de grupos é definido pelo pesquisador, ou seja, particionar n objetos em k grupos.

- Busca coesão interna (semelhança) e isolamento (separação) dos grupos formados.

- Iniciam-se : (1) partição inicial dos itens em grupos ou (2) uma semente inicial de pontos, que formarão os núcleos dos agrupamentos

Algoritmo não hierárquico: K-means

Um dos mais populares algoritmos K-médias (K-means): aloca objetos ao grupo com centróide mais próximo.

Procedimento:

(1) Particione os itens em k grupos iniciais arbitrariamente;

(2) Percorra a lista de itens e calcule as distâncias (usualmente a euclidiana) de cada um deles para o centróide (médias) dos grupos. Faça a realocação do item para o grupo em que ele apresentar mínima distância, obviamente se não for o grupo ao qual este pertença. Recalcule os centróides dos grupos que ganharam e perderam item.

(3) Repita o passo 2 até que nenhuma alteração seja feita.

Algoritmo não hierárquico: K-means

Exemplo:

Utilizando 4 itens (A, B, C e D) e 2 variáveis (X1 e X2) dividir em k=2 grupos, pelo método das k-médias.

	Observação	
Objeto	x_1	x_2
A	2	0
B	5	2
C	1	4
D	8	4

(1) particionar os itens arbitrariamente em 2 grupos, como por exemplo AD e BC. Calcular a média (centróide)de cada grupo.

	Centróide	
Objeto	$\bar{X}_1$	$\bar{X}_2$
AD	(2+8)/2=5	(0+4)/2=2
BC	(1+5)/2=3	(2+4)/2=3

Algoritmo não hierárquico: K-means

(2) Neste passo a distância de cada item será computada em relação ao centróide de cada grupo e se necessário, os objetos serão realocados para o grupo mais próximo.

$$d^2_{A(AD)} = (2-5)^2 + (0-2)^2 = 13$$

$$d^2_{A(BC)} = (2-3)^2 + (0-3)^2 = 10$$

Neste caso há necessidade de realocação de A para o grupo BC, sendo que os centróides dos grupos devem ser recalculados.

Objeto	Centróide	
	$\bar{X}_1$	$\bar{X}_2$
D	8	4
ABC	2,667	2

(5+2+1)/3

(0+4+2)/3

Algoritmo não hierárquico: K-means

Recalculando as distâncias dos objetos para o centróide dos grupos e checando a possibilidade de realocação, tem-se:

$$d^2_{A,D} = 52 \qquad d^2_{B,D} = 13 \qquad d^2_{C,D} = 49$$

$$d^2_{A,(ABC)} = 4,44 \qquad d^2_{B,(ABC)} = 5,44 \qquad d^2_{C,(ABC)} = 6,77$$

Nenhuma realocação deve ser realizada, pois os objetos têm menor distância para os respectivos grupos aos quais eles pertencem.

	Item (distância quadrática p/ centróide)			
Grupo	A	B	C	D
D	52,0	13,0	49,0	0,0
ABC	4,4	5,4	6,8	32,4

Para realizar uma checagem da estabilidade de a partição alcançada é recomendável executar novamente o algoritmo com uma nova partição inicial.

Escolha do método: Hierárquico ou Não-hierárquico?

- Não há uma resposta certa: depende da situação
- Hierárquico exige menos capacidade de processamento e evidencia *outliers*
- Não-hierárquicos exigem escolhas de sementes, mas são menos suscetíveis a *outliers*, à escolha de medidas e à variáveis inapropriadas
- Tomar um conjunto de dados com uma estrutura de grupos conhecida e ver se o algoritmo é capaz de reproduzir essa estrutura.

Escolha da quantidade de grupos e interpretação

Escolha da quantidade de agrupamentos formados interpretação e validação dos agrupamentos.

Não existe um procedimento padrão.

Sugestões:

- Olhar várias medidas de similaridade
- Fazer julgamento teórico
- Fazer análise para diferentes amostras
- Utilizar novas variáveis que teoricamente comprovam os clusters

Segmentação de mercado (Lattin, p217)

Segmentação de mercado

É difícil projetar produtos ou planejar campanhas publicitárias quando os indivíduos no mercado alvo diferem com respeito a suas necessidades e suas reações comportamentais. Os publicitários tentam resolver esse problema *segmentando* o mercado – isto é, dividindo o mercado em grupos menores que são mais homogêneos e, portanto, mais facilmente servidos por um tipo específico de produto ou uma campanha promocional específica.

Esses segmentos corresponderem aos "agrupamentos naturais" é menos importante do que o fato de representarem "grupos" de consumidores que são relativamente mais similares com respeito às suas necessidades e aos seus desejos de consumo.

Já fornecemos um exemplo da segmentação de preferência no contexto do escalonamento multidimensional utilizando as preferências medidas de 32 estudantes para 10 marcas diferentes de cerveja (veja o Capítulo 7). Naquele exemplo, encontramos considerável heterogeneidade na preferência entre os estudantes. Retornamos agora àquele exemplo, usando a análise de agrupamentos para reduzir essa heterogeneidade, dividindo a amostra em subgrupos menores e mais homogêneos. Os dados de preferência apresentados na Tabela 8.1 são medidos em uma escala de nove pontos (onde 9 = mais preferida). Uma vez que estamos interessados, principalmente, nas preferências relativas, podemos centrar na média as classificações de cada estudante para remover diferenças entre estudantes em suas preferências gerais por categoria de produto. (Menos de 10% dos estudantes no estudo mais amplo indicaram não ter preferência por nenhuma marca de cerveja; nenhum desses respondentes está incluído nessa subamostra específica.) Usamos, então, o agrupamento K–means para dividir os dados.

Segmentação de mercado

Tabela 8.1 Preferências expressas por 32 estudantes por 10 marcas diferentes de cerveja (medidas em uma escala de 9 pontos)

	Anchor Steam	Bass Ale	Beck's	Corona	Gordon--Biersch	Guinness	Heineken	Pete's Wicked Ale	Sam Adams	Sierra Nevada
S001	5	9	7	1	7	6	6	5	9	5
S008	7	5	6	8	8	4	8	8	7	7
S015	7	7	5	6	6	1	8	4	7	5

A divisão em dois agrupamentos de nossa análise $K-$means de agrupamento é apresentada na Tabela 8.2 (também examinamos soluções de três e quatro agrupamentos, mas não encontramos qualquer evidência de que fossem superiores à partição em dois agrupamentos). Os perfis de agrupamentos, que são as preferências médias dentro de cada agrupamento para cada uma das 10 marcas de cerveja, ajudam-nos a entender as diferenças entre os agrupamentos. No primeiro agrupamento (que constitui aproximadamente 60% da amostra), os estudantes expressam grande preferência por Corona e Heineken (as classificações médias são 7,0 e 6,8, respectivamente), duas cervejas claras relativamente suaves. Esses estudantes geralmente expressam pouca preferência pelas cervejas escuras no estudo (Guinness Stout e Pete's Wicked Ale). Em contraste, estudantes do segundo agrupamento expressam muita preferência por Anchor Steam e Sierra Nevada, duas cervejas de gosto relativamente mais forte (a última é, de fato, uma *pale ale*, cerveja clara de sabor mais amargo). Esses estudantes geralmente inclinam-se menos favoravelmente às cervejas claras (isto é, eles atribuem 4,3 à Corona e 4,2 à Heineken).

Segmentação de mercado

Tabela 8.2 Perfis de agrupamento da análise de agrupamento K-means de dados de preferência por cervejas (para K = 2)

	Perfis de agrupamentos	
	Agrupamento 1 (n = 20)	Agrupamento 2 (n = 12)
Anchor Steam	4,2	7,4
Bass Ale	5,3	5,9
Beck's	4,6	4,3
Corona	7,0	4,3
Gordon-Biersch	5,9	6,2
Guinness	4,5	6,3
Heineken	6,8	4,2
Pete's Wicked Ale	4,7	6,2
Sam Adams	6,1	6,3
Sierra Nevada	5,5	7,4

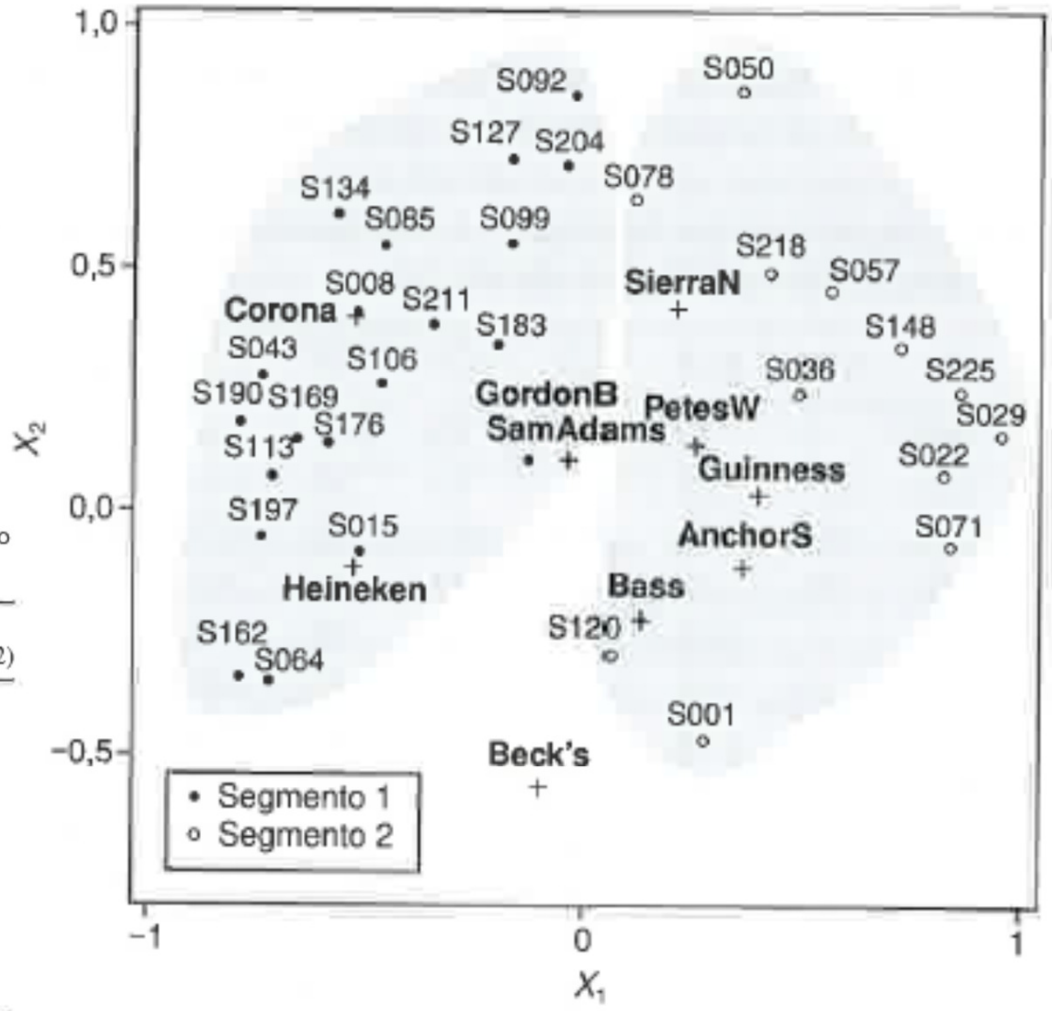


Figura 8.2 Mapa MDPREF de preferências dos estudantes por cerveja mostrando segmentação em dois agrupamentos.