Insper

Análise de Agrupamento Hierárquico e Escalonamento Multidimensional

Aula 4

Magno Severino
PADS - Aprendizagem Estatística de Máquina II

Objetivos de aprendizagem

Ao final dessa aula você deverá ser capaz de

- reproduzir o algoritmo do método de agrupamento hierárquico (hclust);
- interpretar o dendrograma gerado pelo método (hclust);
- compreender a diferença entre distância e dissimilaridade;
- utilizar o escalonamento multidimensional (MDS) para obter coordenadas para dados a partir de uma medida de dissimilaridade.

Introdução

- Desvantagem do *k*-means: requer que o número *k* de clusters seja pré-definido;
- O **agrupamento hierárquico** é uma alternativa que não requer uma pré-definição do valor de *k*;
- No **agrupamento hierárquico**, o resultado pode ser exibido através de uma representação baseada em árvores chamada de *dendrograma*.

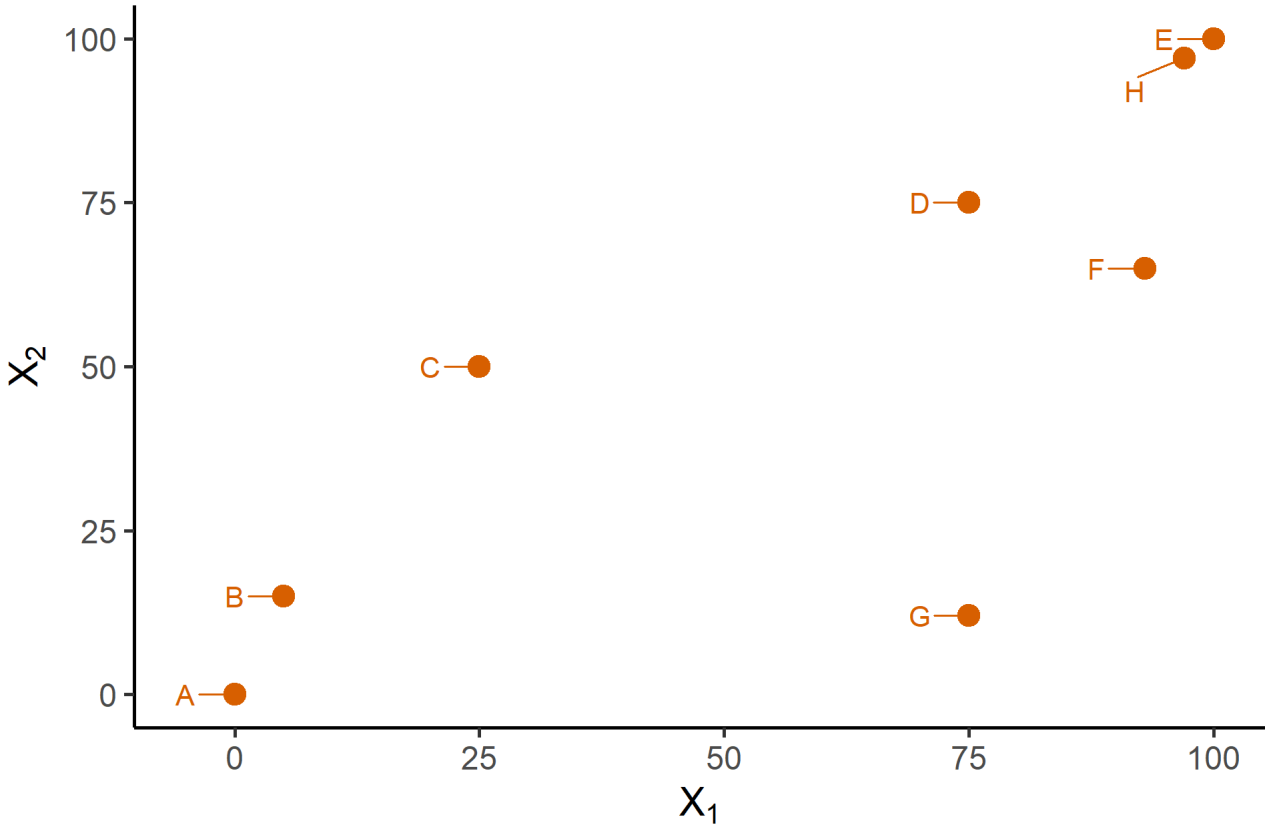
Exemplo inicial

Considere os seguintes dados sobre dois indicadores socioeconomicos em oito cidades brasileiras.

rotulo	x1	x2
A	0	0
B	5	15
C	25	50
D	75	75
E	100	100
F	93	65
G	75	12
H	97	97

Exemplo inicial

Considere os seguintes dados sobre dois indicadores socioeconomicos em oito cidades brasileiras. O nosso objetivo é agrupar observações (neste caso, cidades) similares.



Como fazer esse agrupamento?

Exemplo inicial

A tabela a seguir mostra a distância Euclidiana entre cada observação no conjunto de dados.

	A	B	C	D	E	F	G	H
A	0.0	15.8	55.9	106.1	141.4	113.5	76.0	137.2
B	15.8	0.0	40.3	92.2	127.5	101.2	70.1	123.2
C	55.9	40.3	0.0	55.9	90.1	69.6	62.8	86.0
D	106.1	92.2	55.9	0.0	35.4	20.6	63.0	31.1
E	141.4	127.5	90.1	35.4	0.0	35.7	91.5	4.2
F	113.5	101.2	69.6	20.6	35.7	0.0	56.0	32.2
G	76.0	70.1	62.8	63.0	91.5	56.0	0.0	87.8
H	137.2	123.2	86.0	31.1	4.2	32.2	87.8	0.0

Agrupamento hierárquico

Algoritmo

1. Inicie com n observações e uma medida de dissimilaridade (como a Euclidiana) e calcule a distância entre todos os pares de observações. Trate cada observação como seu próprio cluster.

2. Para $i = n, n - 1, \ldots, 2$:

   a. Examine todos pares de distância intra-cluster entre os i clusters e identifique o par que tem menos dissimilaridade (ou seja, mais similar). Junte esses dois em um único grupo.

   b. Calcule as novas dissimilaridades intra-cluster entre os $i - 1$ demais grupos.

Problema!!!

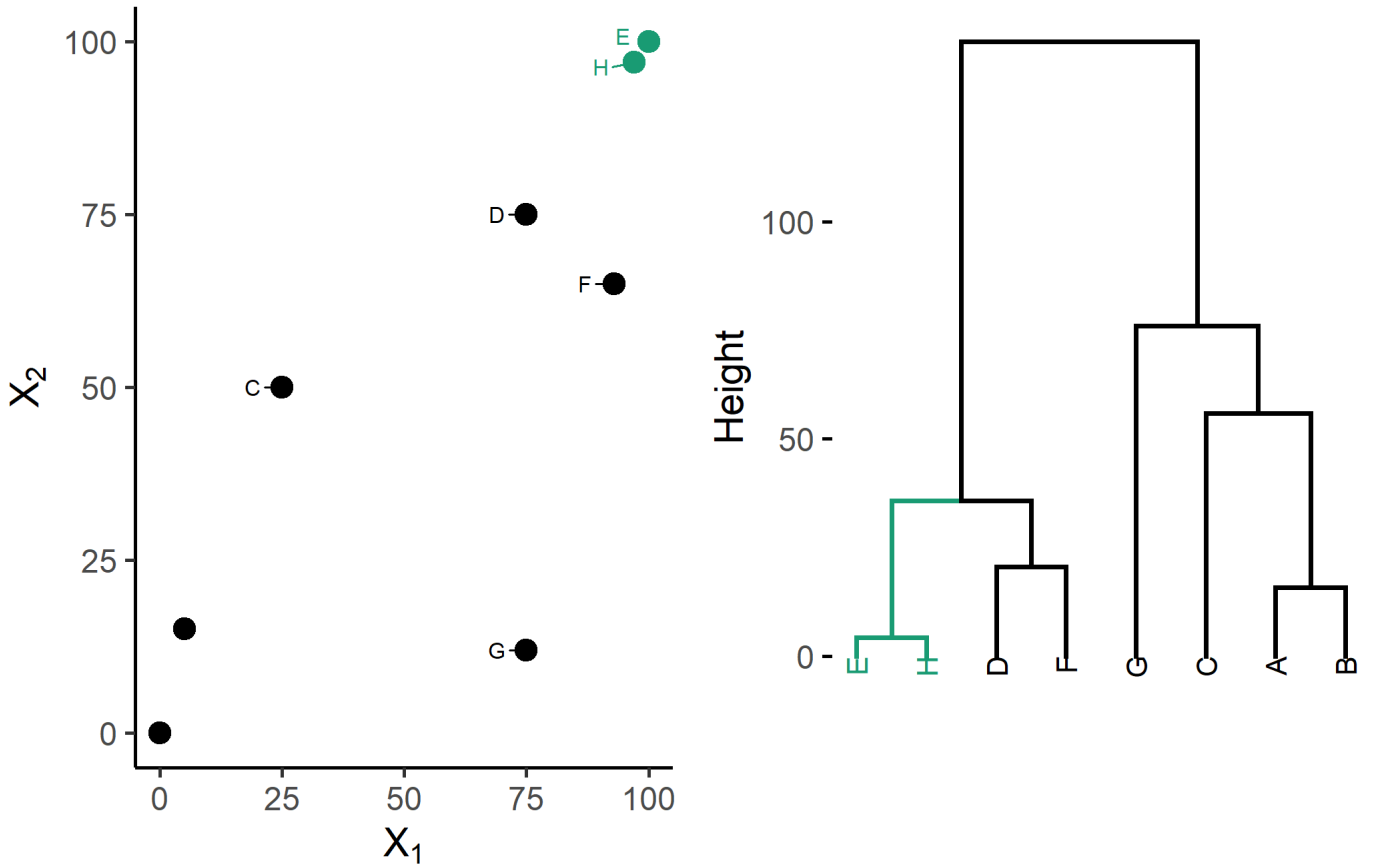
Como calcular distância entre clusters?

Tipos de *linkage*

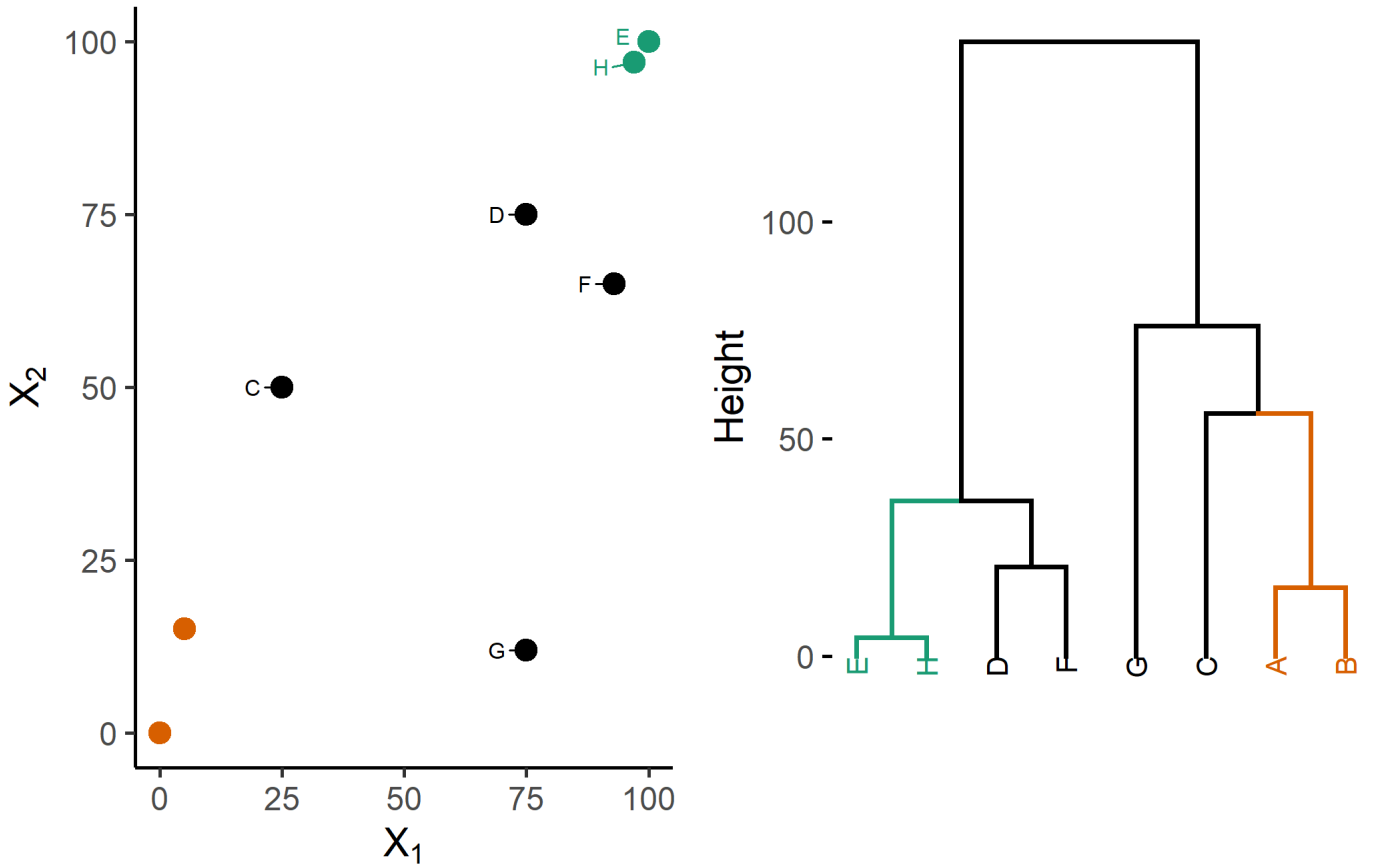
O conceito de distância entre observações é bem definido. Entretanto, precisamos definir como calcular a dissimilaridade entre grupos. Então, o conceito de distância deve ser estendido para pares de grupos de observações. Essa extensão é obtida desenvolvendo a noção de *linkage*, que define a distância entre grupos de observações. As quatro mais comuns estão listadas abaixo.

- **Complete**: a *maior* das distâncias entre todos os pares de observações pertencentes aos dois clusters;
- **Single**: a *menor* das distâncias entre todos os pares de observações pertencentes aos dois clusters.
- **Average**: a *média* das distâncias entre todos os pares de observações pertencentes aos dois clusters;
- **Centroid**: a distância entre os *centróides* dos dois clusters.

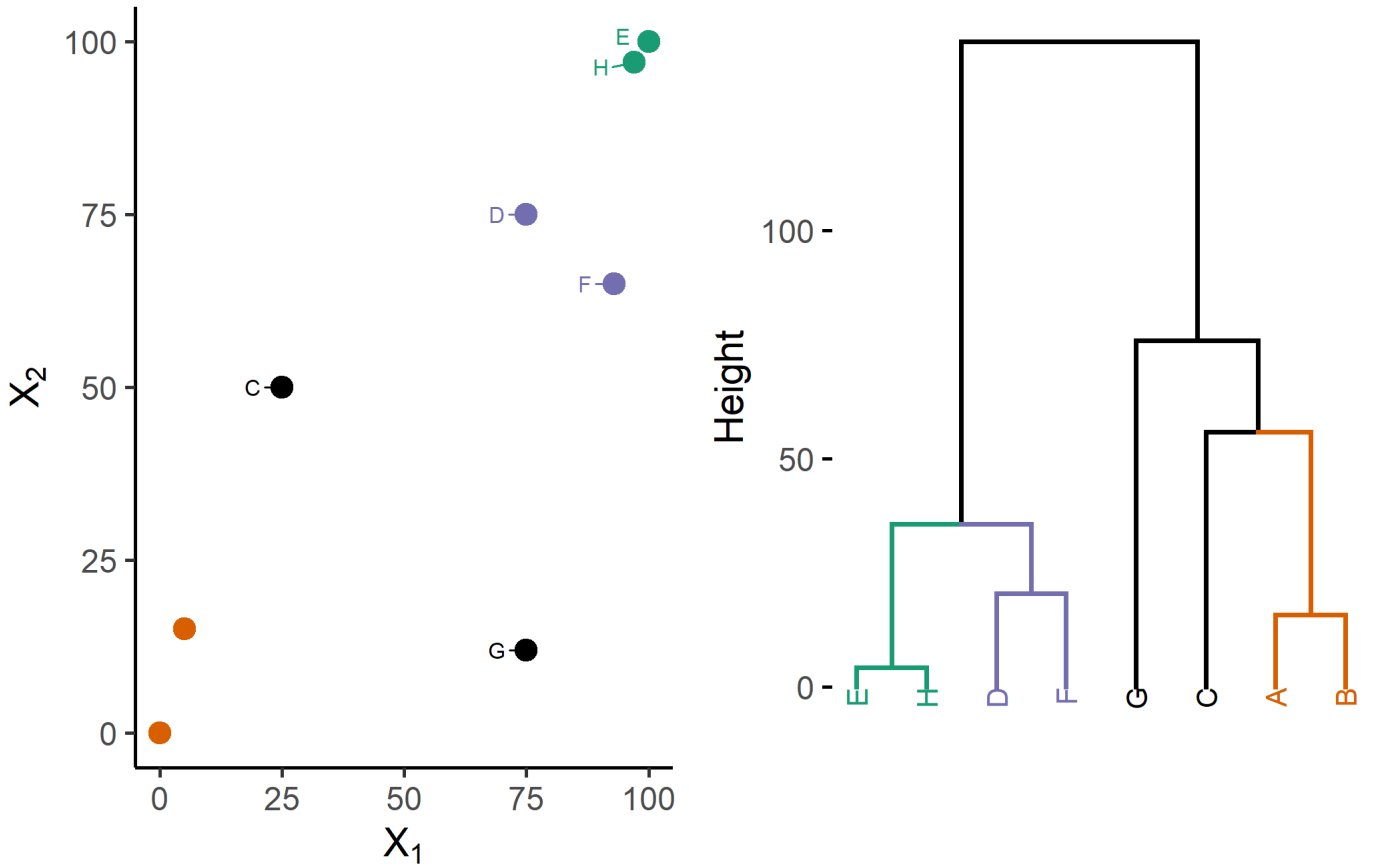
Exemplo inicial (complete)



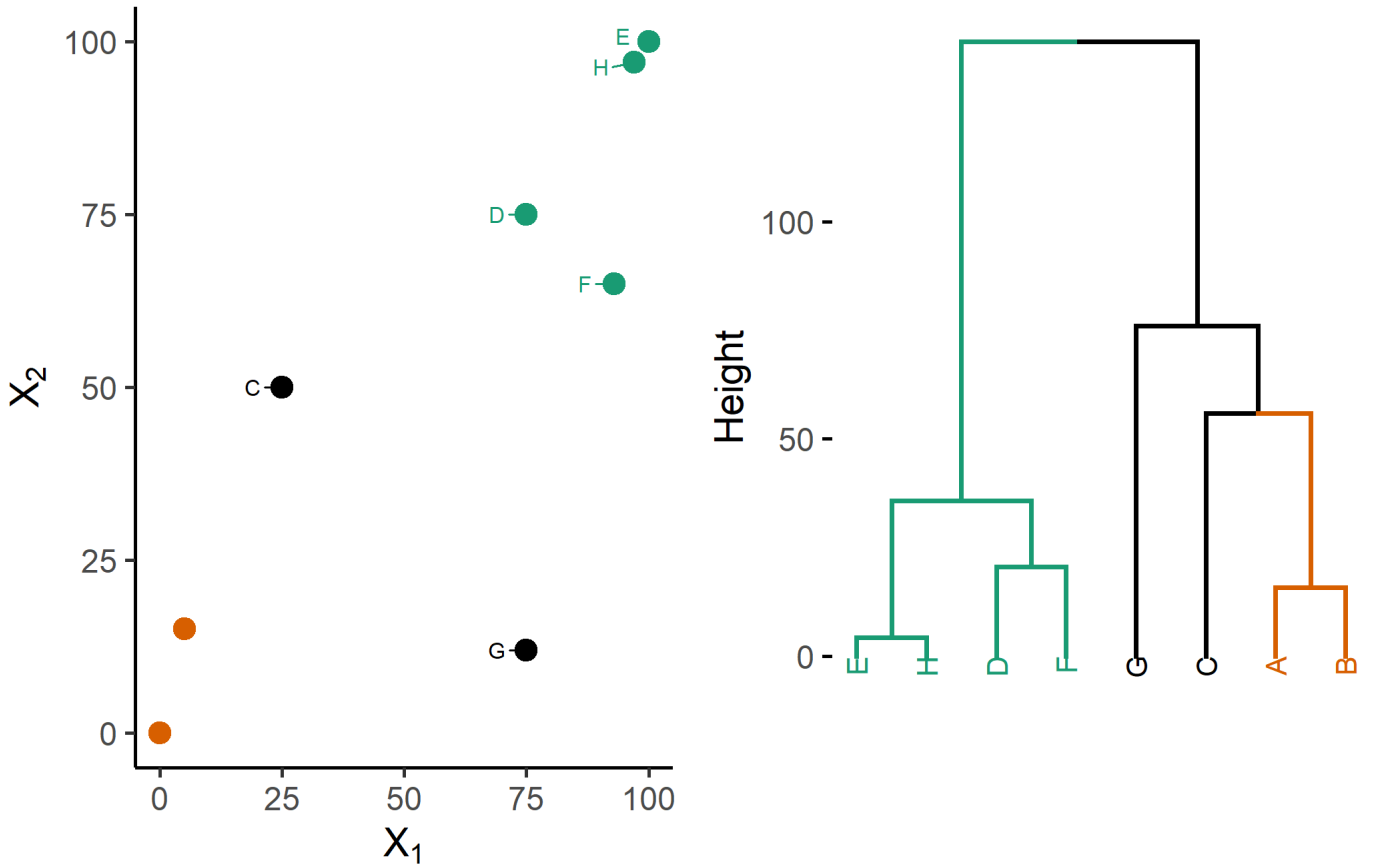
Exemplo inicial (complete)



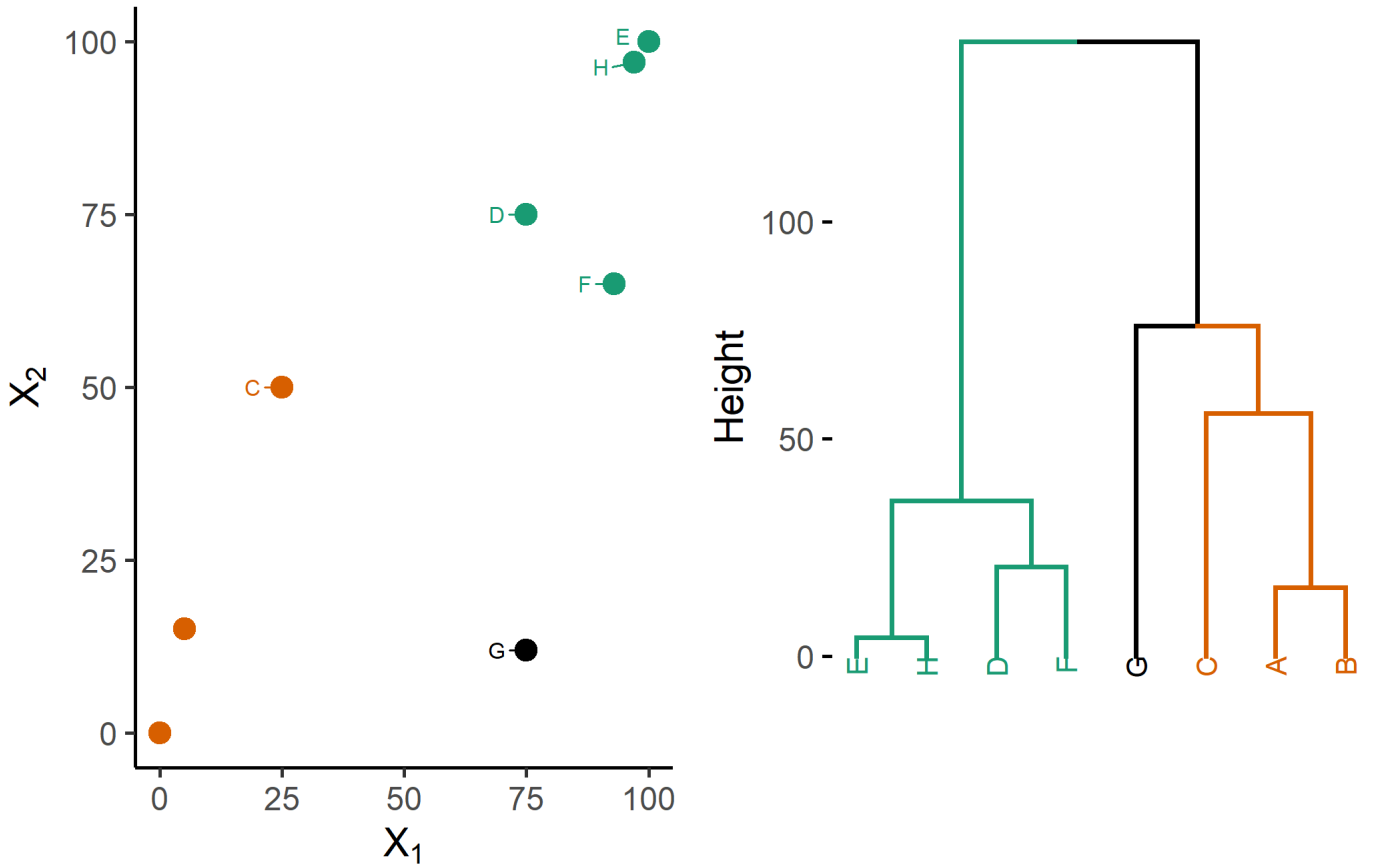
Exemplo inicial (complete)



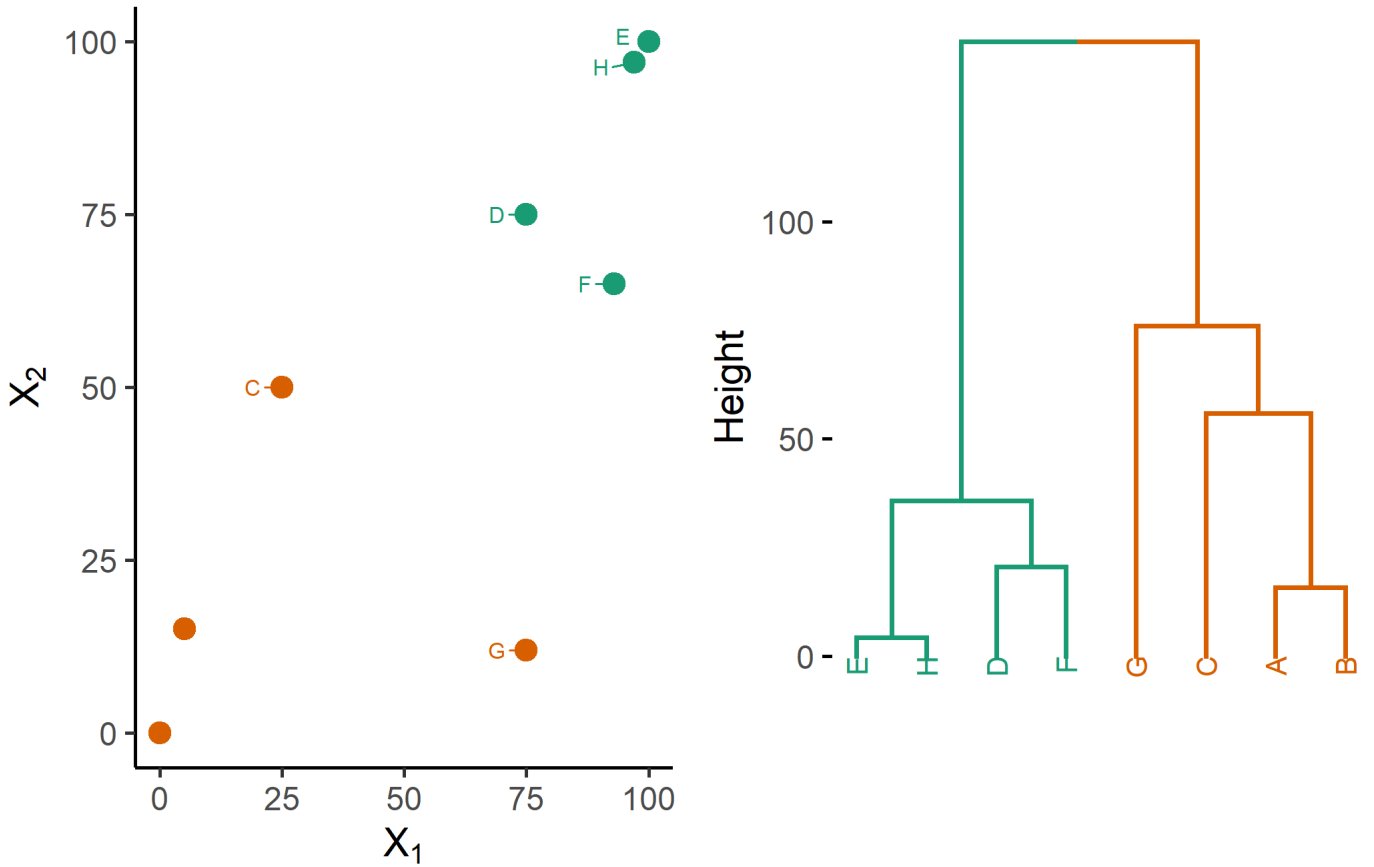
Exemplo inicial (complete)



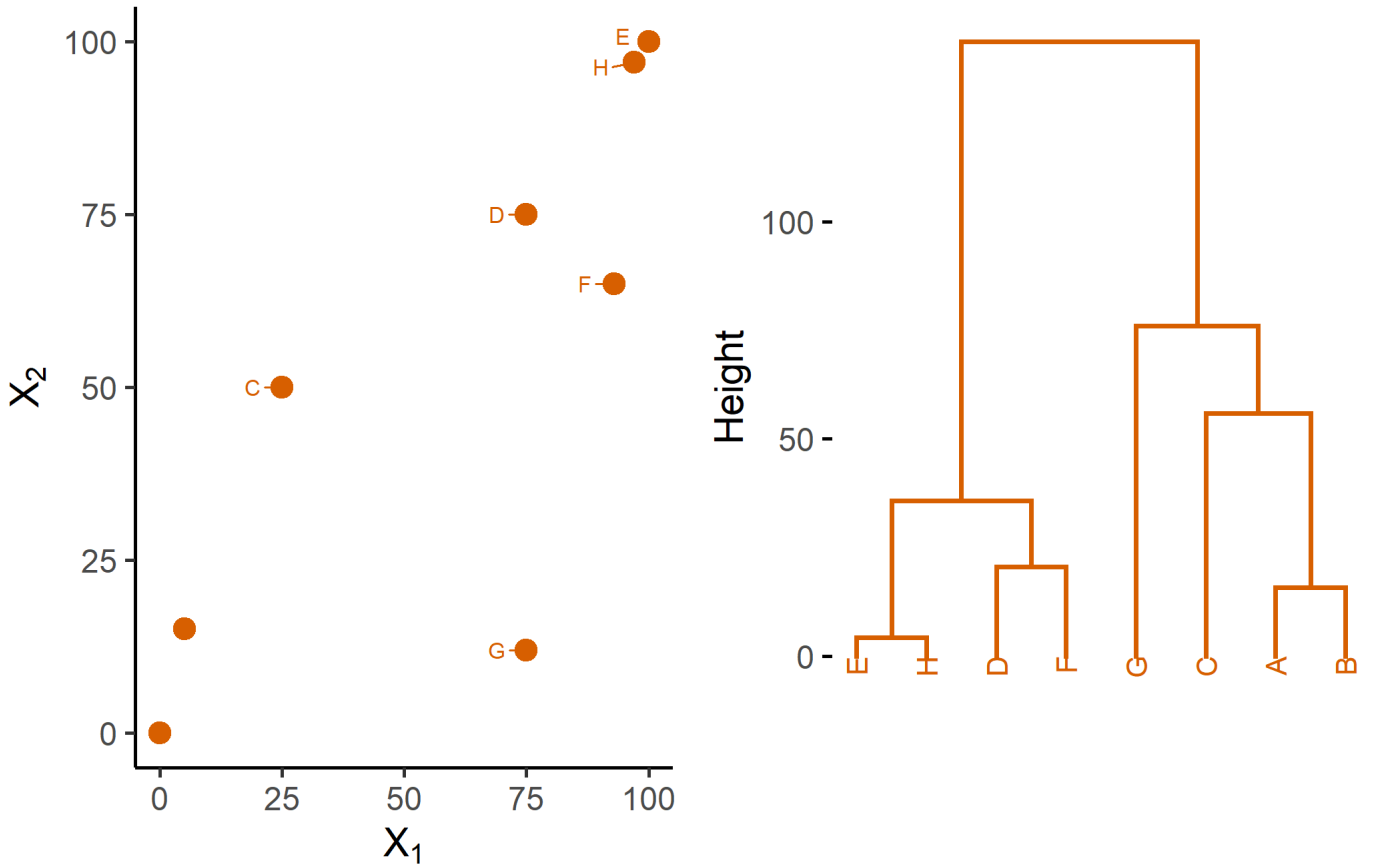
Exemplo inicial (complete)



Exemplo inicial (complete)



Exemplo inicial (complete)



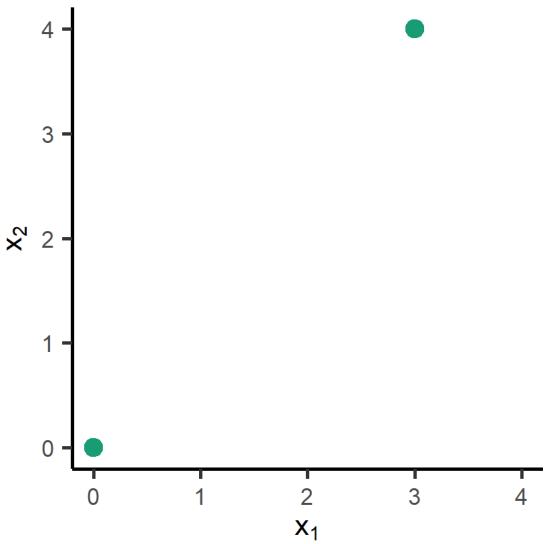
Distâncias

Distância Euclidiana

A distância Euclidiana entre dois pontos A e $B \in \mathrm{R}^p$ é

$$d_{AB} = \sqrt{\sum_{\ell=1}^{p} (x_{A\ell} - x_{B\ell})^2}.$$

Considere os pontos $x_A = (0, 0)$ e $x_B = (3, 4)$.



Distância Euclidiana

A distância Euclidiana entre dois pontos A e $B \in \mathrm{R}^p$ é

$$d_{AB} = \sqrt{\sum_{\ell=1}^{p} (x_{A\ell} - x_{B\ell})^2}.$$

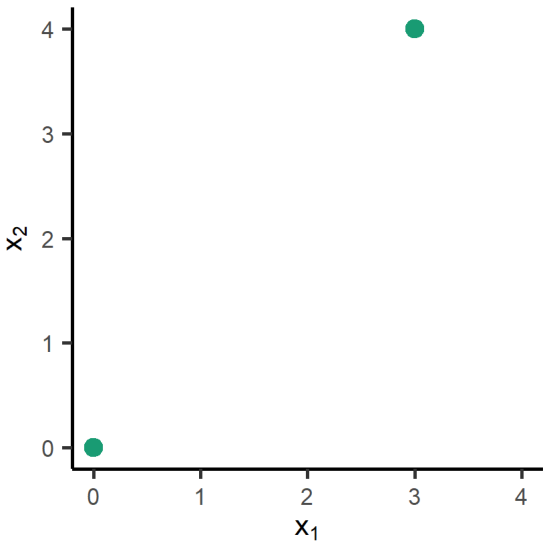
Considere os pontos $x_A = (0, 0)$ e $x_B = (3, 4)$.

Distância Manhattan

A distância Manhattan (ou do taxi) entre dois pontos A e $B \in \mathrm{R}^p$ é

$$d_{AB} = \sum_{\ell=1}^{p} |x_{A\ell} - x_{B\ell}| .$$

Considere os pontos $x_A = (0, 0)$ e $x_B = (3, 4)$.

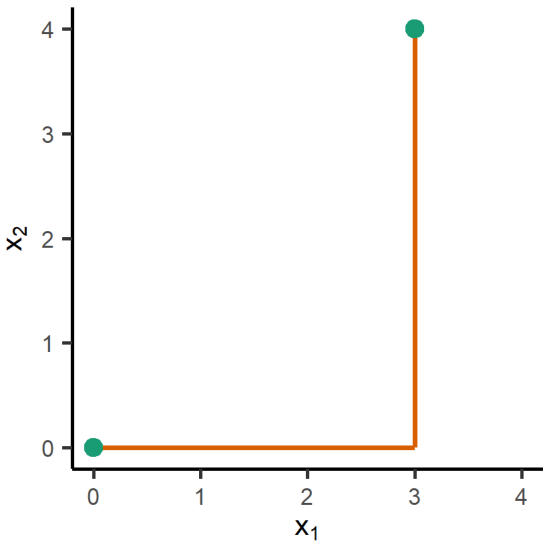


Distância Manhattan

A distância Manhattan (ou do taxi) entre dois pontos A e $B \in \mathrm{R}^p$ é

$$d_{AB} = \sum_{\ell=1}^{p} |x_{A\ell} - x_{B\ell}| .$$

Considere os pontos $x_A = (0, 0)$ e $x_B = (3, 4)$.

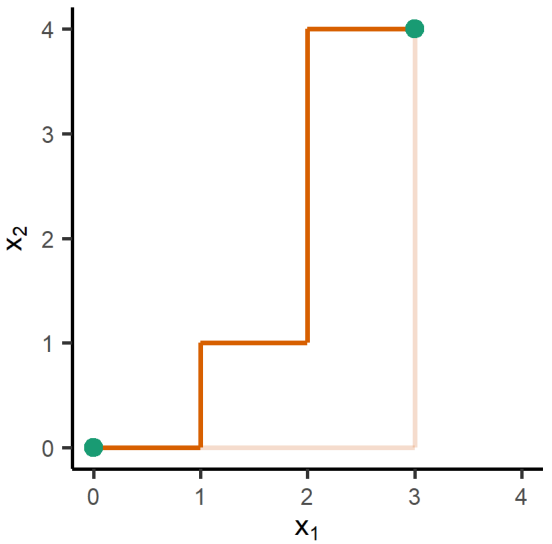


Distância Manhattan

A distância Manhattan (ou do taxi) entre dois pontos A e $B \in \mathrm{R}^p$ é

$$d_{AB} = \sum_{\ell=1}^{p} |x_{A\ell} - x_{B\ell}| .$$

Considere os pontos $x_A = (0, 0)$ e $x_B = (3, 4)$.

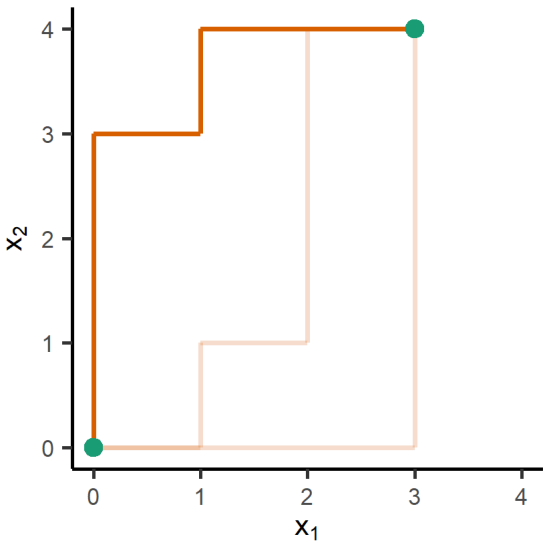


Distância Manhattan

A distância Manhattan (ou do taxi) entre dois pontos A e $B \in \mathrm{R}^p$ é

$$d_{AB} = \sum_{\ell=1}^{p} |x_{A\ell} - x_{B\ell}| .$$

Considere os pontos $x_A = (0, 0)$ e $x_B = (3, 4)$.



Distância de Jaccard

Considere que existam n pessoas e m marcas e as seguintes variáveis binárias:

$$x_{ij} = \begin{cases} 1 & \text{se a pessoa } i \text{ consome a marca } j \\ 0 & \text{caso contrário,} \end{cases}$$

para $i = 1, \ldots, n$ (pessoas) e $j = 1, \ldots, m$ (marcas).

A distância de Jaccard entre as marcas k e ℓ é dada por

$$d_{k\ell} = 1 - \frac{\text{número de pessoas que consomem } k \text{ e } \ell}{\text{número de pessoas que consomem } k \text{ ou } \ell}.$$

Em notação de teoria de conjuntos, se K e L são os conjuntos das pessoas que consomem os produtos k e ℓ, respectivamente, podemos escrever a distância de Jaccard como

$$d_{k\ell} = 1 - \frac{|K \cap L|}{|K \cup L|},$$

em que $|K \cap L|$ indica o número de elementos presentes em K e L, e $|K \cup L|$ indica o número de elementos presentes em K ou L.

Distância de Jaccard

Considere uma situação com cinco pessoas que indicam se consomem 4 marcas de um mesmo produto. Se a pessoa consome, recebe o valor 1. Caso contrário, recebe o valor 0. Assim, temos

id	A	B	C	D
Cliente 1	1	0	1	0
Cliente 2	1	0	1	0
Cliente 3	0	1	1	1
Cliente 4	0	1	1	1
Cliente 5	1	0	0	0
Cliente 6	1	1	1	1

$$d_{AB} = 1 - \frac{1}{6} = \frac{5}{6} \qquad d_{BC} = 1 - \frac{3}{5} = \frac{2}{5}$$

$$d_{BD} = 1 - \frac{3}{3} = 0 \qquad d_{CD} = 1 - \frac{3}{5} = \frac{2}{5}$$

Distância de Jaccard

Então, obtemos a seguinte matriz de distâncias:

	A	B	C	D
A	0	5/6	1/2	5/6
B	5/6	0	2/5	0
C	1/2	2/5	0	2/5
D	5/6	0	2/5	0

Podemos obter o dendrograma utilizando o seguinte código (considerando como *linkage* o método *complete*):

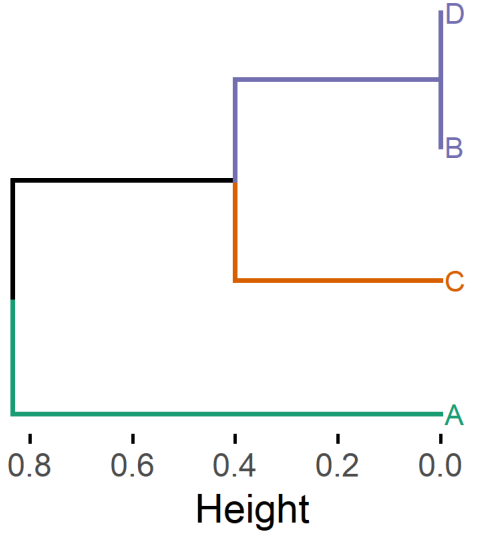
```
hc <- hclust(as.dist(matriz),
             method = "complete")

fviz_dend(hc,
          k = 3,
          labels_track_height = 0.1,
          color_labels_by_k = TRUE,
          horiz = TRUE)
```

Distância de Jaccard

Matriz de distâncias.

	A	B	C	D
A	0	5/6	1/2	5/6
B	5/6	0	2/5	0
C	1/2	2/5	0	2/5
D	5/6	0	2/5	0

Resultado do agrupamento hierárquico.



Mercado de Whisky

Em uma pesquisa, os entrevistados (2218 consumidores) indicaram quais uísques consomem regularmente em uma lista com um total de 21 marcas. O valor 1 indica que a marca é consumida regularmente, enquanto 0 indica que a marca não é consumida regularmente.

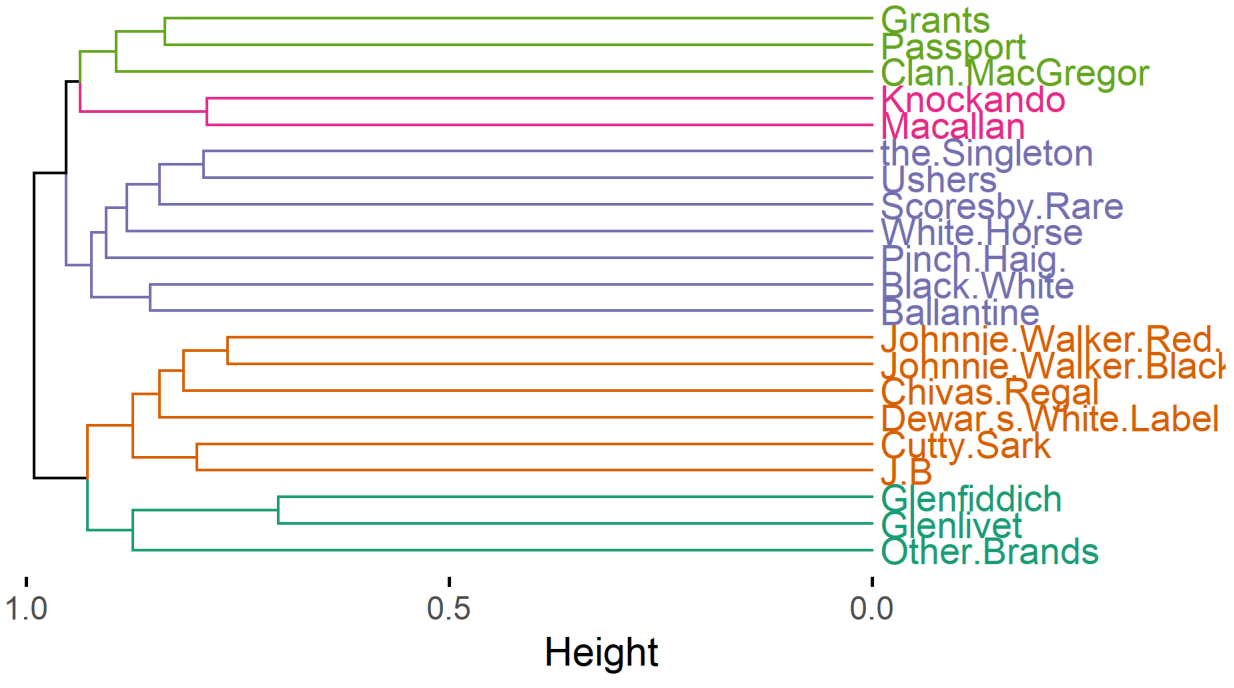
Chivas.Regal	Dewar.s.White.Label	Johnnie.Walker.Black.Label	J.B
1	0	0	0
0	0	1	0
0	0	0	0
1	0	1	0
1	0	1	0
0	0	0	0
0	0	0	0
0	1	0	1
0	0	0	0
1	1	0	0
0	0	0	0
0	0	0	1
0	0	1	0

Mercado de Whisky

A matriz de distância de Jaccard entre as marcas está na tabela abaixo.

	Chivas.Regal	Dewar.s.White.Label	Johnnie.
Chivas.Regal	0.0	0.8	
Dewar.s.White.Label	0.8	0.0	
Johnnie.Walker.Black.Label	0.8	0.8	
J.B	0.8	0.9	
Johnnie.Walker.Red.Label	0.8	0.8	
Other.Brands	0.9	0.9	
Glenlivet	0.9	0.9	
Cutty.Sark	0.8	0.9	
Glenfiddich	0.9	0.9	
Pinch.Haig.	0.9	0.9	
Clan.MacGregor	1.0	0.9	
Ballantine	0.9	0.9	
Macallan	1.0	1.0	
Passport	1.0	1.0	
Black.White	0.9	0.9	
Scoresby.Rare	1.0	1.0	

Mercado de Whisky



Quais possíveis aplicações?

- sistemas de recomendação;
- formação de estoque;
- organização de prateleiras em pontos de venda.

Seria possível visualizar estes dados em um gráfico de dispersão bidimensional?

Com escalonamento multidimensional, sim!

Distância *vs* dissimilaridade

Pela definição, uma **distância** d_{ij} entre os pontos i e j satisfaz 4 propriedades:

1. $d_{ij} \geq 0$
2. $d_{ij} = 0 \iff i = j$
3. $d_{ij} = d_{ji}$
4. $d_{ij} \leq d_{i\ell} + d_{\ell j}$ (desigualdade triangular)

Uma medida de **dissimilaridade** é uma "distância fraca", pois não satisfaz pelo menos uma das 4 propriedades acima.

Distância *vs* dissimilaridade

Mercado automobilístico

Um total de $n = 150$ consumidores recebeu 30 cartões com modelos de diferentes carros. Foi solicitado que cada cliente agrupasse os cartões de modo que dois carros alocados no mesmo grupo fossem considerados equivalentes (substituíveis no momento da compra). Com este processo, definiu-se uma medida de dissimilaridade dada por

$$\begin{aligned}\delta_{ij} &= \frac{n - (\text{quantas vezes } i \text{ e } j \text{ foram alocados no mesmo grupo})}{n} \\ &= 1 - \frac{(\text{quantas vezes } i \text{ e } j \text{ foram alocados no mesmo grupo})}{n}.\end{aligned}$$

Qual propriedade da definição de distância $\Delta = (\delta_{ij})$ não satisfaz?

Observação: toda distância é uma dissimilaridade, a recíproca é falsa.

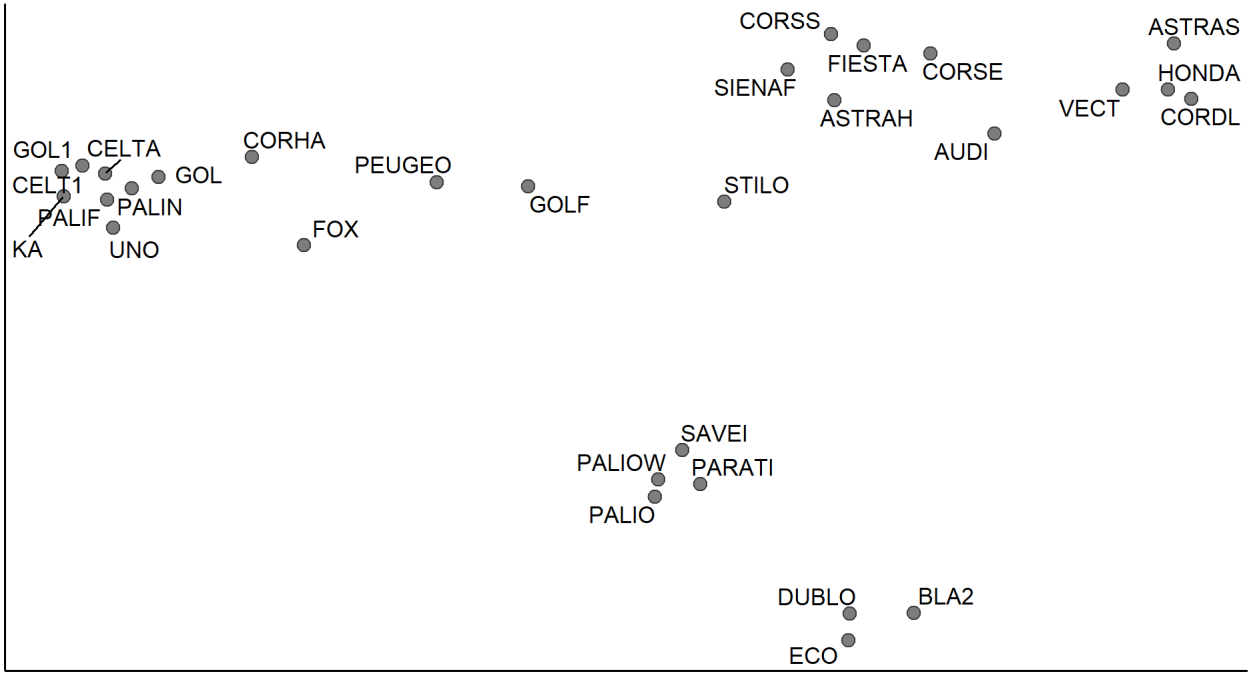
Escalonamento multidimensional (MDS)

- Não temos as coordenadas dos dados $x_1, x_2, \ldots, x_n \in \mathrm{R}^p$;
- Temos apenas a matriz de dissimilaridades $\Delta = (\delta_{ij})$;
- Queremos obter coordenadas $z_1, z_2, \ldots, z_n \in \mathrm{R}^d$, usualmente com $d < p$ (representação em baixa dimensão dos dados) baseadas na matriz Δ.
- Em outras palavras, queremos encontrar pontos $z_1, z_2, \ldots, z_n \in \mathrm{R}^d$ com $d < p$ de forma a minimizar

$$\sum_{i,j} \left(\delta_{ij} - \sqrt{\sum_{\ell=1}^{d} (z_{i\ell} - z_{j\ell})^2} \right)^2.$$

- A solução deste problema de otimização não é trivial (veja Fernandez e Yohai, 2014).
- Veja que para este método, precisamos apenas das medidas de dissmilaridade ou distância entre as observações e não dos dados!

Mercado automobilístico



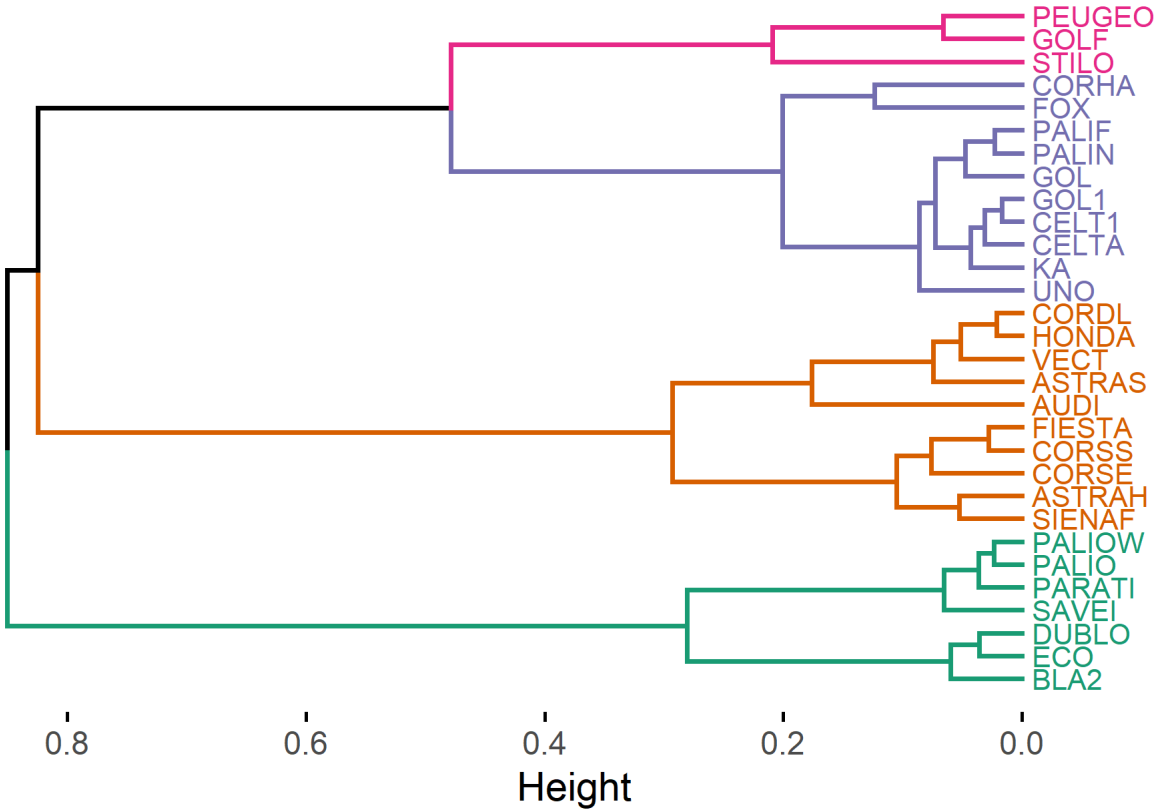
Através deste sistema de coordenadas obtido pelo MDS, podemos utilizar qualquer método de agrupamento (cluster hierárquico ou *k*-médias, por exemplo).

Mercado automobilístico

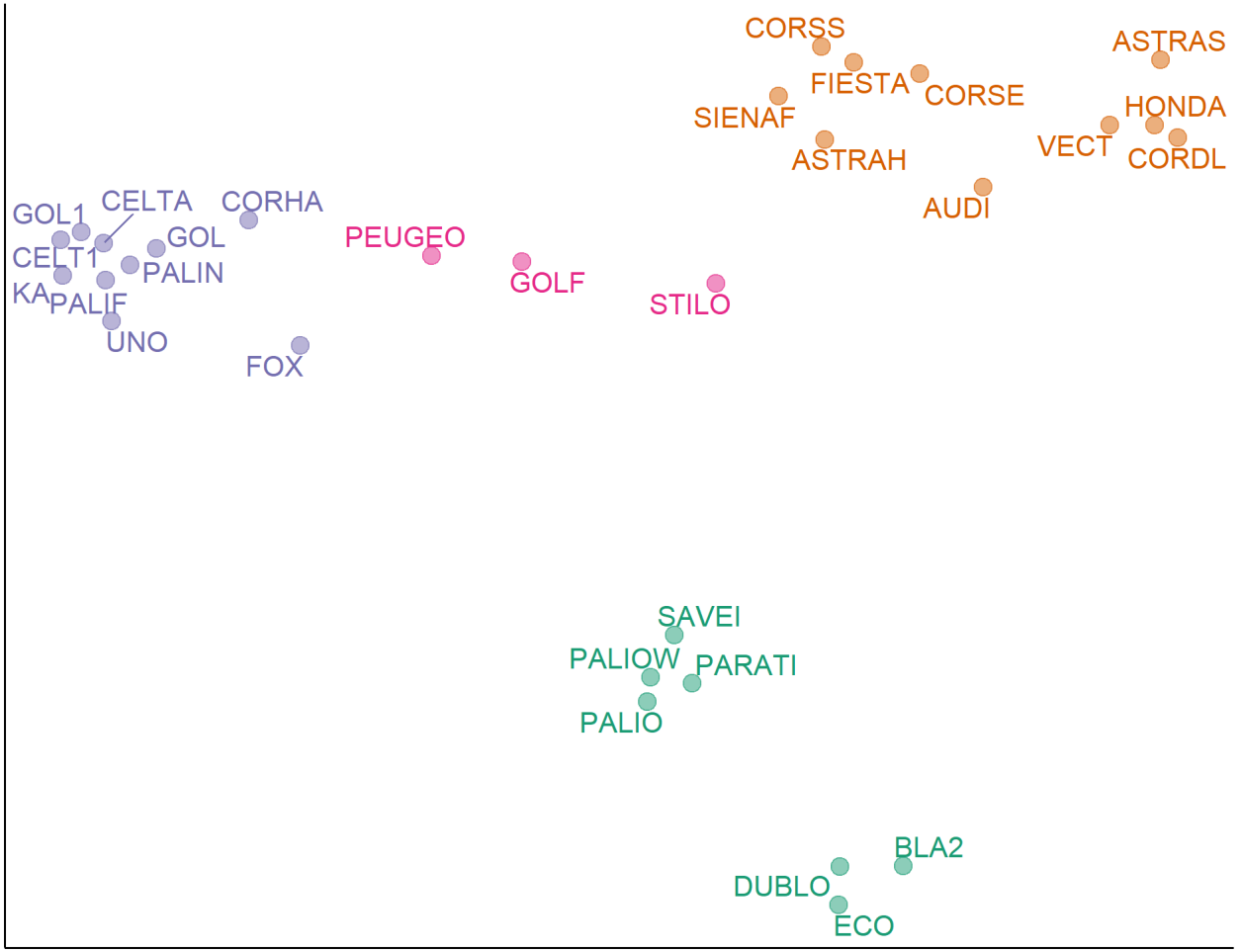
- Note que cada carro/modelo é um objeto complexo, cujos atributos poderiam ser representados por várias colunas, cada uma com um atributo (uma representação em um espaço de dimensão elevada);
- No entanto, não temos acesso às coordenadas de cada carro neste espaço hipotético;
- Tudo o que temos é a informação sobre quão dissimilares são cada um dos pares de carros participantes na pesquisa, do ponto de vista os consumidores entrevistados.
- O escalonamento multidimensional, atribui a cada carro um ponto em um espaço de duas dimensões, de maneira que carros mais próximos neste espaço sejam considerados mais semelhantes pelos consumidores entrevistados

Mercado automobilístico

Resultado do agrupamento hierárquico, considerando um corte no dendrograma que gere $k = 4$ grupos.



Mercado automobilístico



Voltando ao exemplo do whisky...

Mercado de Whisky

Discussão: questões práticas em modelos de agrupamento

- Deve-se padronizar as os dados?
- `hclust`: Qual medida de dissimilaridade deve ser usada? Que tipo de *linkage* deve ser usada? Onde devemos cortar o dendrograma para obter clusters?
- `kmeans`: quantos clusters devemos procurar nos dados?
- Cada uma dessas decisões podem gerar um grande impacto no resultado.
- Na prática, tentamos várias escolhas diferentes e procuramos aquela com a solução mais útil ou interpretável. Com esses métodos, não há uma única resposta correta.

Discussão: validação do resultado apresentado

- Os métodos de agrupamento sempre encontrarão clusters em um conjunto de dados, mas nem sempre eles representam subgrupos reais.
- É difícil saber se os clusters encontrados são resultado do agrupamento do ruído ou representam subgrupos reais nos dados.
- K-means e clusterização hierárquica atribuem cada observação a um cluster, o que pode não ser apropriado se algumas observações pertencerem a pequenos subgrupos desconhecidos e houver outliers que não pertencem a nenhum cluster.
- Modelos de mistura são uma abordagem melhor para acomodar a presença de outliers, pois são uma versão suave da clusterização K-means.
- Os métodos de clusterização não são robustos às perturbações nos dados, e a clusterização novamente após a remoção de um subconjunto aleatório das observações pode resultar em clusters diferentes, indicando a falta de estabilidade nos resultados da clusterização.
- A presença de outliers pode distorcer os resultados da clusterização, e é importante usar técnicas apropriadas para lidar com eles.

Discussão: interpretação dos resultados

- Clustering pode ser uma ferramenta estatística útil e válida se usada corretamente.
- Pequenas decisões podem ter grande efeito nos resultados.
- Realizar clustering com diferentes escolhas de parâmetros e olhar para o conjunto completo de resultados.
- É importante ter cuidado com a forma como os resultados da análise de clustering são relatados.
- Esses resultados não devem ser considerados como a verdade absoluta sobre um conjunto de dados, mas sim como um ponto de partida para o desenvolvimento de uma hipótese e estudos posteriores.

Resumindo

- O **agrupamento hierárquico** é uma técnica de *clusterização* que não requer pré definição do número de grupos;
- No **agrupamento hierárquico**, o resultado pode ser exibido através de uma representação baseada em árvores chamada *dendrograma*;
- Existe diferença entre distância e dissimilidade;
- A partir de uma medida de dissimilaridade, podemos obter coordenadas (através do MDS) para as observações e utilizar um método de agrupamento para criar grupos homogêneos.

Obrigado!

magnotfs@insper.edu.br