

Análise de Clusters II

ACH2036 – Métodos Quantitativos Aplicados à Adm. de Empresas I
Prof. Regis Rossi A. Faria
2º sem. 2020

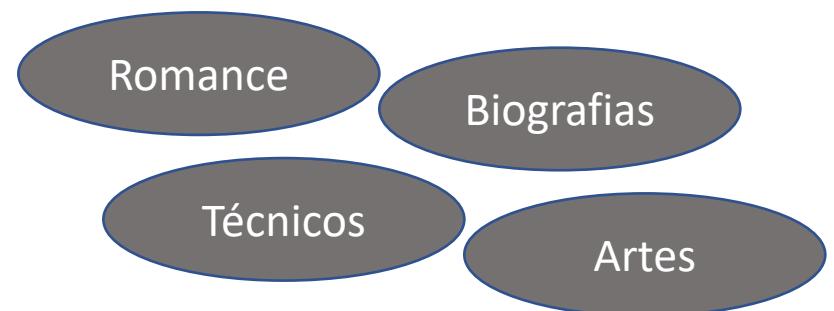
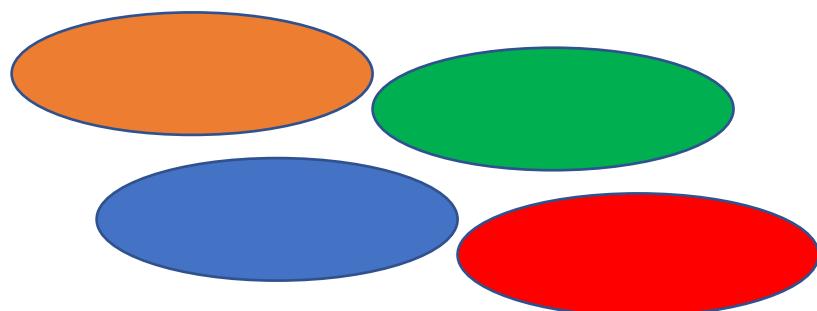


Sumário

- Introdução à técnica segundo abordagem vista no livro de referência *Everitt, B., Hothorn, T. An Introduction to Applied Multivariate Analysis with R. Springer, 2011.*
- Procedimento hierárquico de agrupamento utilizando o R
- Exemplos e Exercício de aplicação utilizando o R

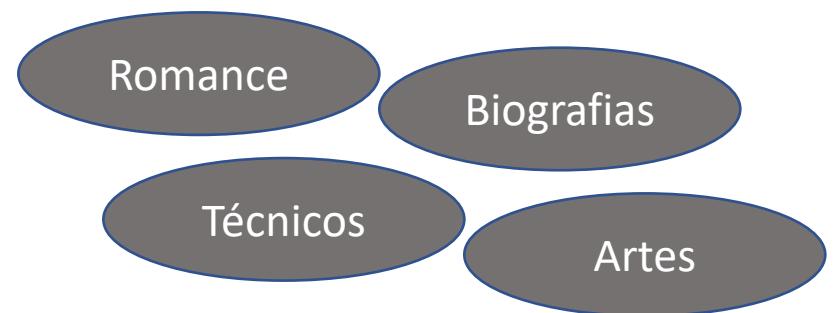
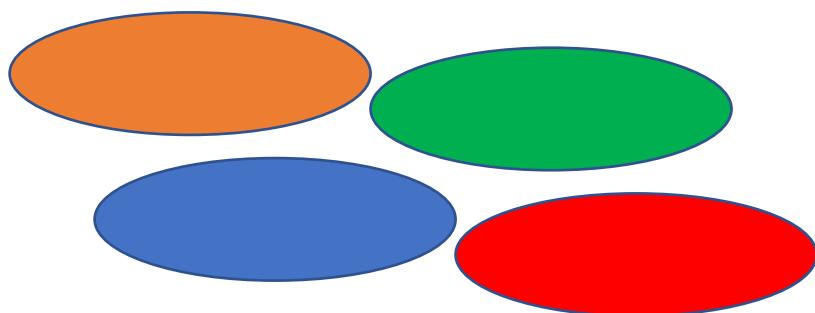
Análise de clusters

Se tivéssemos que classificar os livros de uma biblioteca em grupos, quais dos exemplos abaixo de classificação seriam mais úteis?



Análise de clusters

Se tivéssemos que classificar os livros de uma biblioteca em grupos, quais dos exemplos abaixo de classificação seriam mais úteis?



→ a escolha vai depender se você trabalha numa
escola fundamental, para crianças, ou se trabalha
numa biblioteca universitária

Análise de clusters

- A *análise de clusters* ou *análise de agrupamentos* ou *análise de conglomerados* é um termo genérico para uma ampla gama de métodos numéricos com o objetivo comum de descobrir grupos de observações que são homogêneas e separadas de outros grupos.
- As técnicas de agrupamento tentam essencialmente formalizar o que os observadores humanos fazem tão bem em duas ou três dimensões.
- Considere, por exemplo, o gráfico de dispersão mostrado na Figura 6.1 do livro de Brian Everitt e Torsten Hothorn (An Introduction to Applied Multivariate Analysis with R)

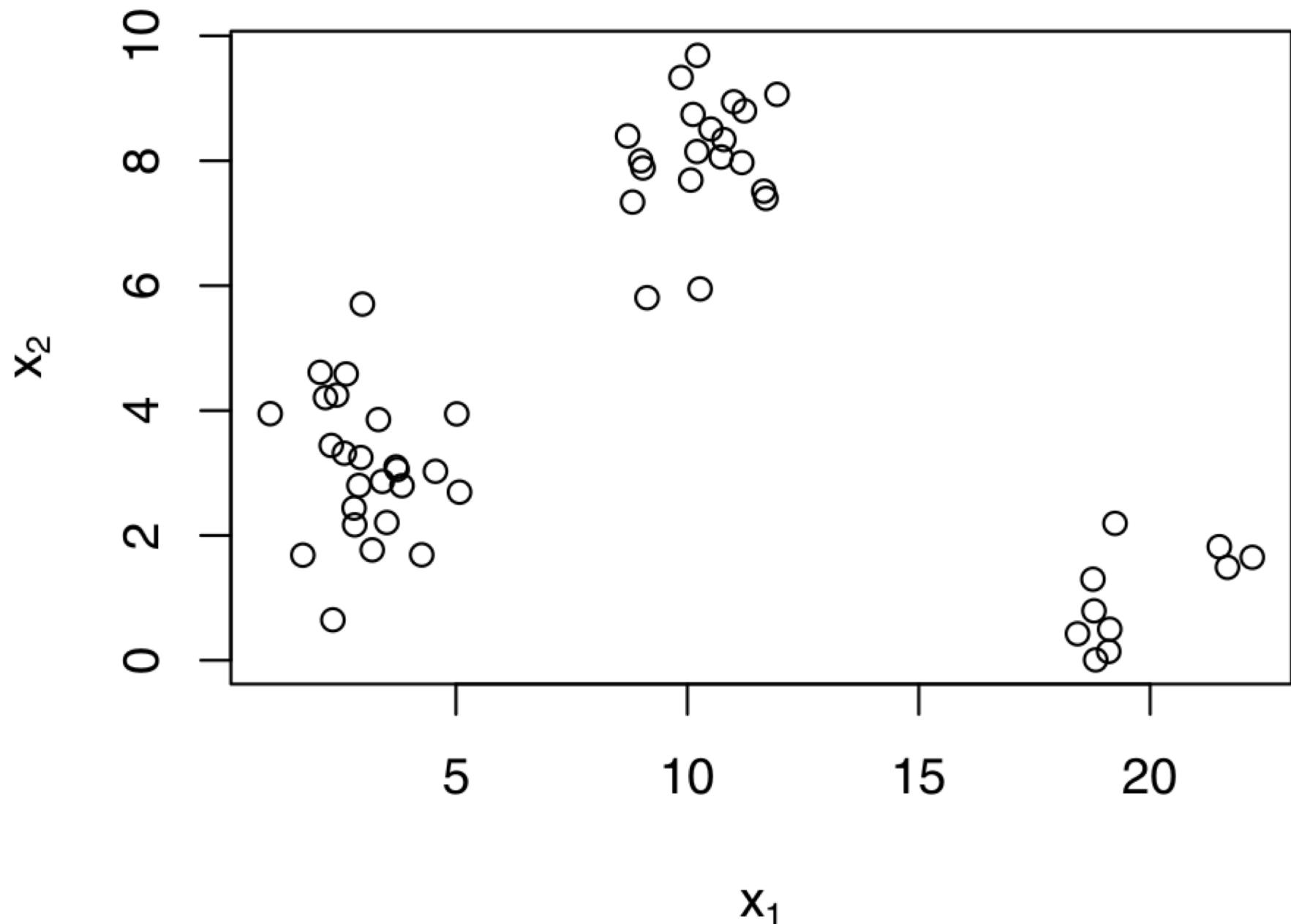


Fig. 6.1. Bivariate data showing the presence of three clusters.

Análise de clusters

- A conclusão de que existem três grupos naturais ou aglomerados de pontos é alcançada sem esforço ou pensamento consciente.
- Os clusters são identificados pela **avaliação das distâncias relativas entre os pontos** e, neste exemplo, a homogeneidade relativa de cada cluster e o grau de separação tornam a tarefa muito simples.
- O exame de gráficos de dispersão com base nos dados originais ou talvez nas primeiras pontuações dos principais componentes dos dados geralmente é uma fase inicial muito útil quando se pretende aplicar alguma forma de análise de cluster a um conjunto de dados multivariados.

Análise de clusters

- As técnicas de análise de clusters são descritas em detalhes em Gordon (1987, 1999) e Everitt, Landau, Leese e Stahl (2011).
- Neste capítulo do livro, apresentamos uma descrição relativamente breve de três tipos de métodos de *clustering*:
 - Técnicas hierárquicas aglomerativas ou procedimento hierárquico de agrupamento
 - Clustering k-means e
 - Clustering baseado em modelo

Análise de clusters

- O valor primário desta técnica está na classificação dos dados, pelo agrupamento natural.
- A técnica é comparável à análise factorial no seu objetivo de determinar a estrutura dos dados, mas a diferença entre elas é que a análise factorial só lida com grupo de variáveis, ao passo que análise de clusters lida com grupo de objetos ou de variáveis.
- O problema que se pretende resolver é:
 - dada uma amostra de n objetos (indíviduos) cada um deles medido segundo p variáveis, procurar um esquema de classificação que agrupe os objetos em g grupos.

Análise de clusters

Questões básicas:

- Como medir a semelhança entre objetos?
- Supondo que se possa medi-la, como colocar objetos semelhantes num mesmo cluster (aglomerado)?
- Após efetuado o agrupamento, como descrever os clusters e saber se eles são reais (têm significado palpável) e não são somente um simples artifício estatístico?

Pressupostos:

- A representatividade da amostra
- O impacto da multicolinearidade entre as variáveis

Análise de clusters

Alguns comentários importantes sobre a técnica:

- É usada precípuamente como técnica exploratória
- As soluções não são únicas (porque os membros de qualquer solução dependem dos critérios adotados)
- A solução obtida é totalmente dependente das variáveis usadas como base de mensuração da similaridade → a adição ou exclusão de variáveis (principalmente as mais relevantes) terá efeito substancial na solução

Análise de clusters

Processo:

1. Pesquisa dos objetivos, de caráter exploratória, dividindo conjunto de dados para formar grupos
 1. Descrição, simplificação de dados, identificação de relações → define objetivos e variáveis a usar
 2. Delineamento da pesquisa → deteção de outliers (decisão de mantê-los), padronização dos dados, medidas de similaridade (medidas correlacionais, medidas de distância)
 3. Pressupostos da análise (representatividade da amostra, impacto da multicolinearidade)
 4. Determinação e avaliação dos grupos (algoritmos de agrupamento, procedimentos hierárquicos de agrupamento)
2. Interpretação dos grupos, suas características, *labeling*
3. Validação da solução encontrada, descrevendo as características de clada cluster

Procedimento hierárquico de agrupamento

Um procedimento de cluster hierárquico aglomerado produz uma série de partições dos dados: P_n, P_{n-1}, \dots, P_1 .

O primeiro, P_n , consiste em n grupos de membros únicos, e o último, P_1 , consiste em um único grupo contendo todos os n indivíduos. A operação básica de todos os métodos é semelhante:

- (INÍCIO) Clusters C_1, C_2, \dots, C_n , cada um contendo um único indivíduo.
- (1) Encontre o par mais próximo de grupos distintos, digamos C_i e C_j , mescle C_i e C_j , exclua C_j , e diminua o número de clusters em um.
- (2) Se o número de clusters for igual a um, então pare; caso contrário, retorne a 1.

- Mas antes que o processo possa começar, uma matriz de distância inter-individual ou **matriz de similaridade** precisa ser calculada.
- Existem muitas maneiras de calcular distâncias ou semelhanças entre pares de indivíduos, mas aqui tratamos apenas de uma medida de distância comumente usada, a **distância euclidiana**, calculada como (ver capítulo 1 do livro):

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

- onde d_{ij} é a distância euclidiana entre o indivíduo i (com valores variáveis $x_{i1}, x_{i2}, \dots, x_{iq}$) e o indivíduo j (com valores variáveis $x_{j1}, x_{j2}, \dots, x_{jq}$). (Detalhes de outras possíveis medidas de distância e medidas de similaridade são apresentados em Everitt et al. 2011).
- As distâncias euclidianas entre cada par de indivíduos podem ser dispostas em uma matriz simétrica porque $d_{ij} = d_{ji}$ e possui zeros na diagonal principal. Essa matriz é o ponto de partida de muitos exemplos de agrupamento, embora o cálculo das distâncias euclidianas dos dados brutos possa não ser sensato quando as variáveis estão em escalas muito diferentes.
- Nesses casos, as variáveis podem ser padronizadas da maneira usual antes do cálculo da matriz da distância, embora isso possa ser insatisfatório em alguns casos (veja Everitt et al. 2011, para detalhes).

Dada uma matriz de distâncias inter-individual, o agrupamento hierárquico pode começar e, em cada estágio do processo, os métodos fundem indivíduos ou grupos de indivíduos formados anteriormente, que são os mais próximos (ou mais semelhantes).

Assim, à medida que os grupos são formados, é necessário calcular a distância entre um indivíduo e um grupo que contém vários indivíduos e a distância entre dois grupos de indivíduos. Como essas distâncias são definidas levará a uma variedade de técnicas diferentes.

Duas medidas inter-grupos simples são

$$d_{AB} = \min(d_{ij}), i \in A \text{ e } j \in B$$

$$d_{AB} = \max(d_{ij}), i \in A \text{ e } j \in B$$

onde d_{AB} é a distância entre dois grupos A e B, e d_{ij} é a distância entre os indivíduos i e j encontrados a partir da matriz de distância inter-individual inicial.

A primeira medida de distância entre grupos acima é a base do *clustering de ligação única*, e a segunda, a base do *clustering completo de ligação*.

Ambas as técnicas têm a propriedade desejável de serem invariantes sob transformações monótonas das distâncias interindividuais originais; ou seja, eles dependem apenas da classificação nessas distâncias, não de seus valores reais.

Uma outra possibilidade para medir a distância ou dissimilaridade entre aglomerados é

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij},$$

onde n_A e n_B são os números de indivíduos nos clusters A e B. Essa medida é a base de um procedimento comumente conhecido como **agrupamento médio por grupo**. Todas as três medidas intergrupos descritas acima estão ilustradas na próxima figura (vide figura 6.2.)

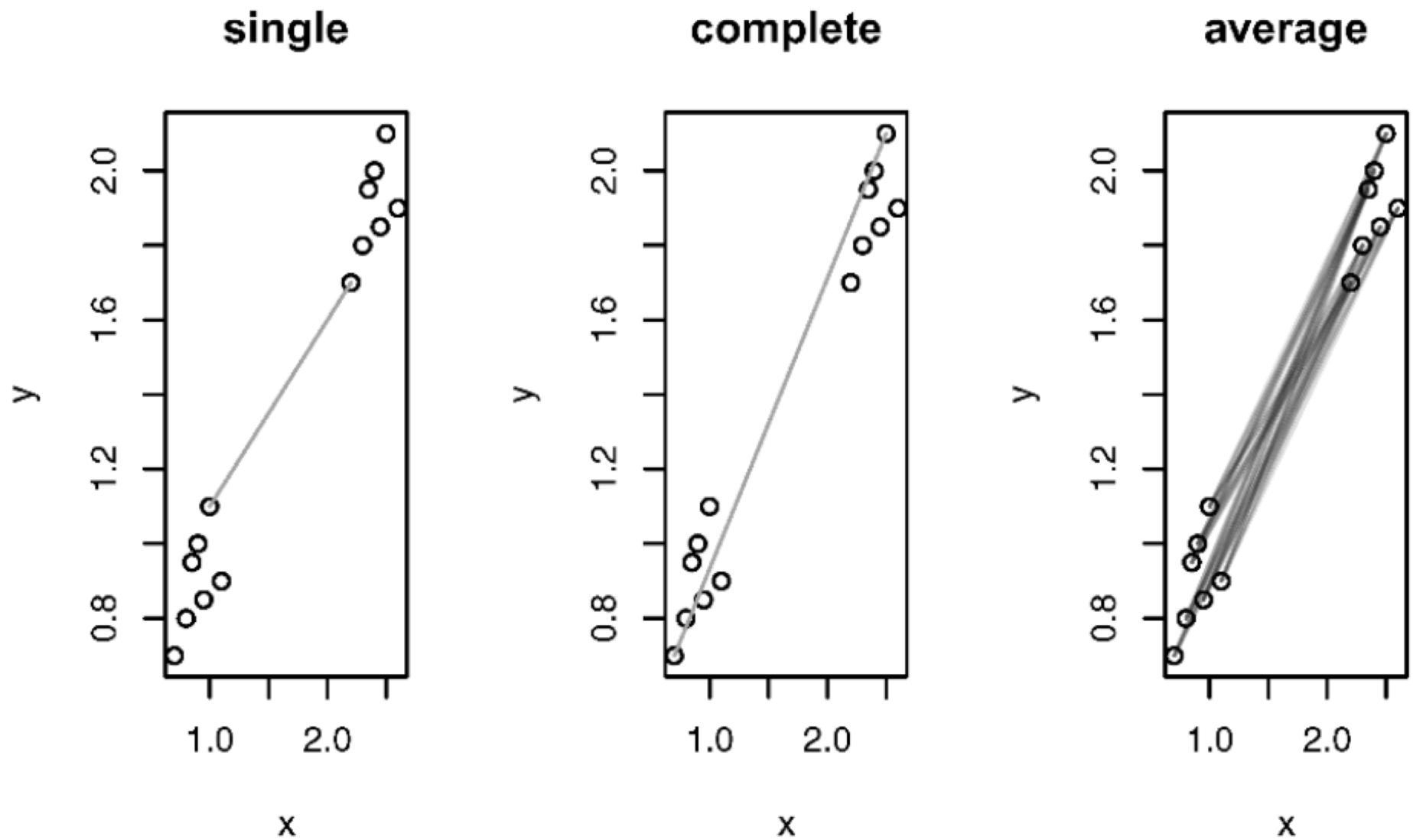


Fig. 6.2. Inter-cluster distance measures.

Como exemplo da aplicação dos **três métodos de agrupamento** (ligação única, ligação completa e média de grupo) cada um será aplicado às medidas de tórax, cintura e quadril de 20 indivíduos, fornecidas no Capítulo 1 do livro (data-set “measures”, vide Tabela 1.2).

As primeiras distâncias euclidianas são calculadas nas medições não padronizadas usando o seguinte código R:

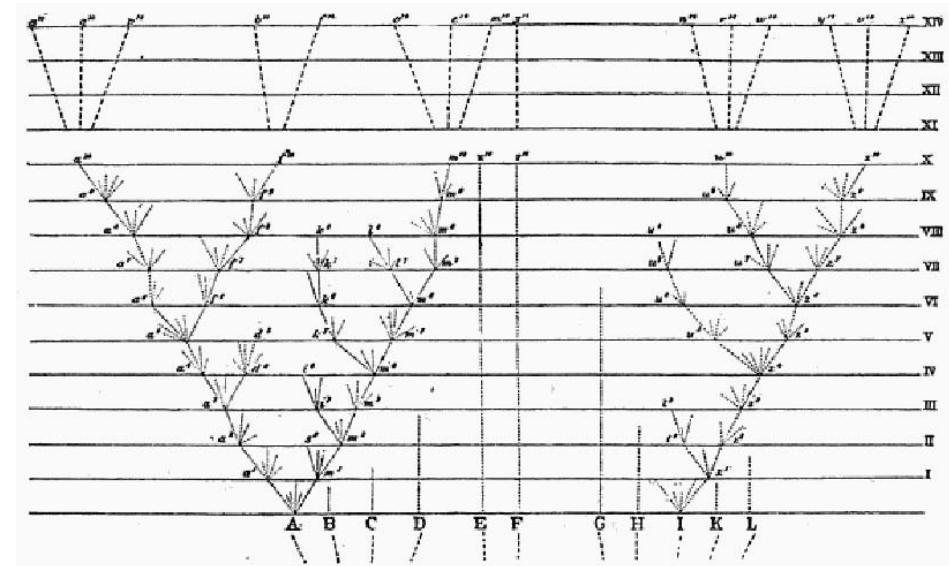
```
R> (dm <- dist(measure[, c("chest", "waist", "hips")]))
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2	6.16																		
3	5.66	2.45																	
4	7.87	2.45	4.69																
5	4.24	5.10	3.16	7.48															
6	11.00	6.08	5.74	7.14	7.68														
7	12.04	5.92	7.00	5.00	10.05	5.10													
8	8.94	3.74	4.00	3.74	7.07	5.74	4.12												
9	7.81	3.61	2.24	5.39	4.58	3.74	5.83	3.61											
10	10.10	4.47	4.69	5.10	7.35	2.24	3.32	3.74	3.00										
11	7.00	8.31	6.40	9.85	5.74	11.05	12.08	8.06	7.48	10.25									
12	7.35	7.07	5.48	8.25	6.00	9.95	10.25	6.16	6.40	8.83	2.24								
13	7.81	8.54	7.28	9.43	7.55	12.08	11.92	7.81	8.49	10.82	2.83	2.24							
14	8.31	11.18	9.64	12.45	8.66	14.70	15.30	11.18	11.05	13.75	3.74	5.20	3.74						
15	7.48	6.16	4.90	7.07	6.16	9.22	9.00	4.90	5.74	7.87	3.61	1.41	3.00	6.40					
16	7.07	6.00	4.24	7.35	5.10	8.54	9.11	5.10	5.00	7.48	3.00	1.41	3.61	6.40	1.41				
17	7.81	7.68	6.71	8.31	7.55	11.40	10.77	6.71	7.87	9.95	3.74	2.24	1.41	5.10	2.24	3.32			
18	6.71	6.08	4.58	7.28	5.39	9.27	9.49	5.39	5.66	8.06	2.83	1.00	2.83	5.83	1.00	1.00	2.45		
19	9.17	5.10	4.47	5.48	7.07	6.71	5.74	2.00	4.12	5.10	6.71	4.69	6.40	9.85	3.46	3.74	5.39	4.12	
20	7.68	9.43	7.68	10.82	7.00	12.41	13.19	9.11	8.83	11.53	1.41	3.00	2.45	2.45	4.36	4.12	3.74	3.74	7.68

Dendogramas

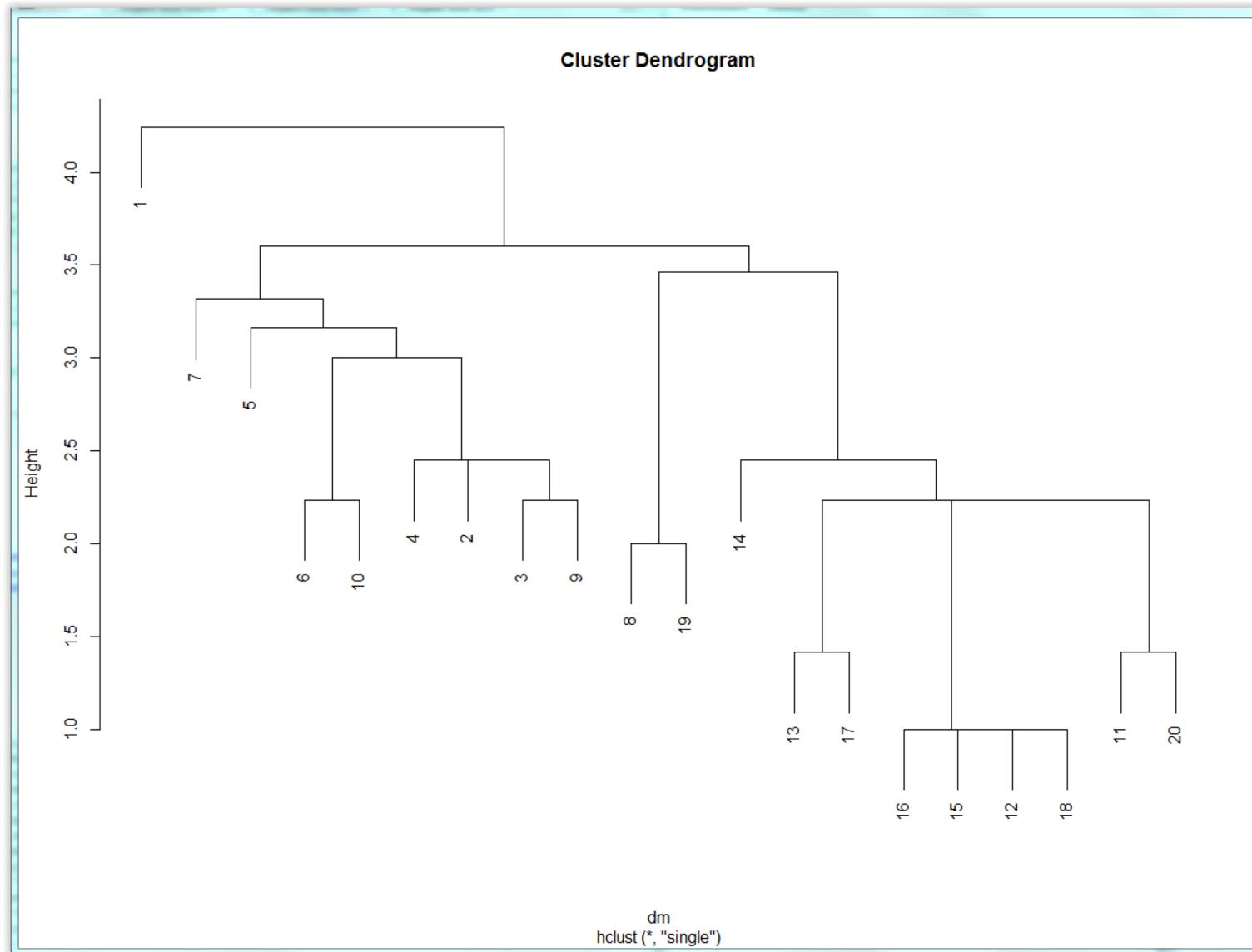
As classificações hierárquicas podem ser representadas por um diagrama bidimensional conhecido como dendrograma, que ilustra as fusões feitas em cada estágio da análise.

Um exemplo desse diagrama é dado na Figura 6.3. A estrutura se assemelha a uma *árvore evolutiva*, um conceito introduzido por Darwin sob o termo “Árvore da Vida” em seu livro *A Origem das Espécies pela Seleção Natural* de 1859.

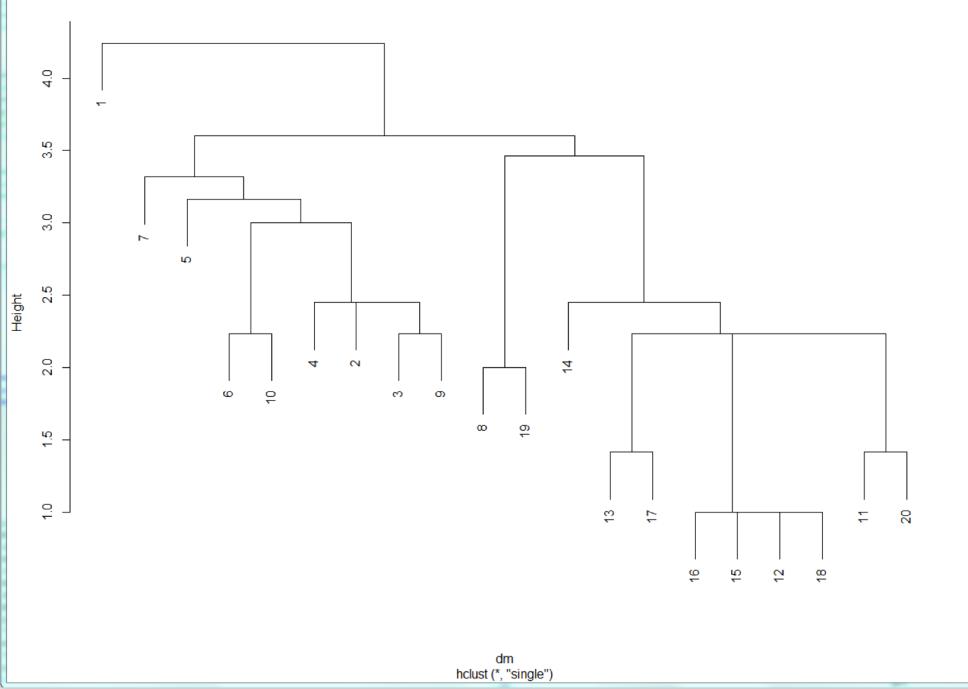


É nas aplicações biológicas que as classificações hierárquicas são mais relevantes e mais justificadas, embora esse tipo de cluster também seja usado em muitas outras áreas.

A aplicação de cada um dos três métodos de agrupamento descritos anteriormente na **matriz de distâncias** e uma **plotagem do dendrograma** correspondente são alcançados usando a função `hclust()`:

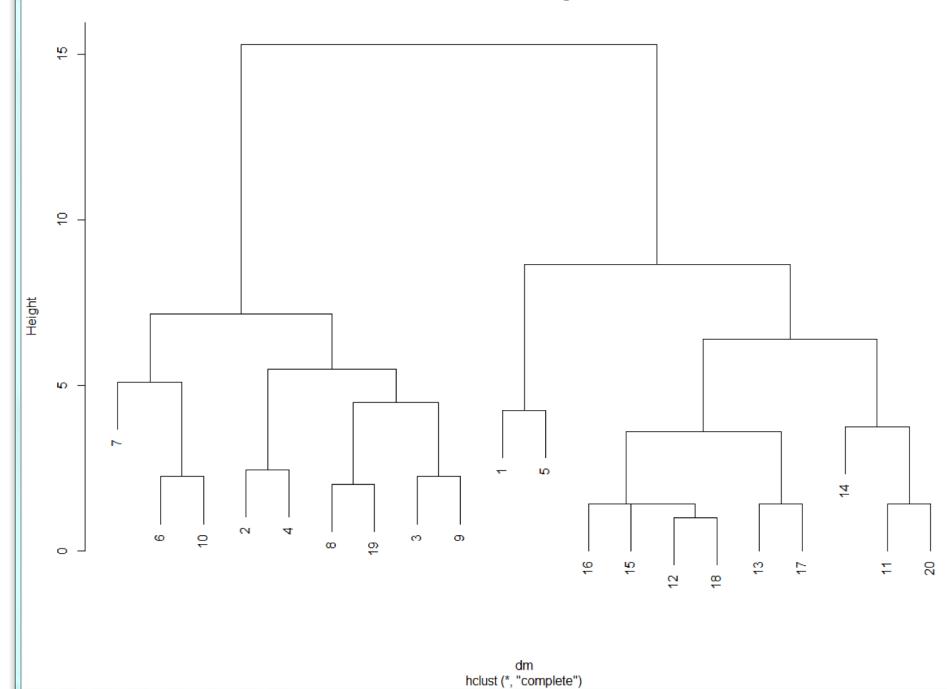


Cluster Dendrogram



```
dm  
hclust(*, "single")
```

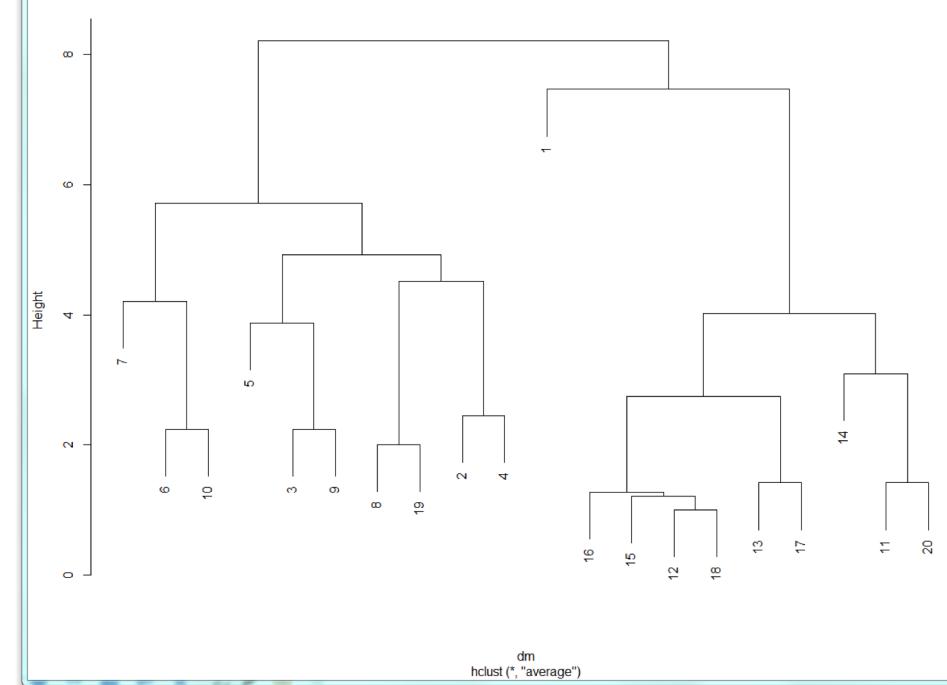
Cluster Dendrogram



```
dm  
hclust(*, "complete")
```

```
R> plot(cs <- hclust(dm, method = "single"))
R> plot(cc <- hclust(dm, method = "complete"))
R> plot(ca <- hclust(dm, method = "average"))
```

Cluster Dendrogram



```
dm  
hclust(*, "average")
```

Agora, precisamos considerar como selecionamos partições específicas dos dados (ou seja, uma solução com um número específico de grupos) a partir desses dendrogramas.

→ A resposta é que "cortamos" o dendrograma a alguma altura, e isso dará uma partição com um número específico de grupos.

Como escolhemos onde cortar ou, em outras palavras, como decidimos sobre um número específico de grupos que, em certo sentido, é ideal para os dados? Esta é uma pergunta mais difícil de responder.

Uma abordagem informal é examinar os tamanhos das mudanças de altura no dendrograma e fazer uma mudança “grande” para indicar o número apropriado de clusters para os dados. (Abordagens mais formais são descritas em Everitt et al. 2011).

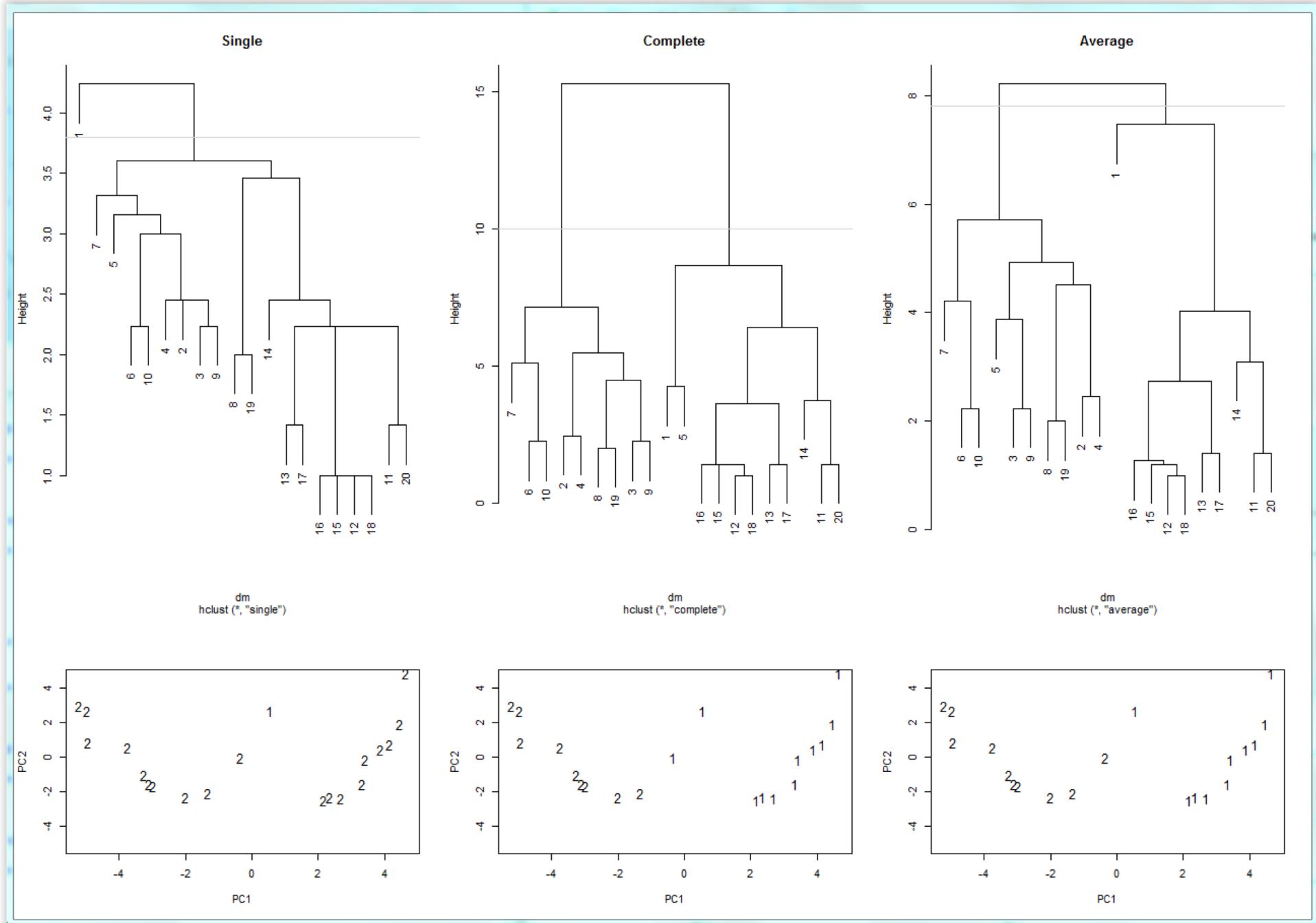
Mesmo usando essa abordagem informal nos dendrogramas da Figura 6.4, não é fácil decidir onde “cortar”.

Então, em vez disso, porque sabemos que esses dados consistem em medições de dez homens e dez mulheres, examinaremos as **soluções de dois grupos para cada método** que são obtidas cortando os dendrogramas em alturas adequadas.

Podemos exibir e comparar as três soluções graficamente, plotando as duas primeiras pontuações dos Componentes Principais dos dados, identificando os pontos para identificar a solução de cluster de um dos métodos usando o seguinte código:

```
R> body_pc <- princomp(dm, cor = TRUE)
R> xlim <- range(body_pc$scores[,1])
R> plot(body_pc$scores[,1:2], type = "n", xlim = xlim, ylim = xlim)
R> lab <- cutree(cs, h = 3.8)
R> text(body_pc$scores[,1:2], labels = lab, cex = 0.6)
```

Os gráficos resultantes são mostrados abaixo (vide Figura 6.4 no livro).



Analisando os dendogramas

As plotagens de dendogramas e dispersões dos componentes principais são combinadas em um único diagrama usando a função `layout()` (consulte a demonstração do capítulo para obter o código R completo).

O gráfico associado à solução de ligação única demonstra imediatamente um dos problemas com o uso desse método na prática, e esse é um fenômeno conhecido como *encadeamento*, que se refere à tendência de incorporar pontos intermediários entre os clusters em um cluster existente, em vez de iniciar um novo.

Como resultado, as soluções de ligação única geralmente contêm clusters "irregulares" longos que não fornecem uma descrição útil dos dados. As soluções de dois grupos de ligação completa e ligação média, também mostradas na Figura 6.4, são semelhantes e, em essência, colocam os homens (observações 1 a 10) juntos em um cluster e as mulheres (observações 11 a 20) no outro.

Fim

(introdução e exemplo no R)

Exercício

(aplicação da técnica no R com geração do *dataset* incluída)

(utilize o roteiro de comandos no R)

Aglomerando aviões de caça

Os dados mostrados na Tabela 6.1, como dados originalmente em Stanley e Miller (1979) e também em Hand et al. (1994) são os valores de seis variáveis para 22 aviões de caça dos EUA. As variáveis são as seguintes:

- FFD (*first flight day*): primeira data do voo, em meses após janeiro de 1940;
- SPR (*specific power*): potência específica, proporcional à potência por unidade de peso;
- RGF (*range factor*): fator de alcance do voo;
- PLF (*payload*): carga útil como uma fração do peso bruto da aeronave;
- SLF (*sustained load factor*): fator de carga sustentada;
- CAR (*carrier*): uma variável binária que assume o valor 1 se a aeronave puder poussar em um porta-aviões e 0 caso contrário.

FFD	SPR	RGF	PLF	SLF	CAR	
82	1.468	3.30	0.166	0.10	no	
89	1.605	3.64	0.154	0.10	no	
101	2.168	4.87	0.177	2.90	yes	
107	2.054	4.72	0.275	1.10	no	
115	2.467	4.11	0.298	1.00	yes	
122	1.294	3.75	0.150	0.90	no	
127	2.183	3.97	0.000	2.40	yes	
137	2.426	4.65	0.117	1.80	no	
147	2.607	3.84	0.155	2.30	no	
166	4.567	4.92	0.138	3.20	yes	
174	4.588	3.82	0.249	3.50	no	
175	3.618	4.32	0.143	2.80	no	
177	5.855	4.53	0.172	2.50	yes	
184	2.898	4.48	0.178	3.00	no	
187	3.880	5.39	0.101	3.00	yes	
189	0.455	4.99	0.008	2.64	no	
194	8.088	4.50	0.251	2.70	yes	
197	6.502	5.20	0.366	2.90	yes	
201	6.081	5.65	0.106	2.90	yes	
204	7.105	5.40	0.089	3.20	yes	
255	8.548	4.20	0.222	2.90	no	
328	6.321	6.45	0.187	2.00	yes	

← Variáveis originais do
dataset de aviões de
caça

Aplicaremos a *ligação completa* aos dados, mas usando apenas as variáveis de dois a cinco, (SPR, RGF, PLF, SLF), que significam na prática:

SPR → potência (power)

RGF → fator de alcance (range)

PLF → carga útil (payload)

SLF → carga sustentada (sustained)

Como as variáveis *estão em escalas muito diferentes*, as **padronizaremos** para terem **variância unitária** antes do agrupamento.

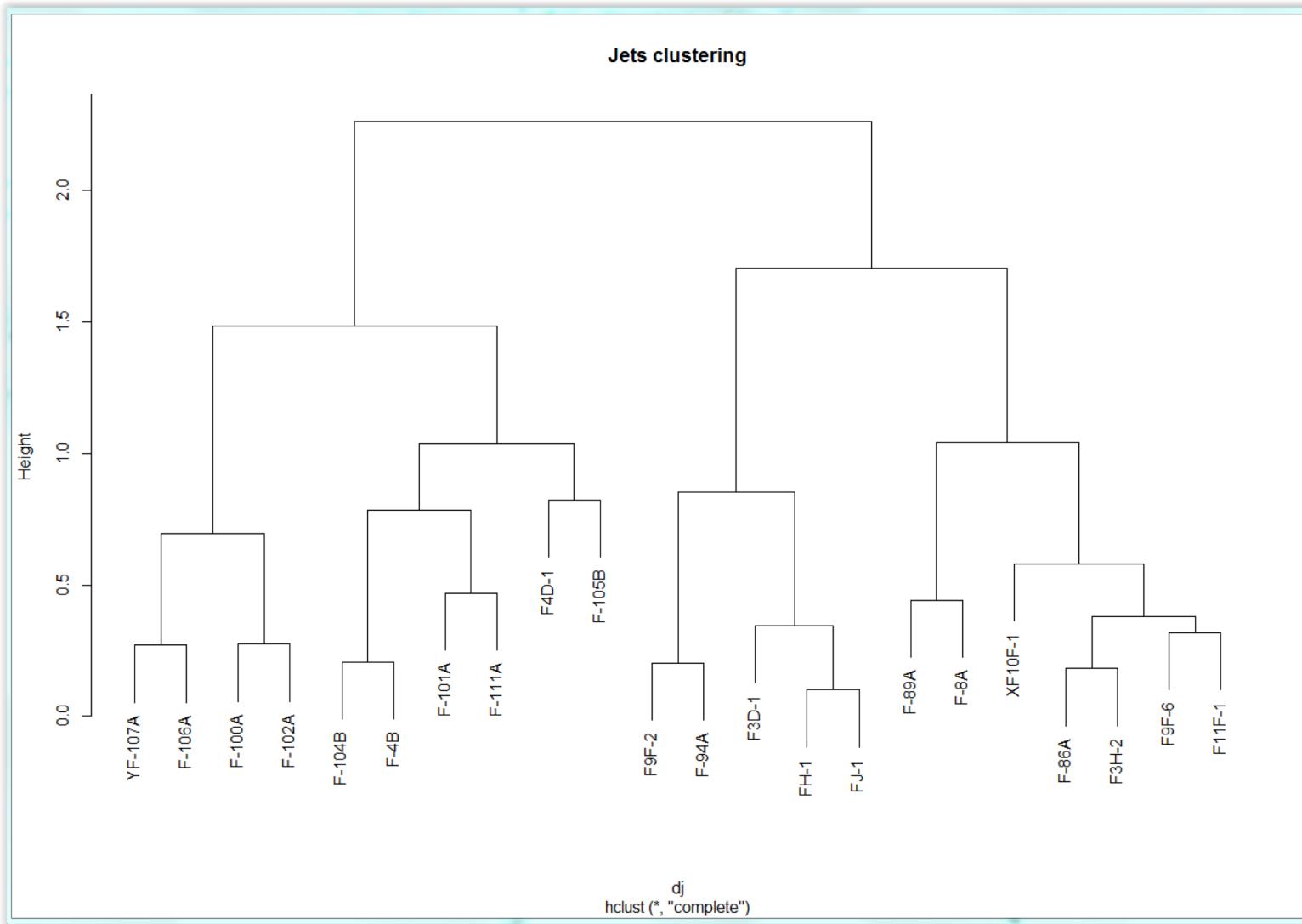
O código R necessário para padronização e armazenamento em cluster é o seguinte:

```
R> X <- scale(jet[, c("SPR", "RGF", "PLF", "SLF")], +  
center = FALSE, scale = TRUE)  
R> dj <- dist(X) ← função calcula distâncias  
R> plot(cc <- hclust(dj), main = "Jets clustering")  
R> cc
```

← função de análise de clusters, agrupamento hierárquico

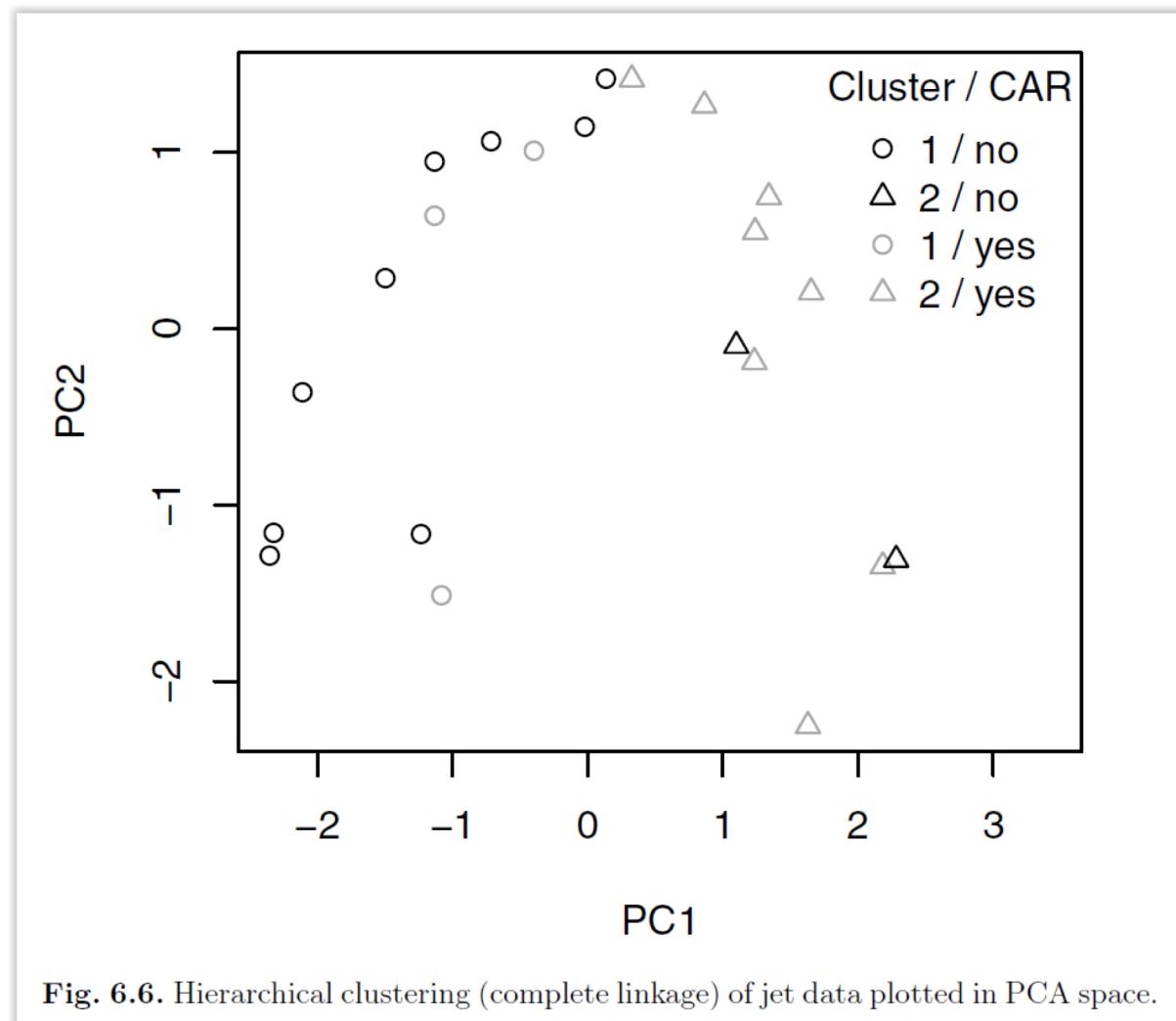
```
Call: hclust(d = dj)  
Cluster method : complete  
Distance : euclidean  
Number of objects: 22
```

O dendrograma resultante abaixo (Figura 6.5) sugere fortemente a presença de dois grupos de caças.



Na figura abaixo os dados são plotados no espaço dos dois primeiros componentes principais da matriz de correlação das variáveis relevantes (SPR para SLF).

A solução de dois grupos corresponde amplamente a aviões que podem e não podem pousar em um porta-aviões.



- Os pontos são rotulados pelo número do cluster para a solução de dois grupos e
- As cores usadas são os valores da variável CAR (que indica se o caça pousa em porta-aviões).

Novamente, cortamos o dendrograma de tal maneira que **dois aglomerados** permanecem e **plotamos** as classes correspondentes no espaço dos dois primeiros componentes principais veja a figura ao lado)

```
R> pr <- prcomp(dj)$x[, 1:2]

R> plot(pr, pch = (1:2)[cutree(cc, k =
2)], + col = c("black",
"darkgrey")[jet$CAR], + xlim =
range(pr) * c(1, 1.5))

R> legend("topright", col = c("black",
"black", + + +
"darkgrey", "darkgrey"), legend = c("1 /
no", "2 / no", "1 / yes", "2 /
yes"), pch = c(1:2, 1:2), title =
"Cluster / CAR", bty = "n")
```

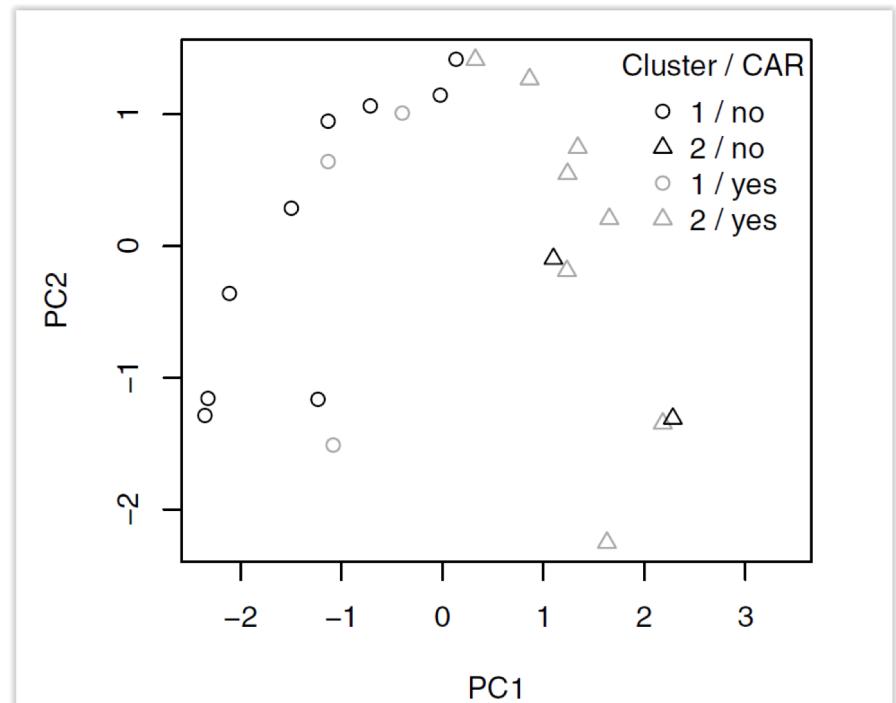


Fig. 6.6. Hierarchical clustering (complete linkage) of jet data plotted in PCA space.

Fim Exercício