XII

Congresso Anual
da Sociedade Portuguesa
de Estatística

UMA INTRODUÇÃO À ANÁLISE DE CLUSTERS

João A. Branco

Évora | 29 de Setembro a 2 de Outubro | 2004

SPE
SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Congresso Anual
da Sociedade Portuguesa
de Estatística

UMA INTRODUÇÃO À ANÁLISE DE CLUSTERS

João A. Branco

Évora | 29 de Setembro a 2 de Outubro | 2004

FICHA TÉCNICA

Título: Uma Introdução à Análise de Clusters

Autor: João A. Branco

Editora: Sociedade Portuguesa de Estatística

Concepção Gráfica da Capa: Maria José Patrocínio

Produção Gráfica e Impressão: Instituto Nacional de Estatística

Tiragem: 400 exemplares

ISBN: 972-98619-9-4

Depósito Legal: nº 213286/04

Agradecemos às seguintes entidades o valioso apoio concedido à realização do XII Congresso Anual da Sociedade Portuguesa de Estatística:

Adega Cooperativa de Borba
Banco BPI
British Council
Câmara Municipal de Évora
Câmara Municipal de Reguengos de Monsaraz
Centro de Estatística e Aplicações da Universidade de Lisboa
Centro de Investigação em Matemática e Aplicações da Universidade de Évora
Centro de Matemática Aplicada à Previsão e Decisão Económica
Centro de Matemática Aplicada do IST
Delta Cafés
Departamento de Matemática da Universidade de Évora
ÉvoraHotel
Embaixada de Espanha em Portugal
Fundação Caloustre Gulbenkian
Fundação Convento da Orada
Fundação para a Ciência e Tecnologia
Governo Civil do Distrito de Évora
Instituto Nacional de Estatística
Livraria Escolar Editora
Núcleo Minerva da Universidade de Évora
PSE-Produtos e Serviços de Estatística, Lda.
Região de Turismo de Évora
Rota dos Vinhos
Timberlake Consultants
Universidade de Évora

Prefácio

O processo que consiste em dividir uma colecção de objectos em grupos naturais, a partir de um conjunto de dados relativos aos objectos, é conhecido por vários nomes sendo o mais popular o de "análise de clusters". Os grupos são designados por clusters e a ideia central no processo de formação de clusters é a ideia de semelhança. Objectos que pertençam ao mesmo cluster são mais semelhantes entre si do que objectos pertencentes a clusters diferentes.

Um dos principais objectivos da análise de clusters é o de induzir uma classificação no conjunto dos objectos. E, como classificar é uma constante da actividade humana e, em particular, de toda a investigação científica, não admira que a prática da análise de clusters seja da maior importância em todas as áreas de trabalho. Por exemplo, em biologia é usada para derivar taxonomias de animais e plantas e agrupar genes de acordo com a sua funcionalidade, e em análise de mercados é usada para identificar nichos de clientes com iguais preferências.

Actualmente devido à grande produção de dados e à emergência de novas áreas de investigação (*data mining* e aprendizagem automática, por exemplo) a análise de clusters atravessa um vigoroso desenvolvimento. A sua prática e o conhecimento dos seus métodos interessam a um vasto leque de estudiosos.

O presente texto, preparado para servir de base ao mini-curso que antecede o XII Congresso da Sociedade Portuguesa de Estatística, é uma tentativa de fornecer, de uma forma breve, uma visão geral dos principais métodos e principais problemas da análise de clusters. Alguns métodos, mencionados no Capítulo 5, realmente aqueles menos usados pela generalidade dos utilizadores, não merecem aqui mais do que uma breve referência. Espera-se que a bibliografia referida seja um guia útil para permitir ir mais além no estudo destes métodos e das várias questões que se colocam em análise de clusters e que aqui não foram incluídas ou foram tratadas de forma leve.

O nível de matemática requerida para compreender o texto é mínimo e por isso tanto estudantes de estatística como utilizadores da análise de clusters em geral não deverão ter muita dificuldade em utilizá-lo. Os exercícios propostos no final de alguns capítulos são uma mistura de pequenos projectos requerendo a utilização do computador, exemplos numéricos e questões teóricas, sendo que estas últimas correspondem muitas vezes a desenvolvimentos de ideias apresentadas no texto.

Quero agradecer aos meus amigos Sérgio Valente que elaborou um texto para descrever o exemplo dos automóveis do Capítulo 6 e no qual a descrição que aqui é feita se inspirou, à Patrícia Ferreira que, com paciência, produziu a Figura 1.1 e as Tabelas 1.1, 2.6 e 2.7 e à Ana Pires que cometeu a loucura de se oferecer, e no fim eu não fui capaz de recusar, para digitar o texto. A sua ajuda foi preciosa, não só porque me libertou da tarefa de digitação, muito morosa para mim, mas também pelas críticas e sugestões que foi fazendo à medida que o texto se ia desenvolvendo.

É claro que os erros e omissões que restam são da inteira responsabilidade do autor que, desde já, agradece a todos os que queiram comunicar falhas que encontrarem ou contribuir com comentários construtivos.

Instituto Superior Técnico,

Lisboa, Julho de 2004

J. A. B.

Índice

1

Introdução

1.1 Classificação

A necessidade de reunir objectos semelhantes para formar grupos, dando origem a uma classificação, tem acompanhado a evolução da humanidade desde os seus primórdios até aos nossos dias. Imagina-se que no princípio esta actividade primitiva seria um recurso para a própria sobrevivência, pois era vital reconhecer os animais ferozes, os frutos comestíveis e identificar as aves avisadoras de perigos e de climas. Mais tarde classificar objectos tornou-se uma prática essencial ao aperfeiçoamento da organização da sociedade e ao progresso da maior parte das áreas científicas. A classificação de animais e plantas está na origem da teoria evolucionista de Darwin, a classificação dos elementos químicos influenciou o conhecimento da estrutura do átomo, a classificação das estruturas, dos fenómenos e dos tempos geológicos é um aspecto fundamental na explicação da origem e evolução do planeta em que vivemos, a classificação dos astros contribuiu para formular as teorias em que se baseia a moderna astronomia e a classificação das palavras permite compreender a estrutura das línguas e saber como elas evoluíram. Enquanto que estas classificações tradicionais estão fortemente dependentes de elaboradas teorias sobre os fenómenos a classificar, modernamente é comum proceder à classificação de fenómenos sobre os quais não existem teorias bem definidas ou até sobre os quais não existe um grande conhecimento, como são os casos, por exemplo, de *data mining* e análise de mercados. Compreende-se assim que a classificação continue a constituir uma ferramenta fundamental na evolução do conhecimento actual. De facto os objectivos finais da classificação são a análise de fenómenos com vista a identificar quais são as suas partes, como é que elas se interligam, e usar este conhecimento assim criado para predizer e controlar o comportamento dos grupos e dos seus membros.

O homem moderno vive numa sociedade que está organizada por grupos. Fazem parte integrante do nosso mundo os países, as escolas, empresas, profissões, partidos políticos, religiões e clubes da mais variada natureza,

uma lista que o destino vai estendendo de maneira inexorável com novos grupos de entidades e ideias que vão surgindo com o passar do tempo e com o evoluir do conhecimento. A necessidade de construir grupos é uma característica da actividade humana e sem dúvida um suporte essencial do método de aprendizagem e do próprio método científico em geral.

Classificação é um conceito científico para o qual é difícil dar uma definição. Boris Mirkin sugere, em Mirkin (1996), uma actualização de uma definição inicial avançada, há longos anos, por J.S. Mill (1806–1873). De acordo com a proposta actualizada,

> Classificação é o verdadeiro ou ideal arranjo em conjunto daqueles que são iguais, e a separação daqueles que são diferentes, sendo que a finalidade deste arranjo é primeiramente:
>
> (i) formar e conservar o conhecimento,
> (ii) analisar a estrutura do fenómeno,
> (iii) relacionar entre si os aspectos do fenómeno em questão.

Uma análise da classificação nas várias ciências revela que para compreender a estrutura do fenómeno e a relação entre os seus diversos aspectos é conveniente basear a análise em cinco dimensões relevantes: história, estrutura, funcionalidade, atitude/acção. Por exemplo, enquanto que na classificação em geologia se destacam as dimensões relativas à estrutura e à história (origem e evolução das camadas rochosas) em biologia interessa considerar a estrutura, a história e a funcionalidade (como o movimento e a reprodução).

Embora a classificação constitua a base para a compreensão das coisas do mundo em que vivemos e a sua prática seja tão longa como a existência do próprio homem, não parece que se tenha constituído numa disciplina académica, coerente e independente, nem que pertença a alguma das disciplinas académicas estabelecidas. A sua esfera de acção abrange todas as áreas do conhecimento e é em cada uma delas que as teorias e métodos de classificação têm florescido.

O acto de classificar assenta numa diversidade grande de conhecimentos, teorias e critérios, e é esta imensa dispersão que possivelmente tem dificultado a montagem de uma estrutura que permita elevar a classificação à categoria de ciência. Esta situação peculiar da classificação vem constituindo motivo de reflexão de grandes cientistas especializados em classificação (Hartigan, 1996 e Mirkin, 1996).

Sem esquecer estas preocupações, algumas de natureza filosófica, muita

da actividade ligada à classificação tem-se concentrado em aspectos matemáticos definidos com o objectivo concreto de construir grupos a partir de um conjunto inicial de indivíduos ou objectos. A frase seguinte, extraída do prefácio em Mirkin (1996), traduz, em minha opinião, o estado em que actualmente se encontra a classificação, enquanto corpo difuso de uma grande variedade de conhecimentos:

> *The science of classification, which deals with the problems of how classifications emerge, function and interact, is still unborn. What we have in hand currently is clustering, the discipline aimed at revealing classifications in observed real-world data.*

1.2 Análise de clusters

A análise de clusters é apenas mais uma maneira de produzir classificações. Enquanto que a classificação é um processo muito geral, actuando sobre fenómenos do mundo real e usando todo o tipo de conhecimento disponível, a análise de clusters actua sobre um conjunto de dados e usa métodos matemáticos geralmente de forma automática.

Partindo de um conjunto de dados relativos a objectos de uma colecção, o seu objectivo consiste em construir grupos ou clusters de tal forma que objectos dentro do mesmo grupo são mais semelhantes do que objectos situados em grupos diferentes, induzindo, desta maneira, uma classificação na colecção inicial. Com grande frequência verificamos que este procedimento é prática corrente em todos os ramos da actividade científica. E é pretensão do analista identificar grupos genuínos e relevantes para os seus estudos sempre que eles de facto existam, desejando ao mesmo tempo não ser ludibriado por revelações arbitrárias de grupos sem qualquer interesse que a análise possa revelar. Interessa conhecer como é que a análise de clusters responde a esta dupla exigência. Para isso é preciso percorrer um longo caminho. O que se segue constitui uma parte de aspectos importantes da análise de clusters. Antes de começar é conveniente lembrar que a análise de clusters assim como todo o processo de classificação em geral se aplica tanto a animais como plantas, rochas, minerais e a todo o tipo de fenómenos, factos e casos. Neste texto usa-se o termo "objectos" para designar as entidades que são elementos de um grupo, quaisquer que elas sejam, podendo, em casos particulares, usar-se a designação que melhor descreva e identifique essa entidade. A formação de grupos a partir da totalidade de elementos de uma colecção inicia-se com o estudo de um ou mais atributos, propriedades ou características dos seus elementos. Para referir qualquer destas designações usa-se em geral o termo "variáveis". Por

vezes a análise tem por objectivo agrupar as características ou variáveis. Nesse caso os objectos da análise são as próprias variáveis.

1.2.1 Objectivos da análise de clusters

Os objectivos da análise de clusters são essencialmente os mesmos da classificação mas como a análise de clusters é aplicada a um conjunto de dados pode dizer-se que esses objectivos são a análise do conjunto de dados com vista a obter uma classificação. Essa análise pode apontar para um ou mais dos seguintes objectivos específicos:

(i) A exploração dos dados
Inicialmente não é sabido se os dados estão ou não estruturados em grupos e espera-se que a análise venha a revelar a estrutura subjacente. Este é o objectivo mais imediato da análise de clusters.

(ii) Redução de dados
Se os grupos obtidos são homogéneos significa que os n objectos podem ser substituídos por um número muito menor de grupos ou seus representantes.

(iii) Geração de hipóteses
A análise pode revelar grupos não esperados pelo utilizador, facto que leva naturalmente a sugerir conjecturas e hipóteses para explicar a estrutura revelada. Uma hipótese muito frequente refere-se ao número de grupos encontrados. Para testar esta hipótese usa-se um novo conjunto de dados.

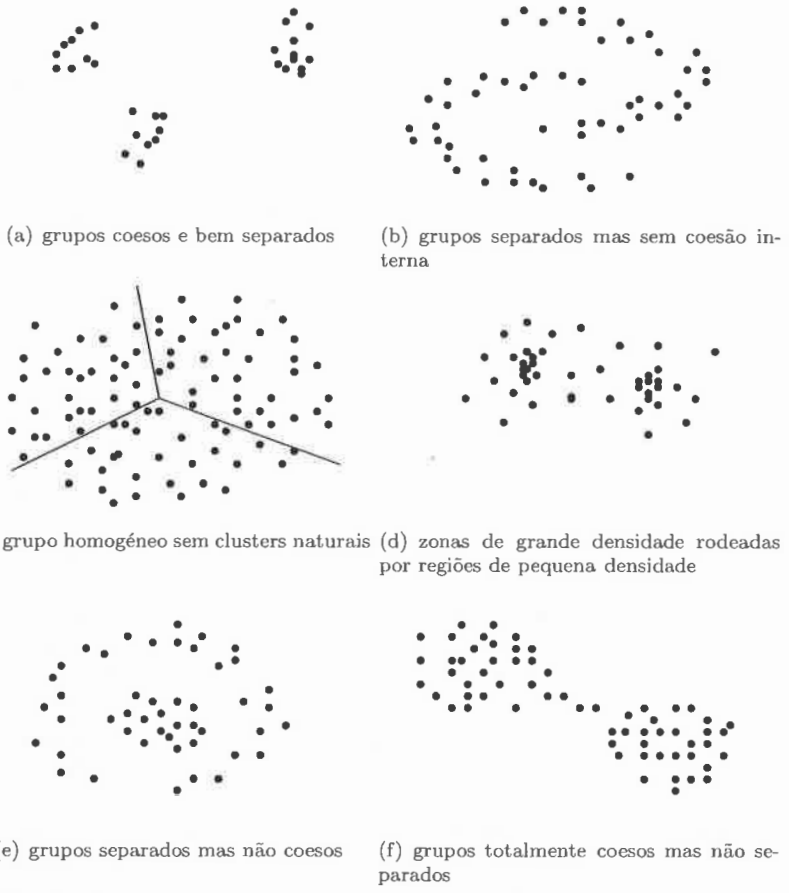
(iv) Predição
Os objectos do mesmo grupo têm propriedades semelhantes. Este facto torna possível predizer propriedades desconhecidas de um objecto, comparando o objecto com os objectos do grupo. Além disso, conhecidos os grupos é possível classificar um novo objecto, isto é, predizer a que grupo ele pertence.

A análise de clusters, com os objectivos que aqui são definidos, é por vezes referida com outros nomes: (i) classificação automática, um termo que aparece geralmente nas publicações em língua francesa; (ii) aprendizagem não supervisionada, linguagem típica da área de reconhecimento de padrões; (iii) taxonomia numérica, muito usada nas ciências da vida; (iv) classificação, termo que aparece também em estatística mas com um sentido diferente, o sentido da análise discriminante, significando a atribuição de objectos a grupos já definidos.

1.2.2 Formas de clusters

Neste texto usam-se indistintamente as palavras grupos e clusters para designar o produto de uma análise de clusters.

Antes de prosseguir interessa reflectir sobre o que é de facto um cluster. Intuitivamente é fácil imaginar ou definir o que é um cluster mas infelizmente não é fácil transpor a definição intuitiva para uma definição formal que seja também operacional. Por isso esta é considerada uma primeira e grande dificuldade da análise de clusters.



(a) grupos coesos e bem separados

(b) grupos separados mas sem coesão interna

(c) grupo homogéneo sem clusters naturais

(d) zonas de grande densidade rodeadas por regiões de pequena densidade

(e) grupos separados mas não coesos

(f) grupos totalmente coesos mas não separados

Figura 1.1 *Formas de clusters mostrando coesão interna e/ou isolamento externo.*

Na Figura 1.1 o caso (a) apresenta três clusters e os casos (b), (d), (e) e (f) podem descrever-se como consistindo em dois clusters cada um. Mas qual deve ser a definição formal de cluster que a ser usada produz precisamente os clusters que aqui se vislumbram?

As ideias que estão por detrás do conceito de cluster são: (i) os objectos de um cluster devem ser homogéneos no sentido de serem mais semelhantes quando comparados entre si do que quando comparados com objectos de outros clusters; (ii) os clusters devem estar bem separados. Cormack (1971) traduziu estas ideias por coesão interna e isolamento externo. Alguns métodos usados na construção de clusters incidem, umas vezes, mais na coesão interna e, outras vezes, mais no isolamento externo.

Os casos (a), (b), (c) e (d) da Figura 1.1 tentam reproduzir várias formas de clusters, apresentadas em Gordon (1999), em que se destaca a coesão interna e/ou o isolamento externo. O caso (c) mostra um grande grupo homogéneo onde não há clusters naturais. A divisão em três clusters que se apresenta é uma divisão artificial e por conveniência, semelhante àquela que consiste em dividir a região de uma cidade em zonas associadas a corporações de bombeiros, para efeitos de combate a fogos.

1.2.3 Alguma literatura relevante

Os primeiros desenvolvimentso da análise de clusters aconteceram principalmente em três áreas: na biologia, o que conduziu à taxonomia numérica, na psicologia, com o agrupamento de variáveis na análise factorial, e no reconhecimento de padrões, com a aprendizagem não supervisionada.

Nos anos sessenta do século XX a análise de clusters tornou-se muito popular absorvendo muitos investigadores interessados na produção, talvez exagerada, como se refere em Cormack (1971), de algoritmos e *software* apropriados para a construção de grupos. Muitos destes algoritmos foram depois integrados nos principais programas de *software* em circulação. Everitt *et al.* (2001) apresentam uma extensa lista de *packages* com programas para realizar análise de clusters. Entre eles estão SPSS, SAS, S-PLUS e Statistica. Além destes *packages* estatísticos de grande abrangência, mas com uma variedade de programas suficiente para servir os utilizadores em geral, a lista contém indicações de ligações a páginas *Web* onde se pode encontrar *software* para análise de clusters que pode ser usado *on-line*. Alguns *packages* como Clustan são especializados só em análise de clusters. Há ainda programas específicos para métodos menos habituais como os métodos para modelos de mistura e os métodos de cobertura, mencionados no Capítulo 5. Outras bibliotecas de programas são desenvolvidas para problemas concretos em áreas específicas, como é o caso do EPCLUST (*Expression Profile data CLUSTering and analysis*) pertencente ao *software*

EP (*Expression Profiler*) para análise de dados de genética. EPCLUST é o módulo especializado em análise de clusters de matrizes de dados relativos à expressão de genes observada em várias condições. Este *software* está disponível na *Web* e permite que diferentes investigadores possam partilhar bases de dados, resultados, análises e experiências de investigação. Não se indicam aqui os sítios da *Web* para acesso a *software on-line*. Os sítios mudam com frequência e é preferível o uso de motores de busca e palavras chave que facilmente conduzem aos sítios de interesse.

A produção de artigos e livros sobre a análise de clusters é muito vasta. Arabie *et al.* (1996) contém uma extensa lista de trabalhos fundamentais que mostram o grande interesse na investigação e utilização da análise de clusters. Entre as várias publicações destacam-se o texto clássico de Sokal and Sneath (1963) que é agora de interesse histórico. Outros textos de interesse são Jardine and Sibson (1971), Anderberg (1973), Sneath and Sokal (1973), Duran and Odell (1974), Romesburg (1984), Jain and Dubes (1988), McLachlan and Basford (1988), Kaufman and Rousseeuw (1990), Gordon (1999) e Everitt *et al.* (2001). No sentido de controlar de forma organizada o grande volume de publicações a "Classification Society of North America" publica a bibliografia anual, de livros e artigos, Classification Literature Automated Search Service, que inclui não só a análise de clusters mas também multidimensional scaling.

1.2.4 Aplicações

A razão pela qual a análise de clusters é tão popular e tão útil é simplesmente porque há necessidade constante de produzir classificações nos vários ramos da actividade científica e da actividade humana em geral. Daí que as novas ideias e que os progressos do próprio tema que é a análise de clusters tenham sido introduzidos por investigadores das mais diversas especialidades.

Os exemplos da aplicação da análise de clusters nos vários domínios constituem uma lista sem fim. Presentemente, com a produção de grandes quantidades de dados, a análise de clusters tornou-se um tópico de investigação muito activo e que está a ter um desenvovimento vigoroso. Os exemplos que a seguir se apresentam pretendem ilustrar a variedade de temas para os quais a análise de clusters é relevante. Incluem-se áreas onde a aplicação da análise de clusters tem fortes tradições e outras, actualmente em grande desenvolvimento, onde essa aplicação é mais recente mas já considerada de grande interesse para a resolução dos problemas que aí se colocam.

Alguns exemplos

- **Arqueologia**
  A identificação de grupos de artefactos semelhantes usados por povos já desaparecidos ajuda a compreender muitos aspectos de civilizações antigas (Hodson, 1971 e Hodson *et al.*, 1966).
- **Sismologia**
  A predição de abalos sísmicos é da maior importância e a análise de clusters tem sido usada com esse objectivo (Davis and Frolich, 1991 e Wardlaw *et al.*, 1991).
- **Medicina**
  Para identificar as causas das doenças e criar os seus tratamentos é preciso separá-las em grupos ou tipos de doenças (Burbank, 1972 e Petrakis and Faloutsos, 1997).
- **Psiquiatria**
  Agrupar as doenças psiquiátricas facilita o seu conhecimento e permite detectar as suas causas e identificar o seu tratamento (Smart *et al.*, 2003).
- **Análise de mercados**
  Segmentos de consumidores ou produtos são em geral clusters que é importante conhecer para perceber a estrutura do mercado (De Sarbo *et al.*, 1993 e Arabie and Hubert, 1994).
- **Dados de *microarrays***
  Dados de *microarrays* resultam da actuação de novas tecnologias em estudos de genética. A expressão genética de vários genes (geralmente em número muito grande) é observada e medida em várias amostras, ou segundo várias condições ou em vários momentos ao longo do tempo. A análise de clusters é usada para agrupar tanto os genes com perfis de expressão semelhantes ao longo das amostras, como as amostras com perfis de expressão semelhantes ao longo dos genes (Parmigiani *et al.*, 2003).
- ***Data mining***
  Data mining é o processo de identificar grupos de registos semelhantes e extrair conhecimento de grandes bases de dados. A análise de clusters constitui um dos primeiros passos do processo de *data mining* (Han and Kamber, 2001).
- **Classificação de documentos**
  A procura de informação em grandes bases de dados, em particular informação existente de forma dispersa na *Web*, fica grandemente facilitada se os documentos estiverem agrupados em clusters (Willet, 1990).

1.3 Estrutura dos dados

A natureza é fértil em oferecer muitos tipos de dados que geralmente são apresentados na forma de tabelas. A diversidade de tabelas relativas a dados é grande mas a análise de clusters opera essencialmente sobre dois tipos de estruturas de dados, identificadas por dois formatos de matrizes.

(i) A matriz de dados ou matriz de perfis, uma matriz $\mathbf{X} = [x_{ij}]$, $i = 1, \ldots, n$ e $j = 1, \ldots, p$, em que x_{ij} representa o valor que a variável j assume quando medida no objecto i. Uma matriz de dados típica encontra-se na Tabela 1.1 onde estão registados valores de várias características (variáveis) dos nove planetas (objectos) do Sistema Solar. Esta matriz inclui:

- Variáveis quantitativas
  As variáveis quantitativas são caracterizadas pelo facto de a cada objecto ser atribuída uma medida ou valor numérico. São variáveis quantitativas as oito primeiras variáveis da Tabela 1.1 (Distância do planeta ao Sol, o seu Diâmetro, Massa, Densidade e Gravidade, os Tempos de Translação e Rotação e o número de Satélites). Entre as variáveis quantitativas distinguem-se as
  - variáveis contínuas, aquelas que podem tomar todos os valores de um intervalo e que estão associadas a um processo de medição. As sete primeiras variáveis quantitativas são variáveis contínuas.
  - Variáveis discretas, aquelas que só podem assumir um número finito ou infinito numerável de valores e que estão associadas a um processo de contagem. O número de Satélites é a única variável discreta no conjunto das oito variáveis quantitativas.
- Variáveis qualitativas
  Com as variáveis qualitativas cada objecto é submetido a uma classificação, sendo-lhe atribuída uma classe ou categoria. As três últimas variáveis da Tabela 1.1 são variáveis qualitativas (existência de Anéis, tipo de Superfície, existência de Campo Magnético global). Nas variáveis qualitativas destacam-se as
  - Variáveis nominais, caracterizadas pelo facto de não ser possível estabelecer uma ordem entre as várias categorias. No caso presente todas as variáveis qualitativas são nominais, sendo as variáveis Anéis e Campo Magnético designadas por binárias ou dicotómicas uma vez que apresentam apenas duas categorias.
  - Variáveis ordinais, aquelas em que é possível estabelecer uma ordem entre as suas categorias. Este tipo de variável

não está contemplado na Tabela 1.1. para dar um exemplo pode pensar-se na variável Satélites e supôr que em vez do número de satélites se indicavam categorias significando que esse número é: nulo (0), pequeno (1, 2), médio (13) e grande (26, 31, 61). Com esta classificação a variável é entendida como uma variável ordinal, na medida em que é possível introduzir uma ordem entre as classes, podendo materializar-se essa ordem com os códigos: 1 - nulo, 2 - pequeno, 3 - médio e 4 - grande.

Finalmente é de salientar que a Tabela 1.1 mostra a existência de um valor omisso, relativo ao Campo Magnético global para Plutão.

(ii) A matriz de dissemelhanças (semelhanças), uma matriz $\mathbf{D} = [d_{ij}]$ ($\mathbf{S} = [s_{ij}]$), $i, j = 1, \ldots, n$, quadrada e em geral simétrica em que d_{ij} (s_{ij}) representa o valor da dissemelhança (semelhança) entre o objecto i e o objecto j. Estes conceitos serão melhor explicados no Capítulo 2. Neste texto faz-se uso da noção de dissemelhança apenas. Com uma transformação simples, como por exemplo, $d_{ij} = k - s_{ij}$ (constante) pode geralmente obter-se uma matriz de dissemelhanças a partir de uma matriz de semelhanças.

Como muitos algoritmos usados em análise de clusters operam sobre matrizes de dissemelhanças isso faz com que uma primeira tarefa da análise de clusters seja a de construir, a partir de uma matriz de dados, a matriz de dissemelhanças à qual os métodos podem então ser aplicados.

Uma matriz de dissemelhanças (objectos por objectos ou variáveis por variáveis) pode ser obtida a partir da transformação de uma matriz de dados (objectos por variáveis). A Tabela 1.2 apresenta a matriz de dissemelhanças construída com base na matriz de dados dos planetas, referida na Tabela 1.1, considerando apenas as três variáveis, diâmetro, massa e densidade do planeta. Nestas condições os nove planetas podem ser representados por nove pontos num espaço de três dimensões. Definiu-se a dissemelhança entre dois planetas como a distância euclidiana entre os dois pontos que os representam. A matriz de dissemelhanças contém 36 dissemelhanças distintas que correspondem aos 36 pares que se podem formar com nove planetas.

Muitas vezes a matriz de dissemelhanças é observada directamente. Como acontece frequentemente em psicometria onde os indivíduos são solicitados a indicar numa escala a dissemelhança entre os elementos de todos os pares de objectos que é possível formar com os objectos disponíveis.

Um exemplo em que as dissemelhanças são observadas directamente envolve 13 objectos identificados com 13 expressões da face de uma mulher, em que cada expressão traduz a reacção a cada um dos 13 cenários descritos

Tabela 1.1 *Algumas características dos planetas do Sistema Solar.*

Planeta	Dist. ao Sol	Diâm.	Massa	Dens.	Grav.	Translação	Rotação	Satélites	Anéis	Superfície	Campo Magnético
Mercúrio	0.387	0.383	0.0553	0.984	0.378	0.241	58.8	0	Não	Sólida	Sim
Vénus	0.723	0.949	0.815	0.951	0.907	0.615	-244	0	Não	Sólida	Não
Terra	1	1	1	1	1	1	1	1	Não	Sólida	Sim
Marte	1.52	0.533	0.107	0.713	0.377	1.88	1.03	2	Não	Sólida	Não
Júpiter	5.20	11.21	317.8	0.240	2.36	11.9	0.415	61	Sim	Líquida	Sim
Saturno	9.58	9.45	95.2	0.125	0.916	29.4	0.445	31	Sim	Líquida	Sim
Urano	19.20	4.01	14.5	0.230	0.889	83.7	-0.720	26	Sim	Mista	Sim
Neptuno	30.05	3.88	17.1	0.297	1.12	163.7	0.673	13	Sim	Líquida	Sim
Plutão	39.24	0.187	0.0021	0.317	0.059	248.0	6.41	1	Não	Sólida	-

Tabela 1.2 *Matriz de dissemelhanças para os planetas (com base nas três primeiras variáveis).*

	Mercúrio	Vénus	Terra	Marte	Júpiter	Saturno	Urano	Neptuno
Vénus	0.950							
Terra	1.128	0.210						
Marte	0.314	0.846	1.048					
Júpiter	317.930	317.152	316.965	317.873				
Saturno	95.580	94.770	94.582	95.512	222.607			
Urano	14.912	14.040	13.853	14.815	303.385	80.883		
Neptuno	17.413	16.558	16.371	17.324	300.789	78.299	2.604	
Plutão	0.697	1.265	1.457	0.536	317.989	95.648	14.994	17.492

Tabela 1.3 *Lista de cenários descritos pelas expressões da face de uma mulher.*

	Cenário
1	Sofrimento pela morte da mãe
2	Saboreando coca-cola
3	Uma surpresa agradável
4	Amor maternal – bebé nos braços
5	Cansaço físico
6	Apercebe-se que há qualquer coisa errada com o avião
7	Acesso de cólera ao ver bater num cão
8	Embaraço – vontade de se esconder
9	Inesperadamente encontra um antigo namorado
10	Mudança súbita de humor
11	Dor intensa
12	Apercebe-se que o avião vai cair
13	Ligeiro descanso

na Tabela 1.3.

A Psicologia dá muita importância ao estudo da expressão facial, uma vez que há grande interesse em saber se é possível identificar as mensagens emocionais a partir das expressões faciais que se observam numa pessoa. Abelson and Sermat (1962) realizaram uma experiência que consistiu em apresentar, a 30 estudantes, 13 fotografias da face de uma mulher com as expressões correspondentes às reacções aos cenários indicados na Tabela 1.3. Os estudantes avaliaram, numa escala de 9 pontos, a dissemelhança entre os dois elementos dos 78 pares que é possível formar com os 13 objectos.

As 30 matrizes de dissemelhanças, correspondentes aos 30 estudantes, foram amalgamadas de forma apropriada tendo produzido a matriz que se encontra na Tabela 1.4.

1.4 Fases de uma análise de clusters

Na condução de uma análise de clusters é preciso tomar decisões e ter em conta aspectos que dependem do problema particular que é objecto do estudo. Contudo é possível indicar uma sequência de passos, como a que se vê na Figura 1.2, que são requeridos pela generalidade das análises.

Os vários passos têm em vista responder às várias questões que geralmente se colocam no decorrer da análise.

Tabela 1.4 *Matriz de dissemelhanças para as 13 fotografias correspondentes aos cenários da Tabela 1.3.*

	1	2	3		5	6	7	8	9	10	11	12
2	4.05											
3	8.25	2.54										
4	5.57	2.69	2.11									
5	1.15	2.67	8.98	3.78								
6	2.97	3.88	9.27	6.05	2.34							
7	4.34	8.53	11.87	9.78	7.12	1.36						
8	4.90	1.31	2.56	4.21	5.90	5.18	8.47					
9	6.25	1.88	0.74	0.45	4.77	5.45	10.20	2.63				
10	1.55	4.84	9.25	4.92	2.22	4.17	5.44	5.45	7.10			
11	1.68	5.81	7.92	5.42	4.34	4.72	4.31	3.79	6.58	1.98		
12	6.57	7.43	8.30	8.93	8.16	4.66	1.57	6.49	9.77	4.93	4.83	
13	3.93	4.51	8.47	3.48	1.60	4.89	9.18	6.05	6.55	4.12	3.51	12.65

1. Selecção de objectos
2. Selecção de variáveis
3. Transformação de variáveis
4. Construção da medida de dissemelhança/semelhança
5. Escolha do método a aplicar aos dados
6. Discussão e apresentação dos resultados

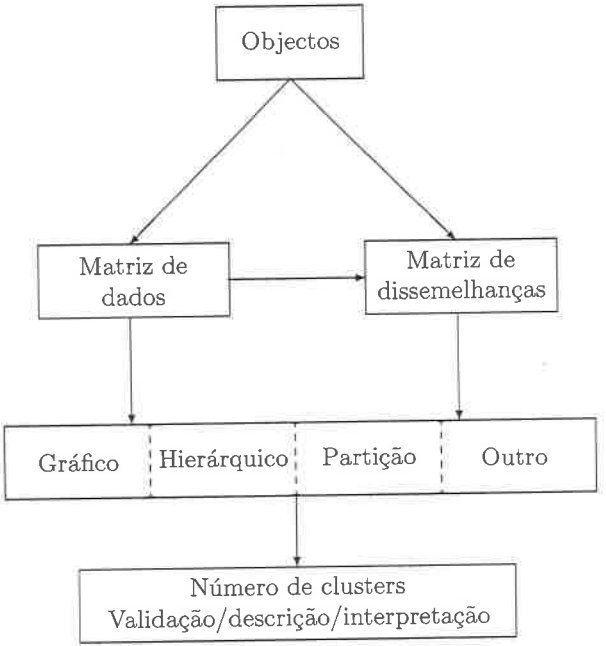


Figura 1.2 *Fases de uma análise de clusters.*

(i) Como seleccionar os objectos?
(ii) Que variáveis devem ser incluídas e a que transformações devem ser submetidas, se for caso disso?
(iii) Havendo que construir uma matriz de dissemelhanças, qual a medida de dissemelhança mais conveniente?
(iv) Considerando que estão disponíveis muitos métodos de análise, qual deles será o mais adequado?
(v) Qual a forma mais clara de apresentar os resultados e como proceder de forma convincente à sua validação?

Uma vez que em cada passo existem várias opções isso significa que os resultados da análise podem ser muito variados. Ao longo do texto são apresentadas muitas dessas opções.

No Capítulo 2 descrevem-se várias medidas de semelhança e dissemelhança e indicam-se métodos para construir essas medidas. Nos Capítulos 3, 4 e 5 apresentam-se os métodos referidos na fase 5 da análise, indicada na Figura 1.2, e no Capítulo 6 faz-se um resumo final dos aspectos a ter em conta para a escolha da opção que melhor se adapte ao problema em estudo.

2

Medidas de proximidade

2.1 Introdução

As ideias subjacentes ao processo de construção de clusters são a ideia de *semelhança* e a ideia de *dissemelhança*, conhecidas por *proximidades* quando a elas nos referimos mas não querendo especificar nenhuma delas em particular. Dois objectos pertencem ao mesmo cluster se são semelhantes e pertencem a clusters diferentes se não são semelhantes ou, dito de forma equivalente, se são dissemelhantes.

Intuitivamente a dissemelhança reflecte o grau de diferença, afastamento ou divergência entre dois objectos. Quanto mais distintos forem os objectos maior é a dissemelhança entre eles.

A semelhança mede o grau de parecença ou proximidade entre os objectos. Quanto mais parecidos forem os objectos maior é a semelhança entre eles.

Para usar os conceitos de semelhança e dissemelhança de forma útil e eficaz é importante eliminar a sua subjectividade criando medidas concretas de proximidade.

Dissemelhança: Dada uma colecção de objectos define-se dissemelhança entre dois objectos da colecção, i e j, como a função dos objectos cujos valores, d_{ij}, satisfazem as seguintes propriedades:

1. $d_{ij} \geq 0, \ \forall_{i,j}$
2. $d_{ii} = 0, \ \forall_i$
3. $d_{ij} = d_{ji}, \ \forall_{i,j}$

Em certos casos a simetria, propriedade 3, não se verifica embora a função possa continuar a ser útil para representar a dissemelhança. Um exemplo desta situação é o caso em que a dissemelhança d_{ij}, entre a cidade i e

a cidade j, é medida pelo número de pessoas que viaja de i para j. É claro que, em geral, $d_{ji} \neq d_{ij}$. A simetria pode ser restabelecida tomando $(d_{ij}+d_{ji})/2$ para valor comum de d_{ij} e d_{ji}. Pode ainda acontecer que $d_{ij} = 0$ e $i \neq j$.

Se além das propriedades anteriores se verifica ainda a propriedade triangular

4. $d_{ij} \leq d_{ik} + d_{kj}, \ \forall_{i,j,k}$

diz-se que a dissemelhança satisfaz as propriedades de uma semi-métrica ou semi-distância. Se a dissemelhança satisfaz também a propriedade

5. $d_{ij} = 0$ se e só se $i = j$

diz-se que a dissemelhança é uma métrica ou uma distância.

Muitas dissemelhanças não satisfazem a propriedade 4 (ver Exercício 2.6). Por outro lado certas dissemelhanças satisfazem uma propriedade mais forte do que a propriedade triangular, a chamada propriedade ultramétrica,

6. $d_{ij} \leq \max\left(d_{ik}, d_{jk}\right), \ \forall_{i,j,k},$

dizendo-se então que as dissemelhanças são ultramétricas.

No entanto, para muitas situações práticas é suficiente que a dissemelhança satisfaça as propriedades 1, 2 e 3.

Semelhança: Em muitas situações a medida de proximidade mais fácil de conseguir é a semelhança entre objectos. Uma definição razoável de semelhança entre os objectos i e j, s_{ij}, deve incluir as seguintes propriedades:

1. $s_{ij} \geq 0, \ \forall_{i,j}$
2. $s_{ij} = s_{ji}, \ \forall_{i,j}$
3. s_{ij} é tanto maior quanto maior for a semelhança entre os objectos.

Vários autores definem semelhança de forma a que os seus valores fiquem no intervalo $[0, 1]$, isto é, $0 \leq s_{ij} \leq 1$, assumindo ainda que $s_{ij} = 1$ se e só se $i = j$. Pode acontecer ainda que $-1 \leq s_{ij} \leq 1$ quando a semelhança depende de grandezas do tipo da correlação.

Na Tabela 2.1 estão registadas as semelhanças, entre 6 universidades, produzidas por 500 estudantes. Foi pedido aos estudantes para indicarem

Tabela 2.1 *Frequências absolutas do número de estudantes que escolheu cada par de universidades.*

	U_1	U_2	U_3	U_4	U_5	U_6
U_1						
U_2	13					
U_3	22	0				
U_4	10	61	18			
U_5	150	25	120	7		
U_6	15	12	5	19	23	

as duas universidades que consideram ser mais semelhantes. Os valores da tabela representam as frequências absolutas do número de estudantes que escolheram o respectivo par de universidades. Querendo reduzir s_{ij} ao intervalo $[0,1]$ basta considerar as frequências relativas em vez das frequências absolutas.

Em geral é possível estabelecer uma relação entre as semelhanças e dissemelhanças dos mesmos objectos. A dissemelhança d_{ij} pode obter-se a partir da semelhança s_{ij}, usando uma função decrescente, como por exemplo, $d_{ij} = k - s_{ij}$, onde k é uma constante adequada. Inversamente, dada a dissemelhança d_{ij}, pode obter-se s_{ij} usando, por exemplo, a transformação

$$s_{ij} = \frac{k}{k + d_{ij}},$$

com k =constante. Note-se que esta última relação é incapaz de recuperar o valor zero para a semelhança pois $0 < s_{ij}$, uma vez que k não pode ser zero. Isto é, a relação não consegue transformar a dissemelhança em semelhança, de acordo com as definições que foram dadas. Embora tanto semelhanças como dissemelhanças possam ser usadas no processo de construção de clusters muito do *software* disponível para análise de clusters usa dissemelhanças, havendo por isso uma fase de conversão de semelhanças em dissemelhanças, se as medidas de proximidade inicialmente disponíveis forem as semelhanças.

Como já foi referido e ilustrado no Capítulo 1 as semelhanças/dissemelhanças são muitas vezes construídas directamente por observadores que examinando os objectos aos pares acabam por indicar, numa escala apropriada, qual é a sua semelhança/dissemelhança. Um outro exemplo que é um clássico, agora com interesse pedagógico e histórico, refere-se à percepção e preferência entre nações.

Os dados foram produzidos por um estudo piloto conduzido em Março de 1968 (Wish, 1971 e Wish *et al.*, 1970). Cada um de 18 estudantes avaliou

numa escala de 1 a 9, as semelhanças entre 12 nações. Cada estudante produziu 66 semelhanças (todos os pares possíveis formados a partir de 12 nações). As 18 matrizes de semelhança foram substituídas por uma única matriz contendo as médias das semelhanças obtidas pelos 18 estudantes. Não foi dito aos estudantes as características das nações que deveriam considerar para construir as semelhanças. De facto o objectivo do estudo era descobrir esta informação em vez de a fornecer. Os resultados da aplicação de multidimensional scaling à matriz de semelhanças final revelou duas direcções (alinhamento político e desenvolvimento económico) que foram interpretadas como as principais características das nações que os estudantes usaram para construir as dissemelhanças.

Porém a observação directa nem sempre é possível e as medidas de proximidade são deduzidas a partir da matriz de dados dos objectos observados.

A seguir indica-se como é, em geral, possível definir dissemelhanças e semelhanças, entre objectos ou entre variáveis, com base nas observações efectuadas em cada um dos objectos.

2.2 Medidas de proximidade entre objectos

Como já se disse é possível, de um modo geral, construir uma medida de dissemelhança a partir de uma semelhança e vice-versa. Como porém a maior parte dos métodos de análise de clusters usam algoritmos que operam sobre dissemelhanças, o objectivo desta secção é construir matrizes de dissemelhança. Quando for mais conveniente começar por construir uma matriz de semelhança significa que se espera que o utilizador execute uma transformação para uma matriz de dissemelhanças.

As medidas de proximidade dependem da natureza das características que são observadas nos objectos. Assim, interessa saber como é possível construir medidas de proximidade quando as variáveis são quantitativas e qualitativas (nominais e ordinais).

Como a seguir se verifica, muitas das medidas de dissemelhança são inspiradas em modelos geométricos em que os objectos são representados por pontos no espaço e a dissemelhança entre dois objectos equivale à distância entre os pontos que os representam.

2.2.1 Variáveis quantitativas

No caso de variáveis quantitativas a medida de dissemelhança mais conhecida é a distância euclidiana.

Tabela 2.2 *Idade e altura de três pessoas.*

Nome	Idade	Altura (cm)
Pedro	18	165
António	19	198
José	20	181

- **Dissemelhanças derivadas da distância euclidiana**

Assumindo que a matriz de dados é representada, como no Capítulo 1, por $\mathbf{X} = [x_{ij}]$, $i = 1,\ldots,n$ e $j = 1,\ldots,p$, a dissemelhança entre o objecto i e o objecto j é definida por

$$d_{ij} = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2\right]^{\frac{1}{2}}, \tag{2.1}$$

ou, em forma vectorial,

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)\right]^{\frac{1}{2}}, \tag{2.2}$$

onde $\mathbf{x}_i$ e $\mathbf{x}_j$ são vectores linha da matriz $\mathbf{X}$, isto é, os vectores de observações relativas aos objectos i e j, respectivamente, ou seja as coordenadas dos dois pontos de $\mathbb{R}^p$ que identificam os dois objectos.

Esta dissemelhança nem sempre é satisfatória, especialmente se as variáveis são medidas em unidades diferentes, se têm variâncias muito diferentes e ainda se são correlacionadas. Nestas condições as variáveis intervêm com pesos diferentes na determinação das dissemelhanças. Além disso a distância euclidiana é sensível a mudanças de escala, no sentido em que, mudando a escala mudam, não só as distâncias, o que é natural, mas também podem mudar, de forma arbitrária, as ordens das distâncias e consequentemente o resultado da análise de clusters (ver Capítulo 6), o que não é conveniente. Na Tabela 2.2 Pedro e António são os mais dissemelhantes $d_{12} = [(18-19)^2 + (165-198)^2]^{1/2} = 33.015$ mas se a *altura* for medida em metros Pedro e José são os mais dissemelhantes, $d_{13} = [(18-20)^2 + (1.65-1.81)^2]^{1/2} = 2.006$.

Para ultrapassar estes inconvenientes da distância euclidiana usam-se várias medidas dela derivadas:

i. Distância euclidiana média

$$d_{ij} = \left[\frac{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}{p}\right]^{\frac{1}{2}}. \tag{2.3}$$

Esta distância goza das mesmas propriedades da distância euclidiana mas apresenta vantagens quando há dados omissos.

ii. Distância euclidiana estandardizada

$$d_{ij} = \left[\sum_{k=1}^{p} \left(z_{ik} - z_{jk} \right)^2 \right]^{\frac{1}{2}}, \tag{2.4}$$

onde

$$z_{rk} = \frac{x_{rk} - \overline{x}_{.k}}{s_k}, \quad r = 1, \ldots$$

e

$$\overline{x}_{.k} = \frac{\sum_{r=1}^{n} x_{rk}}{n} \quad \text{e} \quad s_k = \left[\frac{\sum_{r=1}^{n} \left(x_{rk} - \overline{x}_{.k} \right)^2}{n-1} \right]^{\frac{1}{2}}$$

representam as estimativas do valor médio e do desvio padrão da variável k. Substituindo em (2.4), z_{ik} e z_{jk} pelas suas expressões tem-se

$$d_{ij} = \left[\sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{s_k} \right)^2 \right]^{\frac{1}{2}}, \tag{2.5}$$

ou, em linguagem matricial,

$$d_{ij} = \left[\left(\mathbf{x}_i - \mathbf{x}_j \right)' \mathbf{D}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j \right) \right]^{\frac{1}{2}}, \tag{2.6}$$

onde $\mathbf{D}$ é a matriz diagonal das variâncias das colunas de $\mathbf{X}$, $\mathbf{D} = \text{diag}(s_1^2, s_2^2, \ldots, s_p^2)$. Este processo consegue eliminar a dependência dos resultados da análise de clusters das unidades de medição. Contudo a estandardização pode tornar a distância dentro dos clusters maior do que a distância entre os clusters, o que torna os resultados pouco claros (Hartigan, 1975, p. 62).

iii. Distância euclidiana ponderada

A fórmula (2.6) pode ser vista como uma maneira de atribuir pesos às variáveis de forma a eliminar os seus efeitos arbitrários, fazendo com que estas contribuam, não de forma diferenciada, mas de forma homogénea para a construção das dissemelhanças. De um modo geral podemos pensar numa matriz de pesos $\mathbf{A}$, e com ela construir a distância euclidiana ponderada entre os objectos i e j,

$$d_{ij} = \left[\left(\mathbf{x}_i - \mathbf{x}_j \right)' \mathbf{A} \left(\mathbf{x}_i - \mathbf{x}_j \right) \right]^{\frac{1}{2}}. \tag{2.7}$$

Esta dissemelhança inclui as distâncias anteriores como casos particulares:

- se $\mathbf{A} = \mathbf{I}$ tem-se a distância euclidiana
- se $\mathbf{A} = \frac{1}{p}\mathbf{I}$ tem-se a distância euclidiana média

Tabela 2.3 *Densidade e gravidade dos planetas.*

Planeta	Dens.	Grav.
Mercúrio	0.984	0.378
Vénus	0.951	0.907
Terra	**1**	**1**
Marte	**0.713**	**0.377**
Júpiter	0.240	2.36
Saturno	0.125	0.916
Urano	0.230	0.889
Neptuno	0.297	1.12
Plutão	0.317	0.059

- se $\mathbf{A} = \mathbf{D}^{-1} = \left[\text{diag}(s_1^2, s_2^2, \ldots, s_p^2)\right]^{-1}$ tem-se a distância euclidiana estandardizada.

Um outro caso que interessa é o caso

- $\mathbf{A} = \mathbf{S}^{-1}$,

onde $\mathbf{S}$ é a estimativa da matriz de covariâncias das p variáveis em estudo. A distância que daqui resulta,

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)\right]^{\frac{1}{2}}, \tag{2.8}$$

é conhecida por distância estatística ou distância de Mahalanobis. A distância de Mahalanobis além de reduzir a dependência das unidades de medição, reduz também a influência da correlação entre variáveis. Por isso a distância de Mahalanobis tende a mascarar ainda mais os resultados da análise de clusters (Hartigan, 1975, p. 63). Outro conjunto de pesos que costuma usar-se deriva de $r_k = \max_{i,j} |x_{ik} - x_{jk}|$, tendo-se $\mathbf{R} = \text{diag}(r_1^2, r_2^2, \ldots, r_p^2)$ e

- $\mathbf{A} = \mathbf{R}^{-1} = \left[\text{diag}(r_1^2, r_2^2, \ldots, r_p^2)\right]^{-1}$.

Apesar da grande variedade de pesos que pode apresentar-se parece não haver solução satisfatória para o problema das unidades de medição.

Considerando os dados dos planetas na Tabela 1.1 e seleccionando apenas as variáveis *densidade* (Dens.) e *gravidade* (Grav.) tem-se a nova Tabela 2.3.

A distância d_{ij} da Terra (i) a Marte (j) tem os seguintes valores , correspondentes às cinco definições de distâncias apresentadas anteriormente.

Seguindo a notação adoptada tem-se sucessivamente, $p = 2$,

$$\mathbf{x}_i' = (1.000, 1.000), \quad \mathbf{x}_j' = (0.713, 0.377), \quad (\mathbf{x}_i - \mathbf{x}_j)' = (0.287, 0.623),$$

$$\mathbf{D}=\begin{bmatrix} 0.134 & 0.000 \\ 0.000 & 0.430 \end{bmatrix},\quad \mathbf{S}=\begin{bmatrix} 0.134 & -0.073 \\ -0.073 & 0.430 \end{bmatrix},\quad \mathbf{R}=\begin{bmatrix} 0.766 & 0.000 \\ 0.000 & 5.295 \end{bmatrix},$$

$$\mathbf{D}^{-1}=\begin{bmatrix} 7.444 & 0.000 \\ 0.000 & 2.324 \end{bmatrix},\quad \mathbf{S}^{-1}=\begin{bmatrix} 8.190 & 1.381 \\ 1.381 & 2.557 \end{bmatrix},$$

e

$$\mathbf{R}^{-1}=\begin{bmatrix} 1.306 & 0.000 \\ 0.000 & 0.189 \end{bmatrix}$$

e a fórmula (2.7) produz as cinco distâncias,

- distância euclidiana ($\mathbf{A}=\mathbf{I}$): 0.686
- distância euclidiana média ($\mathbf{A}=\frac{1}{2}\mathbf{I}$): 0.485
- distância euclidiana ponderada ($\mathbf{A}=\mathbf{D}^{-1}$): 1.231
- distância euclidiana ponderada ($\mathbf{A}=\mathbf{S}^{-1}$): 1.470
- distância euclidiana ponderada ($\mathbf{A}=\mathbf{R}^{-1}$): 0.425.

Muitas outras semelhanças podem ser construídas com base na chamada família de métricas de Minkowski.

- **Dissemelhanças usando métricas de Minkowski**

A família de métricas de Minkowski é dada pela fórmula geral

$$d_{ij}=\left[\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{r}\right]^{\frac{1}{r}}, \tag{2.9}$$

com $r\geq 1$. Variando r obtém-se uma infinidade de dissemelhanças, algumas delas já aqui apresentadas. Por exemplo, se $r=2$ tem-se a distância euclidiana, ou métrica L_2 e se $r=1$ obtém-se a métrica do quarteirão, também conhecida por métrica L_1 e por outras designações em língua inglesa (*city-block, taxicab e Manhattan*). Esta última métrica é conhecida pelo seu comportamento robusto relativamente a *outliers*. Quando r tende para infinito tem-se a métrica de Chebychev, métrica L_∞, ou métrica do supremo, ou seja,

$$\lim_{r\to\infty}\left[\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{r}\right]^{\frac{1}{r}}=\sup_{k=1,\ldots,p}\left|x_{ik}-x_{jk}\right|. \tag{2.10}$$

A Figura 2.1 é uma representação geométrica num espaço bidimensional, $p=2$, dos pontos que estão à distância unitária de um ponto fixo O, usando as métricas L_1, L_2 e L_∞.

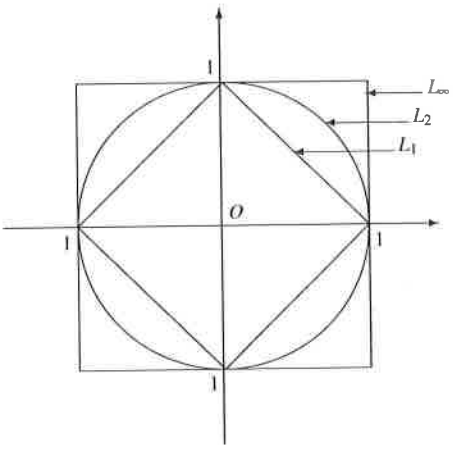
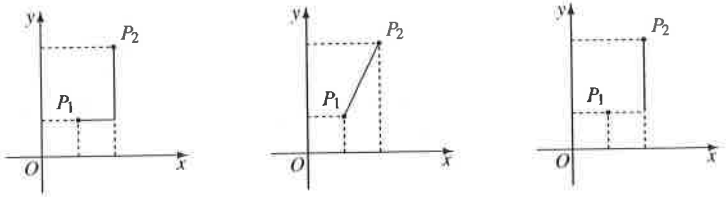


Figura 2.1 *Posição relativa de pontos à distância unitária de um outro ponto O, segundo as métricas L_1, L_2 e L_∞.*

Quando $1 < p < 2$ os lugares geométricos correspondentes a estas métricas ficam entre L_1 e L_2 e quando $p > 2$ os lugares geométricos das respectivas métricas ficam entre L_2 e L_∞. De entre a infinidade de métricas de Minkowski, são as métricas L_1 e L_2 as que mais interesse prático despertam.

Uma outra ilustração da interpretação geométrica das três métricas é aquela que se apresenta na Figura 2.2, onde o comprimento do traço contínuo representa a distância entre os dois pontos, P_1 e P_2, num sistema de coordenadas (x, y), segundo cada uma daquelas métricas.



(a) Métrica L_1 (b) Métrica L_2 (c) Métrica L_∞

Figura 2.2 *Interpretação geométrica das métricas L_1, L_2 e L_∞.*

Para reduzir o efeito resultante das diferenças nas escalas de medição e nas variâncias e ainda o efeito da correlação entre características dos objectos pode pensar-se em aplicar pesos às métricas de Minkovski, à semelhança do que se fez para a distância euclidiana ponderada.

Assim, tem-se a métrica de Minkowski ponderada

$$d_{ij} = \left[\sum_{k=1}^{p} \omega_k \left| x_{ik} - x_{jk} \right|^r \right]^{\frac{1}{r}}, \tag{2.11}$$

onde os pesos ω_k que geralmente interessa considerar são $\omega_k = 1$, $\omega_k = 1/p$, $\omega_k = s_k^{-1}$, $\omega_k = r_k^{-1}$.

- **Outras dissemelhanças**

O aparecimento de novas dissemelhanças e semelhanças é motivado pelo estudo de problemas práticos em muitas áreas de trabalho. Listas de medidas de dissemelhança e semelhança podem ser encontradas em muitos livros e artigos relacionados com a análise de clusters como, por exemplo, Cormack (1971), Anderberg (1973), Sneath and Sokal (1973), Späth (1980) e Legendre and Legendre (1982). A seguir indicam-se mais algumas medidas de uso frequente.

- Métrica de Camberra

$$d_{ij} = \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{x_{ik} + x_{jk}},$$

  com $d_{ij} = 0$ se $x_{ik} = x_{jk} = 0$.
- Coeficiente de Bray-Curtis

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

Para as duas dissemelhanças anteriores interessa que os elementos da matriz de dados sejam não negativos pois só assim se garante que as dissemelhanças resultam não negativas. Por este facto nenhuma das fórmulas deve ser usada com dados estandardizados.

- Coeficiente de Sokal e Sneath

$$d_{ij} = \left[\frac{1}{p} \sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}} \right)^2 \right]^{\frac{1}{2}}.$$

- Coeficiente de Soergel

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

- Métrica de Gower

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

onde r_k é a amplitude de variação da variável k, já definida anteriormente. Trata-se da distância *city-block* ponderada, onde $\omega_k = r_k^{-1}$.

- Coeficiente de separação angular (ou coseno)

$$c_{ij} = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\left(\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2\right)^{\frac{1}{2}}}.$$

- Coeficiente de correlação

$$r_{ij} = \frac{\sum_{k=1}^{p} \left(x_{ik} - \overline{x}_{i\cdot}\right)\left(x_{jk} - \overline{x}_{j\cdot}\right)}{\left[\sum_{k=1}^{p} \left(x_{ik} - \overline{x}_{i\cdot}\right)^2 \sum_{k=1}^{p} \left(x_{jk} - \overline{x}_{j\cdot}\right)^2\right]^{\frac{1}{2}}},$$

onde

$$\overline{x}_{s\cdot} = \frac{\sum_{k=1}^{p} x_{sk}}{p}.$$

Os dois últimos coeficientes são semelhanças e só depois de transformados podem ser usados como dissemelhanças, $d_{ij} = (1 - c_{ij})/2$, $d_{ij} = (1 - r_{ij})/2$. Em ambos os casos tem-se $-1 \leq c_{ij} \leq 1$ e $-1 \leq r_{ij} \leq 1$. Os valores $c_{ij} = -1$ e $r_{ij} = -1$ indicam que a semelhança é mínima (dissemelhança máxima) e $c_{ij} = 1$ e $r_{ij} = 1$ indicam que a semelhança é máxima (dissemelhança mínima).

O coeficiente c_{ij} é igual ao coseno do ângulo formado pelas semi-rectas que unem a origem dos eixos coordenados com os pontos que representam os objectos nesse sistema de eixos.

O uso do coeficiente de correlação r_{ij} neste contexto é tema de controvérsia (Fleiss and Zubin, 1969 e Cronbach and Gleser, 1953). O problema reside no facto de os dados associados a cada objecto se referirem a diferentes características e por isso a média e variância desses mesmos dados não fazem sentido. Pode no entanto fazer-se uma interpretação semelhante à que se fez para o coeficiente c_{ij}, só que agora o ângulo que interessa é formado por semi-rectas partindo da média das observações, isto é, do centróide do conjunto de pontos do espaço de dimensão p.

2.2.2 Variáveis qualitativas

Quando se pretende descobrir a semelhança entre dois objectos é frequente centrar a observação em características qualitativas (os objectos têm a

Tabela 2.4 *Duas universidades observadas em 10 características.*

	Variáveis												
Univ.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
i	1	0	1	1	1	0	1	1	1	1	0	1	0
j	1	1	0	0	1	0	1	1	0	1	1	0	0

mesma cor, a mesma forma, a mesma composição?). Interessa analisar os casos de variáveis binárias, variáveis nominais com mais de dois níveis e variáveis ordinais.

- **Variáveis nominais com dois níveis (binárias ou dicotómicas)**

Suponhamos que o problema é investigar a proximidade entre universidades (objectos) e que as características a observar (variáveis com valores (1) se a característica está presente, e (0) se não está presente) são:

X_1 – mais de cinco faculdades
X_2 – faculdade de medicina
X_3 – departamento de música
X_4 – *campus* fora da cidade
X_5 – posto médico
X_6 – creche
X_7 – banco
X_8 – residências para estudantes
X_9 – residências para professores visitantes
X_{10} – pavilhão desportivo
X_{11} – piscina
X_{12} – grupo de teatro
X_{13} – grupo coral

Duas universidades, i e j foram comparadas observando aquelas características, tendo-se obtido os resultados da Tabela 2.4.

A distância euclidiana é

$$d_{ij} = \left[\sum_{i=1}^{13} (x_{ik} - x_{jk})^2\right]^{\frac{1}{2}} = 2.450$$

e a distância euclidiana média é

$$d_{ij} = \left[\frac{1}{13}\sum_{i=1}^{13} (x_{ik} - x_{jk})^2\right]^{\frac{1}{2}} = 0.680.$$

Tabela 2.5 *Número de pares* (1,1), (1,0), (0,1) *e* (0,0) *para variáveis binárias.*

		objecto j 1	0	
objecto i	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$p=a+b+c+d$

Dados deste tipo costumam organizar-se numa tabela de dupla entrada, como a Tabela 2.5, em que se mostra o número de vezes que ocorrem os pares (1,1), (1,0), (0,1) e (0,0).

Vê-se então que a distância euclidiana média é

$$d_{ij} = \left(\frac{b+c}{a+b+c+d}\right)^{\frac{1}{2}}.$$

Note-se que $(b+c)/(a+b+c+d)$ representa a proporção de características que não são comuns aos dois objectos. Isso significa que de facto d_{ij} é tal que $0 \leq d_{ij} \leq 1$ e tem as propriedades de uma dissemelhança pois quanto maior é d_{ij} maior é aquela proporção e portanto mais diferentes são os objectos. É claro que a razão $(a+d)/p$ é um coeficiente de semelhança. E embora a Tabela 2.5 sirva para construir coeficientes de dissemelhança e semelhança os investigadores começam em geral por usar os números de pares, a, b, c e d da Tabela 2.5 para deduzir coeficientes de semelhança. As Tabelas 2.6 e 2.7 listam alguns desses coeficientes de semelhança.

O estudo das suas propriedades, a sua utilidade e a sua interpretação encontram-se na literatura da especialidade, nomeadamente nas referências indicadas em 2.2.1 e ainda em Cheetham and Hazel (1969), Baroni-Urbani and Buser (1976) e Romesburg (1984). Perante tamanha lista de coeficientes é natural querer saber o que leva os cientistas a produzir tantos coeficientes, alguns aparentemente muito parecidos. O segredo parece estar no tipo de problema, na especificidade da área de trabalho e nos objectivos a atingir. O cientista terá de ter conhecimentos profundos do assunto em estudo e usá-los de forma eficiente e criteriosa para moldar um coeficiente que integre as suas hipóteses de trabalho e vá de encontro aos objectivos que pretende. Por exemplo o coeficiente de Jacard e o de Sorenson excluem os pares (0,0) tornando-os irrelevantes nos cálculos e dando sim importância aos atributos que pertencem a ambos os objectos, sendo que o coeficiente de Sorenson pesa duas vezes mais essa importância. No caso do coeficiente de concordância simples é dada também importância aos atributos que nenhum dos objectos possui.

Tabela 2.6 *Lista de alguns coeficientes de semelhança*

Nome	Coeficiente de Semelhança	Intervalo de Variação
Jacard	$\frac{a}{a+b+c}$	$[0,1]$
Distância Binária de Sokal	$\left(\frac{b+c}{a+b+c+d}\right)^{1/2}$	$[0,1]$
Concordância simples	$\frac{a+d}{a+b+c+d}$	$[0,1]$
Rogers e Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	$[0,1]$
Sokal e Sneath	$\frac{2(a+d)}{2(a+d)+b+c}$	$[0,1]$
	$\frac{a}{a+2(b+c)}$	$[0,1]$
	$\frac{a+d}{b+c}$	$[0,+\infty[$
Russel e Rao	$\frac{a}{a+b+c+d}$	$[0,1]$
Sorenson/Dice/ Czekanowski	$\frac{2a}{2a+b+c}$	$[0,1]$
Ochiai	$\frac{a}{[(a+b)(a+c)]^{1/2}}$	$[0,1]$
Baroni-Urbani-Buser	$\frac{a+(ad)^{1/2}}{a+b+c+(ad)^{1/2}}$	$[0,1]$
Hamann	$\frac{(a+d)-(b+c)}{a+b+c+d}$	$[-1,1]$
Yule	$\frac{ad-bc}{ad+bc}$	$[-1,1]$
ϕ	$\frac{ad-bc}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$	$[-1,1]$

Tabela 2.7 *Lista de alguns coeficientes de semelhança (cont.)*

Nome	Coeficiente de Semelhança	Intervalo de Variação
Ochiai II	$\frac{ad}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$	$[0,1]$
	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	$[0,1]$
	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c}\right)$	$[0,1]$
Diferença de Padrões	$\left(\frac{bc}{(a+b+c+d)^2}\right)^{1/2}$	$[0,1]$
Variância	$\frac{b+c}{4(a+b+c+d)}$	$[0,\frac{1}{4}]$
Dispersão	$\frac{ad-bc}{(a+b+c+d)^2}$	$[-1,1]$
Forma	$\frac{(a+b+c+d)(b+c)-(b-c)^2}{4(a+b+c+d)^2}$	$[0,\frac{1}{4}]$
Lance e Williams	$\frac{b+c}{2a+b+c}$	$[0,1]$
Y de Yule	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$[-1,1]$
Q de Yule	$\frac{ad-bc}{ad+bc}$	$[-1,1]$
Kulezynski	$\frac{a}{b+c}$	$[0,+\infty[$
	$\frac{a/(a+b)+a/(a+c)}{2}$	$[0,1]$

Aplicando estes coeficientes ao exemplo das universidades ($a = 5$, $b = 4$, $c = 2$, $d = 2$ e $p = 13$) tem-se:

Coeficiente de Jacard: $s_{ij} = \dfrac{a}{a+b+c} = \dfrac{5}{5+4+2} = 0.45$,

Coeficiente de Sorenson: $s_{ij} = \dfrac{2a}{2a+b+c} = \dfrac{10}{10+4+2} = 0.62$,

Coeficiente de concordância simples:

$$s_{ij} = \frac{a+d}{a+b+c+d} = \frac{5+2}{5+4+2+2}$$

O coeficiente de Jacard revelou-se o mais pequeno pois dá menos importância ao que há em comum e não dá importância ao que não há em comum.

Note-se que quando $d = 0$, isto é, quando não há pares $(0,0)$, os coeficientes de Jacard e concordância simples têm os mesmos valores conduzindo portanto aos mesmos resultados da análise de clusters. E como é natural muitos dos coeficientes das Tabelas 2.6 e 2.7 são correlacionados, uma vez que partilham os valores a, b, c e d. Por exemplo o coeficiente de Hamann é perfeitamente correlacionado com o coeficiente de concordância simples (ver Exercício 2.9). Uma análise de clusters sobre um conjunto de dados, usando os dois coeficientes, conduz a dendrogramas idênticos, excepto no que diz respeito às respectivas distâncias entre clusters que são proporcionais entre si. A mesma situação acontece quando a correlação entre coeficientes é alta. Se este facto for esquecido pode cair-se na tentação enganadora de pensar que a semelhança de dendrogramas é consequência da forte estrutura dos dados, quando na verdade resulta da existência daquela correlação. É preciso, por isso, investigar se existe essa correlação entre os coeficientes que foram usados nas análises.

- **Variáveis nominais com mais de dois níveis**

Quando as variáveis nominais têm mais de dois níveis a estratégia geralmente adoptada consiste em decompôr cada variável em variáveis binárias, tantas quantos os níveis dessa variável, e construir, a partir do vector de variáveis binárias que resulta da aplicação deste procedimento, um coeficiente de semelhança, da forma que já foi indicada.

Considere-se uma situação concreta para esclarecer o procedimento a seguir com este tipo de variáveis. Suponhamos que se pretende construir uma matriz de semelhanças para um grupo de homens na base de três características nominais:

(i) cor do cabelo, com quatro níveis (preto – P, castanho – C, louro – L e ruivo – R),

Tabela 2.8 *Três variáveis nominais e 10 variáveis binárias.*

		Variáveis nominais									
		cor do cabelo				altura			aparência		
Níveis		*P*	*C*	*L*	*R*	*B*	*M*	*A*	*C*	*R*	*M*
Variáveis binárias	sim	1	1	1	1	1	1	1	1	1	1
	não	0	0	0	0	0	0	0	0	0	0
Homens	*A*	1	0	0	0	0	1	0	0	1	0
	B	0	0	1	0	0	0	1	0	1	0

(ii) altura, com três níveis (baixa – *B*, média – *M* e alta *A*),
(iii) aparência, com três níveis (cuidada – *C*, razoável – R e má – *M*).

Como cada nível dá origem a uma variável binária obtêm-se, neste caso, 10 variáveis binárias. Na Tabela 2.8 apresentam-se as variáveis nominais com os respectivos níveis, as variáveis binárias e os valores observados em dois homens, um homem *A* (de cabelo preto, de altura média, com aparência razoável) e um homem *B* (de cabelo louro, alto, com aparência razoável).

Os valores da Tabela 2.5 para este exemplo são

		homem *B*	
		1	0
homem *A*	1	1	2
	0	2	5

Calculando o valor dos três coeficientes de semelhança referidos no exemplo das universidades tem-se

Jacard: $s_{AB} = \dfrac{1}{1+2+2} = 0.2,$

Sorenson: $s_{AB} = \dfrac{2}{2+3+2} = 0.33,$

Concordância simples: $s_{AB} = \dfrac{1+5}{10} = 0.6.$

O coeficiente de concordância simples é três vezes maior do que o de Jacard e quase o dobro do de Sorenson. Isto é devido à abundância do par $(0,0)$ cujo número aumenta com o número de níveis das variáveis, enquanto que os pares que envolvem o valor 1 só aparecem, em cada objecto, uma vez em cada variável.

Uma outra maneira de construir um coeficiente de semelhança para variáveis nominais com mais de dois níveis é fazer $s_{ij} = c/p$, onde p é o número total de variáveis nominais e c é o número dessas variáveis em que os objectos i e j assumem o mesmo nível. Para o exemplo em discussão $s_{AB} = 1/3$ pois há três variáveis mas só numa delas os dois homens têm o mesmo nível (ambos têm aparência razoável).

O inconveniente deste último coeficiente é que trata igualmente todas as variáveis, quer elas tenham muitos ou poucos níveis. Este desequilíbrio pode ser corrigido fazendo intervir no cálculo do coeficiente o número de níveis de cada variável. Assim, suponhamos que há p variáveis, $Y_1,\ldots,Y_p$, com $l_1,\ldots,l_p$ níveis, respectivamente. Outra maneira de definir s_{AB} é fazer

$$s_{AB} = \frac{\sum_{k=1}^{p} l_k\, I(y_k(A), y_k(B))}{\sum_{k=1}^{p} l_k} = \sum_{k=1}^{p} \omega_k\, I(y_k(A), y_k(B)),$$

onde

$$\omega_k = \frac{l_k}{\sum_{m=1}^{p} l_m}$$

e I é a função indicatriz dos níveis dos dois objectos, A e B, na variável k, isto é,

$$I(y_k(A), y_k(B)) = \begin{cases} 1 & \text{se} \quad y_k(A) = y_k(B) \\ 0 & \text{se} \quad y_k(A) \neq y_k(B) \end{cases},$$

em que $y_k(A)$ e $y_k(B)$ são os níveis de A e de B na variável k.

Quando os objectos A e B são os homens do exemplo tem-se

$$s_{AB} = \frac{4\times 0 + 3\times 0 + 3\times 1}{4+3+3} = \frac{3}{10},$$

um valor ligeiramente mais baixo daquele (1/3) que foi obtido sem a intervenção dos pesos.

Uma outra proposta referida em Späth (1980) é

$$s_{AB} = \frac{\sum_{k=1}^{p} \ln l_k\, I(y_k(A), y_k(B))}{\sum_{k=1}^{p} \ln l_k},$$

em que os pesos são atenuados pelo cálculo do logaritmo. No caso do exemplo o cálculo do coeficiente reduz-se a

$$s_{AB} = \frac{\ln 3}{\ln 36} = 0.307\,.$$

- **Variáveis ordinais**

A ordem existe implícita no caso de variáveis quantitativas mas também pode surgir com naturalidade entre as classes de uma variável qualitativa. No caso da variável nominal cor do cabelo não faz sentido estabelecer uma ordem entre os seus níveis, mas isso já faz sentido e tem significado no caso das três classes em que foi caracterizada a altura, apresentadas na Tabela 2.8. Atribuindo códigos às classes, 1 para B, 2 para M e 3 para A elimina-se a arbitrariedade de classificação criando uma sequência ordenada, com significado e útil.

Para construir um coeficiente de semelhança basta encarar as variáveis como variáveis nominais apenas e aplicar o procedimento anterior, isto é, decompor cada variável em tantas variáveis binárias quantos os níveis dessa variável. Mas este procedimento acaba por desprezar a ordem que é a propriedade que distingue estas variáveis das variáveis puramente nominais. A maneira de proceder com variáveis ordinais é exemplificada em Bassab *et al.* (1990). Este autor apresenta um exemplo em que um objecto que possui um certo nível de uma variável possui todos os níveis inferiores, de acordo com a ordem estabelecida entre os níveis da variável. A variável ordinal é a escolaridade de uma pessoa, admitindo-se a existência de quatro níveis:

1. Analfabeto
2. Básico (completo ou não)
3. Secundário (idem)
4. Universitário (idem)

Considera-se que uma pessoa que tem um certo nível de escolaridade tem naturalmente os níveis inferiores.

Suponhamos que há duas pessoas, A e B, e que A tem o ensino básico e que B tem o ensino secundário. As quatro variáveis binárias permitem definir os vectores associados às duas pessoas

A	1	1	0	0
B	1	1	1	0

de onde se obtém

		B	
		1	0
A	1	2	0
	0	1	1

e os coeficientes

$$\text{Jacard: } s_{AB} = \frac{2}{2+0+1} = 0.67,$$

$$\text{Sorenson: } s_{AB} = \frac{2\times 2}{2\times 2+0+1} = 0.8,$$

$$\text{Concordância simples: } s_{AB} = \frac{2+1}{2+0+1+1} = 0.75.$$

Outro método que pode ser usado para derivar um coeficiente de semelhança para variáveis ordinais com l níveis começa por associar os códigos $1,2,\ldots,l$ aos níveis ordenados. Se o objecto A tem o nível r e o objecto B tem o nível s introduz-se a dissemelhança $d_{AB} = |r-s|/l$ e a partir dela constrói-se a semelhança

$$s_{AB} = 1 - \frac{|r-s|}{l}, \tag{2.12}$$

tendo-se, no caso $s_{AB} = 1 - |2-3|/4 = 0.75$, precisamente o coeficiente de concordância simples (ver Exercício 2.11).

- **Variáveis de diferentes tipos**

Se as variáveis que fazem parte da matriz de dados são de natureza diferente pode não ser fácil deduzir uma medida de semelhança a partir desse conjunto misto de variáveis. Saber como obter uma medida de semelhança no caso de variáveis de diferentes tipos é muito importante pois esta situação ocorre na prática com muita frequência. Detalhes sobre este tema podem ser encontrados em Estabrook and Rodgers (1966), Gower (1971), Romesburg (1984), Lerman (1987) e em várias outras publicações.

Há várias estratégias que podem ser adoptadas:

i. **Estratégia de Romesburg**
   Romesburg (1984) sugere que a maneira mais simples de enfrentar o problema de variáveis do tipo misto é esquecer a natureza das variáveis e considerar todas elas do tipo quantitativo, codificando as que forem qualitativas. Depois é só usar um coeficiente apropriado para variáveis quantitativas, como, por exemplo, a distância euclidiana. Embora pareça absurdo o autor afirma que o método funciona. É claro que a interpretação dos coeficientes de semelhança é difícil pois fica dependente das codificações que se adoptarem para as variáveis qualitativas. Mas é de aproveitar esta simplicidade e usar o método para o cálculo de uma primeira medida de semelhança na fase exploratória dos dados.

ii. **Realizar análises separadas**
Construir uma medida de semelhança para cada grupo de variáveis do mesmo tipo e efectuar análises de clusters separadas para cada grupo. Se os resultados revelarem concordância significa que esta estratégia pode ser adoptada, mas se isso não acontecer significa que a solução não está à vista e que portanto é preferível pensar em processar os dados conjuntamente com vista a realizar uma única análise de clusters.

iii. **Reduzir todas as variáveis a variáveis binárias**
Este procedimento já foi ilustrado para o caso de variáveis nominais e é fácil de implementar para variáveis quantitativas, bastando dividir o domínio de cada variável em dois blocos e aplicar a regra

$$\begin{aligned} &\text{Se } y_{ij} < c_j, \text{ então } x_{ij} = 0 \\ &\text{Se } y_{ij} \geq c_j, \text{ então } x_{ij} = 1, \end{aligned}$$

onde y_{ij} é o valor da variável original j no objecto i, c_j é o valor crítico que divide o domínio da variável j em dois e x_{ij} é o valor que a variável binária criada assume no objecto i. A desvantagem deste procedimento está na perda de informação que resulta de reduzir os dados completos a dados binários.

iv. **Construir um coeficiente de semelhança combinado**
Calcular coeficientes de semelhança para cada grupo de variáveis do mesmo tipo. Usar esses coeficientes, de forma combinada, para construir um único coeficiente de semelhança. De acordo com esta estratégia o coeficiente de semelhança combinado para os objectos i e j é

$$s_{ij} = \omega_1 s_{ij}^q + \omega_2 s_{ij}^n + \omega_3 s_{ij}^o,$$

onde s_{ij}^q, s_{ij}^n e s_{ij}^o são os coeficientes de semelhança calculados para as variáveis quantitativas, nominais e ordinais, e ω_k, $k = 1,2,3$, são os pesos associados.
Bassab *et al.* (1990) trabalham com detalhe um exemplo usando este procedimento para construir uma matriz de semelhanças combinada. Uma forma mais elaborada do coeficiente de semelhança combinado é apresentada em Gower (1971),

$$s_{ij} = \frac{\sum_{k=1}^{p} \omega_{ijk} s_{ijk}}{\sum_{k=1}^{p} \omega_{ijk}}, \tag{2.13}$$

onde s_{ijk} é a semelhança entre os objectos i e j com base na variável k. Geralmente o peso ω_{ijk} toma o valor um ou o valor zero conforme a

comparação dos objectos i e j, na variável k, é ou não é válida. Além disso o valor de ω_{ijk} é fixado em zero se o valor da variável k é omisso em pelo menos um dos dois objectos i e j. Quando as variáveis são binárias ou do tipo nominal com mais de dois níveis os coeficientes s_{ijk} tomam o valor um se os dois objectos têm o mesmo valor na variável k e tomam o valor zero no caso contrário. Para variáveis contínuas Gower (1971) sugere o uso do coeficiente de semelhança

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k},$$

construído com base na métrica *city-block* estandardizada para a variável k.

O exemplo que se segue ilustra como é que um coeficiente de semelhança pode ser construído a partir de variáveis de diferentes tipos. No Capítulo 6 apresenta-se um caso mais extenso que envolve dados reais.

Exemplo 2.1. Suponha que um grupo de doentes de proveniência variada, sofrendo de artrite reumatóide, é examinado antes de iniciar um tratamento com base num novo medicamento. De entre as características observadas em cada doente incluem-se:

- variáveis quantitativas
  - X_1 – idade (anos)
  - X_2 – peso (kg)
- variáveis qualitativas
  - X_3 – sexo (M, F), binária
  - X_4 – raça (branca, negra, outra), nominal
  - X_5 – intensidade da dor (forte, moderada, fraca), ordinal

Sabendo que dois doentes A e B apresentam os seguintes dados nas características observadas

	X_1	X_2	X_3	X_4	X_5
A	58	71	M	branca	forte
B	45	63	F	negra	moderada

Como calcular a semelhança entre os dois doentes?

Uma possibilidade é recorrer ao coeficiente de semelhança combinado

$$s_{AB} = \omega_1 s^q_{AB} + \omega_2 s^n_{AB} + \omega_3 s^o_{AB},$$

e usar, por exemplo, a distância euclidiana para calcular s^q_{AB}, o coeficiente de concordância simples para obter s^n_{AB} e a fórmula (2.12) para calcular s^o_{AB}.

Tem-se sucessivamente:

1. Idade e peso

$$d_{AB} = \left[(58-45)^2 + (71-63)^2\right]^{\frac{1}{2}} = 15.264,$$

$$s^q_{AB} = \frac{1}{1+d_{AB}} = 0.061$$

2. Sexo e raça
A partir das variáveis binárias associadas aos níveis de X_3 e X_4 obtém-se

	Sexo		Raça		
	M	F	B	N	O
A	1	0	1	0	0
B	0	1	0	1	0

e

		B	
		1	0
	1	0	2
A			
	0	2	1

e o coeficiente de concordância simples dá

$$s^n_{AB} = \frac{0+1}{0+2+2+1} = 0.2$$

3. Intensidade da dor
Para esta variável ordinal com três níveis, forte (1), moderada (2) e fraca (3) tem-se

$$s^o_{AB} = 1 - \frac{|1-2|}{3} = 0.667$$

4. Coeficiente de semelhança combinado
Considerando os pesos proporcionais ao número de características usadas na construção de cada coeficiente tem-se

$$s_{AB} = \frac{2\times 0.061 + 2\times 0.2 + 1\times 2/3}{5} = 0.238.$$

2.3 Medidas de proximidade entre variáveis

Até agora considerou-se o agrupamento de objectos. Em certas aplicações o interesse do analista é o agrupameto das variáveis. E há casos em que tanto o agrupamento de objectos como o de variáveis é importante para informar sobre a estrutura dos dados. Em genética, por exemplo, há interesse em agrupar os genes (variáveis) e agrupar condições (objectos). No âmbito das Ciências Sociais é comum encontrar matrizes de dados em que as observações são respostas às várias questões de um dado inquérito. Há então interesse em agrupar as pessoas (objectos) que respondem e as questões (variáveis) que são colocadas.

Para agrupar variáveis basta transpor a matriz de dados $\mathbf{X}_{n\times p}$ e efectuar a análise de clusters sobre as linhas de $\mathbf{X}'_{p\times n}$. As variáveis tomam o lugar dos objectos e as medidas de proximidade necessárias para efectuar a análise de objectos podem servir para a análise das variáveis. Contudo, as medidas de proximidade mais adequadas para variáveis são, em geral, medidas de correlação e associação. Para duas variáveis i e j, encontrada a semelhança s_{ij} pode obter-se a dissemelhança d_{ij} fazendo, por exemplo, $d_{ij} = \sqrt{1 - s_{ij}}$.

2.3.1 Variáveis quantitativas

- Coeficiente de separação angular (ou coseno)

$$s_{ij} = \frac{\sum_{k=1}^{n} x_{ki}x_{kj}}{\left(\sum_{k=1}^{n} x_{ki}^2 \sum_{k=1}^{n} x_{kj}^2\right)^{\frac{1}{2}}} = \cos\alpha,$$

  onde α é o ângulo entre os vectores representativos das variáveis i e j, $(x_{1i},\ldots,x_{ni})'$ e $(x_{1j},\ldots,x_{nj})'$.

- Coeficiente de correlação

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_{\cdot i})(x_{kj} - \bar{x}_{\cdot j})}{\left[\sum_{k=1}^{n} (x_{ki} - \bar{x}_{\cdot i})^2 \sum_{k=1}^{n} (x_{kj} - \bar{x}_{\cdot j})^2\right]^{\frac{1}{2}}},$$

  que é o coeficiente de correlação observado de Pearson.

2.3.2 Variáveis qualitativas

- **Variáveis nominais com dois níveis (binárias ou dicotómicas)**

A classificação dos n objectos segundo as variáveis i e j conduz a

		j		
		1	0	
i	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$a+b+c+d$

e as medidas anteriores tomam a forma

$$s_{ij} = \frac{a}{\sqrt{(a+b)(a+c)}} = \cos\alpha$$

e

$$r_{ij} = \frac{ad-bc}{[(a+b)(c+d)(a+c)(b+d)]^{\frac{1}{2}}}. \tag{2.14}$$

- **Variáveis nominais com mais de dois níveis**

Considerem-se agora duas variáveis g e h. Neste caso, se a variável g tem r categorias e a variável h tem s categorias, a classificação dos n objectos é geralmente representada por uma tabela de contingência:

		h				
		1	2	$\cdots$	s	
g	1					
	2					
	$\vdots$			n_{ij} (f_{ij})		$n_{i\cdot}$ $(f_{i\cdot})$
	r					
				$n_{\cdot j}$ $(f_{\cdot j})$		n (1)

onde n_{ij}, $n_{i\cdot}$ e $n_{\cdot j}$ são frequências absolutas e f_{ij}, $f_{i\cdot}$ e $f_{\cdot j}$ são frequências relativas.

Agresti (1981) apresenta várias medidas de associação para variáveis nominais e variáveis ordinais. A medida mais comum é o Qui-quadrado de Pearson e as outras que a seguir se apresentam são derivadas do Qui-quadrado.

- O Qui-quadrado

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij}-f_{i\cdot}f_{\cdot j})^2}{f_{i\cdot}f_{\cdot j}}.$$

- Coeficiente de contingência quadrático

$$\phi^2 = \frac{\chi^2}{n}.$$

  Esta medida é uma correcção do χ^2 que deixa de ser múltiplo de n, uma desvantagem do Qui-quadrado.
- Coeficiente de contingência de Pearson

$$P = \left(\frac{\phi^2}{1+\phi^2}\right)^{\frac{1}{2}}.$$

- Coeficiente de Tschuprow

$$T = \left[\frac{\phi^2}{(r-1)(s-1)}\right]^{\frac{1}{2}}.$$

- Coeficiente de Cramer

$$C = \frac{\phi^2}{\min(r-1, s-1)}.$$

- **Variáveis ordinais**

Existem várias medidas de associação entre variáveis ordinais e uma das medidas que é frequentemente usada é o coeficiente de correlação ordinal de Spearman, dado por

$$r_s = 1 - \frac{6\sum_{k=1}^{n} d_k^2}{n(n^2-1)},$$

onde d_k é a diferença entre as ordens (*ranks*) dos valores que o objecto k assume nas duas variáveis i e j. Trata-se do coeficiente de correlação de Pearson r_{ij} entre as ordens dos valores assumidos por cada uma das variáveis i e j.

Outra medida de dissemelhança comum, está relacionada com o coeficiente de correlação de Spearman, é conhecida como o τ (tau) de Kendall, descrito em Kendall (1955).

Tanto o coeficiente de correlação de Pearson, como o coeficiente de correlação de Spearman e o τ de Kendall, têm valores no intervalo $[-1, 1]$.

2.4 Considerações de ordem prática

Quando os dados a analisar têm subjacente uma estrutura de grupos muito pronunciada os resultados das possíveis análises de clusters têm tendência

a confirmar essa estrutura. Mas se essa estrutura é débil então é preciso ponderar vários aspectos dos dados que podem influenciar o resultado final. Este cuidado faz todo o sentido porque a análise de clusters põe à disposição do analista uma variedade grande de soluções que dependem dos objectos e variáveis seleccionados, do facto de se usarem dados brutos ou transformados, da definição de dissemelhança, do método de análise escolhido e ainda de outros aspectos. Em seguida alerta-se para os cuidados a ter em conta em alguns destes aspectos.

2.4.1 Selecção de objectos

A escolha dos objectos depende de certo modo dos objectivos da análise. Em casos em que os dados são o produto de análises anteriores pode ser necessário escrutiná-los de forma a expurgar objectos, que sem relevância para os objectivos, podem causar ruído e perturbar os resultados.

Muitas vezes o que se pretende é estudar e classificar e obter conclusões para um determinado conjunto de objectos, sem a pretensão de estender as conclusões a objectos que não pertençam ao conjunto analisado. Neste caso o analista deve certificar-se que não há objectos importantes que fiquem fora do conjunto.

Outras vezes o conjunto de objectos é uma amostra de uma população mais extensa e generalizar as conclusões da amostra à população é uma tentação a que o analista deve resistir pelo menos nos casos em que não tiverem sido respeitados os princípios de escolha aleatória da amostra. Ao adoptar o princípio da escolha aleatória espera-se que os grupos que possam existir na população estejam adequadamente (proporcionalmente) representados na amostra. Mas a escolha aleatória pode trair a análise de clusters no caso de pequenos grupos que acabarão por estar pouco representados na amostra e ser absorvidos pelos grandes clusters produzidos pela análise.

A análise de clusters é usada essencialmente com fins descritivos e qualquer generalização de conclusões obtidas a partir da amostra é mais razoável ser feita com base em analogia do que nos princípios da inferência estatística.

2.4.2 Selecção de variáveis

As variáveis são as características dos objectos e são elas que de facto identificam os objectos. A escolha do número e natureza das variáveis é provavelmente um dos aspectos que mais influencia os resultados de uma análise de clusters. Por uma questão de parsimónia parece razoável pensar num número tão pequeno quanto possível de variáveis a escolher e aten-

der não só à relevância (Milligan, 1980) das variáveis para o objectivo do estudo como ao seu poder discriminatório. O sexo é uma variável decisiva na divisão de um grupo de pessoas em homens e mulheres mas não parece nada relevante se o objectivo é saber se no conjunto em análise há diferentes grupos de pessoas, identificados pelo interesse que revelam por um determinado conjunto de temas culturais.

A escolha do número de variáveis é também uma tarefa vital para a análise mas é polémica e diferentes autores revelam opiniões diversas, como Hands and Everitt (1987) que afirmam que aumentando o número de variáveis se obtém uma melhor identificação dos clusters, e Price (1993) que, nas mesmas circunstâncias, sustenta que se obtém uma fraca identificação dos clusters.

2.4.3 Estandardização

A estandardização é justificada principalmente por três razões:

(i) as variáveis são medidas em unidades diferentes,
(ii) as variáveis têm variâncias muito diferentes,
(iii) as variáveis são de diferentes tipos.

A estandardização elimina os efeitos arbitrários que as variáveis não estandardizadas têm na construção dos índices de semelhança, fazendo com que a sua constribuição para a construção dos coeficientes de semelhança seja mais equilibrada.

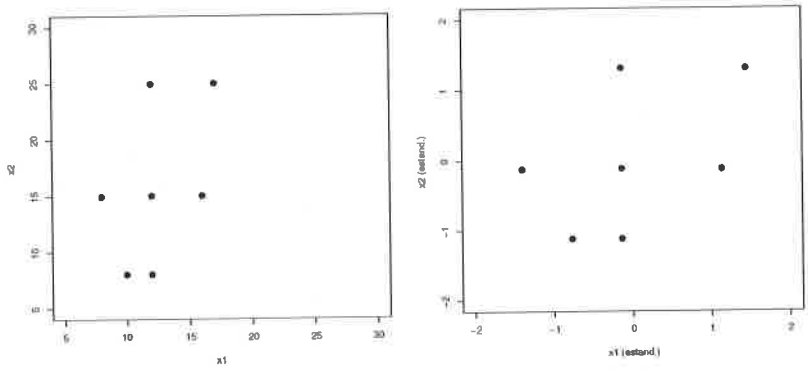
Dois procedimentos comuns para efectuar a estandardização são dividir as observações pelo desvio padrão ou pela amplitude das observações. Em Milligan and Cooper (1988) mostra-se, com base num estudo de simulação, que a divisão pela amplitude é um método de estandardização que supera não só o do desvio padrão como outros experimentados nesse estudo.

A Tabela 2.9 mostra um pequeno conjunto de dados artificiais. A representação gráfica destes dados apresenta-se na Figura 2.3 (a). A Figura 2.3 (b) mostra o gráfico dos mesmos dados depois de estandardizados e serve para ilustrar como a estandardização pode de facto alterar a razão de semelhança entre os objectos iniciais. No Capítulo 6 vê-se, observando a Figura 6.4, como essa alteração se reflecte nos resultados da análise de clusters.

O problema da estandardização levanta alguma polémica e há autores que consideram tratar-se de um comodismo cego que foge ao trabalho de usar maneiras apropriadas de tratar dados heterogéneos.

Tabela 2.9 *Conjunto de dados artificiais.*

Objectos	x_1	x_2
1	10	8
2	12	8
3	8	15
4	12	15
5	16	15
6	12	25
7	17	25



(a) não estandardizados (b) estandardizados

Figura 2.3 *Diagramas de dispersão para os dados da Tabela 2.9.*

A estandardização é um caso especial de um processo mais geral que consiste em pesar as variáveis de forma a homogeneizar a sua contribuição na construção de indíces de semelhança. A escolha de pesos é uma tarefa difícil e requer um conhecimento profundo das variáveis e do seu papel no problema em análise. A fórmula (2.7) dá sugestões para pesos a atribuir às variáveis.

Para os dados relativos a 20 alimentos apresentados na Tabela 3.1 uma maneira sugestiva de atribuir pesos às variáveis é dividir cada uma delas pela quantidade correspondente às necessidades diárias dessa mesma variável.

O procedimento geral para controlar a acção das variáveis é a transformação, o que engloba os procedimentos anteriores. Uma transformação muito comum é o logaritmo ($Y_i = \log X_i$), usada de preferência quando a variável em causa apresenta valores muito grandes em relação às outras.

2.4.4 Escolha da medida de proximidade

Muitos dos métodos usados na análise de clusters requerem uma matriz de dissemelhanças entre as entidades a agrupar, objectos ou variáveis. Nas Secções 2.2 e 2.3 refere-se um grande número de coeficientes de semelhança e dissemelhança e a lista existente parece não ter fim. Por isso, apesar das sugestões dadas em Gower and Legendre (1986), a escolha é difícil, mesmo tendo presente que se conhecem as propriedades mais importantes desses coeficientes (Baulieu, 1989).

Em geral para agrupar objectos usam-se coeficientes de dissemelhanças, muitos deles baseados em distâncias, e o agrupamento de variáveis assenta em medidas de correlação ou associação.

O que se pode dizer é que não se conhece uma indicação clara para escolher a medida de proximidade que melhor se adapte ao problema que se pretende resolver. A natureza das variáveis, as transformações que estas possam exigir e o método de análise de clusters a usar condicionam a escolha, mas igualmente importante é um profundo conhecimento do assunto que está a ser investigado. É por isso que apesar de tanta oferta os investigadores continuam a propor novos coeficientes.

2.4.5 Dados omissos

Uma situação que ocorre muitas vezes quando se observam dados multivariados é a existência de valores omissos que surgem porque há observações que se perderam ou que não puderam ser efctuadas, o que ocorre em muitas áreas como, por exemplo, medicina, sociologia e arqueologia. Everitt *et al.* (2001) e Gordon (1999) dão indicações sobre como se deve proceder com dados omissos no caso da análise de clusters.

A eliminação dos objectos com dados omissos não é aconselhável quando o número de objectos nestas condições é grande, por razões óbvias.

O habitual método de imputação (estimação) baseado no cálculo de uma estatística a partir dos dados completos não é recomendado em análise de clusters, uma vez que o cálculo da estatística deveria ser feito com base nos dados completos relativos ao grupo a que o objecto com o valor omisso pertence. E isso não é possível uma vez que antes da análise não são conhecidos os grupos.

Uma estratégia que se aconselha é o uso do coeficiente de Gower apresentado na Secção 2.2. O peso ω_{ijk} na fórmula de Gower (2.13) é zero quando a variável k é omissa em pelo menos um dos indivíduos i ou j. Isso não perturba o processo uma vez que a semelhança entre i e j é obtida efectuando a média pesada com as restantes variáveis.

Exercícios

2.1 Dois objectos, i e j, têm as mesmas medidas relativamente às p variáveis que neles são observadas.

(a) Qual o valor da dissemelhança d_{ij}?

(b) Mostre que a dissemelhança entre o objecto i e um qualquer objecto h é igual à dissemelhança entre j e h, isto é, $d_{ih} = d_{jh}$, qualquer que seja o objecto h.

2.2 Dois objectos, i e j, são observados em p características. Mostre que a distância euclidiana, d_{ij}, entre os dois objectos se pode obter a partir da distância de cada um deles ao seu centróide, d_{ic} e d_{jc}. Escreva a relação entre os quadrados das três distâncias envolvidas.

2.3 Seja d_{ij}^2, o quadrado da distância euclidiana entre os objectos i e j, θ, o ângulo entre os vectores das coordenadas dos perfis dos dois objectos e r_{ij} o coeficiente de correlação daqueles vectores.

(a) Estabeleça as condições em que se verificam as igualdades

$$d_{ij}^2 = 2(1 - \cos\theta),$$

e

$$d_{ij}^2 = 2(1 - \cos r_{ij}).$$

(b) Quando $r_{ij} = 1$ tem-se que $d_{ij}^2 = 0$. Pode então concluir-se que os objectos são iguais, no sentido de terem perfis coincidentes? Discuta este caso em que r_{ij} é extremo.

2.4 Considere a família de métricas de Minkowski dada pela fórmula geral (2.9) e mostre que qualquer dos seus membros é uma métrica.

2.5 Mostre que a métrica do supremo se pode obter como limite da métrica de Minkowski, isto é, demonstre a igualdade (2.10).

2.6 Considere o quadrado da distância euclidiana entre dois objectos quaisquer e mostre que esta medida preserva a ordem das distâncias mas não satisfaz a propriedade triangular (recorra a um contra-exemplo), isto é, trata-se de uma dissemelhança que não é uma métrica. Este facto ajuda a compreender porque é que a propriedade triangular não está incluída nas propriedades da dissemelhança. O quadrado da distância euclidiana não seria uma dissemelhança, o que não é compatível com a grande importância que a distância euclidiana tem na construção de medidas de proximidade.

2.7 A semelhança para comparar dois objectos i e j é tal que $0 < s_{ij} \leq 1$. Mostre que $d_{ij} = 1 - s_{ij}$ e $d_{ij}^* = -\log s_{ij}$ são dissemelhanças.

2.8 Para comparar as espécies animais:

Tigre, Cão, Golfinho, Tubarão, Homem, Macaco,

foram considerados os atributos binários

- come outros animais
- come vegetais
- desloca-se sobre quatro patas
- vive na água
- tem pêlo.

Construa as matrizes de semelhanças entre os animais com base no coeficiente de Jacard e no coeficiente de concordância simples.

2.9 Recorde o coeficiente de semelhança de Hamann,

$$h_{ij} = \frac{(a+b)-(b+c)}{a+b+c+d},$$

e o coeficiente de concordância simples,

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

entre dois objectos i e j. Mostre que $h_{ij} = 2s_{ij} - 1$. Diga o que pode concluir quanto ao coeficiente de correlação entre os dois coeficientes de semelhança. Como se relacionam os resultados de uma análise de clusters sobre o mesmo conjunto de objectos usando separadamente aqueles coeficientes?

2.10 Duas variáveis binárias, X_1 e X_2, foram observadas em n indivíduos, tendo-se registado o número de concordâncias e discordâncias obtidas:

		X_2	
		1	0
X_1	1	a	b
	0	c	d

onde $a+b+c+d=n$. Um coeficiente de semelhança habitualmente sugerido para comparar variáveis é o coeficiente de correlação. Mostre que no caso de variáveis binárias se obtém a expressão (2.14).

2.11 Mostre que o coeficiente de semelhança para variáveis ordinais dado pela fórmula (2.12) e em que todo o objecto que tem o nível m tem todos os níveis inferiores, é precisamente o coeficiente de concordância simples.

3

Métodos gráficos

3.1 Introdução

A representação gráfica constitui um procedimento essencial em análise de dados multivariados. A inspecção visual dos gráficos produzidos pode revelar aspectos dos dados que sejam úteis para ajudar a perceber a sua estrutura e a escolher o tipo de método a usar na sua análise. Muitos livros de análise multivariada como, por exemplo, Krzanowski (1988) e Johnson (1998), dedicam atenção a este tópico. Outras referências importantes são, Everitt (1978), Wang (1978), Flury e Riedwyl (1981), Chambers *et al.* (1983), Tufte (1983) e de Toit *et al.* (1986).

Presentemente existe abundante *software* que produz ilustrações gráficas de dados univariados com formas e cores variadas muito agradáveis de observar e analisar. Embora no caso de dados multivariados a representação gráfica seja muito mais complicada, há actualmente *software* que permite ao utilizador produzir gráficos atraentes com animação e com cores, e actuar sobre esses gráficos interactivamente.

Do ponto de vista geométrico a análise de clusters é muito simples, uma vez que o que se pretende é visualizar os clusters a partir de uma representação geométrica dos objectos ou das variáveis.

Dispondo de um conjunto de n objectos observados em p características ou variáveis cada objecto pode ser representado por um ponto no espaço p-dimensional. Por sua vez cada variável pode ser representada por um vector num espaço n-dimensional. Fazer análise de clusters resume-se então a observar pontos ou vectores num espaço apropriado. Em termos de visualização isso é particularmente útil em espaços de dimensão inferior ou igual a três.

O espaço geralmente preferido é o espaço a duas dimensões onde o olho humano parece ser capaz de explorar a posição relativa dos pontos.

Na Figura 3.1 estão representados sete objectos cuja observação poderá

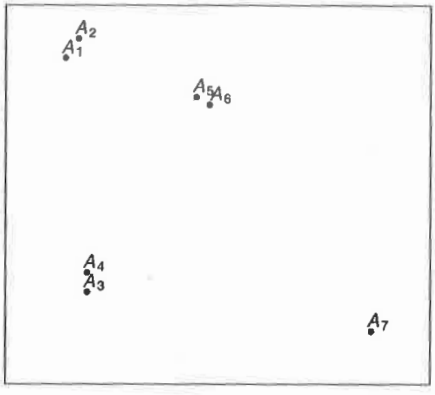


Figura 3.1 *Sete objectos e três clusters.*

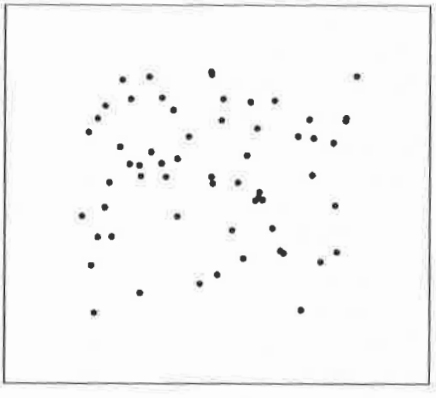


Figura 3.2 *Um conjunto de objectos sem estrutura aparente de grupos.*

levar a concluir que existem três grupos, A_1A_2, A_3A_4 e A_5A_6, e que o objecto A_7 está isolado. Porém, mesmo em duas dimensões a observação visual pode ficar complicada como é o caso da Figura 3.2 onde não é fácil discernir uma estrutura de grupos.

As dificuldades de uma análise de clusters feita com base na visualização da representação gráfica residem no facto de ser um procedimento subjectivo que geralmente só funciona quando é possível representar os objectos em espaços de dimensão não superior a três e, de preferência, quando o número de objectos não é muito grande. A subjectividade resulta do facto de não se conhecer o mecanismo que as pessoas usam para agrupar os pontos num gráfico e reconhecer os clusters (Feldman, 1995). Como é sabido as pessoas geralmente concordam na análise de estruturas onde os clusters aparecem bem demarcados, mas quando isso não acontece as análises

diferem de pessoa para pessoa e há mesmo quem veja clusters em dados onde eles não existem.

O que é preciso é então dispôr de procedimentos analíticos e automáticos que consigam identificar grupos de pontos, qualquer que seja o seu número e qualquer que seja a dimensão do espaço em que estão. Alguns destes procedimentos serão estudados nos capítulos seguintes. Entretanto, tentando explorar esta capacidade visual do olho humano, mostra-se a seguir como é que se podem representar objectos medidos em uma, duas ou mais dimensões. Consideram-se dois processos: representação gráfica directa dos objectos e representação gráfica indirecta dos objectos.

3.2 Representação gráfica directa

3.2.1 Uma e duas variáveis

Quando os objectos são medidos numa só variável, o que é raro, a representação gráfica mais comum é o histograma. A forma do histograma é geralmente informativa sobre o modo como se distribuem os objectos de acordo com a variável observada. A existência de várias modas é em geral reveladora da existência de clusters, cada um deles associado a uma moda. Esta interpretação é particularmente útil quando se dispõe de um grande número de objectos.

O conjunto de dados, descritos na Tabela 3.1, relativos à composição e valor calórico de 100 gramas de cada um de 20 alimentos seleccionados, embora relativamente moderado no que diz respeito ao número de objectos e ao número de variáveis é adequado para ilustrar as várias técnicas gráficas usadas na análise de dados multivariados. Usando a variável *Ferro* e olhando para o correspondente histograma da Figura 3.3 vê-se que o *feijão* está isolado, que a *cenoura* e *espinafres* estão juntos e que os restantes alimentos, com baixo teor de ferro, podem considerar-se agrupados num só cluster. Este é o tipo de análise que se pode fazer quando se trabalha com uma única variável. Neste contexto, o estudo de todas as variáveis passa por construir tantos histogramas quantas as variáveis. A análise da Figura 3.3, onde se encontram todos os cinco histogramas correspondentes às cinco variáveis, juntamente com a análise dos dados da Tabela 3.1, mostra que é muito complicado interpretar globalmente o resultado deste procedimento, pois que, em geral, cada histograma pode sugerir uma classificação diferente, havendo variáveis relativamente às quais não parece haver qualquer divisão em grupos. Estas dificuldades prendem-se com duas importantes questões inerentes a toda a análise de clusters:

Tabela 3.1 *Dados relativos a 20 alimentos seleccionados (por 100 gramas).*

	Energia (kcal)	Proteínas (g)	Lípidos (g)	Cálcio (mg)	Ferro (mg)
Azeite	900	0	100	0.1	0.05
Manteiga	770	0	85	13	0.2
Pescada	85	19	1	25	0.9
Vaca	208	18	15	12	1.5
Frango	158	20	8.5	18	1.8
Leite	57	3	3	126	0.1
Iogurte	59	3.2	3.2	125	0.2
Q. flamengo	316	26	23.2	800	0.8
Q. serra	392	26	32	800	1.2
Arroz	350	7.5	0.5	10	0.5
Pão	258	7	0.6	24	1.6
Feijão	290	20	1.2	170	6.5
Açúcar	400	0	0	15	1
Massas	365	10	0.5	20	1
Alface	22	1.8	0.2	70	1.5
Cebola	22	0.9	0.2	31	0.5
Espinafres	22	2.6	0.9	104	3.6
Cenoura	22	0.6	0	104	3.6
Batata	90	2.5	0	9	0.2
Couve	30	2.9	0.5	234	1.8

(i) qual o objectivo da análise de clusters

(ii) que variáveis devem ser usadas na análise com vista a atingir aquele objectivo.

Além do histograma há outros métodos para representação gráfica de dados univariados que são igualmente informativos. Entre muitas outras possibilidades destacam-se os métodos que produzem gráficos de barras, gráficos circulares e gráficos caule-e-folhas.

Quando há duas variáveis que são observadas em cada objecto também é possível construir um histograma bidimensional. À excepção de casos especiais o histograma bidimensional não se revela útil.

A representação mais corrente e mais natural para duas variáveis é o diagrama de dispersão. Supondo que as duas variáveis são *Lípidos* e *Cálcio*, o diagrama de dispersão, que se encontra na Figura 3.4 no cruzamento das duas variáveis, destaca o agrupamento *azeite* e *manteiga*, o grupo dos *queijos*, ficando aparentemente todos os outros elementos num só grupo.

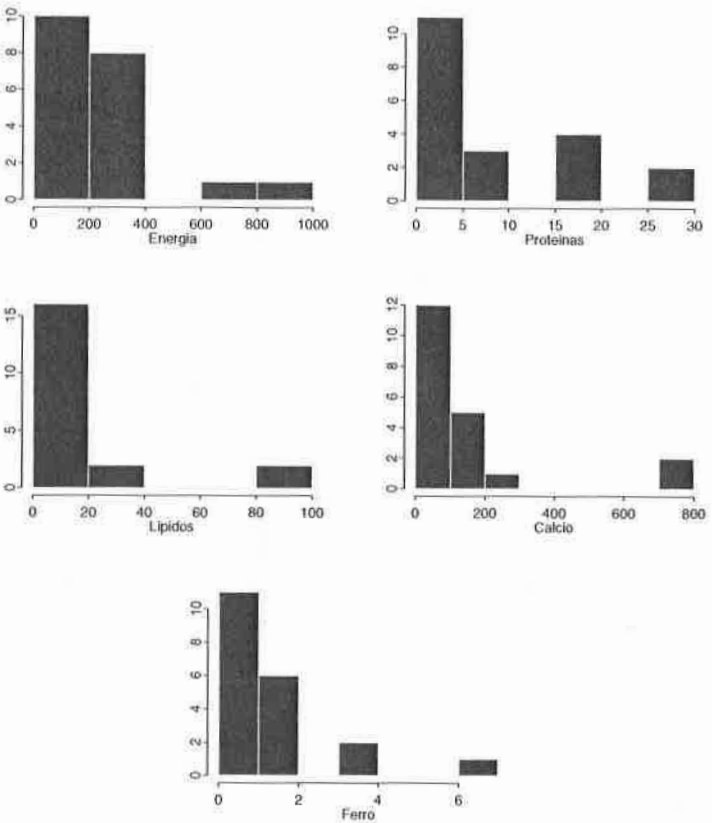


Figura 3.3 *Histogramas para cada uma das variáveis da Tabela 3.1.*

A consideração de todos os pares de variáveis é uma tentativa de análise global mas resulta em geral difícil e confusa, sobretudo quando o número de variáveis é grande. Mas como todas as contribuições, mesmo diminutas, podem ser úteis para a análise e porque o *software* actualmente disponível produz estes gráficos com grande facilidade, vale sempre a pena olhar para eles com atenção e perspicácia. O diagrama *Energia* contra *Lípidos* destaca bem o grupo *azeite* e *manteiga* e além disso revela que para um grande grupo de alimentos parece existir uma relação linear positiva entre as duas variáveis. Esse grupo é constituído por todas as observações à excepção dos farináceos – *arroz, pão, feijão, açúcar* e *massas*. A correlação entre as variáveis *Energia* e *Lípidos* é 0.863, com todas as observações, e 0.992, sem os farináceos.

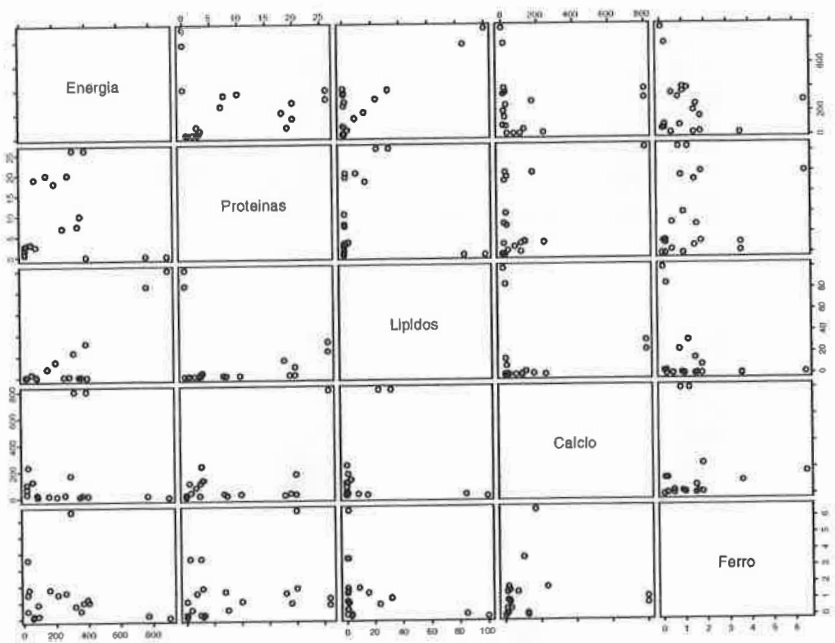


Figura 3.4 *Matriz de diagramas de dispersão para os dados dos alimentos.*

3.2.2 Três ou mais variáveis.

Quando a observação dos objectos é feita em três ou mais variáveis, o histograma e o diagrama de dispersão não podem ser usados directamente, uma vez que estes não podem envolver todas as variáveis simultaneamente. Exceptua-se o caso de três variáveis que permite a construção de diagramas de dispersão tridimensionais, mas a leitura e análise destes diagramas é em geral difícil.

A Figura 3.5 apresenta esse diagrama para as variáveis *Proteínas*, *Lípidos* e *Cálcio*.

Como se viu a aplicação do histograma a cada uma das variáveis e do diagrama de dispersão a cada um dos pares de variáveis separadamente, como mostram as Figuras 3.3 e 3.4, respectivamente, é uma forma indirecta de usar estes gráficos para realizar a análise global com todas as variáveis, como se pretende. Existem, no entanto, outros métodos. As Figuras 3.6, 3.7 e 3.8 mostram três maneiras engenhosas de representar graficamente dados multivariados (caras de Chernoff, estrelas e curvas de Andrews). Os

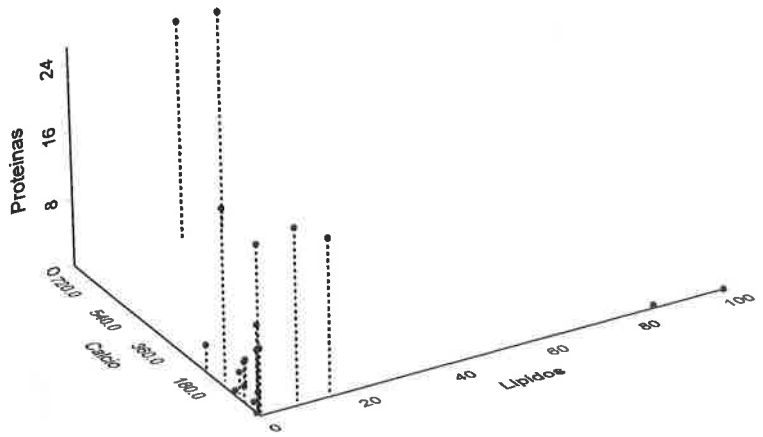


Figura 3.5 *Um diagrama de dispersão tridimensional para os dados dos alimentos.*

dois primeiros gráficos foram obtidos usando S-PLUS e o terceiro usando uma função[1] escrita para o *software* R.

- **Caras de Chernoff**

Cada variável é associada a um aspecto particular da face de uma pessoa (Chernoff, 1973). Com p variáveis, $X_1,\ldots,X_p$, pode associar-se, por exemplo, X_1 ao tamanho global da cara (área da cara), quanto maior o tamanho maior o valor de X_1. O tamanho do nariz pode associar-se a X_2, a distância entre os olhos a X_3 e assim sucessivamente para outros aspectos da cara.

A Figura 3.6 mostra a representação gráfica dos dados da Tabela 3.1 usando caras de Chernoff. O gráfico permite identificar vários clusters, confirmando as expectativas que se têm como resultado do conhecimento geral sobre o valor calórico e a composição dos alimentos. É instrutivo comparar os clusters que resultam da análise desta representação gráfica

[1] Disponível em http://math.usu.edu/~minnotte/research/software/Andrews.r

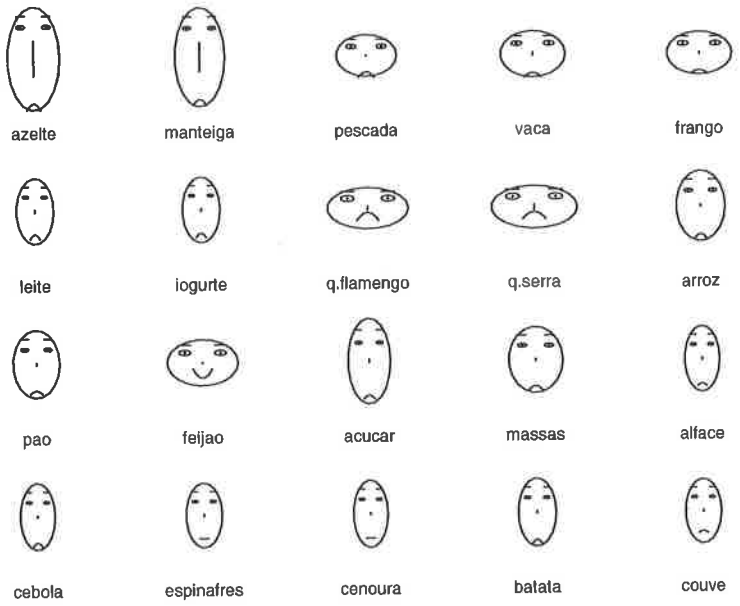


Figura 3.6 *Caras de Chernoff para os 20 alimentos.*

com os clusters produzidos por uma verdadeira análise de clusters aplicando os métodos dos Capítulos 4 e 5. No Capítulo 6 pode ver-se o resultado que se obtém usando análise de clusters.

Uma dificuldade das caras de Chernoff prende-se com o facto da associação das variáveis às características da cara ser subjectiva. Além disso, diferentes maneiras de associar as variáveis às características conduzem a gráficos de diferentes aspectos e podem levar a diferentes conclusões em termos do número e composição dos clusters.

- **Estrelas** (ou polígonos ou raios de sol)

Cada objecto é associado a um círculo de raio constante e o valor das variáveis é indicado ao longo dos raios do círculo. Ao ligar as extremidades dos raios obtém-se um polígono ou estrela. Na Figura 3.7 encontra-se a representação gráfica das estrelas associadas aos alimentos da Tabela 3.1. O número e a composição dos clusters sugeridos por esta representação é idêntico àqueles que são fornecidos pelo uso das caras de Chernoff. O *feijão* aparece como um *outlier*, o que era visível também na Figura 3.6.

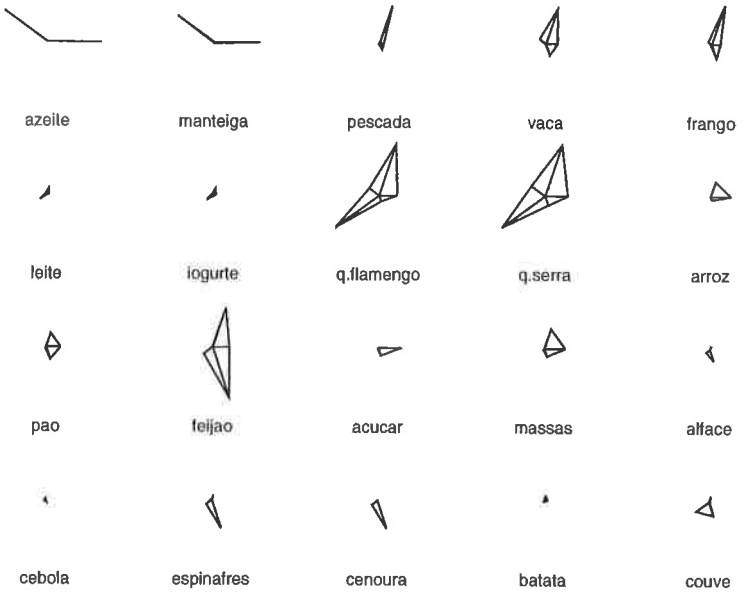


Figura 3.7 *Estrelas para os 20 alimentos.*

- **Curvas de Andrews**

Um procedimento comum usado para representar dados multivariados é associar cada um dos objectos do estudo, isto é, cada observação multivariada, a um objecto conhecido e que nos seja familiar no nosso dia a dia. Chernoff associou objectos a caras. Outra possibilidade é associar os objectos a entidades matemáticas. Andrews (1972) propôs associar o objecto com observações $\mathbf{x}_r' = (x_{r1}, \ldots, x_{rp})$, à função harmónica

$$f_r(t) = \frac{x_{r1}}{\sqrt{2}} + x_{r2}\,\mathrm{sen}\,t + x_{r3}\cos t + x_{r4}\,\mathrm{sen}\,(2t) + x_{r5}\cos(2t) + \cdots,$$

onde t é tal que $-\pi < t < \pi$. Representando graficamente a função harmónica no intervalo $(-\pi, \pi)$ fica feita a associação entre os objectos em estudo e as chamadas curvas de Andrews. A função $f_s(t)$ goza de várias propriedades interessantes (ver Exercício 3.4) e aquela que mais útil se revela no contexto da representação gráfica refere que a função preserva a distância euclidiana, isto é, a distância euclidiana entre dois objectos i e j é proporcional à distância euclidiana entre as respectivas funções $f_i(t)$ e $f_j(t)$.

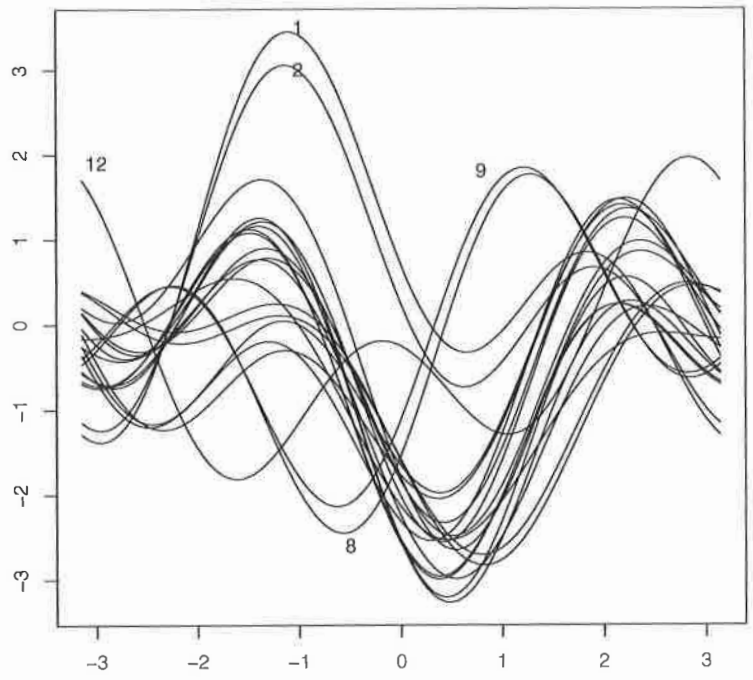


Figura 3.8 *Curvas de Andrews para os 20 alimentos (com os números dos objectos mais destacados).*

As dificuldades das curvas de Andrews são basicamente as mesmas dificuldades das caras de Chernoff. A interpretação é difícil quando há muitos objectos e a representação muda quando muda a ordem das variáveis que figuram na função. Em ambos os casos, caras de Chernoff e curvas de Andrews, estandardizar os dados revela-se útil quando as variáveis são medidas em unidades diferentes.

As curvas de Andrews para os dados dos 20 alimentos estão representadas na Figura 3.8. Ressalta imediatamente o isolamento do *feijão* (curva 12) e as associações do *azeite* e *manteiga* (curvas 1 e 2) e dos queijos (curvas 8 e 9). No feixe das curvas restantes não é fácil distinguir outros clusters como acontece no caso das caras de Chernoff e estrelas em que há outros clusters visíveis.

Além dos métodos aqui brevemente referidos existem muitos outros (grifos, caixas, bolhas, vários tipos de perfis e outros gráficos engenhosos como o diagrama de contornos que se apresenta na Figura 3.9) que estão descritos na bibliografia indicada.

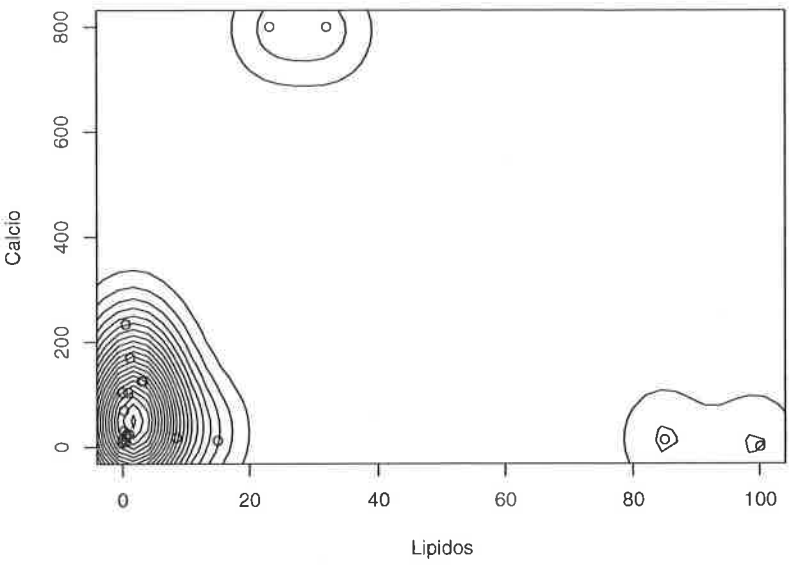


Figura 3.9 *Diagrama de contornos para os 20 alimentos.*

3.3 Representação gráfica indirecta

Muitos dos métodos de análise multivariada assentam em procedimentos que levam à redução do número de dimensões do espaço inicial de trabalho. Este resultado é muito útil porque se à partida o número de variáveis é grande, na sequência da análise esse número pode ficar bastante reduzido, facilitando a interpretação da estrutura dos dados e facilitando análises subsequentes. Uma vantagem específica que resulta desta redução da dimensão traduz-se na possibilidade de os objectos serem representados graficamente em espaços de pequena dimensão, eventualmente em espaços de duas dimensões, onde os objectos podem ser então visualizados.

Para ilustrar este subproduto de natureza gráfica aplicam-se a seguir três métodos de análise multivariada (componentes principais, análise factorial e multidimensional scaling) a três conjuntos de dados. Não se dá mais do que uma breve explicação dos métodos usados e remete-se o leitor para algum dos muitos livros de análise multivariada como, por exemplo, Johnson and Wichern (2002), Jobson (1992) e Everitt and Dunn (2001). Aqui o leitor

encontra não só explicações suficientes sobre os aspectos mais importantes dos métodos multivariados como também bibliografia especializada sobre os mesmos métodos.

Os mesmos dados são novemente explorados no Capítulo 6 usando métodos específicos de análise de clusters. Os resultados aí obtidos são confrontados depois com os resultados desta análise gráfica indirecta.

• Componentes Principais

Este método de análise multivariada parte de p variáveis iniciais, $X_1, \ldots, X_p$, observadas num conjunto de n objectos, e encontra p combinações lineares, $Y_1, \ldots, Y_p$, onde

$$Y_i = a_{i1}X_1 + \cdots + a_{ip}X_p,$$

que são não correlacionadas entre si. A inexistência de correlação significa que as novas variáveis medem diferentes dimensões da estrutura dos dados, sendo que muitas vezes é possível interpretar essas dimensões atribuindo-lhes significado físico. As novas variáveis chamam-se Componentes Principais e são construídas de tal forma que aparecem ordenadas segundo a magnitude da sua variância, isto é

$$\operatorname{var} Y_1 \geq \operatorname{var} Y_2 \geq \cdots \geq \operatorname{var} Y_p .$$

Ao efectuar uma análise de componentes principais espera-se que as variâncias de muitas das novas variáveis, ou componentes principais, sejam tão pequenas que possam ser desprezadas sem que a variabilidade total do sistema inicial de p variáveis fique grandemente reduzida. Se acontecer o que se espera, a variação total dos dados pode então descrever-se por um número reduzido de componentes principais, portadoras de uma percentagem elevada da variação total. Os resultados mais interessantes ocorrem quando as variáveis originais são altamente correlacionadas.

Os casos que mais interessam à representação gráfica são aqueles em que apenas duas componentes principais são suficientes para descrever o sistema de variáveis iniciais. Podem então calcular-se os valores, ou *scores*, dos objectos em cada uma das duas componentes principais retidas e em seguida representá-los graficamente. Por exemplo, as coordenadas do objecto i nas duas primeiras componentes principais, Y_1 e Y_2, respectivamente y_{i1} e y_{i2}, são

$$\begin{aligned} y_{i1} &= a_{11}x_{i1} + a_{12}x_{i2} + \cdots + a_{1p}x_{ip} \\ y_{i2} &= a_{21}x_{i1} + a_{22}x_{i2} + \cdots + a_{2p}x_{ip} , \end{aligned}$$

onde $i = 1, \ldots, n$ e $x_{i1}, \ldots, x_{ip}$ são os valores das observações do objecto i nas variáveis originais, $X_1, \ldots, X_p$.

Tabela 3.2 *Componentes principais associadas às variáveis da Tabela 3.1.*

	Componentes principais				
Variáveis	1	2	3	4	5
Energia	−0.633	0.170	−0.302	0.155	0.675
Proteínas	0.158	0.682	0.017	0.692	−0.175
Lípidos	−0.654	0.142	−0.194	−0.163	−0.699
Cálcio	0.044	0.678	0.312	−0.645	0.159
Ferro	0.381	0.164	−0.879	−0.233	−0.024
Variância	2.030	1.648	0.829	0.374	0.119
% de variância acumulada	40.6	73.5	90.2	97.7	100.0

A aplicação do método das componentes principais à matriz de correlações dos dados dos alimentos produziu os resultados que se encontram na Tabela 3.2.

Como se observa as duas primeiras componentes principais são responsáveis por 73.5% da variância total e por isso é razoável usar estas duas componentes para explicar o sistema inicial. De acordo com os critérios usuais de interpretação a primeira componente principal define um contraste entre *Energia* e *Lípidos*, por um lado, e as restantes variáveis, por outro. A segunda componente pode ser vista como uma média das cinco variáveis observadas.

A representação gráfica dos 20 alimentos, apresentada na Figura 3.10, permite visualizar os objectos e identificar clusters que possam existir. Neste caso o gráfico das duas primeiras componentes principais revela essencialmente a mesma estrutura de agrupamentos já captada pelos vários métodos de representação gráfica directa descritos na Secção 3.2.

- **Multidimensional Scaling** (MDS)

Dados um conjunto de objectos e uma matriz de dissemelhanças associadas a esses objectos, o método MDS tem por objectivo construir uma configuração do conjunto dos objectos num espaço de dimensão reduzida. Entende-se por configuração um conjunto de pontos do referido espaço determinados de tal forma que a cada ponto corresponde um objecto e que a ordem das distâncias euclidianas entre pontos da configuração respeita a ordem das dissemelhanças entre os objectos correspondentes. Isto significa que, dados dois objectos muito dissemelhantes, os dois pontos da configuração que os representam estão muito distantes. Um exemplo muito comum que esclarece bem o objectivo do método MDS consiste em usar as distân-

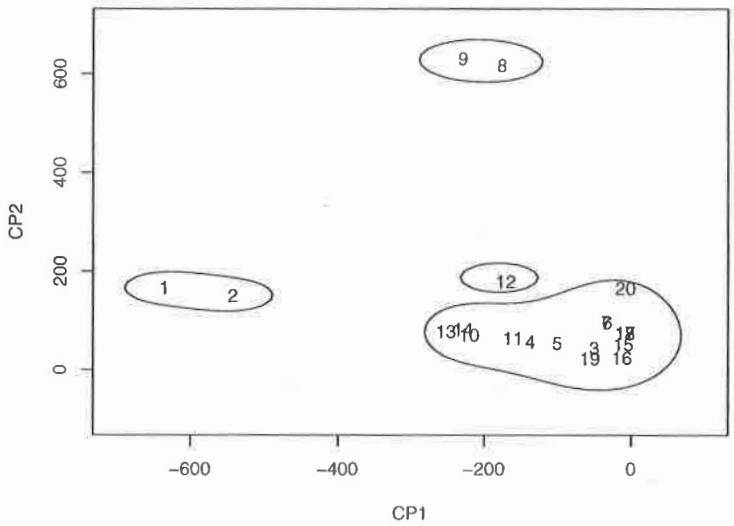


Figura 3.10 *Representação dos 20 alimentos no plano das duas primeiras componentes principais.*

cias kilométricas entre as cidades de um país (matriz de dissemelhanças) e aplicar o MDS para reconstruir o mapa das cidades (configuração).

Como em outros métodos de análise multivariada espera-se que o produto final (aqui é a configuração) esteja situado num espaço de baixa dimensão, mas, do ponto de vista gráfico, o que realmente interessa é um espaço de dimensão dois ou, no máximo, de dimensão três.

O método MDS foi aplicado à matriz de dissemelhanças (Tabela 1.4) entre expressões da face de uma mulher e a configuração em duas dimensões, apresentada na Figura 3.11, revela três grupos de cenários: um primeiro que inclui os cenários agradáveis (2, 3, 4, 8, 9), um segundo onde estão os cenários desagradáveis (1, 5, 10, 11, 13) e um terceiro contendo os cenários 6, 7 e 12. Esta interpretação simplista resulta apenas de uma tentativa de classificar os clusters que foram identificados à vista. Uma interpretação mais rigorosa e completa não dispensa conhecimentos da psicologia da expressão facial.

- **Análise Factorial**

O objectivo da análise factorial é semelhante ao da análise de compo-

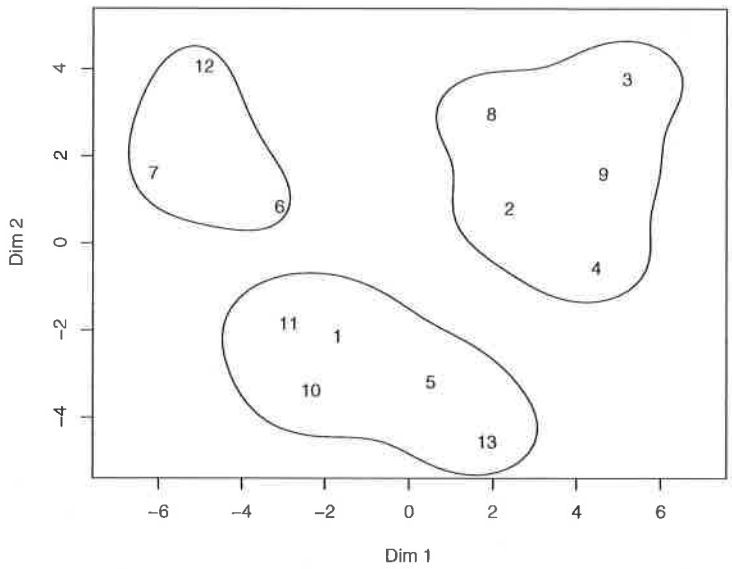


Figura 3.11 *Configuração MDS dos objectos com a matriz de dissemelhanças da Tabela 1.4.*

nentes principais. A ideia básica é descrever um conjunto de p variáveis $X_1, X_2, \ldots, X_p$ em termos de um pequeno número de factores (ou índices), esperando que este processo venha a elucidar sobre a relação entre as variáveis originais.

A grande diferença entre os dois métodos é que a análise factorial é baseada num verdadeiro modelo estatístico, enquanto que as componentes principais não.

O modelo de análise factorial pode ser expresso em termos algébricos por um conjunto de equações lineares

$$\begin{aligned} X_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + e_1 \\ X_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + e_2 \\ &\vdots \\ X_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pk} f_k + e_p \end{aligned} \tag{3.1}$$

em que cada variável observada, X_i, é uma soma ponderada de k factores, ou variáveis latentes, não observadas directamente, mais uma variável residual

específica da variável X_i. Os factores $f_1, \ldots, f_k$ são geralmente conhecidos por factores comuns, as variáveis $e_1, \ldots, e_p$ por factores específicos e os pesos $\lambda_{11}, \ldots, \lambda_{pk}$ por *loadings*.

O modelo pode ser usado meramente para explorar a estrutura dos dados e investigar a relação entre variáveis observadas e factores.

Evitando entrar na complexidade deste modelo, nas hipóteses em que assenta e nos detalhes dos métodos de estimação que estão disponíveis para estimar os seus parâmetros, pode afirmar-se que a resolução do sistema de equações 3.1 é equivalente à seguinte decomposição da matriz de covariâncias das variáveis observadas, $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi},$$

em que $\mathbf{\Lambda}$ é a matriz dos pesos e $\mathbf{\Psi}$ é a matriz diagonal das variâncias de $e_1, \ldots, e_p$. Em vez da matriz de covariâncias pode ser usada a matriz de correlações das variáveis observadas.

Estimados os parâmetros, podem calcular-se os *scores* dos objectos nos k factores do modelo. Com base nestes valores (*scores*) podem representar-se graficamente os objectos no espaço dos factores, o que é particularmente útil no caso de dois factores. Este processo de representação gráfica dos objectos é semelhante ao que foi usado na análise de componentes principais.

A análise factorial está porém vocacionada para representar graficamente as variáveis. Para exemplificar esta capacidade analisou-se a matriz de correlações apresentada na Tabela 3.3. Os dados, referidos em Harman (1976), consistem em oito variáveis físicas medidas em 305 raparigas dos sete aos dezassete anos de idade.

A análise directa da matriz de correlações sugere a existência de dois grupos de variáveis: (i) as quatro primeiras que medem a "esbelteza" e (ii) as quatro últimas que medem a "robustez física".

O resultado da análise factorial produziu as estimativas dos *loadings* que se apresentam na Tabela 3.4.

A Figura 3.12 faz a representação gráfica das variáveis num sistema em que os eixos são os factores. As coordenadas de cada variável são os seus *loadings* nos dois factores. O gráfico mostra claramente a existência dos dois grupos de variáveis já sugeridos pela análise directa da matriz de correlações. Fica assim evidenciado o interesse da análise factorial na pesquisa de agrupamentos de variáveis, aspecto que constitui o único objectivo desta breve e incompleta introdução.

Os três métodos aqui referidos a propósito da representação gráfica indirecta de objectos e variáveis não esgotam a oferta que a análise multivariada nos dá. Dois outros métodos muito comuns e que servem estes

Tabela 3.3 *Matriz de correlações (oito características físicas de 305 raparigas).*

Variável	1	2	3	4	5	6	7	8
1. Altura	1.000							
2. Envergadura	0.846	1.000						
3. Antebraço	0.805	0.881	1.000					
4. Tíbia	0.859	0.826	0.801	1.000				
5. Peso	0.473	0.376	0.380	0.436	1.000			
6. Anca	0.398	0.326	0.319	0.329	0.762	1.000		
7. Peito-c	0.301	0.277	0.237	0.327	0.730	0.583	1.000	
8. Peito-d	0.382	0.415	0.345	0.365	0.629	0.577	0.539	1.000

Tabela 3.4 *Estimativas dos loadings correspondentes à análise factorial de oito variáveis físicas.*

	Factores	
Variáveis	1	2
1	0.856	−0.324
2	0.848	−0.410
3	0.809	−0.409
4	0.831	−0.342
5	0.746	0.563
6	0.632	0.496
7	0.570	0.513
8	0.608	0.353

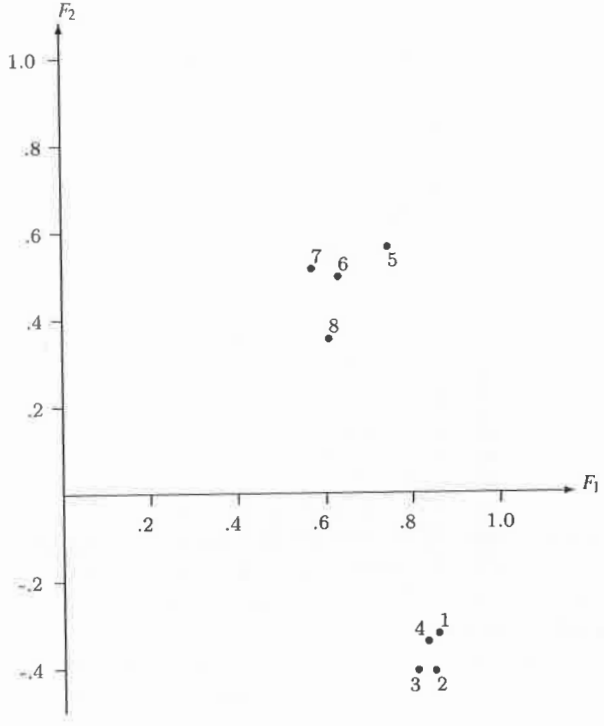


Figura 3.12 *Análise factorial de oito variáveis físicas.*

objectivos gráficos são a análise de correspondências e o método que assenta na representação gráfica *biplot*. Uma introdução a estes métodos pode ser encontrada na bibliografia referida no ínicio desta secção.

Exercícios

3.1 Considere os dados dos planetas apresentados no Capítulo 1 e as sete primeiras variáveis da Tabela 1.1. Faça, recorrendo a *software* adequado, a representação gráfica dos dados usando

(a) caras de Chernoff
(b) estrelas
(c) curvas de Andrews.

3.2 Repita o Exercício 3.1 trabalhando com os dados estandardizados.

3.3 Compare os gráficos produzidos no Exercício 3.1 e no Exercício 3.2 e identifique aquele que lhe parece mais útil. Justifique a sua escolha.

3.4 Mostre que as curvas de Andrews preservam a distância euclidiana. Isto é, usando o quadrado da distância euclidiana entre os objectos i e j, tem-se que

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

é proporcional a

$$\int_{-\pi}^{\pi} [f_i(t) - f_j(t)]^2 \, dt \, .$$

Recorde que

$$\int_{-\pi}^{\pi} \operatorname{sen} kt \operatorname{sen} lt \, dt = 0, \;\; k \neq l \, , \quad \int_{-\pi}^{\pi} \cos kt \cos lt \, dt = 0, \;\; k \neq l \, ,$$

e

$$\int_{-\pi}^{\pi} \operatorname{sen} kt \cos lt \, dt = 0 \, .$$

3.5 Considere os dados da temperatura apresentados na Tabela 3.5.

(a) Apresente o diagrama de dispersão a três dimensões para as três variáveis.
(b) Apresente os diagramas de dispersão para todos os pares de variáveis. Acha que algum destes gráficos suporta a ideia das variáveis estarem relacionadas?
(c) Faça o estudo gráfico dos dados usando
- caras de Chernoff
- estrelas
- curvas de Andrews

Tabela 3.5 *Dados relativos às temperaturas médias diárias em 33 cidades da Europa no mês de Janeiro de 2004.*

	Máxima (^oC)	Mínima (^oC)	Média (^oC)
Amesterdão	9.2	-3.6	3.9
Atenas	15.6	0.3	8.8
Belgrado	8.2	-9.0	-1.2
Berna	10.7	-6.5	1.1
Bratislava	5.4	-9.9	-2.2
Bruxelas	10.2	-5.7	3.8
Bucareste	2.1	-10.4	-3.3
Budapeste	5.7	-7.8	-2.5
Copenhaga	3.2	-7.6	-1.4
Dublin	10.4	-0.8	5.5
Estocolmo	1.9	-12.7	-3.3
Helsínquia	-0.3	-16.4	-7.2
Kiev	1.6	-11.2	-4.3
Lisboa	15.8	8.4	12.5
Londres	10.4	0.6	6.2
Madrid	12.0	1.4	6.0
Minsk	-0.1	-15.3	-6.9
Moscovo	1.2	-16.4	-7.0
Munique	7.1	-10.8	-1.0
Oslo	0.3	-16.8	-6.0
Paris	11.8	-2.2	5.3
Praga	3.3	-16.6	-3.4
Pristina	7.2	-8.2	-1.1
Reiquejavique	6.0	-9.4	-0.1
Riga	1.1	-12.1	-5.6
Roma	13.3	0.4	6.7
Skopje	8.4	-5.4	-0.3
Sófia	8.6	-10.7	-2.9
Tblisi	8.4	-1.2	3.4
Tirana	14.2	-2.2	6.7
Varsóvia	2.8	-12.1	-5.0
Viena	5.7	-10.9	-1.8
Zagreb	9.7	-5.8	0.0

Identifique os clusters se os houver. Altere a ordem das variáveis e repita as três representações gráficas anteriores. Compare as representações. O que pode concluir?

(d) Efectue uma mudança de escala apresentando as temperaturas em graus Fahrenheit. Em seguida responda à alínea (c). Comente os resultados obtidos.

4

Métodos hierárquicos

4.1 Introdução

Nos métodos hierárquicos os grupos formam uma hierarquia caracterizada pelo facto de dados dois grupos, quaisquer que eles sejam, os grupos ou são disjuntos ou um deles está contido no outro. Para aplicar os métodos hierárquicos recorre-se geralmente a dois tipos de procedimentos ou algoritmos:

(i) os algoritmos *aglomerativos* ou ascendentes que actuam a partir dos n objectos iniciais, encarados como grupos com um só objecto, formando novos grupos por aglutinação sucessiva de grupos formados anteriormente;

(ii) os algoritmos hierárquicos *divisivos* ou descendentes que actuam a partir de um grupo inicial, formando novos grupos por divisão sucessiva de grupos anteriores até chegar a n grupos singulares de um só objecto.

A estrutura hierárquica proveniente destes procedimentos, tanto aglomerativos como divisivos, costuma representar-se por um gráfico a duas dimensões chamado *dendrograma*, também conhecido por *diagrama de árvore* ou *árvore hierárquica* e ainda por *fenograma* que é a designação usada em taxonomia numérica. Em Kaufman and Rousseeuw (1990) referem-se outras maneiras de representar graficamente a informação contida num dendrograma. Trata-se do gráfico em bandeira (*banner plot*), do gráfico *icicle* e do gráfico tipo *Ward*.

O gráfico da Figura 4.1 representa um dendrograma que configura o esquema de uma árvore em posição invertida, com a raiz para cima e os ramos para baixo. Os nós internos representam os clusters e a altura dos troncos indica a distância a que os clusters se ligam. Alturas pequenas indicam que a aglutinação é feita entre clusters razoavelmente homogéneos. Existem diversas variantes deste dendrograma e o formato gráfico em que

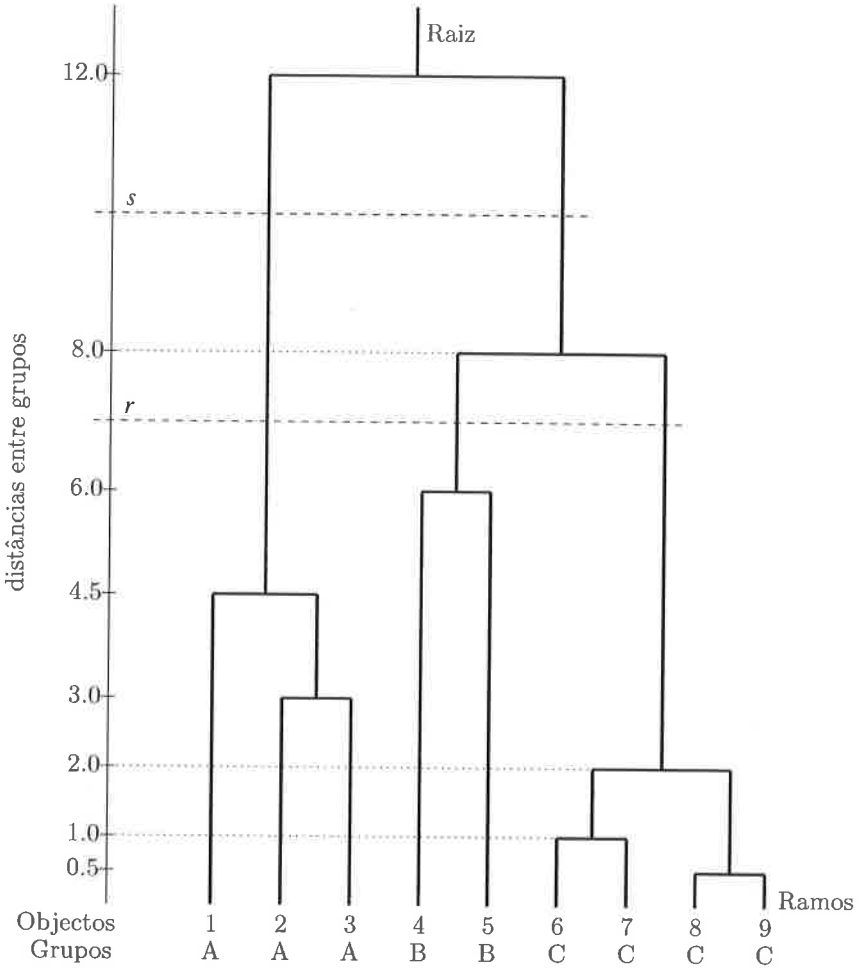


Figura 4.1 *Exemplo de dendrograma.*

é apresentado é também variado, dependendo do *software* que o produz.

Por vezes a árvore aparece em posição horizontal com os ramos à esquerda e a raíz à direita. Esta simples representação faz lembrar o tipo de estruturas de classificação em famílias, espécies e subespécies encontradas em zoologia e botânica. A grande virtude do dendrograma é mostrar como os sucessivos grupos efectivamente se vão formando ao longo do processo hierárquico, quer subindo quer descendo a árvore. No entanto o dendrograma

não mostra a informação relativa às dissemelhanças iniciais. O dendrograma serve também para ver quantos grupos se devem considerar. Neste exemplo há 9 objectos e a partir deles podem obter-se três grupos (A,B,C) cortando a árvore pelo segmento de recta r e obtêm-se dois grupos se a árvore for cortada pelo segmento de recta s. Note-se que embora os objectos tenham sido colocados por ordem não é essa a prática visto que a ordem dos objectos (ramos) é arbitrária embora dependente do algoritmo usado na construção dos clusters. De facto o dendrograma não é rigído e deve interpretar-se como um esquema móvel em torno dos eixos de ligação. Assim podemos colocar 2 na primeira posição e 1 na terceira, o que corresponde a efectuar uma rotação em torno do eixo de ligação vertical mais à esquerda. Efectuando a rotação do segundo eixo de ligação mais à esquerda consegue-se trocar 2 com 3 e colocar os objectos que definem o cluster A pela ordem 3, 2, 1. Na verdade, usando o mesmo método hierárquico e o mesmo conjunto de dados, podem obter-se 2^{n-1} dendrogramas de aparência diferente, dependendo da ordem pela qual se dispõem os objectos. Na Figura 4.1 encontra-se ainda um eixo vertical designado por "distâncias entre grupos" com uma escala que mede a distância a que os clusters se fundem. Dada uma matriz de dissemelhanças existem muitas maneiras de definir, a partir destas dissemelhanças, a distância entre grupos. E uma vez escolhida a definição apropriada pode então determinar-se a distância entre quaisquer dois clusters ou objectos. A distância entre dois objectos, d^*, é a distância dada pelo nível mínimo, nível crítico, ou nível de fusão a que os objectos se ligam para formar um novo cluster. No caso da Figura 4.1 as distâncias entre os objectos 6 e 7, 6 e 8, 6 e 5, são: $d^*_{67} = d_{67} = 1$, $d^*_{68} = 2$, $d^*_{65} = 8$. As novas distâncias d^*, chamadas distâncias críticas (*threshold*), satisfazem as habituais propriedades das dissemelhanças e gozam ainda da desigualdade ultramétrica, uma propriedade que a maior parte das dissemelhanças não satisfaz. De acordo com a desigualdade ultramétrica tem-se

$$d^*_{ij} \leq \max(d^*_{ik}, d^*_{kj}), \tag{4.1}$$

para todos os objectos i, j, k, onde d^*_{ij} representa a distância crítica entre os objectos i e j.

A propriedade ultramétrica surgiu pela primeira vez em 1967, aparecendo simultaneamente em três artigos, Hartigan (1967), Jardine *et al.* (1967) e Johnson (1967).

Intuitivamente a desigualdade ultramétrica implica que dados três objectos quaisquer, ou as três distâncias entre eles são iguais, como acontece num triângulo equilátero, ou, como se verifica mais frequentemente, uma das distâncias é menor do que as outras duas, que são iguais entre si, como acontece num triângulo isósceles.

A desigualdade ultramétrica é mais forte do que a desigualdade triangular (ver Exercício 4.2) e por isso as matrizes de dissemelhança não satisfazem, em geral, a propriedade ultramétrica. Porém, a matriz de distâncias críticas, isto é, a matriz resultante da transformação de uma matriz de dissemelhanças inicial numa estrutura hierárquica, satisfaz a propriedade ultramétrica. De facto, a condição necessária e suficiente para que uma matriz de dissemelhanças possa ser representada exactamente por um dendrograma é que satisfaça a desigualdade ultramétrica. A propriedade ultramétrica garante ainda que à medida que os clusters se vão formando, as distâncias correspondentes vão surgindo de maneira monótona não decrescente.

4.2 Procedimentos aglomerativos

Como já se disse, na construção de uma hierarquia, podem usar-se dois tipos de algoritmos e um primeiro passo é decidir qual o tipo de algoritmo a usar, se aglomerativo, que parte de n grupos singulares e termina num grupo de n objectos, se divisivo que parte de um único grupo e termina em n grupos singulares. A escolha tem caído largamente nos algoritmos aglomerativos, com o argumento de que os algoritmos do tipo divisivo são geralmente muito mais exigentes do ponto de vista computacional. Enquanto que o primeiro passo do processo aglomerativo inclui a construção de dissemelhanças que envolvem $C_2^n = n(n-1)/2$ objectos, o primeiro passo do processo divisivo envolve um mínimo de grupos ou partições igual a $2^n - 1$.

Até ao presente os algoritmos aglomerativos têm sido os mais populares e até os únicos existentes para muitos utilizadores. O procedimento aglomerativo descreve-se em poucos passos.

Passo 1: Considerar os n objectos iniciais como n grupos singulares. A dissemelhança entre os grupos coincide com a matriz de dissemelhanças entre os objectos, $\mathbf{D} = [d_{ij}]$, onde d_{ij} é a dissemelhança entre o objecto i e o objecto j.

Passo 2: Identificar o elemento mais pequeno da matriz $\mathbf{D}$, o que equivale a identificar os dois grupos mais semelhantes, digamos A e B, e a sua dissemelhança, que se representa por d_{AB}.

Passo 3: Unir os grupos A e B à distância crítica d_{AB}. Actualizar a matriz $\mathbf{D}$ eliminando as linhas e as colunas correspondentes aos grupos A e B e introduzindo uma nova linha e coluna com as dissemelhanças calculadas entre o novo grupo (AB) e cada um dos restantes grupos. Com esta operação a ordem da matriz baixa de uma unidade.

Passo 4: Repetir os Passos 2 e 3 num total de $n-1$ vezes até obter um único grupo que, desta maneira, incluirá todos os objectos.

É fácil pôr em operação este algoritmo fazendo uso do computador. Mas antes é indispensável definir a dissemelhança, requerida no Passo 3 do algoritmo, entre dois grupos, um dos quais, pelo menos, pode ter mais do que um objecto.

Há muitas maneiras de definir a dissemelhança entre dois grupos e uma maneira que parece natural é considerar um qualquer tipo de média das dissemelhanças entre os indivíduos ou a distância entre os seus centróides. Cada uma delas está associada a um método hierárquico aglomerativo. De entre os vários métodos possíveis destacam-se aqueles cuja utilização é mais frequente. Considera-se ainda que dados dois grupos genéricos A e B com n_A e n_B objectos, respectivamente, d_{AB} designa o seu nível crítico, ou seja, a distância entre os dois grupos. Definir uma distância equivale a determinar um método hierárquico aglomerativo que lhe fica associado.

4.2.1 Métodos hierárquicos aglomerativos mais comuns e seu funcionamento

• **Ligação simples** (ou método do vizinho mais próximo)

No método da ligação simples a dissemelhança entre os dois grupos é a menor das $n_A n_B$ dissemelhanças entre cada elemento de A e cada elemento de B

$$d_{AB} = \min\left\{d_{ij} : i \in A,\ j \in B\right\}. \tag{4.2}$$

Este é um dos métodos hierárquicos mais simples e ao mesmo tempo muito geral, na medida em que pode detectar grupos com forma muito variada. Começou por se tornar muito popular na área da taxonomia numérica, passando a ser muito conhecido depois da publicação do trabalho de Sokal and Sneath (1963).

É possível identificar várias propriedades do método que é importante conhecer para perceber o seu funcionamento. Embora haja muitos objectos envolvidos, a dissemelhança entre dois grupos é determinada pelos objectos mais próximos (os vizinhos mais próximos), o que significa que uma única ligação é suficiente para juntar os grupos. Deste modo acontece que cada vez que um objecto é adicionado a um grupo, as distâncias do novo grupo aos restantes grupos tornam-se menores ou ficam inalteradas. Há assim uma tendência para que os grupos vão crescendo e se vão juntando para formarem grupos maiores, deixando os objectos isolados firmes na sua posição. Isto pode entender-se como um ponto positivo que joga a favor do método, uma vez que traduz a sua capacidade de detectar *outliers*. O tipo de distância entre dois grupos baseia-se na mais forte medida de proximidade entre os objectos nos dois grupos e isto faz com que o método não

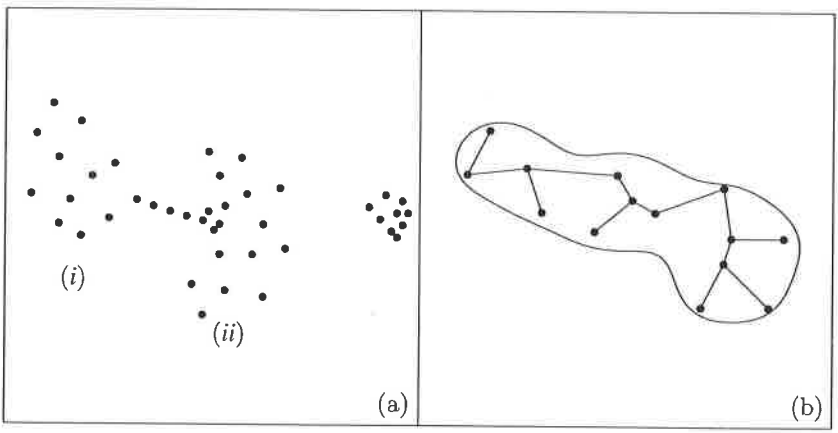


Figura 4.2 *Grupos de ligação simples.*

seja capaz de isolar grupos cuja separação não seja muito nítida. Como se observa na Figura 4.2 (a) os grupos (i) e (ii) não são claramente discerníveis no dendrograma e a sua ligação é feita a um nível crítico. A inclusão de um ou dois objectos entre dois grupos bem definidos pode ser suficiente para levar à junção dos grupos.

Este facto, conhecido como o efeito de cadeia, é característico do método da ligação simples e tende a produzir grupos alongados. Este comportamento pode ser visto como não robusto, no sentido de que a adição de dados, mesmo que em pequena quantidade, pode alterar radicalmente o resultado final. Contudo o efeito de cadeia, geralmente considerado como defeito, pode ser útil em certas áreas. É o caso da taxonomia numérica onde a existência de objectos fazendo a ligação entre dois grupos bem definidos pode trazer a explicação necessária para compreender a verdadeira relação entre os dois grupos, isto é, o mecanismo da cadeia evolutiva. Além disso o efeito de cadeia faz com que o método seja capaz de delinear grupos de forma não elíptica, o que geralmente não acontece com outros métodos, como mostra a Figura 4.2 (b).

Uma propriedade única do método de ligação simples é o que se pode chamar a sua indiferença em relação a casos de empate. Isto é, se houver duas dissemelhanças que sejam iguais e menores que todas as outras, $d_{AB} = d_{CD}$, e tendo que escolher uma das duas para produzir um novo grupo e prosseguir a análise, o resultado final não se altera seja qual for a opção tomada, d_{AB} ou d_{CD} (ver Exercício 4.1).

Pensando no caso de variáveis contínuas este comportamento pode ser

classificado como um comportamento robusto, na medida em que é muito possível que pequenas perturbações das distâncias (pode dizer-se que neste caso praticamente não há empates perfeitos) não alterem o padrão geral da estrutura dos grupos.

Finalmente refere-se que o resultado de uma análise de clusters com base na ligação simples não é alterado por qualquer transformação monótona das distâncias usadas por este método. Esta propriedade é partilhada também pelo método da ligação completa que é introduzido a seguir.

Exemplo 4.1. Para ilustrar o funcionamento do algoritmo aglomerativo aplicado ao método da ligação simples, construiu-se a seguinte matriz de dissemelhanças, relativa a cinco objectos hipotéticos, designados por 1, 2, 3, 4, 5.

$$\mathbf{D} = [d_{ij}] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7 & 0 & & & \\ 4 & 2 & 0 & & \\ 8 & 5 & 8 & 0 & \\ 3 & 10 & 9 & \textcircled{1} & 0 \end{bmatrix} \end{array} \tag{4.3}$$

Passo 1: Considera-se que cada objecto é um grupo singular e procuram-se os dois objectos mais próximos, identificados pelo elemento de menor valor da matriz $\mathbf{D}$, $d_{ij} = \min\{d_{ij} : i, j = 1, \ldots, 5\} = d_{45} = 1$. Os dois objectos, 4 e 5, fundem-se, ao nível crítico $d_{45} = 1$, para formar um novo grupo (45). Passa-se então à construção da nova matriz de dissemelhanças, $\mathbf{D}_1$. Primeiro obtêm-se as dissemelhanças $d_{(45)i}$ entre o grupo (45) e os restantes grupos singulares ou objectos, $i = 1, 2, 3$

$$d_{(45)1} = \min(d_{41}, d_{51}) = \min(8, 3) = 3$$

$$d_{(45)2} = \min(d_{42}, d_{52}) = \min(5, 10) = 5$$

$$d_{(45)3} = \min(d_{43}, d_{53}) = \min(8, 9) = 8$$

A nova matriz, $\mathbf{D}_1$, obtém-se a partir de $\mathbf{D}$ eliminando as linhas e as colunas correspondentes aos objectos 4 e 5 e acrescentando a nova linha e coluna correspondentes ao grupo (45)

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & (45) \end{array} \\ \begin{bmatrix} 0 & & & \\ 7 & 0 & & \\ 4 & \textcircled{2} & 0 & \\ 3 & 5 & 8 & 0 \end{bmatrix} \end{array}$$

Passo 2: O menor elemento da matriz $\mathbf{D}_1$ é $d_{23} = 2$ e por isso os objectos 2 e 3 são aglutinados, ao nível crítico $d_{23} = 2$, para formar o grupo (23). Calculando as distâncias do grupo (23) aos restantes,

$$d_{(23)1} = \min(d_{21}, d_{31}) = \min(7,4) = 4$$

$$d_{(23)(45)} = \min\left(d_{2(45)}, d_{3(45)}\right) = \min(5,8) = 5$$

obtém-se a matriz

$$\mathbf{D}_2 = \begin{array}{c} \\ 1 \\ (23) \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (23) & (45) \end{array} \\ \begin{bmatrix} 0 & & \\ 4 & 0 & \\ \textcircled{3} & 5 & 0 \end{bmatrix} \end{array}$$

Passo 3: O menor elemento de $\mathbf{D}_2$ é definido pela distância entre o objecto 1 e o grupo (45), ao nível crítico $d_{1(45)} = 3$, para formar o grupo (145). Uma vez que a distância entre os dois únicos grupos é $d_{(145)(23)} = \min\left(d_{1(23)}, d_{(45)(23)}\right) = \min(4,5) = 4$, tem-se a nova matriz

$$\mathbf{D}_3 = \begin{array}{c} \\ (23) \\ (145) \end{array} \begin{array}{c} \begin{array}{cc} (23) & (145) \end{array} \\ \begin{bmatrix} 0 & \\ \textcircled{4} & 0 \end{bmatrix} \end{array}$$

Passo 4: A única possibilidade que resta é aglutinar os dois grupos, ao nível crítico igual a 4, para formar um só grupo contendo os elementos (12345).

A sequência dos vários passos e o processo de aglutinação com os níveis críticos está descrita no dendrograma da Figura 4.3, produzido pela função agnes implementada no *software* S-PLUS 2000. *D* representa a distância entre grupos. O resultado gráfico fornecido por este *software* inclui não só o dendrograma como também o gráfico em bandeira que se mostra na Figura 4.4. Em muitas situações este gráfico faz lembrar uma bandeira com o mastro à direita, motivo porque é assim chamado. O gráfico contém exactamente a mesma informação do dendrograma e os seus defensores, Kaufman and Rousseeuw (1994), clamam que a sua interpretação é simples e intuitiva, permitindo analisar com facilidade a estrutura dos dados, tanto globalmente como a qualquer nível de dissemelhança que se considere. Contudo o gráfico está longe de atingir a popularidade do dendrograma. O gráfico é lido da esquerda para a direita da mesma forma que o dendrograma é lido de baixo para cima.

Na Figura 4.4 vêm-se claramente os vários níveis de fusão e pode obter-se, como no dendrograma, a estrutura de grupo, isto é, o número de grupos, a um nível qualquer. Cortando o gráfico por uma linha vertical aos níveis

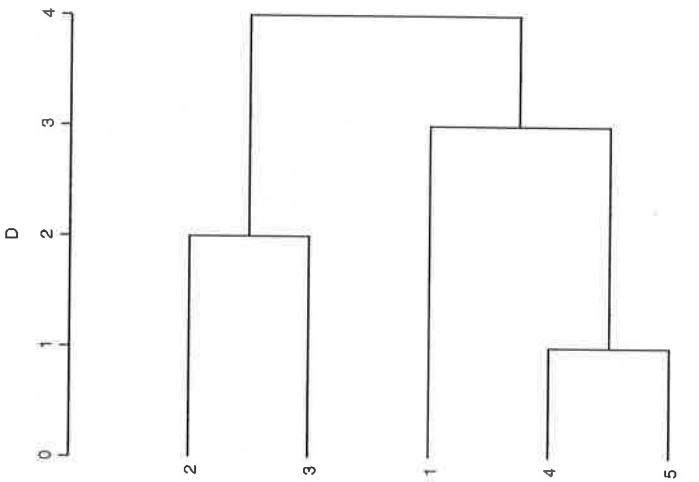


Figura 4.3 *Dendrograma para cinco objectos (método da ligação simples).*

3.5, 2.5 e 1.5 obtêm-se, respectivamente, dois, três e quatro grupos. A parte escura é usada para dar uma ideia do nível de estrutura encontrado nos dados pelo algoritmo. Quando há estrutura natural nos dados as dissemelhanças entre clusters tornam-se maiores do que dentro dos clusters e as faixas tornam-se mais longas. Uma medida da magnitude da estrutura existente é dada pelo coeficiente aglomerativo que é também fornecido com o gráfico.

Para cada objecto i define-se $m(i)$ como a dissemelhança entre i e o primeiro cluster com o qual i é aglutinado, dividida pelo maior nível de fusão. O coeficiente aglomerativo, AC, é dado pela média de $1-m(i)$, $i=1,\ldots,n$, ou seja,

$$AC = \frac{\sum_{i=1}^{m}(1-m(i))}{n}. \tag{4.4}$$

Os dois casos extremos acontecem quando:

- $AC = 1$, isto é, $m(i) = 0$, $i = 1,\ldots,n$. Neste caso os grupos, contendo

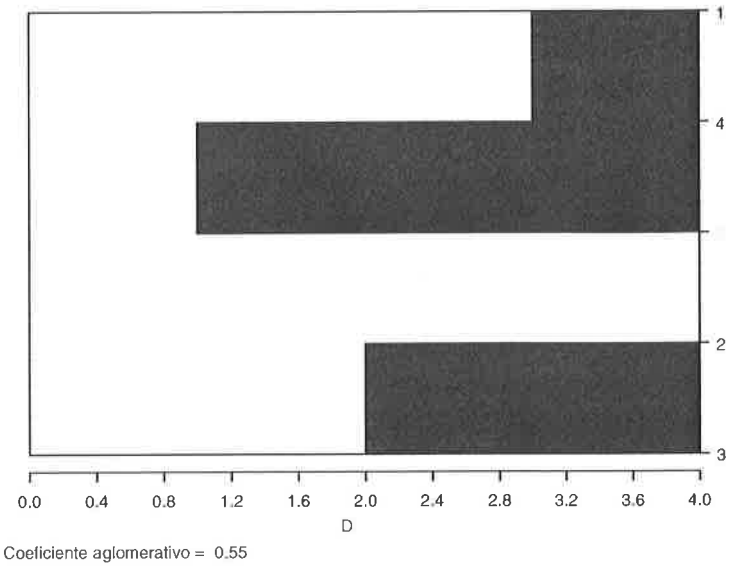


Figura 4.4 *Gráfico em bandeira para cinco objectos (método da ligação simples).*

objectos coincidentes, estão bem separados. A zona escura atinge o seu máximo possível. A Figura 4.5 ilustra uma destas situações em que há dois grupos, um com três objectos representados por três pontos coincidentes e outro, à distância dez do primeiro, formado por dois pontos coincidentes.

- $AC = 0$, isto é, $m(i) = 1$, $i = 1, \ldots, n$, os objectos formam um único grupo e não há faixa escura no gráfico. Basta pensar nos cinco pontos e admitir que a matriz de dissemelhanças que lhe está associada tem todos os elementos fora da diagonal principal iguais a dez, por exemplo. Os objectos mantêm-se separados até ficarem ligados ao nível 10.

Esquecendo estes dois casos extremos que raramente surgem na prática, o que se pode dizer é que AC próximo de 1 é indicação da existência de uma estrutura natural nos dados. Se AC é próximo de zero pode concluir-se que o algoritmo não encontrou uma estrutura natural, isto é, os dados formam apenas um único e grande grupo. AC é por assim dizer uma medida

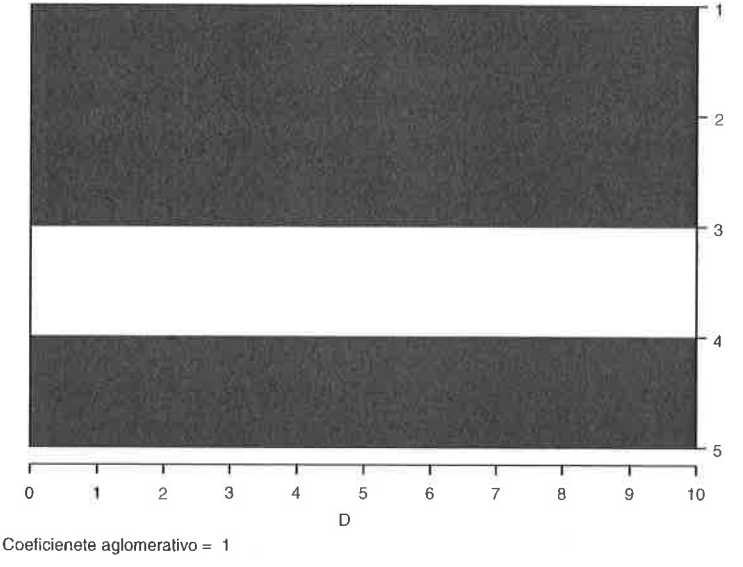


Figura 4.5 *Gráfico em bandeira para dois grupos isolados,* $AC = 1$.

global da dissemelhança dos dados. Note-se que AC tende a aumentar com o número de objectos, o que desaconselha a sua utilização para comparar estruturas de dados com tamanhos muito diferentes.

Também se verifica que o coeficiente aglomerativo geralmente aumenta quando um *outlier* é adicionado ao conjunto de dados, o que leva a que o utilizador não deva embarcar em conclusões precipitadas perante um AC grande. Felizmente o gráfico tem a grande vantagem de mostrar bem os *outliers*, concluindo-se que uma análise correcta é aquela que tem em conta o valor de AC e o exame do gráfico em bandeira.

Na Tabela 4.1 mostram-se os cálculos que permitem confirmar o valor de $AC = 0.55$ para a estrutura produzida pelo método da ligação simples aplicado aos cinco objectos.

Exemplo 4.2. A análise que se segue envolve um conjunto de dados reais e ilustra o uso do método da ligação simples na prática.

Os dados referem-se ao consumo de carne nos países da União Europeia

Tabela 4.1 *Cálculo de AC para os cinco objectos.*

Objectos	Dissemelhanças	$1-m(i)$
1	3	$1-\frac{3}{4}$
2	2	$1-\frac{2}{4}$
3	2	$1-\frac{2}{4}$
4	1	$1-\frac{1}{4}$
5	1	$1-\frac{1}{4}$

$$AC = \frac{11/4}{5} = 0.55$$

Tabela 4.2 *Consumo de carne nos países da UE em 2001.*

	vaca	porco	carneiro	aves	outra
Áustria	18	56	1	18	1
Bélg.+Lux.	20	46	2	18	4
Dinamarca	22	63	1	21	1
Finlândia	12	32	0	15	3
França	25	37	4	26	6
Alemanha	10	54	1	19	2
Grécia	19	32	13	20	1
Holanda	19	43	1	22	0
Irlanda	17	39	5	31	2
Itália	23	38	2	18	5
Portugal	15	44	3	32	3
Espanha	13	66	6	27	3
Suécia	21	35	1	13	3
Reino Unido	19	25	6	29	0

em 2001. Os dados, fornecidos pelo Eurostat, encontram-se na Tabela 4.2, onde se nota que dois países, Bélgica e Luxemburgo, foram aglutinados. Portanto em vez dos quinze países que formavam a UE em 2001 trabalha-se apenas com catorze objectos.

A partir da matriz de dados pode obter-se a matriz de distâncias entre os objectos usando algumas das várias definições propostas no Capítulo 2. Calculando a distância euclidiana entre elementos de cada par de objectos

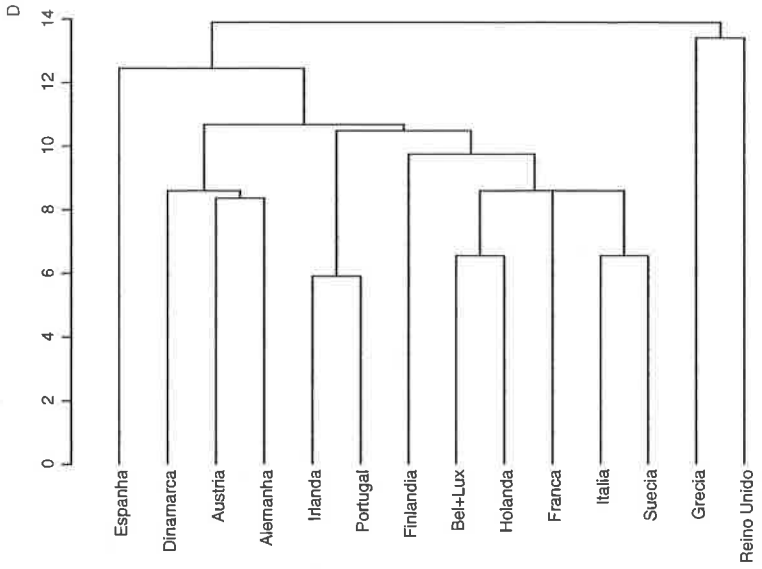


Figura 4.6 *Dendrograma para países da UE (método da ligação simples).*

obtém-se a matriz da Tabela 4.3.

O método manual seguido no Exemplo 4.1 é agora impraticável, visto que o número inicial de distâncias a analisar é de 91, contra 10 distâncias do Exemplo 4.1. É por isso inevitável o recurso ao uso do computador. Usando novamente a função **agnes** produziu-se o dendrograma da Figura 4.6 e o gráfico em bandeira que se apresenta na Figura 4.7.

• **Ligação completa** (ou método do vizinho mais afastado)

Este método é muito semelhante ao método da ligação simples, excepto que a dissemelhança entre dois grupos é a maior das $n_A n_B$ dissemelhanças entre cada elemento de A e cada elemento de B

$$d_{AB} = \max\left\{d_{ij} : i \in A,\ j \in B\right\}. \tag{4.5}$$

O método funciona ao contrário da ligação simples pois serve-se dos dois elementos mais afastados, mais diferentes um em cada grupo, para derivar a

$$\mathbf{D} = \begin{bmatrix}
0.00 & 10.67 & 8.60 & 25.01 & 22.53 & 8.36 & 26.92 & 13.67 & 21.81 & 19.13 & 18.89 & 15.32 & 21.88 & 33.30 \\
10.67 & 0.00 & 17.66 & 0.00 & 13.34 & 13.03 & 18.19 & 6.55 & 15.49 & 8.60 & 15.06 & 23.38 & 12.20 & 24.39 \\
8.60 & 17.66 & 0.00 & 33.19 & 27.27 & 15.16 & 33.39 & 20.27 & 26.79 & 25.53 & 23.21 & 12.44 & 29.20 & 39.28 \\
25.01 & 16.55 & 33.19 & 0.00 & 18.43 & 22.49 & 15.71 & 15.13 & 18.86 & 13.19 & 21.23 & 36.56 & 9.74 & 18.41 \\
22.53 & 13.34 & 27.27 & 18.43 & 0.00 & 24.24 & 14.24 & 11.53 & 10.48 & 8.60 & 13.96 & 31.60 & 14.38 & 15.13 \\
8.36 & 13.03 & 15.16 & 22.49 & 24.24 & 0.00 & 26.66 & 14.66 & 20.83 & 20.88 & 17.29 & 15.58 & 22.78 & 32.41 \\
26.92 & 18.19 & 33.39 & 15.71 & 14.24 & 26.66 & 0.00 & 16.43 & 15.45 & 13.89 & 20.19 & 35.97 & 14.49 & 13.41 \\
13.67 & 6.55 & 20.27 & 15.13 & 11.53 & 14.66 & 16.43 & 0.00 & 11.00 & 9.11 & 11.40 & 24.97 & 12.56 & 19.94 \\
21.81 & 15.49 & 26.79 & 18.86 & 10.48 & 20.83 & 15.45 & 11.00 & 0.00 & 14.96 & 5.91 & 27.62 & 19.31 & 14.45 \\
19.13 & 8.60 & 25.53 & 13.19 & 8.60 & 20.88 & 13.89 & 9.11 & 14.96 & 0.00 & 17.34 & 31.38 & 6.55 & 18.62 \\
18.89 & 15.06 & 23.21 & 21.23 & 13.96 & 17.29 & 20.19 & 11.40 & 5.91 & 17.34 & 0.00 & 22.84 & 21.95 & 20.09 \\
15.32 & 23.38 & 12.44 & 36.56 & 31.60 & 15.58 & 35.97 & 24.97 & 27.62 & 31.38 & 22.84 & 0.00 & 35.29 & 41.59 \\
21.88 & 12.20 & 29.20 & 9.74 & 14.38 & 22.78 & 14.49 & 12.56 & 19.31 & 6.55 & 21.95 & 35.29 & 0.00 & 19.84 \\
33.30 & 24.39 & 39.28 & 18.41 & 15.13 & 32.41 & 13.41 & 19.94 & 14.45 & 18.62 & 20.09 & 41.59 & 19.84 & 0.00
\end{bmatrix}$$

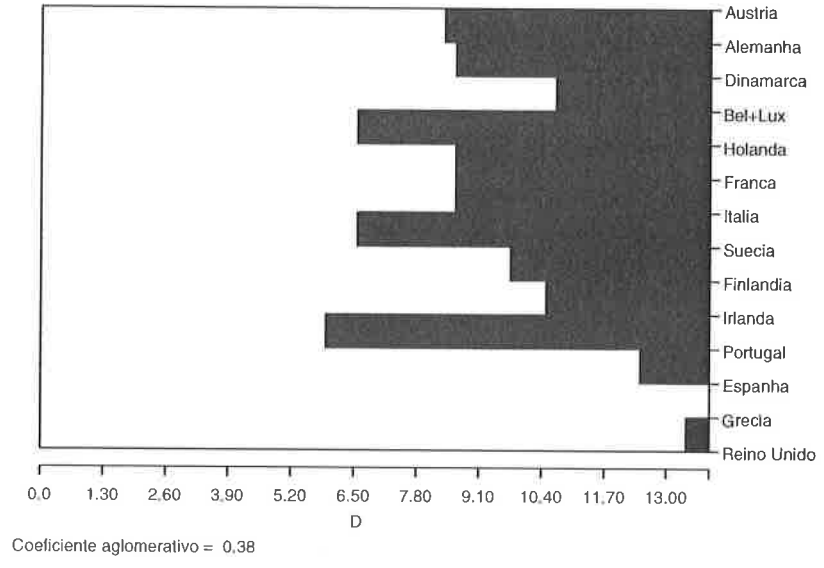


Figura 4.7 *Gráfico em bandeira para países da UE (método da ligação simples).*

medida de proximidade entre os grupos. Quando um objecto é acrescentado a um grupo a distância do novo grupo aos restantes aumenta ou então fica inalterada e por isso há tendência para grupos grandes não crescerem mais. Em dados não estruturados o método da ligação completa tende a formar grupos pequenos que depois são aglutinados para formar grupos maiores. Se os dados apresentam grupos naturais mas de tamanho diferente o método da ligação completa tenderá em primeiro lugar a reunir os grupos mais pequenos e só depois cuidará dos grandes grupos.

Se em vez de detectar clusters naturais o objectivo da análise é dividir um conjunto de dados não estruturados, em grupos convenientes, o método da ligação completa produz geralmente grupos de tamanho razoavelmente equilibrado e mostra-se mais vantajoso do que o método da ligação simples que não costuma conduzir a resultados com interesse.

Exemplo 4.3. Para ilustrar o método da ligação completa usa-se a matriz de dissemelhanças para cinco objectos (4.3), o que vai permitir a comparação dos resultados aqui obtidos com os resultados do Exemplo 4.1. Segue-

se a lista dos passos percorridos pelo algoritmo, tendo-se resumido a sua descrição. Observa-se em primeiro lugar a matriz

$$\mathbf{D} = [d_{ij}] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7 & 0 & & & \\ 4 & 2 & 0 & & \\ 8 & 5 & 8 & 0 & \\ 3 & 10 & 9 & \textcircled{1} & 0 \end{bmatrix} \end{array}$$

Passo 1:

- menor distância em **D**: 1
- aglutinar os objectos 4 e 5 para formar o grupo (45), ao nível de fusão 1
- distâncias ao grupo (45)

$$d_{(45)1} = \max(d_{41}, d_{51}) = \max(8,3) = 8$$

$$d_{(45)2} = \max(d_{42}, d_{52}) = \max(5,10) = 10$$

$$d_{(45)3} = \max(d_{43}, d_{53}) = \max(8,9) = 9$$

- matriz actualizada

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & (45) \end{array} \\ \begin{bmatrix} 0 & & & \\ 7 & 0 & & \\ 4 & \textcircled{2} & 0 & \\ 8 & 10 & 9 & 0 \end{bmatrix} \end{array}$$

Passo 2:

- menor distância em $\mathbf{D}_1$: 2
- aglutinar os objectos 2 e 3 para formar o grupo (23), ao nível de fusão 2
- distâncias ao grupo (23)

$$d_{(23)1} = \max(d_{21}, d_{31}) = \max(7,4) = 7$$

$$d_{(23)(45)} = \max\left(d_{2(45)}, d_{3(45)}\right) = \max(10,9) = 10$$

- matriz actualizada

$$\mathbf{D}_2 = \begin{array}{c} \\ 1 \\ (23) \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (23) & (45) \end{array} \\ \begin{bmatrix} 0 & & \\ \textcircled{7} & 0 & \\ 8 & 10 & 0 \end{bmatrix} \end{array}$$

Passo 3:

- menor distância em $\mathbf{D}_2$: 7
- aglutinar (23) e 1 para formar o grupo (123), ao nível de fusão 7
- distâncias ao grupo (123)

$$d_{(123)(45)} = \max\left(d_{1(45)}, d_{(23)(45)}\right) = \max(8, 10) = 10$$

- matriz actualizada

$$\mathbf{D}_3 = \begin{array}{c} \\ (123) \\ (45) \end{array} \begin{array}{c} \begin{array}{cc} (123) & (45) \end{array} \\ \begin{bmatrix} 0 & \\ 10 & 0 \end{bmatrix} \end{array}$$

Passo 4: Aglutinar os grupos (123) e (45) para formar um só grupo final (12345), sendo o nível de fusão igual a 10, única distância não nula na matriz $\mathbf{D}_3$.

A Figura 4.8 mostra o dendrograma produzido pelo método da ligação completa. Comparando os passos seguidos neste exemplo com os passos do Exemplo 4.1 pode apreciar-se o funcionamento dos dois métodos. Comparem-se, em particular, as sucessivas matrizes e níveis de fusão. Observando as Figuras 4.8 e 4.3 verifica-se que os dois dendrogramas diferem apenas porque a fusão do objecto 1 foi feita com (45) no método da ligação simples e com (23) no caso do método da ligação completa.

- **ligação média** (ou UPGMA, *unweighted pair-group method using the average approach*)

Neste método a dissemelhança entre dois grupos é a média das dissemelhanças entre todos os pares de objectos, formados com um objecto de cada grupo,

$$d_{AB} = \frac{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij}}{n_A n_B}. \tag{4.6}$$

Trata-se de um compromisso entre as duas situações extremas, traduzidas pelos métodos da ligação simples e ligação completa. Esta ideia de dissemelhança entre grupos é muito natural e por isso não admira que este tenha sido um dos primeiros métodos hierárquicos a ser construído. O método é considerado adequado para isolar grupos de forma arredondada, esférica ou mesmo elipsoidal, uma vez que é relativamente robusto. É recomendado

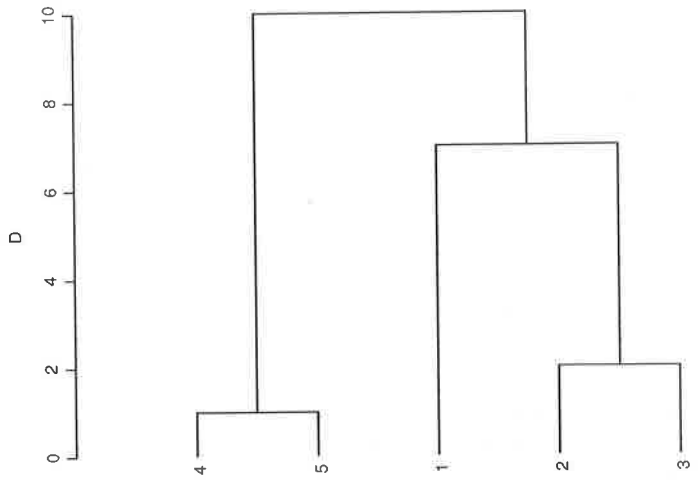


Figura 4.8 *Dendrograma para cinco objectos (método da ligação completa).*

por muitos autores que o consideram superior aos métodos da ligação simples e da ligação completa, o que vai de encontro às conclusões do estudo comparativo realizado por Cunningham and Ogilvie (1972), onde o método é vencedor no confronto com outros sete métodos hierárquicos.

Exemplo 4.4. Partindo da matriz inicial

$$\mathbf{D} = [d_{ij}] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7 & 0 & & & \\ 4 & 2 & 0 & & \\ 8 & 5 & 8 & 0 & \\ 3 & 10 & 9 & \textcircled{1} & 0 \end{bmatrix} \end{array}$$

percorrem-se os diversos passos.

Passo 1:

- menor distância em **D**: 1
- aglutinar os objectos 4 e 5 para formar o grupo (45), ao nível de fusão 1

- distâncias ao grupo (45)

$$d_{(45)1} = \frac{d_{41} + d_{51}}{2 \times 1} = \frac{8+3}{2} = 5.5$$

$$d_{(45)2} = \frac{d_{42} + d_{52}}{2 \times 1} = \frac{5+10}{2} = 7.5$$

$$d_{(45)3} = \frac{d_{43} + d_{53}}{2 \times 1} = \frac{8+9}{2} = 8.5$$

- matriz actualizada

$$\mathbf{D}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & (45) \end{array} \\ \begin{bmatrix} 0 & & & \\ 7 & 0 & & \\ 4 & \textcircled{2} & 0 & \\ 5.5 & 7.5 & 8.5 & 0 \end{bmatrix} \end{array}$$

Passo 2:

- menor distância em $\mathbf{D}_1$: 2
- aglutinar os objectos 2 e 3 para formar o grupo (23), ao nível de fusão 2
- distâncias ao grupo (23)

$$d_{(23)1} = \frac{d_{21} + d_{31}}{2 \times 1} = \frac{7+4}{2} = 5.5$$

$$d_{(23)(45)} = \frac{(d_{24} + d_{25}) + (d_{34} + d_{35})}{2 \times 2} = \frac{(5+10)+(8+9)}{4} = 8$$

- matriz actualizada

$$\mathbf{D}_2 = \begin{array}{c} \\ 1 \\ (23) \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (23) & (45) \end{array} \\ \begin{bmatrix} 0 & & \\ \textcircled{5.5} & 0 & \\ \textcircled{5.5} & 8 & 0 \end{bmatrix} \end{array}$$

Passo 3:

- menor distância em $\mathbf{D}_2$: 5.5
- aqui estamos numa situação de empate, pois há duas dissemelhanças iguais e podemos optar por uma das duas possibilidades que se oferecem

(i) se optarmos por aglutinar (45) e 1 para formar o grupo (145) as distâncias ao grupo (145) resumem-se a

$$d_{(145)(23)} = \frac{(d_{12}+d_{13})+(d_{42}+d_{43})+(d_{52}+d_{53})}{3\times 2} = \\ = \frac{(7+4)+(5+8)+(10+9)}{6} = 7.17$$

e a matriz actualizada é

$$\mathbf{D}_3 = \begin{matrix} \\ (23) \\ (145) \end{matrix} \begin{matrix} \begin{matrix} (23) & (145) \end{matrix} \\ \begin{bmatrix} 0 & \\ 7.17 & 0 \end{bmatrix} \end{matrix}$$

que conduz a um dendrograma idêntico ao do método da ligação simples, excepto no que diz respeito aos níveis de fusão que são diferentes.

(ii) se optarmos por aglutinar (23) a 1 para formar o grupo (123) tem-se

$$d_{(123)(45)} = \frac{(d_{14}+d_{15})+(d_{24}+d_{25})+(d_{34}+d_{35})}{3\times 2} = \\ = \frac{(8+3)+(5+10)+(8+9)}{6} = 7.17$$

e a matriz actualizada é

$$\mathbf{D}_3 = \begin{matrix} \\ (123) \\ (45) \end{matrix} \begin{matrix} \begin{matrix} (123) & (45) \end{matrix} \\ \begin{bmatrix} 0 & \\ 7.17 & 0 \end{bmatrix} \end{matrix}$$

que, por sua vez, é a solução do método da ligação completa, excepto no que diz respeito aos níveis de fusão que são diferentes.

Passo 4: Seguindo (i) ou (ii) obtém-se o grupo final (12345).

Como se verifica, no caso de empate o resultado do método não é indiferente à escolha da dissemelhança, contrariamente ao que acontece com o método da ligação simples (ver Exercício 4.1). O dendrograma da Figura 4.9 refere-se à opção (i) e por isso coincide com o dendrograma da Figura 4.3. Em geral o *software* existente para análise de clusters não produz soluções múltiplas, possivelmente porque em situações reais não é frequente a existência de empates.

Os três métodos que acabam de ser descritos e ilustrados usam uma matriz de dissemelhanças como ponto de partida mas o critério de distância

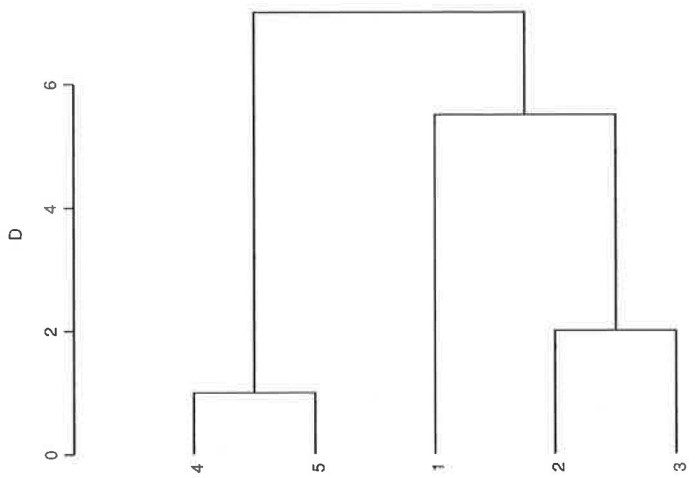


Figura 4.9 *Dendrograma para cinco objectos (método da ligação média).*

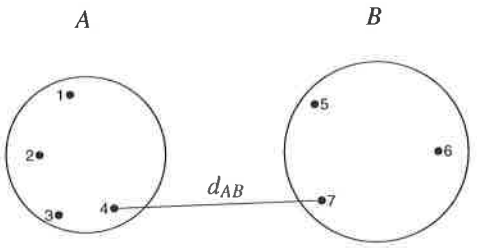
entre os grupos é diferente de método para método. A Figura 4.10 permite apreciar graficamente como é que essas distâncias são construídas e compará-las entre si.

Os três métodos hierárquicos já apresentados partem de uma matriz de proximidades. Para operar com estes métodos basta usar procedimentos aritméticos simples. Os métodos que a seguir se introduzem são mais complicados do ponto de vista algébrico e operam sobre a matriz de dados, $\mathbf{X}_{n\times p}$, dos n objectos observados em p variáveis, que nos métodos anteriores pode ser eliminada ou pode mesmo não existir.

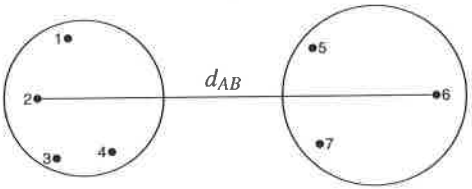
- **centróide** (ou UPGMC, *unweighted pair-group method using the centroid approach*)

Um procedimento muito comum em estatística é o de representar as populações pelas suas médias ou centróides e comparar as populações comparando os seus representantes, ou seja, as suas médias. O método do centróide parece inspirar-se neste facto, uma vez que, de acordo com este método, a distância entre dois grupos, A e B, é a distância entre os seus centróides, isto é

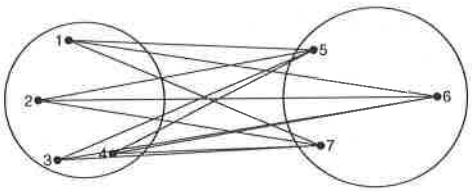
$$d_{AB} = d(\overline{\mathbf{x}}_A, \overline{\mathbf{x}}_B), \tag{4.7}$$



(*a*) Ligação simples



(*b*) Ligação completa



(*c*) Ligação média

$$d_{AB} = \frac{1}{12}\left[(d_{15}+d_{16}+d_{17})+(d_{25}+d_{26}+d_{27})+(d_{35}+d_{36}+d_{37})+(d_{45}+d_{46}+d_{47})\right]$$

Figura 4.10 *Algumas dissemelhanças entre grupos.*

onde $\overline{\mathbf{x}}_A$ e $\overline{\mathbf{x}}_B$ são os centróides dos grupos A e B, respectivamente, isto é,

$$\overline{\mathbf{x}}_A = \frac{\sum_{i\in A}\mathbf{x}_i}{n_A} \quad \text{e} \quad \overline{\mathbf{x}}_B = \frac{\sum_{i\in B}\mathbf{x}_i}{n_B}$$

e $\mathbf{x}_i$ é o vector das p observações do objecto i.

Em cada passo do algoritmo os grupos a aglutinar são aqueles cujos centróides estão mais próximos de acordo com a distância entre centróides que foi definida. Um inconveniente do método é o facto da distância de

fusão de dois grupos poder aumentar e diminuir de passo para passo, o que torna a interpretação difícil. A distância entre clusters pode ser qualquer medida de proximidade, como por exemplo o coeficiente de correlação ou a distância euclidiana, mas o quadrado da distância euclidiana é a medida com mais sucesso em termos de facilidade de aplicação e da clareza dos resultados que produz.

- **mediana** (ou WPGMC, *weighted pair-group method using the centroid approach*)

O critério da mediana é semelhante ao do centróide excepto que ao aglutinar dois grupos A e B os seus centróides, $\bar{\mathbf{x}}_A$ e $\bar{\mathbf{x}}_B$, recebem pesos iguais antes de produzirem o centróide do novo cluster resultante da aglutinação. O novo centróide, $\bar{\mathbf{x}}$, fica a meio dos centróides dos grupos aglutinados, $\bar{\mathbf{x}} = (\bar{\mathbf{x}}_A + \bar{\mathbf{x}}_B)/2$. A ideia é evitar que o grupo com maior número de objectos absorva o grupo com menor número de objectos, acabando este por perder a sua identidade, o que pode distorcer o processo. Isto é particularmente grave se de facto se tratar de grupos genuínos e os números de objectos observados em cada grupo não forem necessariamente representativos das respectivas frequências relativas. A mediana aqui referida não é a mediana estatística mas sim a mediana de um triângulo, isto é, o segmento de recta que liga um vértice do triângulo ao ponto médio do lado oposto (ver Exercício 4.5).

- **Ward**

De acordo com o método de Ward (Ward, 1963) o critério de fusão de dois grupos A e B é baseado no incremento da soma dos quadrados que ocorre quando os clusters A e B são aglutinados. Esse incremento é

$$SSW_c - (SSW_A + SSW_B),$$

onde

$$SSW_A = \sum_{i \in A} \sum_{j=1}^{p} (x_{ijA} - \bar{x}_{jA})^2$$

é a soma dos quadrados dentro do grupo A,

$$SSW_B = \sum_{i \in B} \sum_{j=1}^{p} (x_{ijB} - \bar{x}_{jB})^2$$

é a soma dos quadrados dentro do grupo B e

$$SSW_C = \sum_{i \in C} \sum_{j=1}^{p} (x_{ijC} - \bar{x}_{jC})^2$$

é a soma dos quadrados dentro do grupo $C = A \cup B$, resultante da aglutinação do grupo A com o grupo B. x_{ijA} (x_{ijB}) é a observação do objecto i do grupo A (B) na variável j, $\bar{x}_{jA}$ e $\bar{x}_{jB}$ são as médias da variável j nos grupos A e B, respectivamente.

O incremento da soma dos quadrados corresponde efectivamente a uma perda de informação. Em cada passo do algoritmo são formados todos os pares possíveis de clusters e calculado o incremento da soma de quadrados resultante da reunião dos clusters de cada par. Os clusters seleccionados para formar um novo cluster são aqueles a que corresponde o menor incremento, ou seja, a menor perda de informação resultante da aglutinação.

Os métodos centróide, mediana e Ward foram aplicados à matriz de dados a que corresponde a matriz de dissemelhanças usada para ilustrar as ligações simples, completa e média. Assim é possível colocar em confronto o resultado destes seis métodos hierárquicos aplicados aos mesmos dados. A Figura 4.11 mostra os seis dendrogramas. Sem esforço é possível observar a existência de três grupos distintos nos dados. Todos os algoritmos conduziram ao mesmo número de clusters e com a mesma composição. A única diferença a salientar é que o nível de fusão dos clusters varia com o algoritmo. A este respeito os pares (centróide e mediana) e (ligação completa e Ward) são os que mais se assemelham. Pode concluir-se que os dados possuem uma estrutura de grupos que é unanimemente revelada pela análise de clusters com base nos seis métodos hierárquicos utilizados.

4.2.2 Fórmula de recorrência de Lance-Williams

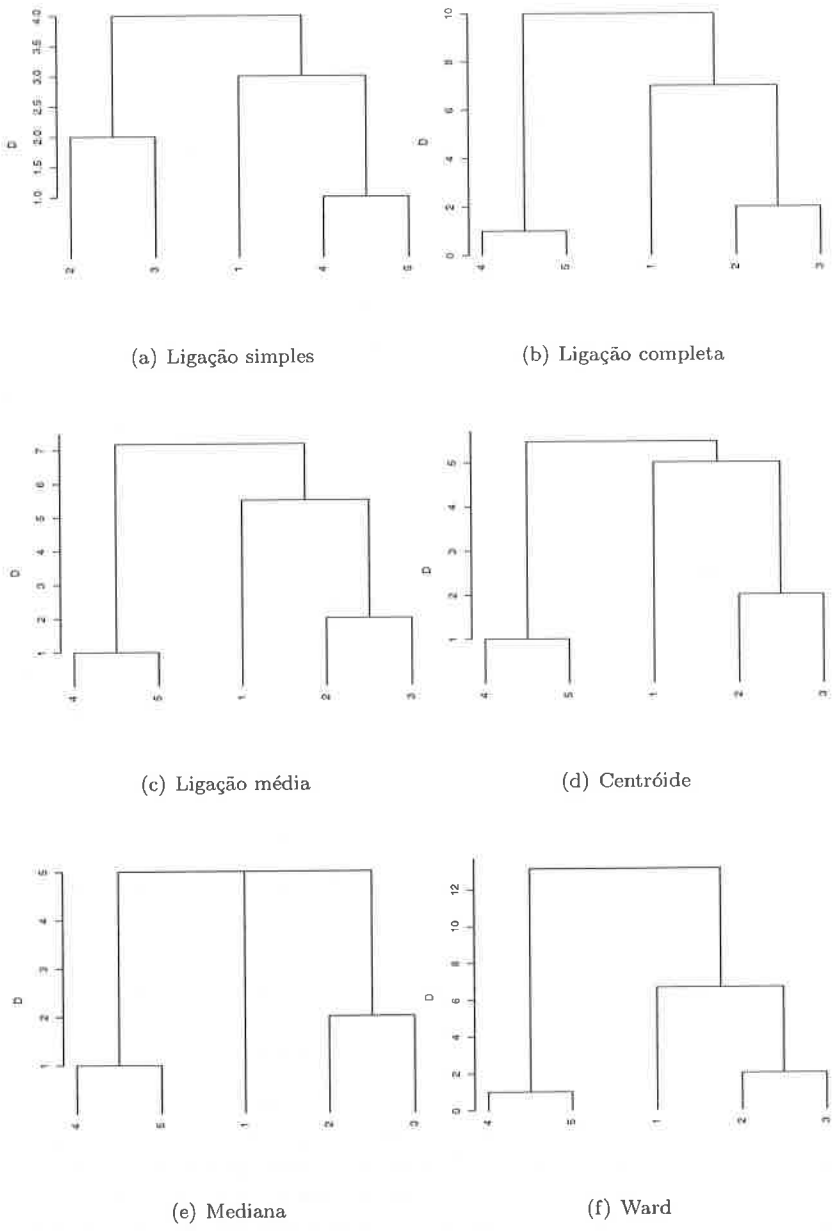
A definição e cálculo de dissemelhanças entre grupos é fundamental em cada um dos métodos hierárquicos estudados anteriormente. Lance and Williams (1967) introduziram uma regra geral para definir essa dissemelhança baseada numa fórmula de recorrência. A fórmula dá a dissemelhança, $d_{C(AB)}$, entre o grupo C e o grupo (AB) formado pela fusão dos grupos A e B,

$$d_{C(AB)} = \alpha_A d_{CA} + \alpha_B d_{CB} + \beta d_{AB} + \gamma |d_{CA} - d_{CB}| , \qquad (4.8)$$

onde d_{IJ} é a dissemelhança entre os grupos I e J. A fórmula depende de quatro parâmetros que ou são constantes ou dependem do número de objectos em cada grupo, n_A, n_B e n_C. Jambu (1978) generalizou a fórmula de Lance-Williams introduzindo mais três parâmetros.

Ao fixar os parâmetros em valores particulares as fórmulas produzem, entre outros, os vários métodos hierárquicos já estudados em 4.2.1.

Lance e Williams analisaram os resultados correspondentes a possíveis



(a) Ligação simples

(b) Ligação completa

(c) Ligação média

(d) Centróide

(e) Mediana

(f) Ward

Figura 4.11 *Seis dendrogramas para os mesmos cinco objectos.*

Tabela 4.4 *Parâmetros de Lance-Williams para vários métodos hierárquicos aglomerativos.*

Método	Parâmetros de Lance-Williams			
	α_A	α_B	β	γ
Ligação simples	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Ligação completa	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Ligação média	$\frac{n_A}{n_A+n_B}$	$\frac{n_B}{n_A+n_B}$	0	0
Centróide	$\frac{n_A}{n_A+n_B}$	$\frac{n_B}{n_A+n_B}$	$-\frac{n_A n_B}{(n_A+n_B)^2}$	0
Mediana	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_C+n_A}{n_C+n_A+n_B}$	$\frac{n_C+n_B}{n_C+n_A+n_B}$	$-\frac{n_C}{n_C+n_A+n_B}$	0
Lance-Williams	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	< 1	0

valores dos parâmetros e concluiram que uma escolha adequada para um grande número de situações é $\alpha_A = \alpha_B$, $\alpha_A + \alpha_B + \beta = 1$, $\beta < 1$ e $\gamma = 0$.

Esta estratégia flexível produz métodos com diferentes características, nomeadamente com diferentes intensidades no que respeita ao efeito de cadeia. Quando $\beta = 1$ tem-se o método de ligação simples onde o efeito de cadeia tem intensidade máxima. À medida que β decresce para zero e se torna negativo, o efeito de cadeia torna-se menor e os grupos produzidos ficam mais homogéneos. Os valores que têm sido sugeridos para β são valores negativos pequenos tais como -0.25 e -0.50.

A Tabela 4.4 mostra os valores particulares dos quatro parâmetros assim como os métodos correspondentes descritos pela fórmula.

Fazendo, por exemplo, $\alpha_A = \frac{1}{2}$, $\alpha_B = \frac{1}{2}$, $\beta = 0$ e $\gamma = -\frac{1}{2}$, a fórmula dá o método da ligação simples, uma vez que estas condições equivalem a (ver Exercício 4.6)

$$d_{C(AB)} = \min\{d_{ij} : i \in C,\ j \in (AB)\}\ .$$

A grande atracção da fórmula de recorrência é de natureza computacional uma vez que dadas as dissemelhanças iniciais entre objectos, e escolhido o método hierárquico, o processameento do algoritmo prossegue automaticamente, com sucessivas actualizações da matriz de dissemelhanças, sem haver necessidade de reter a informação inicial relativa aos objectos, depois de estes terem sido incorporados em grupos maiores. Mas, pelo facto de haver assim acesso a muitos métodos e soluções, a disponibilidade da fór-

mula pode trazer desvantagens, por poder conduzir a situações de indecisão na escolha da solução final. Há que reflectir bem e decidir sobre o método a usar antes de efectuar a análise.

4.3 Procedimentos divisivos ou de desagregação

Estes procedimentos actuam no sentido contrário ao dos procedimentos aglomerativos, movendo-se da raiz do dendrograma para os ramos em vez de se moverem dos ramos para a raiz como no caso aglomerativo. O processo começa com um único grupo contendo todos os objectos. Esse grupo é dividido em dois grupos distintos com base em algum critério de dissemelhança. O próximo passo consiste em repetir o processo anterior em cada um dos dois novos clusters e continuar até obter n clusters, sendo n o número total de objectos.

Os métodos divisivos são muito exigentes em termos computacionais. De facto em cada passo é preciso construir $2^k - 1$ dissemelhanças, correspondentes à divisão dos k objectos em dois grupos distintos. É claro que no primeiro passo o número de dissemelhanças é $2^n - 1$.

Apesar disso os procedimentos divisivos, se computacionalmente viáveis, podem ter vantagens sobre os algoritmos aglomerativos, uma vez que podem fornecer grandes grupos ao fim dos primeiros passos do processo e os grandes grupos são o que geralmente interessa ao investigador, em vez de uma lista longa de pequenos grupos.

A função `diana` do S-PLUS faz análise de clusters usando procedimentos divisivos.

O uso dos procedimentos divisivos é menos comum do que o uso dos procedimentos aglomerativos e por isso o seu estudo fica aqui limitado a estes breves comentários.

Exercícios

4.1 Considere a situação em que há duas dissemelhanças iguais e menores do que todas as outras, por exemplo entre os clusters A, B e C, verifica-se que $d_{AB} = d_{AC}$.

(a) Mostre que para o método da ligação simples é indiferente unir A com B ou unir A com C, pois qualquer que seja a opção o resultado final não se altera.

(b) Diga, justificando, se acha que este resultado ainda é válido para os outros métodos hierárquicos.

4.2 Mostre que a desigualdade ultramétrica implica a desigualdade triangular mas que a implicação inversa não é verdadeira.

4.3 (a) Construa a matriz das distâncias críticas d^*_{ij} entre os objectos associadas ao dendrograma da Figura 4.1 e verifique que essas distâncias satisfazem a desigualdade ultramétrica, escolhendo ao acaso três dos objectos em estudo.

(b) Mostre que a distância crítica associada a um qualquer dendrograma satisfaz de facto a desigualdade ultramétrica.

4.4 Considere a seguinte matriz de dissemelhanças entre 5 objectos:

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 6 & 0 & & & \\ 3 & 6 & 0 & & \\ 6 & 5 & 6 & 0 & \\ 3 & 6 & 2 & 6 & 0 \end{bmatrix} \end{array}$$

(a) Desenhe o dendrograma resultante da aplicação do método da ligação simples. Verifique que a matriz de distâncias **U** construída a partir daquele dendrograma, coincide com **D** e portanto **D** é ultramétrica, isto é, os seus elementos satisfazem a desigualdade ultramétrica.

(b) Mostre que $\mathbf{D} = \mathbf{U}$ se e só se **D** é ultramétrica.

4.5 No método da mediana a dissemelhança $d_{A(BC)}$ entre o grupo A e o grupo BC resultante da reunião dos grupos B e C é tal que verifica a seguinte igualdade

$$d_{A(BC)} = \frac{1}{2}(d_{AB} + d_{AC}) - \frac{1}{4}d_{BC} \,.$$

Mostre que se $d_{A(BC)}$ é o quadrado da distância euclidiana então o seu valor coincide com o quadrado do comprimento da mediana do triângulo de vértices A, B, C que passa no vértice A.

4.6 (a) Mostre que a fórmula de recorrência de Lance e Williams para $\alpha_A = \frac{1}{2}$, $\alpha_B = \frac{1}{2}$, $\beta = 0$ e $\gamma = -\frac{1}{2}$ reproduz de facto o método da ligação simples.

(b) Mostre ainda que a fórmula dá o método da ligação média para os valores adequados dos parâmetros.

4.7 A matriz

$$\mathbf{A} = \begin{bmatrix} 1.0 & 7.0 \\ 1.5 & 7.5 \\ 2.0 & 8.5 \\ 3.0 & 7.0 \\ 4.0 & 4.0 \\ 2.0 & 1.0 \end{bmatrix}$$

resultou da medição de dois objectos relativamente a duas das suas características.

(a) Obtenha e compare os dendrogramas da análise da matriz **A**, servindo-se da distância euclidiana, para os métodos da ligação simples, ligação completa, ligação média, Ward, mediana e centróide.

(b) Represente as duas variáveis graficamente e observe as posições relativas dos pontos representando os seis objectos para tentar explicar os resultados dos diferentes métodos.

4.8 Uma análise de clusters sobre cinco objectos é realizada a partir da matriz de dissemelhanças

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.00 & & & & \\ 18.03 & 0.00 & & & \\ 20.62 & 14.14 & 0.00 & & \\ 22.36 & 11.18 & 5.00 & 0.00 & \\ 8.60 & 17.00 & 25.08 & 25.15 & 0.00 \end{bmatrix} \end{array}$$

e no último passo do processo o grupo (15) é unido com o grupo (234).

(a) Apresente a matriz de dissemelhanças actualizada, referente ao último passo, considerando que são usados os métodos da ligação simples, completa e média.

(b) Ao tentar efectuar a análise de clusters um analista argumentou que os grupos (23) e (45) devem ser aglutinados já que a sua dissemelhança é pequena.

- Calcule essa dissemelhança.
- Diga porque é que essa aglutinação não foi considerada pelos métodos da alínea (a).

4.9 Os n objectos de uma população foram observados em p características.

(a) Estabeleça a relação entre a distância euclidiana e a distância euclidiana média entre dois objectos da população.

(b) Suponha que aplica um mesmo método de análise de clusters usando as duas distâncias. Como compara os resultados das duas análises no que se refere ao número e composição dos clusters obtidos?

(c) A matriz de dados relativa a quatro objectos e seis características

é

	Características					
Objectos	C_1	C_2	C_3	C_4	C_5	C_6
O_1	17	5	10	15	16	11
O_2	13	9	7	8	10	14
O_3	17	–	–	15	16	11
O_4	13	9	7	–	–	14

Calcule as distâncias euclidiana e euclidiana média entre O_1 e O_2 e entre O_3 e O_4. Note que $O_1 \equiv O_3$ e $O_2 \equiv O_4$, excepto nos valores omissos que foram introduzidos de propósito. O que pode concluir quanto aos resultados da análise de clusters usando as duas distâncias, quando há e quando não há valores omissos?

4.10 Verifique as conclusões a que chegou no Exercício 2.8 realizando uma análise de clusters sobre os dados de cinco objectos observados em seis variáveis binárias, aplicando h_{ij} e s_{ij} separadamente.

	Características					
Objectos	V_1	V_2	V_3	V_4	V_5	V_6
O_1	1	0	1	1	0	1
O_2	0	1	1	0	1	0
O_3	1	1	0	1	0	1
O_4	0	0	1	1	1	1
O_5	1	1	1	0	0	1

4.11 A Tabela 4.5 mostra as frequências relativas de genes para quatro tipos de sangue, A_1, A_2, B e O, numa amostra de grande dimensão recolhida em quatro populações: Esquimó, Bantu, Inglesa e Coreana. Num estudo com vista a caracterizar aquelas populações Cavalli-Sforza and Edwards (1967) chegaram à conclusão que as quatro populações formam dois grupos distintos, Bantu–Inglesa e Esquimó–Coreana.

(a) Recorra aos métodos da ligação simples e da ligação completa para analisar os dados.

(b) Qual das propostas, ligação simples ou ligação completa, lhe parece mais adequada para analisar este tipo de dados?

(c) Acha que se confirma a existência dos dois grupos encontrados por aqueles autores?

Tabela 4.5 *Frequências de genes para quatro tipos de sangue.*

Tipo de sangue	Populações			
	Esquimó	Bantu	Inglesa	Coreana
A_1	0.2914	0.1034	0.2090	0.2208
A_2	0.0000	0.0866	0.0696	0.0000
B	0.0316	0.1200	0.0612	0.2069
O	0.6770	0.6900	0.6602	0.5723

5

Métodos não hierárquicos

Como já se viu a classe dos métodos hierárquicos fornece algoritmos para muitos tipos de problemas. Os métodos hierárquicos produzem hierarquias, aplicam-se tanto a variáveis como a objectos, usam matrizes de proximidade e sempre que um objecto é atribuído a um cluster não mais abandona esse cluster. Existem muitos outros métodos que assentam em diferentes princípios e cujos resultados não constituem hierarquias. Distinguem-se em primeiro lugar os métodos de partição.

5.1 Métodos de partição

Os métodos de partição aplicam-se a objectos, operam sobre uma matriz de dados e, contrariamente ao que acontece com os métodos hierárquicos, exigem que o número de grupos seja fixado à partida.

O problema concreto consiste em construir, a partir de um conjunto de dados, uma partição, isto é, uma colecção de grupos disjuntos de objectos cuja reunião constitui o conjunto de objectos inicial. Os grupos devem satisfazer propriedades básicas de coesão interna e isolamento dos grupos, às quais se associam em geral critérios de homogeneidade e heterogeneidade que servem de guia à formação dos clusters.

Uma ideia ingénua seria a de construir todas as partições e analisá-las todas com vista a seleccionar a melhor. Mas este número de partições é tão grande que se torna impraticável a observação de todas elas. Basta ver que a partir de n objectos é possível construir $P(n,k)$ partições, cada uma com k grupos, onde

$$P(n,k) = \left[k^n - \sum_{i=1}^{k-1} \frac{k!}{(k-i)!} P(n,i) \right] \Big/ k!,$$

número geralmente muito grande, mesmo para valores moderados de n e k. Por exemplo $P(16,3) \simeq 14 \times 10^6$ e já se tinha visto a propósito dos procedimentos divisivos que $P(n,2) = 2^{n-1} - 1$.

O problema reduz-se então a examinar algumas partições de forma a encontrar a melhor partição, o que é feito optimizando algum critério de formação dos clusters que tenha sido fixado.

Os métodos de partição usam procedimentos que em geral seguem os passos seguintes:

1. Seleccionar uma partição inicial dos n objectos em k grupos.
2. Considerar todas as deslocações de cada objecto do seu próprio grupo para cada um dos outros e registar a alteração produzida no critério de formação de clusters utilizado.
3. Efectuar a deslocação correspondente ao maior valor da melhoria verificada no valor do critério.
4. Repetir os passos 2 e 3 até se verificar que a deslocação de qualquer objecto não produz melhoria no valor do critério.

Alguns autores têm sugerido alterações de vários aspectos deste procedimento geral, com vista a obter algoritmos mais eficientes para os métodos de partição:

- **Escolha da partição inicial**

A selecção da partição inicial pode ser feita de várias maneiras. Os k grupos da partição inicial podem ser o resultado da aplicação prévia de outro método de análise, podem ser definidos com base no conhecimento do problema em estudo ou podem mesmo ser escolhidos ao acaso.

No caso em que os objectos são identificados com pontos de um espaço euclidiano podem tomar-se k pontos (sementes) como centróides dos k grupos e a sua escolha pode ser feita de forma variada, inclusivamente de maneira aleatória

- **Deslocação de objectos para grupos**

Aqui há várias possibilidades mas os procedimentos mais comuns consistem em deslocar um objecto de cada vez ou grupos de objectos simultaneamente.

- **Critérios de formação de clusters**

Há muitos critérios para formação de clusters pertencentes a uma partição. Entre eles destacam-se os critérios usados para análise de uma matriz de dados do tipo contínuo, $\mathbf{X}_{n\times p}$, que assentam na conhecida equação

$$\mathbf{T} = \mathbf{W} + \mathbf{B},$$

onde **T**, **W** e **B** são as matrizes associadas à variação total dos dados, à variação dentro dos grupos e entre os grupos, respectivamente, e são dadas por

$$\mathbf{T} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}) (\mathbf{x}_{ij} - \overline{\mathbf{x}})', \tag{5.1}$$

onde $\mathbf{x}_{ij}$ é o vector de observações do objecto j no grupo i,

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbf{x}_{ij}}{\sum_{i=1}^{k} n_i},$$

com $\sum_{i=1}^{k} n_i = n$, é o vector das médias de cada uma das p variáveis nos n objectos,

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i) (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)' \tag{5.2}$$

e

$$\mathbf{B} = \sum_{i=1}^{k} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}}) (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'. \tag{5.3}$$

A equação $\mathbf{T} = \mathbf{W} + \mathbf{B}$ tem uma interpretação muito intuitiva e clara no caso $p = 1$. Trata-se da partição da soma total de quadrados na soma de quadrados dentro (W) e entre (B) grupos que é fundamental na análise de variância. Nesta situação é então natural usar como critério de construção de clusters de uma partição uma destas somas de quadrados. A melhor partição é aquela em que W é mínimo ou, de forma equivalente, B é máximo, isto é, quanto maior for a homogeneidade interna e maior a separação entre grupos.

(i) Traço de **W**

O critério sugerido para o caso univariado generaliza-se ao caso multivariado, embora aqui a interpretação já não seja tão intuitiva. No caso multivariado minimizar a soma dos quadrados dentro dos grupos para as p variáveis equivale a minimizar o traço da matriz **W**, tr**W**, ou maximizar tr**B**. Everitt *et al.* (2001) refere que minimizar tr**W** é equivalente a minimizar a soma dos quadrados das distâncias euclidianas entre os objectos e as médias dos respectivos grupos, isto é,

$$\mathrm{tr}\,\mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)' (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d_{ij,i}^2$$

onde $d_{ij,i}$ é a distância euclidiana do objecto j do grupo i à média do grupo i.

(ii) Determinante de $\mathbf{W}$

$|\mathbf{W}|$ é outra medida de variabilidade global dos dados. Um novo critério para obter uma partição óptima consiste em minimizar $|\mathbf{W}|$.

(iii) Traço de $\mathbf{BW}^{-1}$

Aqui o critério é maximizar $\mathrm{tr}\mathbf{BW}^{-1}$.

Os casos (ii) e (iii) surgem no contexto da análise de variância multivariada (MANOVA) e correspondem a situações em que as médias dos grupos são diferentes. Quanto menor é $|\mathbf{W}|$ e quanto maior é $\mathrm{tr}\mathbf{BW}^{-1}$ mais diferentes são as médias dos grupos.

A questão que se coloca é a de saber qual é o melhor critério. Friedman and Rubin (1967) e Marriott (1971) preferem $|\mathbf{W}|$ pela sua sensibilidade à estrutura dos dados e porque leva em conta as correlações entre as variáveis. Este critério tende a produzir clusters com a mesma forma elíptica. Contudo o critério baseado em $\mathrm{tr}\mathbf{W}$ é o mais popular por ser simples e fácil de tratar computacionalmente. O critério é não invariante a mudanças de escala e isto é um inconveniente sério, uma vez que o recurso à estandardização, frequente em análise de clusters, conduz a uma solução diferente daquela que é obtida a partir dos dados não estandardizados.

Como se viu o critério que consiste em minimizar $\mathrm{tr}\mathbf{W}$ é equivalente a minimizar a soma dos quadrados das distâncias euclidianas entre os objectos e os centróides dos respectivos grupos. Isto equivale a dizer que a deslocação do objecto (referida no passo 2 do procedimento de construção de uma partição) é feita para o grupo cujo centróide está mais próximo do objecto. Este algoritmo é conhecido por **algoritmo das k-médias** e a sequência de passos é:

1. Seleccionar uma partição inicial.
2. Deslocar os objectos de forma a que cada objecto fique colocado no grupo da partição que tem o centróide mais próximo.
3. Recalcular os centróides dos novos grupos assim formados.
4. Repetir os passos 2 e 3 até não ser possível efectuar mais deslocações.

Para mostrar o funcionamento do algoritmo faz-se a seguir a sua aplicação a um conjunto de dados artificiais.

Exemplo 5.1. Para ver como funciona o método das k-médias considere-se o conjunto de dados artificiais da Tabela 5.1. Suponhamos que o objectivo é construir dois clusters a partir dos 5 objectos agora designados por A, B, C, D e E.

1. O processo é iniciado considerando dois clusters arbitrários. Sejam

Tabela 5.1 *Dados artificiais relativos a cinco objectos.*

	Variáveis	
Objectos	x_1	x_2
A	2	8
B	5	1
C	4	12
D	15	4
E	16	5

AB e CDE esses clusters.

2. Em seguida são calculados os centróides dos clusters e as distâncias dos objectos aos centróides (usou-se d^2, o quadrado da distância euclidiana):

	Centróides		d^2				
Clusters	$\bar{x}_1$	$\bar{x}_2$	A	B	C	D	E
AB	3.5	4.5	14.5	14.5	56.5	132.75	156.5
CDE	11.67	7	94.51	80.49	83.83	20.09	22.75

Estes resultados mostram que C deve sair de CDE e juntar-se a AB, obtendo-se os novos clusters ABC e DE.

3. Observando as distâncias dos objectos aos centróides dos novos clusters

	Centróides		d^2				
Clusters	$\bar{x}_1$	$\bar{x}_2$	A	B	C	D	E
ABC	3.67	7	3.79	37.77	25.11	137.77	156.03
DE	15.5	4.5	194.5	122.5	188.75	0.5	0.5

verifica-se que não é possível deslocar qualquer dos objectos do cluster a que pertence e portanto ABC e DE são os dois clusters encontrados usando o procedimento das k-médias.

Uma desvantagem do algoritmo das k-médias é a sua falta de robustez devida em parte ao facto de, por um lado, usar o quadrado das distâncias euclidianas e, por outro lado, se basear na média. A presença de *outliers* pode distorcer grandemente a distribuição dos dados. O **algoritmo k-medóides** tenta resolver esta dificuldade usando para representante do grupo um objecto do próprio grupo, em vez da média do grupo. No k-medóides o representante de um grupo é o objecto mais central relativamente aos outros

objectos do grupo. A estratégia usada por este algoritmo consiste em dividir os n objectos em k clusters procurando em primeiro lugar um objecto representativo (medóide) para cada cluster. Cada um dos outros objectos é então colocado junto com o medóide que lhe é mais semelhante para formar um cluster. Kaufman and Rousseeuw (1990) desenvolveram o programa *PAM (partitioning around medoids)* que está disponível em S-PLUS. Aqueles autores apresentam também a representação gráfica do resultado dos k-medóides que é conhecida por *silhueta* (*silhouette plot*). O gráfico silhueta permite avaliar a qualidade da partição e distinguir, dentro da partição, clusters bem definidos.

Aplicando o procedimento dos k-medóides aos dados artificiais da Tabela 5.1 são obtidos os mesmos clusters das k-médias, no entanto os representantes de cada cluster são distintos:

	Centróides		Medóides	
Clusters	$\bar{x}_1$	$\bar{x}_2$	x_1	x_2
ABC	3.67	7	2	8
DE	15.5	4.5	15	4

5.2 Outros métodos

Os métodos descritos neste capítulo e em capítulos anteriores constituem de longe os métodos mais utilizados na prática. Mas além destes métodos a literatura apresenta muitos outros. Gordon (1999) tem um capítulo e Everitt *et al.* (2001) dois capítulos dedicados a estes métodos especiais. A seguir indicam-se algumas das categorias em que podem agrupar-se estes métodos. Note-se que alguns deles integram ideias de outros métodos, incluindo princípios dos métodos descritos anteriormente, e por isso nem sempre é fácil situá-los numa categoria. Não se dão aqui explicações suficientes que permitam compreender os objectivos e funcionamento dos métodos e remete-se o leitor para os livros indicados onde encontrará algumas dessas explicações e a sugestão de bibliografia para estudos mais completos.

- **Métodos baseados em modelos**

Aqui supõe-se a existência de uma estrutura com k grupos e que há um modelo subjacente responsável por ter gerado cada um dos grupos. As hipóteses que se colocam com mais frequência são:

(i) O vector de observações $\mathbf{x}$ tem função de densidade de probabilidade (f.d.p.) $f_i(\mathbf{x};\boldsymbol{\theta}_i)$ se provém do grupo i, $i = 1,\ldots,k$, onde $\boldsymbol{\theta}_i$ representa um vector de parâmetros desconhecidos.

(ii) O vector de observações $\mathbf{x}$ tem f.d.p.

$$f(\mathbf{x};\mathbf{p},\boldsymbol{\theta}) = \sum_{i=1}^{k} p_i f_i(\mathbf{x};\boldsymbol{\theta}_i)\,,$$

correspondente a uma mistura finita das densidades $f_i(\mathbf{x};\boldsymbol{\theta}_i)$ de cada um dos clusters. O peso p_i é a probabilidade associada a cada componente da mistura, sendo $\sum_{i=1}^{k} p_i = 1$.

O problema resume-se então a estimar, a partir das observações, os parâmetros em cada um dos modelos, depois do que os grupos ficam identificados. A situação mais comum é assumir que as densidades envolvidas são de distribuições normais multivariadas.

- **Pesquisa de densidades**

Neste método os objectos são identificados como pontos de um espaço euclidiano e os grupos são encarados como regiões com alta densidade de pontos, separados uns dos outros por regiões de baixa densidade, representando algum tipo de ruído.

- **Métodos difusos**

A análise de clusters designada por difusa (*fuzzy*) é uma generalização da ideia de partição. Numa partição cada objecto pertence a um e um só cluster, assumindo-se que esta divisão é feita com clareza e não havendo dúvidas na decisão. Ora na prática ocorrem muitas situações em que a decisão de colocar um objecto no grupo A ou no grupo B não é óbvia, mas sim rodeada de ambiguidades. O procedimento *fuzzy* associa a cada objecto um vector cujas componentes representam o grau de ligação (ou pertença) do objecto a cada um dos grupos *fuzzy*. E em consequência cada grupo fica identificado por um vector de coeficientes que representam o grau de pertença de cada um dos objectos a esse mesmo grupo. No caso dos métodos hierárquicos e de partição o vector associado a um objecto é um vector do tipo $(0,\ldots,0,1,0,\ldots,0)$ em que a posição do 1 identifica o cluster a que o objecto pertence. É claro que o vector que identifica um cluster é do tipo $(0,1,1,0,1,\ldots,1,0,1)$ em que o número de valores iguais a 1 traduz o número de objectos que pertencem ao cluster.

O método de ajustamento de mistura de densidades pode considerar-se como uma análise de clusters *fuzzy* em que os pesos p_i são os coeficientes de pertença.

A análise *fuzzy* tem a vantagem de fornecer mais informação sobre a estrutura dos dados do que os métodos hierárquicos ou de partição. No entanto o método exige algoritmos complicados, é dispendioso em termos de tempo de cálculo e havendo muitos objectos o *output* que produz é geralmente volumoso, tornando-se por isso difícil de interpretar.

- **Métodos de sobreposição**

Os métodos até agora considerados produzem clusters que não se sobrepõem ou intersectam, isto é, clusters que são disjuntos. Um objecto pertence a um e um só cluster. Há no entanto situações em que é mais significativo permitir que o mesmo objecto pertença a mais do que um grupo. Por exemplo, em linguística, a mesma palavra pode ter significados diferentes e pertencer a categorias gramaticais diferentes. Um atleta pode estar em vários grupos por praticar várias modalidades e um indivíduo pode pertencer a diferentes grupos de consumidores.

Não deve confundir-se o método de sobreposição com o método *fuzzy*. Reconhecer que um dado objecto pertence a vários clusters sem qualquer dúvida é diferente da situação em que há dúvidas sobre qual o cluster a que o objecto pertence.

Duas abordagens à sobreposição, ADCLUS (*additive clustering*) e pirâmides, estão descritas em Shepard and Arabie (1979) e Diday (1986).

Além destas classes de métodos, cada uma com diferentes abordagens, existem outras que podem ser encontradas na bibliografia já referida. Mirkin (1996) apresenta uma tipologia de algoritmos para análise de clusters onde se incluem muitos dos métodos não *standard*.

5.3 Considerações de ordem prática

5.3.1 Escolha do método e do algoritmo

A escolha é difícil porque há muitos métodos e algoritmos. Qual dos tipos de métodos, hierárquicos ou não hierárquicos, se deve escolher? Tendo optado por uma destas famílias de métodos, que algoritmo usar? Não é possível dar uma resposta clara e o que se pode dizer é que a selecção depende muito do objectivo da investigação e das propriedades dos vários métodos. Em Punj and Stewart (1983) apresentam-se resumos das propriedades dos algoritmos e referem-se as conclusões de vários estudos empíricos levados a cabo para comparação desses algoritmos.

Os métodos hierárquicos não requerem o conhecimento prévio do número de clusters, o que parece ser uma vantagem sobre os métodos não hierárquicos. Mas também apresentam desvantagens:

(i) sempre que um objecto é atribuído a um cluster não mais pode sair desse cluster,
(ii) os algoritmos disponíveis são numerosos pois dependem das dissemelhanças entre os objectos e do critério de agregação dos clusters.

Uma sugestão é operar com vários métodos, comparar os resultados, verificar se são consistentes e no final seleccionar a solução cuja interpretação seja mais fácil.

Os métodos não hierárquicos de partição apresentam uma grande exigência que é o conhecimento do número de clusters *a priori*. Por outro lado exigem também que se defina uma partição inicial, o que nem sempre é óbvio uma vez que o número de possibilidades é muito grande. Além do mais sabe-se que a partição inicial influencia os resultados.

Uma sugestão que conduz muitas vezes a resultados satisfatórios consiste em produzir uma solução hierárquica para em seguida ser usada como partição inicial nos métodos de partição. A experiência e bom senso consideram que os métodos devem funcionar de forma complementar. A opção por uma só escolha não é aconselhável, a menos que a escolha seja claramente natural. Usar vários métodos e algoritmos tem a vantagem destes poderem revelar diferentes aspectos da estrutura.

5.3.2 Quantos clusters há nos dados?

Um dos grandes problemas da análise de clusters é a escolha do número de clusters. Os métodos hierárquicos adiam a decisão para o final da análise. Onde cortar o dendrograma? Nos métodos não hierárquicos de partição um dos primeiros passos da análise é saber quantos grupos deve ter a partição.

Um método simples e informal é a análise gráfica. No caso dos métodos hierárquicos representa-se o índice de fusão contra o número de clusters. A situação é semelhante à construção do gráfico usado em componentes principais para decidir sobre o número mínimo de componentes a reter. No caso da análise de clusters grandes alterações no nível de fusão correspondem à junção de grupos muito diferentes e podem sugerir o número de clusters que interessa. Geralmente a zona de cotovelo do gráfico dá indicação do número de clusters.

Para os métodos não hierárquicos pode usar-se o mesmo procedimento representando graficamente o valor do critério de optimização usado, por

exemplo $\mathrm{tr}\,\mathbf{W}$, contra o número de clusters. Examinando o gráfico e procurando a zona de cotovelo obtém-se uma sugestão sobre o número de clusters adequado.

Pode também construir-se um índice com base na razão entre a soma dos quadrados entre clusters, correspondente a k clusters, e a soma total de quadrados

$$R_k^2 = \frac{\mathrm{tr}\,\mathbf{B}_k}{\mathrm{tr}\,\mathbf{T}} = 1 - \frac{\mathrm{tr}\,\mathbf{W}_k}{\mathrm{tr}\,\mathbf{T}}.$$

Para $k = n$ clusters tem-se $\mathrm{tr}\,\mathbf{W} = 0$ e $R_n^2 = 1$. À medida que o número de clusters diminui de n até 1, os clusters vão ficando mais separados, e um decréscimo grande no valor de $R_k^2 = 1$ é sinal de clusters bem distintos e indicação do número de clusters que se procura.

O problema do número de clusters tem sido abordado por muitos autores e o número de propostas de solução é grande. Milligan and Cooper (1985) apresentam uma revisão de muitas dessas propostas. Gordon (1994) apresenta também uma discussão detalhada do assunto.

5.3.3 Validação

Dado que a análise de clusters produz sempre uma classificação e dado que diferentes métodos de análise podem conduzir a diferentes resultados, é imprescindível avaliar esses resultados para ver se constituem um resumo útil dos dados ou se, pelo contrário, representam apenas uma estrutura inconsistente que está a ser imposta sobre os dados.

A maneira natural de proceder à validação de resultados é efectuar uma análise confirmatória, em que os mesmos procedimentos são usados noutros dados, como é habitual fazer em estatística e paralelamente efectuar testes sobre o modelo usado. Porém isso não tem acontecido muito em análise de clusters, em parte devido à natureza dos dados e ao tipo de amostra de trabalho.

A maior parte das vezes os dados são descritos por variáveis de tipo misto impedindo que seja imposta a hipótese de normalidade e a construção de testes estatísticos. Muitas vezes o conjunto de objectos a analisar é a população total e seria inapropriado dividir a população para usar uma parte na exploração e a outra na confirmação dos resultados obtidos na primeira.

Jain and Dubes (1988) apresentam um conjunto de estratégias de validação identificando três tipos de testes ou critérios:

(i) Criérios externos que medem o desempenho dos resultados analisando a utilidade da estrutura, a sua capacidade preditiva e, se possível,

a sua consistência em diferentes amostras. No fundo estes critérios tentam comparar a estrutura com informação exterior não utilizada na análise.

(ii) Critérios internos que comparam a estrutura obtida com os dados iniciais. No caso dos métodos hierárquicos é habitual usar o coeficiente de correlação cofenético para comparar a matriz de proximidades associada aos dados e a matriz cofenética associada ao dendrograma. Os elementos da matriz cofenética são as distâncias ultramétricas entre cada dois objectos como elementos do dendrograma. O coeficiente de correlação cofenético é o habitual coeficiente de correlação de Pearson entre os $n(n-1)/2$ pares de dissemelhanças das duas matrizes. Quanto maior é o coeficiente de correlação cofenético maior é a concordância entre os dados e o dendrograma, o que significa que ambas as estruturas revelam a mesma informação.

(iii) Critérios relativos que comparam diferentes estruturas construídas a partir dos mesmos objectos. Neste caso, dispondo de duas partições, uma possibilidade é construir uma tabela de contingência em cujas células ficam os objectos pertencentes simultaneamente às duas classificações. Com base na tabela de contingência pode então avaliar-se a magnitude da associação entre as duas partições. É claro que se espera poder decidir sobre qual é a melhor partição no sentido de ser mais estável e mais apropriada aos dados.

Todos os critérios de validação pressupõem a existência de uma estrutura de grupos. Nem sempre essa hipótese é verdadeira e por isso é importante começar a análise por testar a completa ausência de estrutura (Gordon, 1999).

5.3.4 Apresentação dos resultados de uma análise de clusters

O resultado de uma análise de clusters, concretizado muitas vezes num diagrama ou numa partição esconde um conjunto de hipóteses de trabalho que o analista não pode deixar de incluir no seu relatório final. É fundamental explicar a teoria subjacente ao problema em estudo, descrever o enquadramento, indicar como foram seleccionados os objectos e as variáveis, que medidas de proximidade foram usadas e que métodos e algoritmos adoptados. É útil indicar também o *software* utilizado pois por vezes *software* de diferentes proveniências pode produzir resultados diferentes embora operando noa mesmos dados e usando os mesmos métodos (Blashfield, 1977). No relatório de apresentação dos resultados deve constar a descrição do critério usado para a determinação do número de clusters e a evidência que seja possível conseguir para suportar a validade da estrutura de grupos produzida pela análise.

Exercícios

5.1 Considere nove pontos do espaço euclidiano bidimensional

$$A_1(1,2), \quad A_2(2,3), \quad A_3(4,1),$$

$$B_1(2,10), \quad B_2(4,7), \quad B_3(6,8),$$

$$C_1(4,4), \quad C_2(8,4), \quad C_3(7,3)$$

e suponha que se pretendem agrupar estes objectos em três clusters.

(a) Aplique o método das k-médias considerando que A_1, B_1 e C_1 são os centros iniciais dos três clusters. Identifique os centros dos clusters depois do primeiro passo. Apresente os três clusters finais e os seus centros.

(b) Responda às questões da alínea (a) tomando para centros iniciais dos clusters os pontos A_2, B_2 e C_2. Compare os dois procedimentos e diga qual a conclusão a que chegou.

5.2 Pretende-se dividir um conjunto de n objectos em k grupos de forma a constituir uma partição. Como se viu na Secção 5.1 três possíveis algoritmos que levam à partição pretendida consistem em

1. minimizar $\mathrm{tr}\,\mathbf{W}$,
2. minimizar $|\mathbf{W}|$,
3. maximizar $\mathrm{tr}\,\mathbf{B}\mathbf{W}^{-1}$,

onde $\mathbf{W}$ e $\mathbf{B}$ são definidas por (5.2) e (5.3), respectivamente. Mostre que no caso $k = 2$ os três métodos produzem a mesma partição.

5.3 Considere os dados da Tabela 5.2 relativos ao consumo de proteínas em países da Europa.

(a) Seleccione dois métodos de análise com vista a produzir uma solução hierárquica e uma solução não hierárquica com base nos dados apresentados. Compare e discuta as soluções encontradas.

(b) Responda ao que se pede na alínea (a) usando os dados estandardizados.

Tabela 5.2 *Dados relativos ao consumo de proteínas em países da Europa.*

	Carne Verm.	Carne Branca	Ovos	Leite	Peixe	Cereais	Amilá-ceos	Frutos Secos	Frutas e Veg.
Albânia	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Áustria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Bélgica	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgária	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Checoslováquia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Dinamarca	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
R.F.A.	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finlândia	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
França	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Grécia	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungria	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Irlanda	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Itália	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Holanda	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Noruega	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Polónia	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Roménia	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Espanha	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Suécia	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Suíça	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
R. Unido	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
U. Soviética	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
R.D.A.	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Jugoslávia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

6
Aplicações

6.1 Introdução

Realizar uma análise de clusters de dados reais não se resume a aplicar cegamente as várias regras e critérios propostos. A especificidade de cada situação prática faz com que, para além das regras e critérios disponíveis, haja necessidade de considerar de forma crítica particularidades que se prendem com aspectos importantes da análise, como sejam a selecção de variáveis, a construção de proximidades e a escolha dos métodos de análise. É esta postura, geralmente adoptada pelo analista ao longo da sequência de fases da análise, que se pretende mostrar com a discussão do exemplo que se apresenta na Secção 6.3. O objectivo da Secção 6.2 é o de apreciar o resultado da aplicação de alguns dos métodos introduzidos nos Capítulos 4 e 5 a dados que se encontram ao longo do texto. Em particular pretende-se comparar estes resultados com os resultados da análise gráfica conduzida no Capítulo 3 sobre os mesmos dados.

6.2 Alguns exemplos anteriores revisitados

6.2.1 Dados dos planetas

A Figura 6.1 mostra o dendrograma obtido a partir das variáveis *diâmetro, massa, densidade* e *gravidade* da Tabela 1.1 usando o método da ligação média e os dados estandardizados. Foram experimentados vários métodos e todos eles identificam Júpiter como *outlier*. Se os dados não estão estandardizados então Saturno, fazendo valer a sua massa, também surge como *outlier*. É interessante ver (Exercício 6.1) como as variáveis seleccionadas, a dissemelhança escolhida, a estandardização e o método de análise influenciam o resultado. A Figura 6.1 parece sugerir a existência de três clusters: {Júpiter}, {Saturno, Urano, Neptuno} e {Mercúrio, Vénus, Terra, Marte, Plutão}. Decidiu-se então aplicar aos mesmos dados os algoritmos k-médias

e k-medóides com $k = 3$. Os clusters obtidos por estes dois métodos foram exactamente os mesmos, o que vem confirmar a estrutura de grupos detectada.

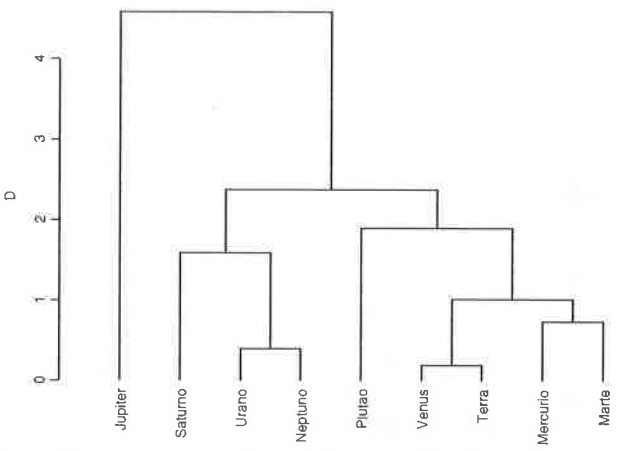


Figura 6.1 *Dendrograma para os planetas do sistema solar com base nas variáveis diâmetro, massa, densidade e gravidade (dados estandardizados, método da ligação média).*

6.2.2 Dados dos cenários faciais

A Figura 6.2 mostra o resultado do método da ligação média aplicado à matriz de dissemelhanças da Tabela 1.4. Como se vê é possível identificar três clusters, precisamente aqueles que já tinham sido detectados por observação visual da Figura 3.11. Esta concordância serve, de certo modo, para confirmar a existência de uma estrutura de grupos subjacente aos dados e por outro lado confirma o valor da representação gráfica como método de pesquisa de clusters num conjunto de dados.

6.2.3 Dados dos alimentos

Esta é a altura de submeter os dados dos alimentos a uma análise de clusters e comparar os resultados com os resultados obtidos graficamente no Capítulo 3. Comparando as Figuras 6.3 (a) (dados não estandardizados) e 6.3 (b) (dados estandardizados) verifica-se que esta última apresenta uma

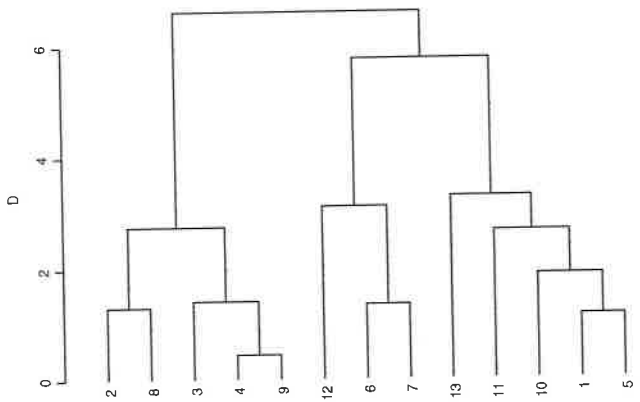
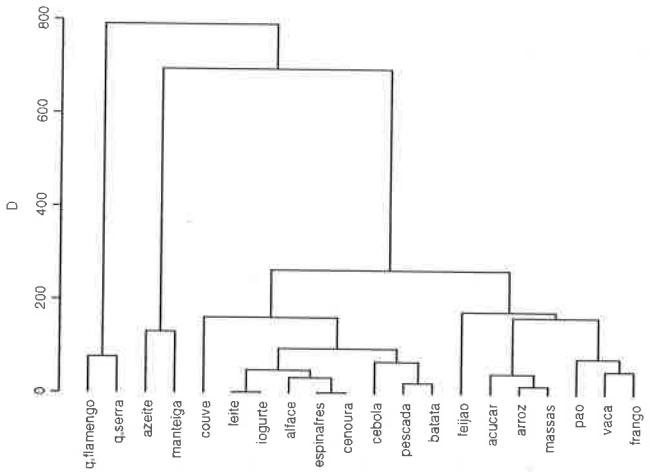


Figura 6.2 *Dendrograma para os objectos correspondentes à matriz de dissemelhanças da Tabela 1.4 (método da ligação média).*

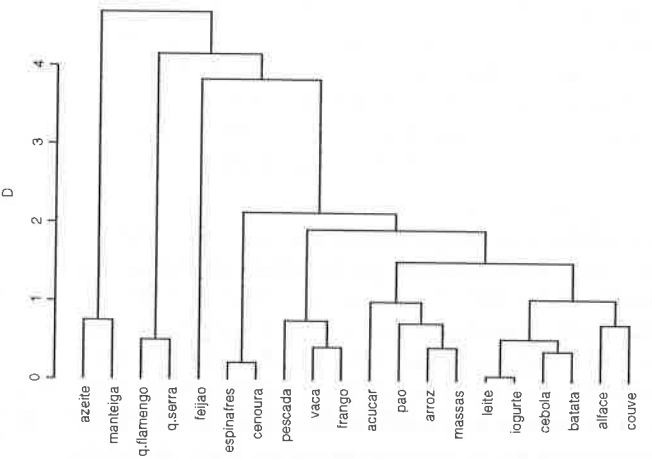
estrutura de grupos mais próxima daquela que é revelada pelos métodos gráficos. Nota-se ainda que a estrutura hierárquica é mais fina do que a dos métodos gráficos, pois consegue construir clusters mais pequenos o que se torna por vezes muito difícil na análise gráfica. Veja-se, por exemplo, que a Figura 6.3 aprsenta dois clusters distintos, {*alface, couve*} e {*espinafres, cenoura*}, que é praticamente impossível identificar nos gráficos das caras de Chernoff, estrelas e curvas de Andrews (Figuras 3.6, 3.7 e 3.8).

Um outro aspecto a reter neste exemplo é o efeito da estandardização. Os dados não estandardizados apresentam um conjunto de grupos que não só se desviam dos resultados da análise gráfica como se afastam das expectativas que se têm da aplicação do senso comum. Com os dados estandardizados o *feijão* não revela a sua condição de *outlier* e a *pescada* junta-se à *batata* e depois à *cebola* sugerindo que o critério de agregação está mais assente nos gostos da cozinha portuguesa do que propriamente na composição dos alimentos. Com dados estandardizados a *pescada* fica junto à *vaca* e ao *frango* que parecem ser os seus companheiros mais adequados.

Para destacar a importância da estandardização aplica-se o método da ligação completa aos dados artificiais da Tabela 2.9. Como se observa na Figura 6.4, realizar a análise sobre dados não estandardizados ou dados estandardizados é uma decisão vital pois os resultados são geralmente diferentes e muitas vezes radicalmente diferentes o que pode ter consequências profundas nas conclusões da investigação que estiver a ser realizada.

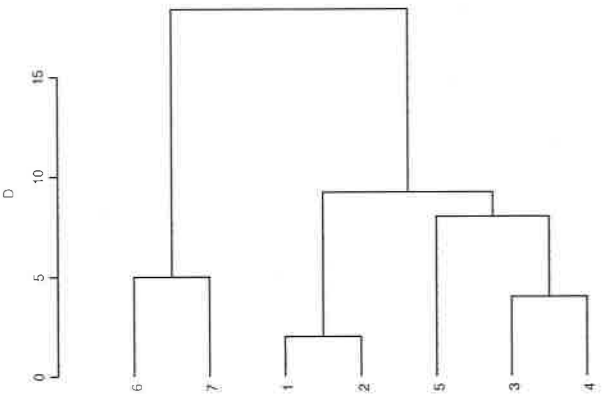


(a) dados não estandardizados

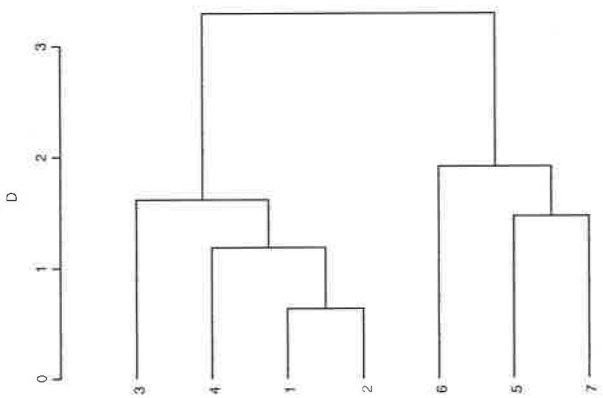


(b) dados estandardizados

Figura 6.3 *Dendrograma para os objectos da Tabela 3.1 (método da ligação média).*



(a) dados não estandardizados



(b) dados estandardizados

Figura 6.4 *Dendrograma para os objectos da Tabela 2.9 (método da ligação completa).*

6.2.4 Dados das características físicas de raparigas

A matriz de correlações destes dados (Tabela 3.3) foi usada para produzir a matriz de dissemelhanças da Tabela 6.1, fazendo $d_{ij} = 1 - r_{ij}$, onde r_{ij} é o coeficiente de correlação entre as variáveis i e j.

O resultado do método da ligação média aplicado a esta matriz de dissemelhanças (Figura 6.5) destaca dois clusters que coincidem precisamente com os clusters visualizados na Figura 3.12. Isto vem corroborar novamente o que já foi dito em 6.2.2 sobre a utilidade dos métodos gráficos e sobre a importância de usar vários métodos de análise para o estudo dos mesmos objectos.

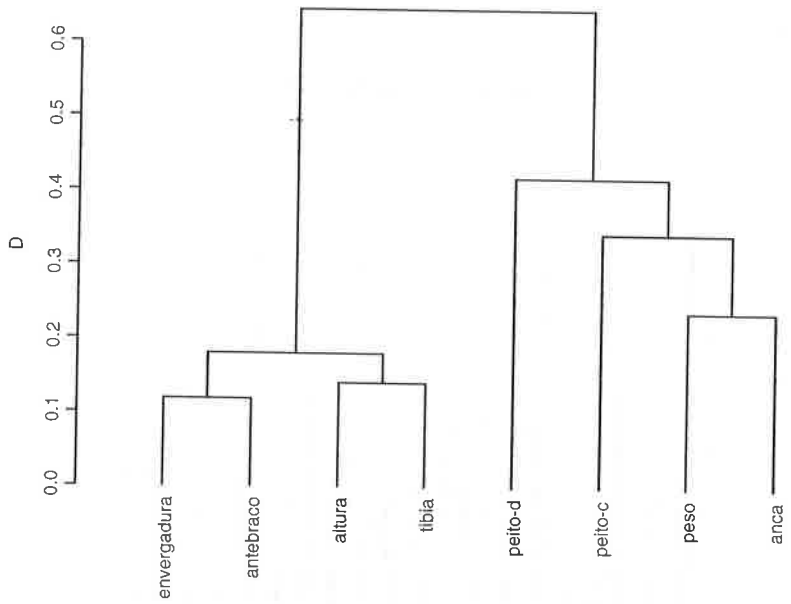


Figura 6.5 *Dendrograma para as variáveis correspondentes à matriz de dissemelhanças da Tabela 6.1 (método da ligação média).*

Tabela 6.1 *Matriz de dissemelhanças entre variáveis (oito características físicas de 305 raparigas).*

Variável	1	2	3	4	5	6	7	8
1. Altura	0.000							
2. Envergadura	0.154	0.000						
3. Antebraco	0.195	0.119	0.000					
4. Tíbia	0.141	0.174	0.199	0.000				
5. Peso	0.527	0.624	0.620	0.564	0.000			
6. Anca	0.602	0.674	0.681	0.671	0.238	0.000		
7. Peito-c	0.699	0.723	0.763	0.673	0.270	0.417	0.000	
8. Peito-d	0.618	0.585	0.655	0.635	0.371	0.423	0.461	0.000

6.3 Um exemplo completo: características de modelos de automóveis

É o desenvolvimento gradual de uma análise de clusters que este exemplo pretende dar a conhecer. Como o número de objectos e de variáveis não é pequeno, o exemplo também permite imaginar o volume e a delicadeza do trabalho a efectuar em situações em que o número de variáveis e o número de objectos são grandes.

O objectivo deste estudo (Branco e Valente, 1996) é agrupar automóveis (objectos), identificados pelas suas marcas e modelos, com base num conjunto de características (variáveis) neles observadas. Os dados foram retirados de uma longa lista de automóveis, com as suas características, publicados na revista AUTOMOTOR de Março de 1996 (38 marcas, 942 modelos e 54 características). Para o estudo foram seleccionados 35 modelos apenas.

Dados sobre automóveis constituem informação muito rica uma vez que o automóvel tem associadas muitas características e que é fácil observar e medir essas características. Esta informação preciosa é procurada, em revistas da especialidade, pelo público em geral que deseja comprar e vender este bem de consumo. Os estatísticos apreciam este tipo de informação e têm-se servido dela para ilustrar o funcionamento tanto de *software* como de uma variedade de métodos estatísticos, como se pode observar em Chambers and Hastie (1992), Lock (1993) e Venables and Ripley (1999).

No estudo da AUTOMOTOR as variáveis são acompanhadas de uma descrição.

Lista de variáveis (originais):

1. *Tipo de carroçaria*
2. *Número de portas*
3. *Preço*
4. *Equipamentos*
   - *ABS*
   - *air-bag condutor*
   - *air-bag passageiro*
   - *alarme antifurto*
   - *imobilizador*
   - *ar condicionado*
   - *regulação de apoio lombar do banco do condutor*
   - *regulação da altura do banco do condutor*

- *regulação eléctrica do banco do condutor*
- *bancos desportivos*
- *bancos em pele*
- *bancos traseiros rebatíveis*
- *computador de bordo*
- *direcção assistida*
- *volante regulável em altura*
- *volante regulável em alcance*
- *elevador eléctrico dos vidros da frente*
- *elevador eléctrico dos quatro vidros*
- *retrovisores eléctricos*
- *fecho centralizado*
- *fecho centralizado com comando à distância*
- *pintura metalizada*
- *faróis antinevoeiro*
- *jantes de liga leve*
- *tecto de abrir manual*
- *tecto de abrir eléctrico*
- *som-pré-equipamento*
- *som-rádio+cassetes*
- *som-rádio+CD*

5. *Consumo*
   - *a 90 km/h* (litros aos 100 km)
   - *a 120 km/h* (litros aos 100 km)
   - *urbano* (litros aos 100 km)
   - *capacidade do depósito* (litros)
   - *autonomia* (km)
6. *Motor*
   - *cilindrada* (cc)
   - *tipo de combustível* (gasolina/gasóleo)
   - *potência* (cv)
   - *binário* (kgm/rpm)
   - *transmissão* (dianteira, traseira, integral)
7. *Performance*
   - *velocidade máxima* (km/h)
   - *aceleração 0-100 km* (segundos)
   - *relação peso/potência*
8. *Peso e dimensões*
   - *peso* (kg)
   - *capacidade da mala* (litros)

- *comprimento* (mm)
- *largura* (mm)
- *altura* (mm)

9. *Custos*
   - *custo por km*
   - *seguro de responsabilidade civil*
   - *seguro de danos próprios*
10. *Anos de garantia*

Selecção e organização das variáveis:

Nem sempre a totalidade das variáveis observadas num determinado estudo é interessante para a análise. Algumas não contribuem para os objectivos do estudo e outras são redundantes na medida em que a característica que representam se encontra bem correlacionada com outras variáveis. Tendo em conta estes argumentos foram eliminadas algumas variáveis.

Variáveis eliminadas:

- *Preço, custo por km, seguro de responsabilidade civil* e *seguro de danos próprios.*
  Estas variáveis estão relacionadas com encargos financeiros. Como o que se pretende comparar são os próprios automóveis, estas variáveis não devem ser tidas em conta.
- *Autonomia* e *relação peso/potência.*
  A variável *autonomia* é função das outras quatro variáveis do item *consumo* e a variável *relação peso/potência* é função das variáveis *peso* e *potência.*

Redução de variáveis binárias no item equipamento:

- Substituir as duas variáveis *elevador eléctrico dos vidros da frente* e *elevador eléctrico dos quatro vidros* pela variável ordinal *elevador eléctrico dos vidros*, com três níveis: sem elevador, elevador só para os vidros da frente e elevador para os quatro vidros.
- Substituir as duas variáveis *fecho centralizado* e *fecho centralizado com comando à distância* (a existência da segunda implica a existência da primeira) pela variável ordinal *fecho centralizado* com três níveis: sem fecho centralizado, com fecho centralizado e com fecho centralizado com comando à distância.

- Substituir as variáveis *tecto de abrir manual* e *tecto de abrir eléctrico* por uma variável ordinal *tecto de abrir* com três níveis: sem tecto de abrir, com tecto de abrir manual e com tecto de abrir eléctrico.
- Substituir as variáveis *som-pré-equipamento*, *som-rádio+cassetes* e *som-rádio+CD* pela variável ordinal *som* com quatro níveis: sem pré-equipamento de som, com pré-equipamento de som, com equipamento de som rádio+cassetes e com equipamento de som rádio+cassetes+CD. Note-se que de uma leitura dos dados confirma-se que todos os automóveis com CD têm também equipamento de rádio+cassetes.

Construção de medidas de proximidade:

Feita a selecção de variáveis os dados ficam reduzidos a 35 modelos e 42 variáveis e o próximo passo é procurar uma medida de proximidade para estes dados. Neste exemplo estão presentes todos os tipos de variáveis já referidos. As variáveis podiam então ser agrupadas de acordo com o seu tipo e em seguida proceder-se à construção de um índice de semelhança combinado. Contudo isso não parece adequado neste caso pois dar-se-ia a mesma importância a todas as variáveis do mesmo tipo. Ora, parece natural pensar que, por exemplo, as variáveis binárias, *tipo de combustível* e *retrovisores eléctricos* não têm a mesma importância. E o mesmo acontece com as variáveis quantitativas *cilindrada* e *comprimento*. Para tornear estas dificuldades adoptou-se outro critério. As variáveis forem agrupadas por itens, determinou-se uma medida de proximidade para cada item e calculou-se uma média ponderada tendo em conta a importância relativa de cada um dos itens. Como é evidente esta ponderação é subjectiva, mas existe sempre muita subjectividade na escolha de uma medida de proximidade.

Lista de itens:

1. Aspecto geral do carro

   Variável qualitativa não binária: *tipo de carroçaria* (Esta variável tem os seguintes níveis: Hatchback (H) – dois volumes com 3 ou 5 portas; Liftback (L) – dois volumes e meio com 5 portas; Sedan (S) – três volumes de 4 portas; Break (B) – carrinha, dois volumes, 5 portas; Coupé (Cp) – dois volumes e meio com 3 portas ou três volumes e 2 portas; Cabrio (Cb) – descapotável derivado de um modelo fechado, com 4 ou 5 lugares e 2 portas; Roadster (R) – descapotável desportivo de 2 lugares e 2 portas; Todo-o-terreno (TT) – dois volumes com 3 ou 5 portas e tracção integral; Van (V) – comercial derivado de hatchback ou de uma break, 2 lugares e 3 portas; Monovolume (M) – um único volume com 5 portas e habitáculo modulável.)

Variáveis quantitativas: *número de portas, peso, capacidade da mala, comprimento, largura, altura.*

2. Equipamento

   Variáveis binárias: *ABS, air-bag condutor, air-bag passageiro, alarme antifurto, imobilizador, ar condicionado, regulação de apoio lombar do banco do condutor, regulação da altura do banco do condutor, regulação eléctrica do banco do condutor, bancos desportivos, bancos em pele, bancos traseiros rebatíveis, computador de bordo, direcção assistida, volante regulável em altura, volante regulável em alcance, retrovisores eléctricos, pintura metalizada, faróis antinevoeiro, jantes de liga leve.*

   Variáveis ordinais: *elevador eléctrico dos vidros, fecho centralizado, tecto de abrir, som.*

3. Consumo

   Variáveis quantitativas: *a 90 km/h, a 120 km/h, urbano, capacidade do depósito.*

4. Motor

   Variáveis quantitativas: *cilindrada, potência, binário.*
   Variável qualitativa não binária: *transmissão.*
   Variável binária: *tipo de combustível.*

5. Performance

   Variáveis quantitativas: *velocidade máxima, aceleração 0-100 km.*

6. Anos de garantia

   Variável quantitativa: *anos de garantia.*

Foi então definido o coeficiente de semelhança para cada tipo de variável.

Variáveis quantitativas: o coeficiente de semelhança de Gower;
Variáveis qualitativas binárias: o coeficiente de concordância simples;
Variáveis ordinais: o coeficiente $1 - |r - s|/l$ (fórmula (2.12) na pág. 36).

Para as variáveis qualitativas *tipo de carroçaria* e *transmissão* não parece adequado decompô-las em variáveis binárias. De facto, a diferença entre um sedan e um coupé é inferior à diferença entre um sedan e um todo-o-terreno, e isso deve reflectir-se no coeficiente. Do mesmo modo, a diferença entre uma transmissão integral e uma transmissão dianteira é maior que a diferença entre uma transmissão dianteira e uma traseira. Assim optou-se por construir matrizes de semelhanças que reflectissem esta situação.

Tipo de carroçaria: considerando os níveis de semelhança.

1. muito diferentes
2. diferentes
3. semelhantes
4. muitos semelhantes

construiu-se a matriz da Tabela 6.2 e que reflecte a opinião dos autores do estudo.

Tabela 6.2 *Semelhanças entre os níveis da variável Tipo de carroçaria.*

	H	L	S	B	Cp	Cb	R	TT	V	M
H										
L	4									
S	3	3								
B	3	3	3							
Cp	4	4	3	3						
Cb	2	2	2	2	2					
R	2	2	2	2	2	4				
TT	1	1	1	1	1	1	1			
V	1	1	1	2	1	1	1	1		
M	1	1	1	2	1	1	1	1	2	

Para transformar estes valores em coeficientes de semelhança com valores entre 0 e 1 podemos depois multiplicá-los por 1/4.

Transmissão: considerando os níveis de semelhança.

1. diferentes
2. semelhantes

construiu-se a matriz seguinte

	D	T	I
D			
T	2		
I	1	1	

a qual pode ser multiplicada por 1/2 para obter coeficientes de semelhança com valores entre 0 e 1.

Ponderações dentro de cada item:

Os pesos associados a cada variável dentro de cada item são os seguintes:

1. Aspecto geral do carro
   Média aritmética entre o coeficiente atribuído à variável *tipo de carroçaria* e o coeficiente atribuído ao conjunto de variáveis quantitativas.
2. Equipamento
   Média entre o coeficiente atribuído ao conjunto de variáveis binárias e o coeficiente atribuído ao conjunto de variáveis ordinais. Esta média será ponderada pelo número de variáveis envolvidas em cada coeficiente (20 binárias e 4 ordinais).
3. Consumo
   Média aritmética entre o coeficiente atribuído à variável *capacidade do depósito* e o coeficiente atribuído ao conjunto de três variáveis que dizem respeito ao consumo propriamente dito (*consumo a 90 km/h, a 120 km/h e urbano*).
4. Motor
   Média ponderada entre o coeficiente atribuído ao conjunto de variáveis quantitativas (*cilindrada, potência, binário*) e o coeficiente atribuído ao conjunto de variáveis qualitativas (*transmissão e tipo de combustível*), no sentido de dar maior importância ao primeiro coeficiente. Pesos: 3 e 1.
5. Performance
   Neste item há apenas duas variáveis quantitativas, por isso não há problemas.
6. Anos de garantia
   Uma só variável.

Pesos a atribuir aos diferentes itens:

Estes pesos refectem a opinião dos autores do estudo sobre a importância relativa dos diferentes itens.

- Aspecto geral do carro – 3
- Equipamento – 2
- Consumo – 2
- Motor – 3
- Performance – 1
- Anos de garantia – 1

Para exemplificar considere-se o cálculo do coeficiente de semelhança entre um Ferrari F355 Spider e um Fiat Punto 55 S com 3 portas.

1. Aspecto geral do carro

	Variáveis						
	t. carro.	*n. portas*	*peso*	*mala*	*comp.*	*larg.*	*alt.*
Ferrari	R	2	1350	220	4250	1945	1170
Punto	H	3	840	275	3760	1625	1450

Coeficiente de semelhança para a variável qualitativa:

$$s_1 = 2 \times \frac{1}{5} = 0.4 .$$

Coeficiente de semelhança para as variáveis quantitativas:

$$s_2 = \frac{1}{6}\left[\left(1 - \frac{|2-3|}{R_1}\right) + \left(1 - \frac{|1350-840|}{R_2}\right) + \cdots \right.$$
$$\left. \cdots + \left(1 - \frac{|1170-1450|}{R_6}\right)\right].$$

Coeficiente de semelhança para o item *aspecto geral do carro*:

$$s_a = \frac{s_1 + s_2}{2}.$$

2. Equipamento

	Variáveis binárias																			
Ferrari	1	1	1	0	0	1	1	0	1	0	1	0	0	1	1	0	1	1	1	1
Punto	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

O que no formato da Tabela 2.5 dá

		Punto	
		1	0
Ferrari	1	0	13
	0	2	5

Utilizando o coeficiente de concordância simples vem

$$s_1 = \frac{0+5}{0+13+2+5} = \frac{1}{4}.$$

	Variáveis ordinais			
	elevador eléctrico	*fecho centralizado*	*tecto abrir*	*som*
	Níveis			
N. total	3	3	3	4
Ferrari	2	2	3	2
Punto	1	1	1	1

pelo que

$$s_2 = \frac{1}{4}\left[\left(1 - \frac{|2-1|}{3}\right) + \left(1 - \frac{|2-1|}{3}\right) + \left(1 - \frac{|3-1|}{3}\right) + \left(1 - \frac{|2-1|}{4}\right)\right] = 0.604.$$

Coeficiente de semelhança para o item *equipamento*:

$$s_e = \frac{20s_1 + 4s_2}{24}$$

O procedimento continua para os quatro itens restantes obtendo-se s_c (*consumo*), s_m (*motor*), s_p (*performance*) e s_g (*anos de garantia*).

Finalmente, o coeficiente de semelhança global entre estes dois automóveis é:

$$s_{FERGT,FIATP} = \frac{3s_a + 2s_e + 2s_c + 3s_m + 1s_p + 1s_g}{12}.$$

Procedendo de igual forma para os restantes pares de automóveis obtém-se a matriz de semelhanças, **S**. Esta foi depois convertida na matriz de dissemelhanças, $\mathbf{D} = \mathbf{1} - \mathbf{S}$, que se apresenta nas Tabelas 6.3 e 6.4.

Uma vez construída a matriz de dissemelhanças foram ensaiados vários métodos de análise tendo-se concluído que três deles (ligação média, ligação completa e Ward) conduzem essencialmente a três clusters bem destacados (ver Figuras 6.6, 6.7 e 6.8) o que leva a concluir que há uma clara estrutura de grupos subjacente aos dados. O resultado do método da ligação simples, também ensaiado, distancia-se muito dos outros três resultados e não revela estrutura.

Tomando o resultado do método Ward como termo de comparação nota-se apenas que o resultado da ligação média isola LROVE num novo cluster que é singular, e que o resultado da ligação completa faz mudar o par LANCD e VOLKG do cluster que se vê à esquerda na Figura 6.8 para o cluster central da Figura 6.7.

Tabela 6.3 *Matriz de dissemelhanças de 35 automóveis (17 primeiras colunas).*

	ARSPI	AUDI4	AUDI8	BMW31	BMW85	CITAX	CITXM	FERGT	FIATP	FIATB	FORDE	FORDS	HONDC	HYUDA	LADAS	LANCY	LANCD
AR145	0.28	0.15	0.43	0.14	0.41	0.14	0.30	0.56	0.11	0.11	0.09	0.32	0.07	0.10	0.11	0.09	0.17
ARSPI		0.21	0.30	0.22	0.39	0.38	0.25	0.45	0.36	0.34	0.34	0.25	0.29	0.33	0.36	0.33	0.22
AUDI4			0.29	0.12	0.39	0.26	0.27	0.51	0.24	0.21	0.20	0.20	0.15	0.20	0.24	0.22	0.20
AUDI8				0.39	0.27	0.55	0.17	0.30	0.52	0.47	0.48	0.17	0.44	0.49	0.52	0.49	0.29
BMW31					0.28	0.22	0.28	0.45	0.20	0.20	0.18	0.29	0.14	0.18	0.20	0.18	0.18
BMW85						0.49	0.30	0.18	0.47	0.47	0.46	0.39	0.42	0.46	0.47	0.46	0.31
CITAX							0.42	0.64	0.04	0.18	0.10	0.42	0.15	0.06	0.04	0.06	0.27
CITXM								0.35	0.40	0.36	0.34	0.19	0.27	0.37	0.39	0.37	0.21
FERGT									0.64	0.62	0.60	0.43	0.56	0.59	0.62	0.61	0.43
FIATP										0.15	0.08	0.41	0.14	0.07	0.04	0.04	0.24
FIATB											0.11	0.35	0.14	0.17	0.17	0.15	0.23
FORDE												0.35	0.09	0.09	0.08	0.08	0.23
FORDS													0.31	0.38	0.39	0.38	0.24
HONDC														0.12	0.13	0.13	0.22
HYUDA															0.07	0.07	0.24
LADAS																0.06	0.24
LANCY																	0.21
LANCD																	
LROVE																	
MEREC																	
MERSL																	
NISSM																	
NISSP																	
OPELC																	
OPELW																	
PEUG8																	
RENCL																	
RENLA																	
ROVER																	
SAAB9																	
SEATI																	
SUZUK																	
TOYTC																	
VOLKG																	

	LROVE	MEREC	MERSL	NISSM	NISSP	OPELC	OPELW	PEUG8	RENCL	RENLA	ROVER	SAAB9	SEATI	SUZUK	TOYTC	VOLKG	VOLVO
AR145	0.54	0.37	0.54	0.13	0.20	0.13	0.32	0.32	0.13	0.27	0.26	0.29	0.28	0.16	0.09	0.14	0.28
ARSPI	0.43	0.15	0.42	0.38	0.31	0.37	0.26	0.26	0.36	0.25	0.19	0.12	0.37	0.39	0.26	0.19	0.27
AUDI4	0.48	0.31	0.40	0.26	0.13	0.25	0.28	0.26	0.24	0.24	0.14	0.23	0.31	0.27	0.15	0.17	0.26
AUDI8	0.27	0.24	0.16	0.54	0.37	0.54	0.19	0.31	0.52	0.22	0.18	0.30	0.56	0.56	0.40	0.31	0.22
BMW31	0.51	0.28	0.48	0.22	0.21	0.21	0.29	0.32	0.21	0.24	0.22	0.23	0.32	0.25	0.14	0.17	0.26
BMW85	0.41	0.28	0.21	0.50	0.48	0.49	0.33	0.42	0.49	0.33	0.36	0.36	0.59	0.53	0.41	0.37	0.31
CITAX	0.67	0.46	0.66	0.02	0.26	0.04	0.44	0.44	0.03	0.39	0.37	0.38	0.27	0.03	0.19	0.26	0.40
CITXM	0.31	0.22	0.31	0.42	0.36	0.41	0.14	0.26	0.41	0.08	0.19	0.24	0.49	0.45	0.25	0.22	0.10
FERGT	0.38	0.36	0.20	0.64	0.62	0.64	0.41	0.50	0.62	0.40	0.45	0.43	0.70	0.66	0.52	0.47	0.37
FIATP	0.64	0.44	0.63	0.03	0.23	0.03	0.41	0.42	0.04	0.36	0.35	0.37	0.26	0.06	0.18	0.23	0.38
FIATB	0.58	0.43	0.59	0.17	0.12	0.16	0.36	0.36	0.17	0.31	0.30	0.33	0.24	0.20	0.17	0.22	0.33
FORDE	0.59	0.42	0.59	0.08	0.20	0.07	0.36	0.36	0.08	0.31	0.31	0.33	0.28	0.12	0.14	0.20	0.32
FORDS	0.33	0.26	0.32	0.41	0.25	0.41	0.18	0.22	0.41	0.17	0.14	0.23	0.46	0.45	0.29	0.23	0.17
HONDC	0.55	0.39	0.55	0.15	0.18	0.14	0.33	0.31	0.14	0.25	0.26	0.29	0.27	0.18	0.05	0.18	0.25
HYUDA	0.63	0.41	0.60	0.07	0.23	0.07	0.39	0.38	0.04	0.34	0.33	0.32	0.24	0.07	0.15	0.22	0.35
LADAS	0.65	0.43	0.63	0.03	0.25	0.03	0.41	0.41	0.04	0.36	0.35	0.35	0.27	0.07	0.17	0.23	0.37
LANCY	0.61	0.41	0.60	0.06	0.22	0.05	0.38	0.39	0.05	0.33	0.33	0.34	0.25	0.09	0.17	0.20	0.37
LANCD	0.44	0.24	0.42	0.27	0.25	0.26	0.22	0.27	0.26	0.16	0.16	0.17	0.37	0.30	0.18	0.07	0.21
LROVE		0.37	0.30	0.67	0.55	0.66	0.32	0.36	0.66	0.35	0.39	0.43	0.65	0.70	0.52	0.44	0.34
MEREC			0.32	0.46	0.40	0.45	0.21	0.31	0.43	0.24	0.23	0.12	0.45	0.47	0.35	0.28	0.24
MERSL				0.65	0.49	0.65	0.34	0.44	0.63	0.38	0.32	0.42	0.67	0.67	0.51	0.43	0.35
NISSM					0.25	0.03	0.44	0.43	0.03	0.38	0.37	0.37	0.26	0.04	0.19	0.25	0.39
NISSP						0.25	0.34	0.33	0.25	0.32	0.21	0.30	0.23	0.28	0.21	0.23	0.34
OPELC							0.43	0.43	0.04	0.38	0.36	0.37	0.25	0.06	0.18	0.25	0.39
OPELW								0.23	0.43	0.14	0.18	0.24	0.44	0.47	0.29	0.23	0.15
PEUG8									0.41	0.25	0.27	0.24	0.38	0.45	0.28	0.28	0.26
RENCL										0.38	0.36	0.35	0.24	0.04	0.18	0.25	0.39
RENLA											0.18	0.22	0.46	0.42	0.22	0.17	0.12
ROVER												0.18	0.42	0.40	0.23	0.16	0.18
SAAB9													0.38	0.39	0.25	0.21	0.23
SEATI														0.27	0.29	0.35	0.48
SUZUK															0.22	0.28	0.42
TOYTC																0.15	0.21
VOLKG																	0.21
VOLVO																	

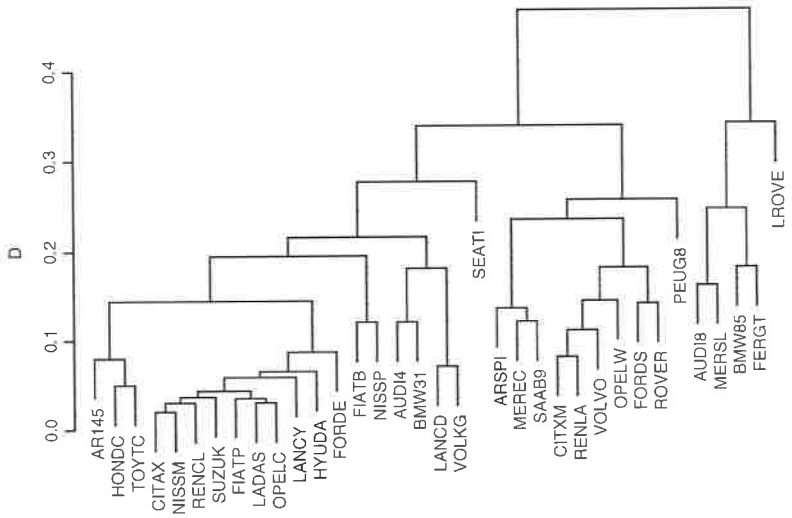


Figura 6.6 *Dendrograma de ligação média para dados de 35 automóveis.*

A característica que parece ter presidido à formação dos grupos é o preço do automóvel. O preço cresce de grupo para grupo, como se observa nos três dendrogramas em análise. Sendo assim percebe-se a deslocação de LANCD (LANCIA Delta HPE 2.0 HF 16v) e VOLKG (VOLKSWAGEN Golf GTI 2.0 16v) entre clusters, uma vez que estes automóveis se encontram na zona intermédia entre preços médios e altos. Percebe-se também a insegurança de LROVE (LAND ROVER Range Rover 4.6 HSE) dentro do seu grupo, pois trata-se de um automóvel que apesar de dispôr de grandes potencialidades difere dos seus pares na funcionalidade e no preço que é, apesar de alto, significativamente mais baixo do que os outros automóveis do mesmo grupo.

O tratamento dos dados usando multidimensional scaling produziu o resultado que se apresenta na Figura 6.9. Foi feita a correspondência da totalidade dos objectos a 35 pontos do espaço de dimensão 42, $\mathbb{R}^{42}$, que em seguida foram amalgamados permitindo a sua representação num espaço de duas dimensões, $\mathbb{R}^2$. Esta representação é aceitável de acordo com os critérios estabelecidos pelo método multidimensional scaling (*stress* = 8%). A Figura 6.9 revela duas dimensões importantes que contribuem para

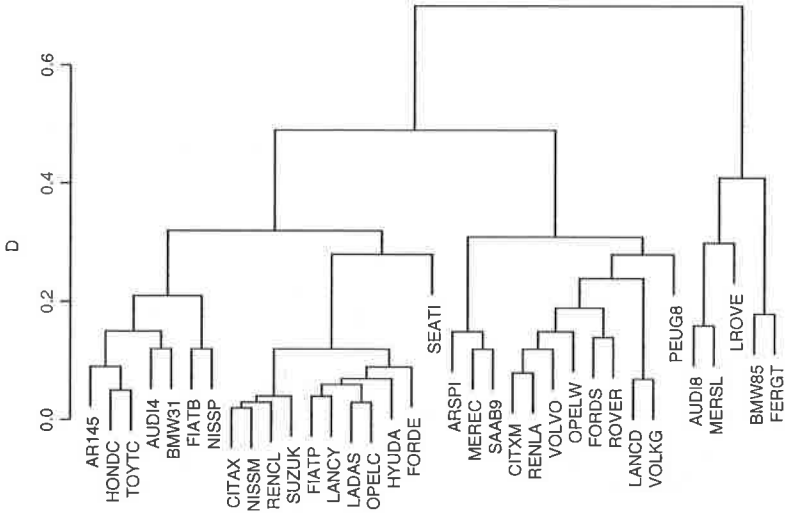


Figura 6.7 *Dendrograma de ligação completa para dados de 35 automóveis.*

a estrutura dos objectos:

(i) preço – baixo, médio, alto e elevado, ao longo do eixo horizontal;
(ii) funcionalidade – desportivos, utilitários e especiais, ao longo do eixo vertical.

Note-se que a análise de clusters tem dificuldades em distinguir o bloco de preços baixos do bloco de preços médios (ver Figuras 6.6, 6.7 e 6.8) e não mostra a dimensão funcionalidade.

Conclusões do estudo:

Resultados

(i) O resultado final não contradiz as expectativas e o conhecimento que em geral se tem deste assunto, o que leva a concluir que a estratégia seguida é adequada.

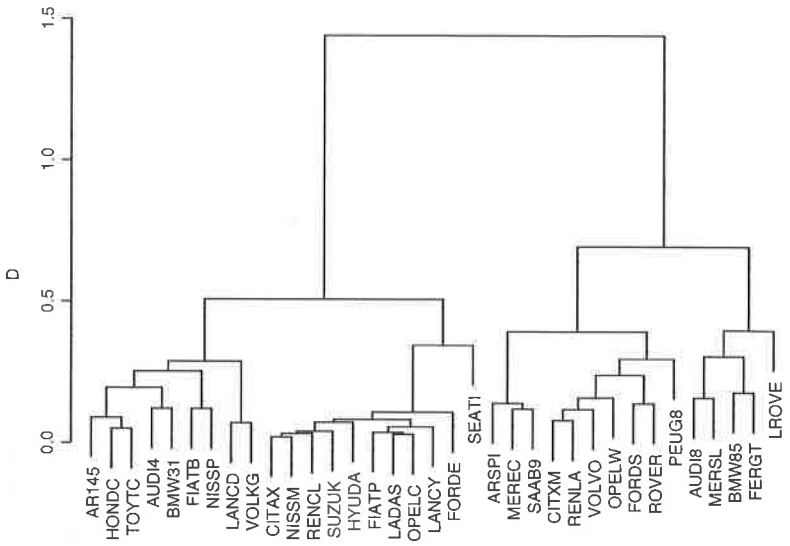


Figura 6.8 *Dendrograma de Ward para dados de 35 automóveis.*

(ii) O resultado final foi enriquecido e clarificado mediante a utilização do método multidimensional scaling.

Recomendações

(i) A orientação que é, em geral, indicada para a construção de semelhanças deve ser tomada como guia e não como receita a aplicar de forma acrítica. Cada caso é um caso e deve ser cuidadosamente escalpelizado por aqueles que conheçam bem o assunto.

(ii) O recurso a várias análises e a utilização de outros métodos multivariados é um processo de validar e clarificar as conclusões.

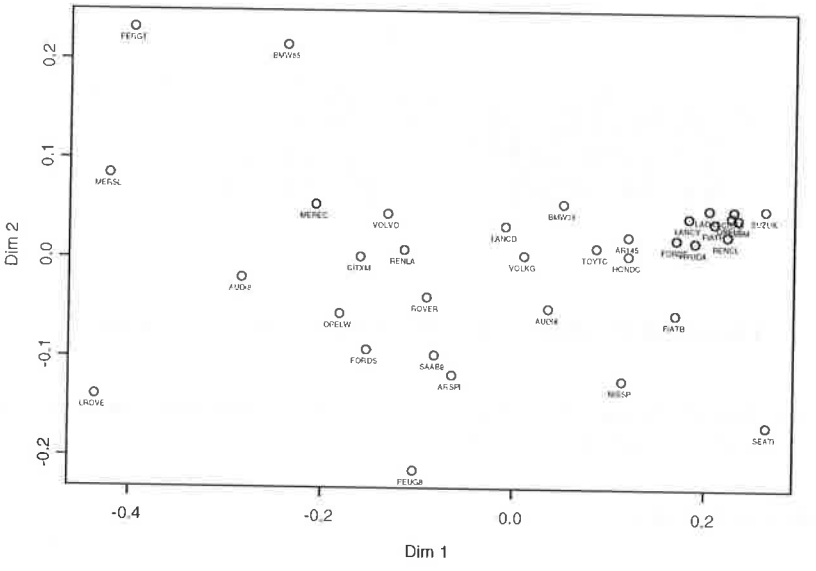


Figura 6.9 *Representação MDS para os dados de 35 automóveis.*

Tabela 6.5 *Dados relativos a 28 países da Europa (referentes a 2001).*

	Área Terrestre (km^2)	População Média (mil hab.)	T. de desemprego (%)	T. de Inflação (%)	PIB per capita (Euros)	Exp. de bens e serviços (%)	Veículos passag. (/mil hab.)	Telef. móveis (/mil hab.)	Util. de Internet (/mil hab.)	Peso da Agric. no VAB (%)
Bélgica	30538	10285	7.3	1.6	25260	85	46	75	28	1.5
Rep. Checa	78866	10283	7.3	1.4	13700	71	34	68	14	4.2
Dinamarca	43094	5359	4.5	2.4	26660	45	35	74	45	2.9
Alemanha	357031	82350	8.2	1.3	24000	35	53	68	37	1.2
Estónia	45227	1364	9.1	3.6	9240	91	30	54	32	5.8
Grécia	131957	10582	10.3	3.9	15020	23	34	56	13	7.0
Espanha	505124	40266	11.4	3.6	19510	30	45	74	18	3.4
França	549087	59191	8.7	1.9	23870	28	48	63	27	2.8
Irlanda	70295	3854	4.4	4.7	27360	98	36	–	23	3.5
Itália	301338	57075	9.1	2.6	23860	28	58	–	28	2.7
Chipre	9251	762	5.3	2.8	17180	47	37	41	20	4.0
Letónia	64589	2355	12.9	2.0	7750	45	25	26	7	4.7
Lituânia	65300	3478	13.1	0.4	8960	50	32	28	7	7.1
Luxemburgo	2586	442	2.4	2.1	44160	152	62	70	34	0.6
Hungria	93030	10188	5.6	5.2	12250	61	24	31	15	4.3
Malta	316	393	7.5	2.2	–	88	50	61	25	2.4
Holanda	35518	16046	2.6	3.9	26670	65	41	–	33	2.7
Áustria	83858	8130	4.1	1.7	25740	52	50	76	32	2.3
Polónia	312685	38638	20.0	1.9	9410	28	27	25	10	3.8
Portugal	91916	10299	5.0	3.7	16059	31	50	78	35	3.6
Eslovénia	20273	1992	6.0	7.5	16210	60	44	76	30	3.1
Eslováquia	49035	5397	19.4	3.3	11200	73	24	40	17	4.6
Finlândia	338150	5188	9.1	2.0	24170	40	41	81	43	3.4
Suécia	449974	8896	4.9	2.0	23700	45	45	81	52	1.9
Reino Unido	244101	60004	5.1	1.3	23530	27	44	75	40	0.9
Bulgária	110910	7910	18.6	5.8	5710	56	26	20	7	13.7
Roménia	238391	22408	8.0	22.5	5560	34	14	20	5	14.6
Turquia	769604	68670	10.4	32.2	5230	34	7	22	4	11.3

Tabela 6.6 *Dados de nutrientes em Carnes e Peixes por cada 3 onças (A=Assado, C=Cozido, D=Defumado, E=Enlatado, F=Frito, FE=Fervido, G=Grelhado).*

Descrição dos alimentos	Energia (Kcal)	Proteína (g)	Gordura (g)	Cálcio (mg)	Ferro (mg)
Carne estufada	340	20	28	9	2.6
Hamburguer	245	21	17	9	2.7
Carne/A	420	15	39	7	2
Bife	375	19	32	9	2.6
Carne/E	180	22	10	17	3.7
Galinha/G	115	20	3	8	1.4
Galinha/E	170	25	7	12	1.5
Coração	160	26	5	14	5.9
Perna de ovelha/A	265	20	20	9	2.6
Mão de ovelha/A	300	18	25	9	2.3
Presunto/D	340	20	28	9	2.5
Porco/A	340	19	29	9	2.5
Porco/FE	355	19	30	9	2.4
Língua	205	18	14	7	2.5
Costeleta de vitela	185	23	9	9	2.7
Bluefish/C	135	22	4	25	0.6
Amêijoas/cru	70	11	1	82	6
Amêijoas/E	45	7	1	74	5.4
Siri/E	90	14	2	38	0.8
Haddock/F	135	16	5	15	0.5
Cavala/G	200	19	13	5	1
Cavala/E	155	16	9	157	1.8
Perca/F	195	16	1	14	1.3
Salmão/E	120	17	5	159	0.7
Sardinha/E	180	22	9	367	2.5
Atum/E	170	25	7	7	1.2
Camarão/E	110	23	1	98	2.6

Exercícios

6.1 Considere os dados dos planetas registados na Tabela 1.1.

(a) Faça uma análise de clusters usando as variáveis *distância ao Sol*, *translação* e *rotação*. Considere vários métodos de análise e use os dados brutos e os dados estandardizados. Comente os resultados obtidos.

(b) Responda ao que é pedido na alínea (a) mas servindo-se das variáveis *massa*, *densidade* e *gravidade*. Compare os resultados obtidos com os resultados da alínea (a) e diga o que pode concluir.

6.2 A Tabela 6.5 contém dados relativos aos antigos países da União Europeia (UE-15), aos países recém chegados à UE (UE-10) e ainda sobre três países, Bulgária, Roménia e Turquia, que se encontram em fase de adesão.

(a) Realize um estudo global, com base em métodos de análise de clusters, que inclua todos os países presentes na Tabela 6.5.

(b) Considere dois blocos de países, os antigos, UE-15, e os UE-10 que aderiram recentemente. Compare estes resultados parciais com o resultado global obtido em (a).

6.3 Considere os dados da Tabela 6.6.

(a) Realize um estudo global dos dados com base em métodos da análise de clusters. Considere métodos hierárquicos e não hierárquicos. Descreva brevemente o trabalho que realizou e faça uma análise comparativa e um comentário crítico às soluções que encontrou.

(b) Repita o estudo usando dados estandardizados.

(c) Faça o estudo pedido em (a) e (b) para os alimentos do grupo da carne.

(d) Faça o trabalho pedido na alínea (c) usando os alimentos do grupo do peixe.

(e) Termine com um comentário final sobre os resultados de todas as análises.

Referências Bibliográficas

Abelson, R.P. and Sermat, V. (1962). Multidimensional scaling of facial expressions. *Journal of Experimental Psychology*, **63**, 546–554.

Agresti, A. (1981). Measures of nominal-ordinal association. *Journal of the American Statistical Association*, **76**, 524-529.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*. Academic Press, New York.

Andrews, D.F. (1972). Plots of high dimensional data. *Biometrics*, **28**, 125–136.

Arabie, P. and Hubert, L.J. (1994). Cluster analysis in marketing research, in *Advanced Methods in Marketing Research* (R.P. Bagozzi, ed.), Blackwell, Oxford.

Arabie, P., Hubert, L.J. and De Soete, G. (1996). *Clustering and Classification*. World Scientific, Singapore.

Baroni-Urbani, C. and Buser, M.W. (1976). Similarity of binary data. *Systematic Zoology*, **25**, 251–259.

Bassab, W. de O., Miazaki, E.S. e de Andrade, D.F. (1990). *Introdução à Análise de Agrupamentos*. Associação Brasileira de Estatística. 9º. Simpósio Nacional de Probabilidade e Estatística, São Paulo.

Baulieu, F.B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, **6**, 223–246.

Branco, J.A. e Valente, S.M. (1996). Análise de proximidades construídas a partir de variáveis de tipos diferentes. Comunicação apresentada no IV Congresso Anual da Sociedade Portuguesa de Estatística, Funchal.

Burbank, F. (1972). A sequential space-time cluster analysis of cancer mortality in the United States: etiological implications. *American Journal of Epidemiology*, **95**, 393–417.

Cavali-Sforza, L.L. and Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution*, **21**, 550–570.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Wadsworth, Belmont.

Chambers, J.M. and Hastie, J. (1992) (eds.). *Statistical Models in S*.

Wadsworth and Brooks, Pacific Grove.

Cheetham, A.H. and Hazel, J.E. (1969). Binary (presence-absence) similarity coefficients. *Journal of Paleontology*, **43**, 1130–1136.

Chernoff, H. (1973). Using faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, **68**, 361–368.

Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society A*, **134**, 321–367.

Cronbach, L.J. and Gleser, G.C. (1953). Assessing the similarity between profiles. *Psychological Bulletin*, textbf50, 456–473.

Cunningham, K.M. and Ogilvie, J.C. (1972). Evaluation of hierarchical grouping techniques: A preliminary study. *Computer Journal*, **15**, 209–213.

Davis, S.D. and Frolich, C. (1991). Single-link cluster analysis of earthquake aftershooks: decay laws and regional variations. *Journal of Geophysical Research*, **96**, 6335–6350.

De Sarbo, W.S., Manrai, A.K. and Manrai. L.A. (1993). Non-spatial tree models for the assessment of competitive market structure: an integrated review of the marketing and psychometric literature, in *Handbook in Operations Research and Managment Science: Marketing* (J. Eliashberg and G. Lilien, eds.), Elsevier, New York.

de Toit, S.H.C., Steyn, A.G.N. and Stumøf, R.H. (1986). *Graphical Exploratory Data Analysis*. Springer-Verlag, New York.

Diday, E. (1986). Orders and overlapping clusters by pyramids, in *Multidimensional Data Analysis* (J. de Leeuw, W. Heiser, J. Meulman and F. Critchley, eds.), DSWO Press, Leiden.

Duran, B.S. and Odell, P.L. (1974). *Cluster Analysis: A Survey*. Springer, New York.

Estabrook, C.G. and Rodgers, D.G. (1966). A general method of taxonomic description for a computed similarity measure. *Bioscience*, **16**, 789–793.

Everitt, B.S. (1978). *Graphical Techniques for Multivariate Data*. North-Holland, New York.

Everitt, B.S. and Dunn, G. (2001). *Applied Multivariate Data Analysis* (2nd ed.). Arnold, London

Everitt, B.S., Landau, S. and Leese, M. (2001). *Cluster Analysis* (4th ed.). Arnold, London.

Feldman, J. (1995). Perceptual models of small dot clusters, in *Partitioning Data Sets* (I. Cox, P. Hansen and B. Julesz, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 19, American Mathematical Society, Providence.

Fleiss, L.L. and Zubin, J. (1969). On the methods and theory of clustering. *Multivariate Behavioral Research*, **4**, 235–250.

Flury, B. and Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymmetric faces. *Journal of the American Statistical Association*, **76**, 757–765.

Friedman, H.P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, **62**, 1159–1178.

Gordon, A.D. (1994). Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, **18**, 561–581.

Gordon, A.D. (1999). *Classification* (2nd ed.). Chapman & Hall/CRC, Boca Raton.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–872.

Gower, J.C. and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficientes. *Journal of Classification*, **5**, 5–48.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques.* Academic Press, San Diego.

Hands, S. and Everitt, B. (1987). A Monte Carlo study of the recover of cluster structure in binary data by hierarchical clustering techniques. *Multivariate Behavioral Research*, **22**, 235–243.

Harman, H.H. (1976). *Modern Factor Analysis.* The University of Chicago Press, Chicago.

Hartigan, J.A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association*, **62**, 1140–1158.

Hartigan, J.A. (1975). *Clustering Algorithms.* Wiley, New York.

Hartigan, J.A. (1996). Introduction, in *Clustering and Classification* (P. Arabie, L.J. Hubert and G. De Soete, eds.), World Scientific, Singapore.

Hodson, F.R. (1971). Numerical typology and prehistoric archaelogy, in *Mathematics in the Archaelogical and Historical Sciences* (F.R. Hodson, D.G. Kendall and P.A. Tauter, eds.), Edimburgh University Press, Edimburgh.

Hodson, F.R., Sneath, P.H.A. and Doran, J.E. (1966). Some experiments in the numerical analysis of archaelogical data. *Biometrika*, **53**, 311-324.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data.* Prentice-Hall, Englewood Cliffs.

Jambu, M. (1978). *Classification Automatique pour l'Analyse des Données.* North-Holland, Amsterdam.

Jardine, C.J., Jardine, N. and Sibson, R. (1967). The structure and construction of taxonomic hierarchies. *Mathematical Bioscience*, **1**, 173–

179.

Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy.* John Wiley, London.

Jobson, J.D. (1992). *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods.* Springer-Verlag, New York.

Johnson, D.E. (1998). *Applied Multivariate Methods for Data Analysts.* Brooks /Cole, Pacific Grove.

Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Stastical Analysis* (5th ed.). Prentice Hall, Upper Saddle River.

Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32**, 241–245.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis.* Wiley, New York.

Kendall, M. G. (1955). *Rank Correlation Methods* (2nd edition). Charles Griffin, London.

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis: A User's Perspective.* Oxford University Press, Oxford.

Lance, G.N. and Williams, W.T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Computer Journal*, **9**, 373–380.

Legendre, L. and Legendre, P. (1982). *Numerical Ecology.* Elsevier, Amsterdam.

Lerman, I.C. (1987). Construction d'un indice de similarité entre objects décrits par des variables d'un type quelconque. Application au probléme du consensus en classification (1). *Revue de Statistique Appliquée*, **25**, 39–60.

Lock, R.H. (1993). 1993 New Car Data. *Journal of Statistical Education.* **1**(1).

Marriott, F.H.C. (1971). Practical problems in a method of cluster analysis. *Biometrics*, **27**, 501–514.

McLachlan, G.J. and Basdorf, K.E. (1988). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

Milligan, G.W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45**, 325–342.

Milligan, G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Milligan, G.W. and Cooper, M.C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, **5**, 181–204.

Mirkin, B. (1996). *Mathematical Classification and Clustering.* Kluwer Academic Publishers, Dordrecht.

Parmigiani, G., Garrett, E.S., Irizarry, R.A. and Zeger, S.L. (2003). The analysis of gene expression data: an overview of methods and software, in *The Analysis of Gene Expression Data: Methods and Software* (G. Parmigiani, E.S. Garrett, R.A. Irizarry and S.L. Zeger, eds.). Springer, New York.

Petrakis, E.G.M. and Faloutsos, C. (1997). Similarity searching in medical image databases. *IEEE Transactions on Knowledge and Data Engineering*, **9**, 435–447.

Price, L.J. (1993). Identifying cluster overlap with NORMIX population membership probabilities. *Multivariate Behavioral Research*, **28**, 235–262.

Punj, G. and Stewart, D.W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*, textbf20, 134–148.

Romesburg, H.C. (1984). *Cluster Analysis for Researchers*. Lifetime Learning Publications, California.

Shepard, R.N. and Arabie, P. (1979). Additive clustering: representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, **86**, 87–123.

Smart, R.G., Asbridge, M., Mann, R.E. and Adlaf, E.M. (2003). Psychiatric distress among road rage victims and perpetrators. *Canadian Journal of Psychiatry*, **48**, 681–688.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

Sokal, R.R. and Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*. Freeman, San Francisco.

Späth, H. (1980). *Cluster Analysis Algorithms*. Ellis Horwood, Chichester.

Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire.

Venables, W.N. and Ripley, B.D. (1999). *Modern Applied Statistics with S-PLUS* (2nd edition). Springer-Verlag, New York.

Wang, P.C.C. (1978). *Graphical Representation of Multivariate Data*. Academic Press, New York.

Ward, J.H. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

Wardlaw, R.L., Frolich, C. and Davis, S.D. (1991). Evaluation of precursory seismic quiescence in sixteen subduction zones using single-link cluster analysis. *Pure and Applied Geophysics*, **134**, 57–78.

Willet, P. (1990). *Parallel Database Processing, Text Retrieval and Cluster Analysis*. Pitman Publishers, London.

Wish, M. (1971). Individual differences in perceptions and preferences

among nations, in *Attitude Research Reaches New Heights* (C.W. King and T. Tigert, eds.), American Marketing Association, Chicago.

Wish, M., Deutsch, M. and Biener, L. (1970). Differences in conceptual structures of nations: an exploratory study. *Journal of Personality and Social Psychology*, **16**, 361–373.

ISBN 972-98619-9-4
Depósito Legal nº 213286/04