Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Lattin, James M.
Análise de dados multivariados / James M. Lattin, J. Douglas Carroll, Paul E. Green ; [tradução Harue Avritscher]. -- São Paulo : Cengage Learning, 2011.

Título original: Analyzing multivariate data.
ISBN 978-85-221-0901-2

1. Análise multivariada I. Carroll, J. Douglas. II. Green, Paul E. III. Título.

10-12603 CDD-519.535

Índice para catálogo sistemático:

1. Análise de dados multivariados : Matemática 519.535

Análise de Dados Multivariados

JAMES M. LATTIN
Pós-Graduação na Business School da Stanford University

J. DOUGLAS CARROLL
Pós-Graduação na Escola de Administração da Rutgers University

PAUL E. GREEN
Wharton School, Pennsylvania State University

Revisão Técnica
FLAVIO SOARES CORRÊA DA SILVA
PhD em Inteligência Artificial pela Edinburgh University, livre-docente e professor associado do Departamento de Ciência da Computação no Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP)

CENGAGE Learning™

Austrália • Brasil • Japão • Coreia • México • Cingapura • Espanha • Reino Unido • Estados Unidos

8

Análise de agrupamentos

8.1 INTRODUÇÃO

Em termos gerais, a *análise de agrupamentos* envolve categorização: dividir um grande grupo de observações em grupos menores para que as observações dentro de cada um deles sejam relativamente similares (isto é, para que elas possuam, na maior parte, as mesmas características) e as observações em diferentes grupos sejam relativamente dissimilares. Em muitos aspectos, a análise de agrupamentos apresenta um relacionamento muito próximo ao escalonamento multidimensional (MDS). Ambos podem ser vistos como metodologias de construção de representações de objetos baseadas em suas similaridades ou dissimilaridades (ou outras medidas de proximidade). A principal diferença é que o MDS constrói representações *espaciais e contínuas*, enquanto o agrupamento constrói representações *não espaciais e discretas* (por exemplo, divisão em conjuntos sobrepostos ou não sobrepostos). O MDS nos fornece variáveis de *escala intervalar* que nos informam sobre coordenadas que definem a localização de cada objeto no espaço; a análise de agrupamentos fornece-nos variáveis de *escala nominal* que indicam se cada objeto pertence ou não pertence a cada um de determinado número de agrupamentos.

A maior parte da análise de agrupamentos é realizada com o objetivo de se tratar da heterogeneidade dos dados. Em vez de lidar com um grupo de observações amplamente divergentes, dividimos explicitamente o grupo em subconjuntos mais homogêneos. No entanto, separar os dados em subgrupos mais homogêneos *não* é a mesma coisa de encontrar agrupamentos que ocorram naturalmente. Por exemplo, com uma série de números uniformemente distribuídos entre zero e um, pode-se tratar da heterogeneidade separando-se os dados em dois intervalos: de zero a 0,5 e de 0,5 a 1. Embora os dados em cada intervalo sejam mais similares (por exemplo, a variância interna do grupo diminui por um fator de quatro), os dois intervalos não correspondem a quaisquer agrupamentos de observações claramente separados. Encontrar agrupamentos que ocorrem naturalmente exige que haja grupos de observações com densidade local relativamente alta (isto é, há muitas outras observações dentro da mesma pequena área) separados por regiões de densidade local relativamente baixa. Nesse caso, os próprios agrupamentos correspondem a uma modalidade de dados, e o *número* de agrupamentos corresponde ao *número* de modas em uma distribuição multimodal de dados. Através do uso de exemplos e aplicações, esperamos tornar clara a distinção entre tratar a heterogeneidade (o que é sempre

possível, embora não necessariamente desejável) e encontrar os agrupamentos naturais, o que somente é possível quando a modalidade dos dados (isto é, o número de modas na distribuição subjacente) é maior que um.

Muitas abordagens diferentes para a análise de agrupamentos têm sido desenvolvidas. Neste capítulo, discutiremos exemplos de dois tipos diferentes: métodos *hierárquicos* (cujo resultado é representado como uma estrutura hierárquica de árvore, em que a solução de k agrupamento é formada pela junção de dois agrupamentos da solução de agrupamento $k + 1$) e métodos de *partição* (que separam as observações em um número determinado de subgrupos, e em que a solução de k agrupamento e a solução de agrupamento $k + 1$ não são necessariamente aninhadas). Olhando para o conjunto de soluções de agrupamento fornecido por esses métodos e focando na solução de k agrupamento, a atribuição resultante de objetos aos agrupamentos é mutuamente exclusiva (isto é, nenhum objeto é atribuído a mais de um agrupamento) e coletivamente exaustiva (isto é, todos os objetos são atribuídos a algum agrupamento). Há também métodos que podem levar a uma solução de k agrupamento com agrupamentos sobrepostos (em que um objeto é atribuído a mais de um agrupamento e os agrupamentos não são necessariamente aninhados para formar uma árvore hierárquica) e agrupamentos indistintos (em que as atribuições de um objeto para um agrupamento é um número entre zero e um); esses métodos não são aqui considerados.

Os métodos hierárquicos geralmente abordam a análise de dados por intermédio de um destes dois modos: de baixo para cima (chamados métodos *aglomerativos*, começando com cada observação em um agrupamento separado e juntando-se os agrupamentos a cada etapa do processo até que reste somente um agrupamento de tamanho n) ou de cima para baixo (chamados métodos *divisivos*, começando com todas as observações em um único agrupamento e dividindo-se um agrupamento em dois a cada etapa do processo até que restem n agrupamentos de tamanho um). Alguns métodos não são nem aglomerativos nem divisivos (por exemplo, várias abordagens que usam mínimos quadrados para ajustar certas estruturas de árvore). Nossa discussão de métodos de agrupamento hierárquico, neste capítulo, foca principalmente o agrupamento aglomerativo. Métodos hierárquicos de agrupamentos dependem em alguma medida da proximidade entre as observações (avaliada diretamente ou derivada de dados atribuídos). Em contraste, a maioria dos métodos de partição requer atributos de dados, porque frequentemente é necessário calcular a proximidade de cada observação ao centroide de cada agrupamento.

Quase todos os problemas de agrupamento de qualquer tamanho apreciável exigem uma solução heurística. Isso se dá porque, à medida que o número de objetos no conjunto de dados aumenta, o número de soluções de agrupamento possíveis cresce espetacularmente. O número de diferentes modos de dividir n objetos em m agrupamentos de tamanho $n_1, n_2, n_3, \ldots, n_m$ é dado por

$$n!/[n_1!n_2!n_3!\ldots n_m!m!]$$

Para se ter uma ideia da magnitude desse número, considere um problema de agrupamento envolvendo 20 objetos (um problema de agrupamento de tamanho moderado, na realidade). O número de maneiras diferentes de dividir esses objetos em quatro agrupamentos de tamanho igual (isto é, $n_1 = n_2 = n_3 = n_4 = 5$) é maior que 488 milhões! Isso sequer começa a explicar todas as outras soluções possíveis de quatro agrupamentos (isto é, há mais de 100 modos de dividir 20 objetos em 4 conjuntos não vazios; por exemplo, $n_1 = 3$, $n_2 = 4$, $n_3 = 6$ e $n_4 = 7$), sem mencionar todas as soluções possíveis de dois agrupamentos, de três agrupamentos, de cinco agrupamentos e assim por diante. Como não é computacionalmente viável buscar através de todas as possíveis partições para que se encontre a "melhor" solução, a análise de agrupamentos adota uma abordagem em grande parte heurística, baseada em algoritmos fáceis de programar, extremamente eficientes e que fornecem soluções alinhadas com nossos objetivos (ainda que nem sempre a melhor solução possível). Neste capítulo, explicamos os métodos heurísticos para os agrupamentos de ligação individual (que exemplifica as abordagens aglomerativas discutidas neste capítulo) e para os agrupamentos de K-means (que constitui uma abordagem possível à partição). Finalmente, discutimos os desafios associados com a validação das soluções de agrupamentos.

8.1.1 Aplicações potenciais

Taxonomia numérica

Alguns dos primeiros exemplos de análise de agrupamentos ocorreram em biologia evolucionária e ecológica, quando os cientistas tentavam identificar e discriminar diferentes espécies e subespécies de plantas e animais de acordo com a similaridade relativa de suas características físicas. Um dos exemplos mais famosos envolveu um conjunto de dados para três diferentes espécies de íris: *Iris setosa*, *Iris versicolor* e *Iris virginica*. Cinquenta plantas de cada uma das três espécies foram coletadas e enumeradas as seguintes medidas: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. Um gráfico do comprimento da pétala *versus* a largura da pétala na Figura 8.1 mostra que as três diferentes espécies formam três agrupamentos que ocorrem naturalmente, com base nessas características físicas específicas. Em particular, está claro que duas das espécies (*Iris versicolor* e *Iris virginica*) estão mais proximamente relacionadas em termos de suas similaridades físicas. A maioria dos trabalhos em taxonomia numérica envolve ajustar árvores taxonômicas (hierárquicas) para descrever essas relações, que podem ser vistas frequentemente como modelos do processo evolucionário.

Segmentação de mercado

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

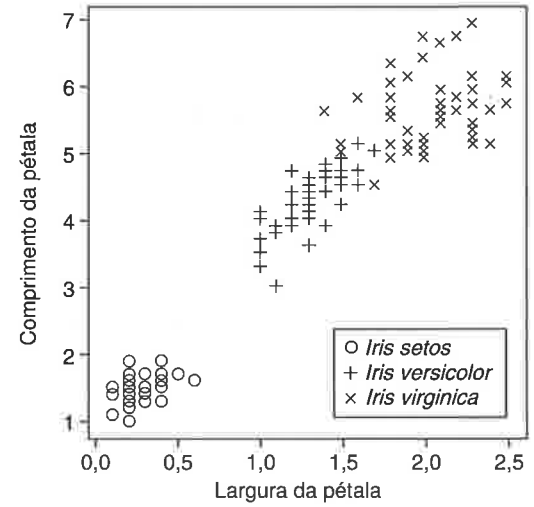


Figura 8.1 Gráfico com dados das íris de Fisher (largura da pétala *versus* comprimento da pétala).

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

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

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

	Anchor Steam	Bass Ale	Beck's	Corona	Gordon--Biersch	Guinness	Heineken	Pete's Wicked Ale	Sam Adams	Sierra Nevada
S001	5	9	7	1	7	6	6	5	9	5
S008	7	5	6	8	8	4	8	8	7	7
S015	7	7	5	6	6	1	8	4	7	5
S022	7	7	5	2	5	8	4	6	8	9
S029	9	7	3	1	6	8	2	7	6	8
S036	7	6	4	3	7	6	6	5	4	9
S043	5	5	5	6	6	4	7	5	5	6
S050	5	3	1	5	5	5	3	5	5	9
S057	9	3	2	6	4	6	1	5	3	6
S064	2	6	6	5	6	4	8	4	4	3
S071	7	7	7	5	7	8	6	7	7	8
S078	8	3	3	9	9	2	1	9	7	8
S085	6	5	3	7	6	5	8	6	7	5
S092	5	6	3	8	6	7	6	7	6	7
S099	4	7	2	8	5	9	8	3	8	8
S106	3	3	4	5	6	5	9	7	5	5
S113	2	4	5	7	6	6	8	1	7	4
S120	9	3	7	4	2	4	6	3	8	6
S127	5	3	4	7	7	7	6	6	6	6
S134	2	4	4	8	5	5	5	4	6	6
S141	5	7	6	7	5	8	8	7	5	7
S148	8	9	6	7	7	8	6	8	8	8
S162	5	6	6	7	5	3	7	3	4	3
S169	5	5	6	7	7	4	6	3	7	6
S176	5	5	7	8	7	6	7	5	4	7
S183	3	5	4	7	3	1	2	6	6	5
S190	4	3	6	8	6	1	8	2	7	7
S197	3	8	4	8	6	2	8	4	6	1
S204	3	5	1	5	5	3	4	6	7	5
S211	3	8	5	8	7	5	5	3	7	8
S218	8	8	5	7	9	9	7	7	6	8
S225	7	6	2	2	6	6	2	7	5	5

indicaram não ter preferência por nenhuma marca de cerveja; nenhum desses respondentes está incluído nessa subamostra específica.) Usamos, então, o agrupamento K–means para dividir os dados.

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

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

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

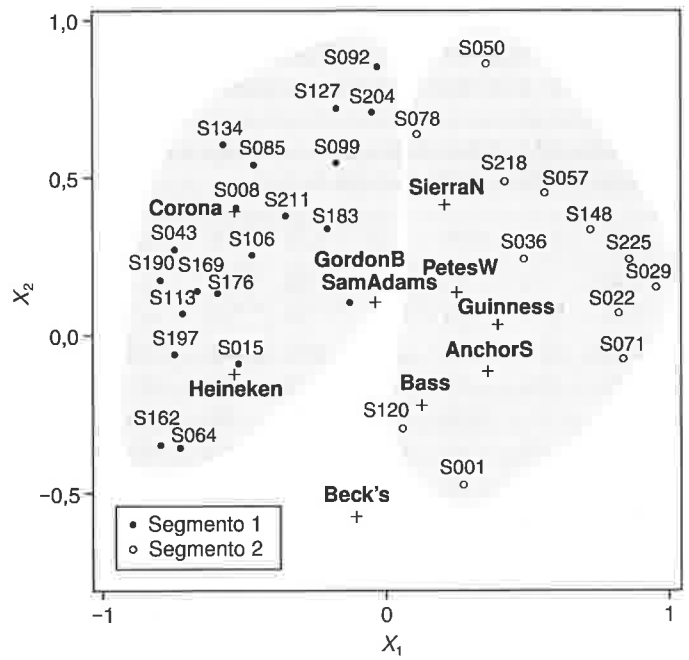


Figura 8.2 Mapa MDPREF de preferências dos estudantes por cerveja mostrando segmentação em dois agrupamentos.

inclinam-se menos favoravelmente às cervejas claras (isto é, eles atribuem 4,3 à Corona e 4,2 à Heineken).

Essas diferenças tornam-se mais claras quando é efetuada a sobreposição da segmentação implicada pela partição de agrupamento no mapa MDPREF da Figura 7.2. A Figura 8.2 mostra a separação entre os dois agrupamentos. À esquerda estão os estudantes cujas preferências vão para as cervejas mais claras, enquanto à direita estão os estudantes que preferem cervejas escuras ou mais amargas.

Análise da estrutura do mercado

As companhias precisam, com frequência, conhecer não somente quais consumidores estão comprando seu produto, mas também quais outros produtos estão no conjunto competitivo. Uma maneira de fazer isso é pela análise da estrutura do mercado e posterior identificação dos grupos de produtos similares, de acordo com medidas competitivas de similaridade. Uma dessas medidas de similaridade competitiva é uma matriz de comutação. A ideia é que observamos mais trocas entre concorrentes próximos (isto é,

alternativas diferentes que têm maior chance de ser escolhidas pelo mesmo consumidor) do que entre produtos não concorrentes.

Bucklin e Lattin (1992) estudaram padrões de competição entre categorias de produtos entre os varejistas. Como parte de seu estudo, examinaram o comportamento de compra de 300 famílias escolhidas aleatoriamente (extraídas de um painel da A. C. Nielsen em Sioux Falls, Dakota do Sul) durante um período de 52 semanas. Essas famílias foram às compras 30.966 vezes no total, em 13 lojas de Sioux Falls. O padrão agregado de mudança de lojas está resumido na matriz da Tabela 8.3.

EXEMPLO *Usando padrões de mudança de lojas para um modelo de competição entre estabelecimentos varejistas (Bucklin e Lattin, 1992) STORE_SWITCH*

Bucklin e Lattin usaram análise de agrupamento para analisar a estrutura do mercado competitivo implicado pela matriz de comutação na Tabela 8.3. Depois de ajustar a matriz de comutação para explicar as diferenças devidas ao tamanho das lojas (com base no número total de compras, as lojas variam de menos de 1.000 visitas até mais de 5.000). Bucklin e Lattin realizaram uma análise de agrupamentos usando o método de Ward (discutiremos esse método mais adiante, neste capítulo). Os resultados, superpostos sobre um mapa da cidade na Figura 8.3, sugerem que os padrões de competição entre as lojas são impulsionados principalmente pela geografia (isto é, pela localização da loja).

Tabela 8.3 A matriz mostra os padrões de mudança (da loja da linha para a loja da coluna) de 300 famílias entre 13 lojas em Sioux Falls, Dakota do Sul

	S03	S04	S07	S16	S18	S21	S24	S26	S29	S36	S10	S43	S45
S03	1428	52	199	5	55	12	19	422	14	167	37	32	14
S04	15	1001	256	227	20	26	33	33	162	23	76	185	95
S07	84	96	2553	124	292	214	655	165	227	122	329	206	130
S16	1	80	29	256	23	27	37	21	82	2	67	206	44
S18	34	11	164	11	910	112	105	98	50	42	102	50	20
S21	13	19	123	10	63	1294	70	59	187	44	60	88	61
S24	17	47	364	7	39	30	803	55	119	1	248	236	126
S26	506	29	125	10	47	37	15	1983	67	132	121	175	30
S29	22	162	219	44	48	87	43	49	1590	35	73	268	139
S36	78	3	29	14	93	100	12	220	48	868	29	20	15
S10	73	82	330	17	52	34	76	52	13	34	441	370	117
S43	113	299	509	102	42	49	123	99	113	36	70	1040	450
S45	62	243	358	54	24	74	97	28	116	15	47	100	455

Fonte: Bucklin e Lattin, 1992.

8.2 OBJETIVOS DA ANÁLISE DE AGRUPAMENTOS

Vale a pena repetir que o objetivo mais comum da análise de agrupamentos talvez seja tratar da heterogeneidade nos dados. O resultado esperado é um pequeno número (administrável) de grupos, cada um consistindo em um número de objetos relativamente homogêneos com uma variação dentro do grupo consideravelmente menor do que o total de variação no conjunto completo de dados. Mas há outro objetivo, ligeiramente diferente, que também motiva o uso da análise de aglomerados: encontrar uma modalidade natural de dados. Nesse caso, usa-se a análise de agrupamentos para determinar se os dados contêm subconjuntos homogêneos de observações que ocorrem naturalmente.

Para ilustrar as diferenças entre esses dois objetivos, introduzimos dois conjuntos de amostra de dados similares na quantidade de heterogeneidade que encarnam, mas diferentes em termos de sua modalidade natural. O primeiro conjunto de dados consiste em 50 observações sobre duas variáveis X_1 e

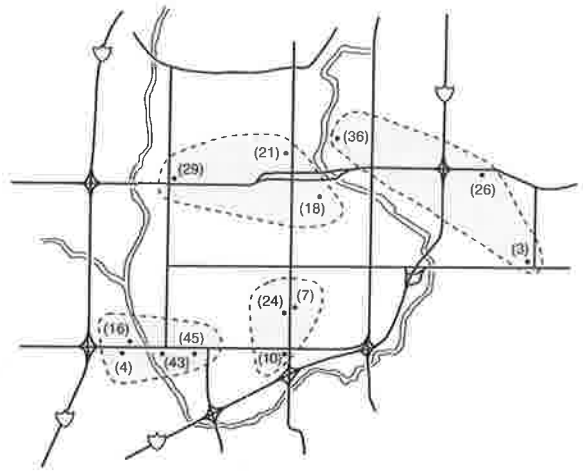


Figura 8.3 Agrupamentos de lojas baseados no comportamento de mudança de lojas dos consumidores (Fonte: Lattin e Bucklin, 1992). Reimpresso com permissão dos autores.

X_2. Da amostra total, são retiradas 25 observações de uma distribuição normal bivariada com centroide (3,3) e matriz de covariância igual à matriz identidade **I** (isto é, X_1 e X_2 são não correlacionadas com variância igual a 1,0); outras 25 observações são retiradas de uma distribuição normal bivariada com centroide (6,6) e a mesma matriz de covariância. O segundo conjunto de dados consiste também em 50 observações sobre X_1 e X_2; neste caso, todas as 50 são retiradas de uma distribuição normal bivariada com centroide (4,5, 4,5) e matriz de covariância igual a

$$\sum = \begin{bmatrix} 2{,}25 & 1{,}50 \\ 1{,}50 & 2{,}25 \end{bmatrix}$$

Os diagramas de dispersão são apresentados nas Figuras 8.4 e 8.5. Em uma inspeção visual rápida, as duas configurações parecem similares. A diferença é que a primeira é retirada de uma população bimodal por natureza (como apresentado pelas linhas pontilhadas do contorno na figura), enquanto a modalidade natural da população subjacente à segunda amostra é unimodal.

Se nosso objetivo for reduzir a heterogeneidade, podemos fazê-lo para cada conjunto de dados dividindo os dados em agrupamentos (nesse caso, uma solução de dois agrupamentos) ao longo das linhas apresentadas na Figura 8.6. Independente da modalidade natural dos dados, ainda terminamos

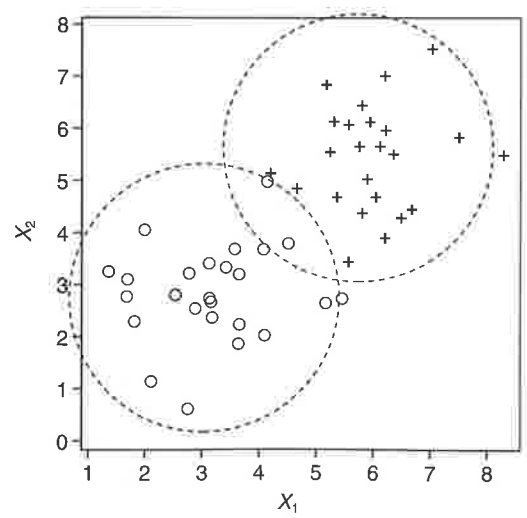


Figura 8.4 Diagrama de dispersão dos dados da amostra de população bimodal (as linhas pontilhadas mostram os contornos da densidade).

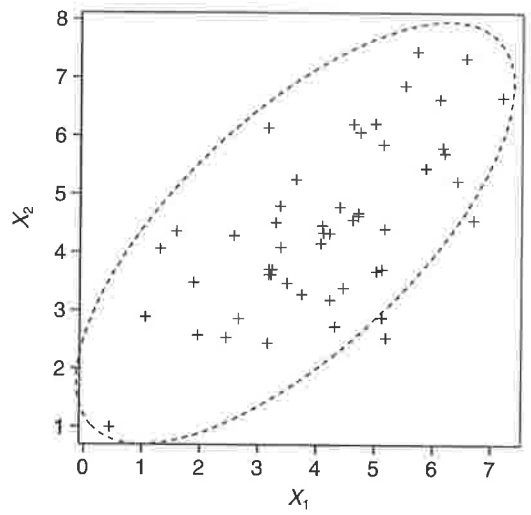
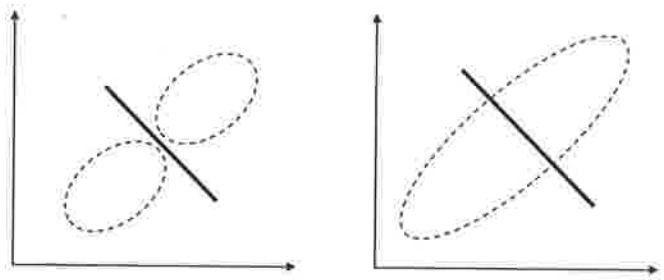


Figura 8.5 Diagrama de dispersão dos dados da amostra de distribuição unimodal (as linhas pontilhadas mostram o contorno da densidade).

com agrupamentos que exibem homogeneidade interna maior. Em qualquer um dos casos, a distância média ao quadrado de cada ponto ao centroide do agrupamento (efetivamente, a variância interna do agrupamento) é muito menor que a distância média ao quadrado da média da amostra (isto é, a variância da amostra total).



(a) Amostra bimodal (b) Amostra unimodal

Figura 8.6 Soluções de dois agrupamentos para amostras bimodal e unimodal.

Se nosso objetivo, no entanto, for encontrar a modalidade natural dos dados, necessitamos então de uma abordagem que dividirá a primeira amostra em dois agrupamentos, mas fará da segunda amostra um grande agrupamento. Com o desenvolvimento da discussão sobre os diferentes métodos de agrupamentos, retornaremos a esse exemplo para determinar como cada método funciona com respeito aos objetivos de reduzir-se a heterogeneidade e de se encontrar a modalidade natural.

8.3 MEDIDAS DE DISTÂNCIA, DISSIMILARIDADE E DENSIDADE

Na análise de agrupamentos, nosso foco está nas observações do conjunto de dados. Quando se usam métodos de agrupamentos hierárquicos aglomerativos, precisamos agrupá-los com base na proximidade mútua ou na similaridade. Essa medida de proximidade pode vir de duas fontes: podemos usar uma avaliação direta da proximidade (como fizemos com o escalonamento multidimensional) ou podemos calcular uma medida derivada das classificações dos atributos para cada observação (isto é, do mesmo tipo de dados que tínhamos à disposição nos componentes principais e na análise fatorial). Há várias opções diferentes quando se trata de calcular uma medida derivada de proximidade ou similaridade.

8.3.1 Medidas de distância

Distância euclidiana

Quando as variáveis em estudo possuem propriedades métricas (isto é, elas são medidas em escalas de razão ou intervalares), um modo óbvio de se refletir a "proximidade" de dois objetos é com uma medida de distância. A mais familiar é a distância euclidiana, que é definida da seguinte maneira:

$$d_{ij} = \left[\sum_k (x_{ik} - x_{jk})\right]^{1/2} \tag{8.1}$$

onde d_{ij} é a distância euclidiana entre os objetos i e j. Como as variáveis X são frequentemente medidas em unidades diferentes, a fórmula da distância na Equação (8.1) acima é aplicada mais comumente para dados padronizados. Com isso, atribui-se peso igual a cada variável padronizada, de maneira que elas sejam igualmente importantes (após a padronização) na determinação da proximidade relativa dos objetos no espaço.

Uma exceção possível a essa regra se dá quando são utilizados componentes principais como contribuição para a análise de agrupamentos. Nesse caso, deve ser concedido ao primeiro componente principal (que explica a maior parte das informações nos dados) um peso maior na determinação da proximidade relativa de quaisquer dois objetos no conjunto de dados. Especificamente, os autovetores que definem os componentes principais devem ser ponderados pelas raízes quadradas de seus autovalores associados para que as variâncias dos respectivos componentes principais sejam iguais à variância considerada para (VAF) por esses componentes nos dados originais.

Métrica p (ou L_p) de Minkowski

Embora a distância euclidiana seja a mais familiar, outras distâncias métricas podem ser mais apropriadas em alguns casos. Uma classe geral conhecida como *métrica p de Minkowski* (ou, às vezes, métrica L_p), é definida pela seguinte equação:

$$d_{ij}(p) = \left[\sum_k |x_{ik} - x_{jk}|^p\right]^{1/2} \tag{8.2}$$

Portanto, a equação para a distância euclidiana é apenas um caso particular da métrica p de Minkowski com $p = 2$. Outros dois casos especiais merecem atenção adicional. Um é o caso em que $p = 1$, que se reduz para:

$$d_{ij}(1) = \sum_k |x_{ik} - x_{jk}| \tag{8.3}$$

Essa medida de distância é, às vezes, chamada de *métrica do quarteirão* porque é como andar de um ponto A a um ponto B em uma cidade disposta em um sistema de "grades" de ruas, todas formando ângulos retos entre si: a distância entre os dois pontos é a soma dos trechos do percurso leste-oeste com o trecho do percurso norte-sul. Outro caso especial da métrica de Minkowski é o caso em que $p = \infty$, algumas vezes chamado de *sup-métrica*:

$$d_{ij}(\infty) = \text{máx}(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, \ldots, |x_{ip} - x_{jp}|) \tag{8.4}$$

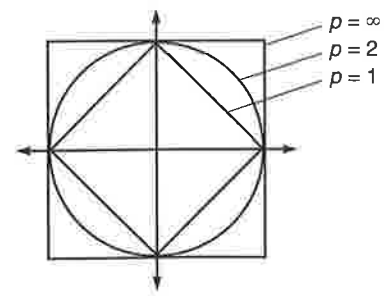


Figura 8.7 Círculos de unidade para três diferentes métricas de Minkowski.

Para ilustrar melhor as diferenças entre essas três distâncias métricas (quarteirão, euclidiana e sup-métrica), a Figura 8.7 mostra o gráfico de um "círculo" de unidade (isto é, o locus dos pontos localizados a uma unidade de distância da origem) em duas dimensões para cada métrica. Observe que o único valor de p para o qual a métrica p de Minkowski é invariante sob rígida rotação é $p = 2$ (isto é, o caso euclidiano).

Distância de Mahalanobis

As distâncias definidas pela métrica euclidiana (ou, nesse caso, qualquer uma das métricas p de Minkowski) não levam em consideração quaisquer padrões de covariância que existam nos dados. A métrica proposta por Mahalanobis é ajustada para a covariância, de acordo com a seguinte equação:

$$D_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j) \quad \textbf{(8.5)}$$

onde Σ é a matriz de covariância da população da matriz de dados $\mathbf{X}$. O resultado é uma medida de distância ao quadrado (euclidiana generalizada).

A interpretação do D^2 de Mahalanobis pode ser mais bem entendida no contexto de dados distribuídos de acordo com uma distribuição normal. A Figura 8.8 mostra um exemplo de pontos retirados de uma distribuição normal bivariada (centrada na origem) com covariância positiva. A elipse na figura representa o locus dos pontos equidistantes da origem (na distância de Mahalanobis, definido como a raiz quadrada do D^2 de Mahalanobis). Isso sugere que o ponto A, localizado no quadrante I do gráfico, está na mesma distância de Mahalanobis a partir da origem como o ponto B no quadrante II, ainda que o ponto B esteja mais próximo em termos da distância euclidiana comum e não ponderada.

O que o D^2 de Mahalanobis capta é o fato de que os pontos A e B têm a mesma probabilidade de terem sido extraídos de uma distribuição normal multivariada com centro em (0,0). Em outras palavras, A e B estão no mesmo contorno de isodensidade. Uma vantagem do D^2 de Mahalanobis é que ele "esfericiza" os dados de maneira eficaz (isto é, efetua uma rotação e reescalona os eixos da coordenada de tal modo que a distribuição multivariada resultante possua matriz de covariância igual à matriz identidade $\mathbf{I}$). Isso, às vezes, é útil quando o objetivo da análise de agrupamentos for o de gerar agrupamentos compactos e convexos.

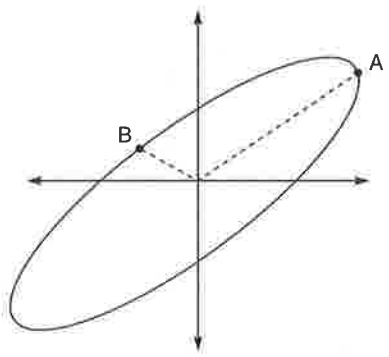


Figura 8.8 Diagrama demonstrando a distância de Mahalanobis: um contorno da isodistância em duas dimensões.

Outras medidas

Medidas de correspondência. Com frequência, lidamos com dados medidos somente em escala nominal, em cujo caso não é apropriado calcular a medida de distância. A abordagem usual para tais dados de escala nominal é baseada em correspondência de atributos. Falando intuitivamente, dois perfis são vistos como similares com base na extensão com que partilham atributos comuns.

Como exemplo, considere quatro refrigerantes: Coca-Cola, Pepsi, Diet Coke e Diet Coke Sem Cafeína avaliados por quatro atributos discretos: é uma cola? Contém cafeína? É uma bebida dietética? É uma marca da Coca-Cola? O perfil de cada refrigerante é apresentado na Tabela 8.4, caracterizados por possuir ou não cada um dos quatro atributos. Uma medida simples de correspondência para cada

Tabela 8.4 Perfis de quatro marcas de refrigerantes em relação a quatro atributos

	Sabor de cola	Cafeína	Diet	Produzido pela Coca-Cola
Coca-Cola	1	1	0	1
Pepsi	1	1	0	0
Diet Coke	1	1	1	1
Diet Coke Sem Cafeína	1	0	1	1

par de objetos (i, j), apresentada na Tabela 8.5, é obtida contando-se o número de correspondências – em que ambos os objetos possuem o atributo ou nenhum objeto possui o atributo – e dividindo-se pelo número total de atributos. Com base nessa medida, Coca-Cola e Pepsi estão relativamente próximos, combinando em três dos quatro atributos (isto é, ambos são colas e contêm cafeína e nenhum dos dois é uma bebida dietética). Por sua vez, a Diet Coke é tão próxima da Coca-Cola quanto a Pepsi, porque também combina com ela (com a Coca-Cola) em três atributos. A diferença é que a Coca-Cola e a Pepsi têm correspondência no atributo dietético, enquanto a Coca-Cola e a Diet Coke possuem correspondência em relação ao atributo marca Coca-Cola. Note que uma modificação posterior da medida de correspondência é possível (por exemplo, peso desigual) se um dos atributos for relativamente mais importante na determinação da proximidade do objeto.

Observe que a medida de correspondência apresentada na Tabela 8.5 é uma medida de similaridade, e não uma medida de dissimilaridade (isto é, quanto maior o valor, maior a similaridade entre os dois objetos).

Tabela 8.5 Medidas de similaridade de quatro marcas de refrigerantes com base em atributos que tenham correspondência

	Coca-Cola	Pepsi	Diet Coke	Diet Coke Sem Cafeína
Coca-Cola				
Pepsi	3/4			
Diet Coke	3/4	2/4		
Diet Coke Sem Cafeína	2/4	1/4	3/4	

Cautela sobre correlação. Observe que um coeficiente de correlação nem sempre é uma medida apropriada de similaridade. A correlação é uma medida de covariância que é também um tipo de proximidade, mas não necessariamente de similaridade. Considere duas observações com os seguintes perfis: (1, 2, 1, 2) e (9, 10, 9, 10). Nesses quatro atributos, as duas observações exibem uma correlação de 1,0 e ainda assim não estão próximas em termos do nível expresso por cada atributo. Considere um par diferente de observações com os perfis (1, 2, 1, 2) e (1, 1, 2, 2): essas duas observações são mais próximas em termos do nível e ainda assim exibem uma correlação de 0,0. Antes de usar uma matriz de correlação como ponto de partida de uma rotina de agrupamentos, é importante verificar se os dados são dimensionados de modo que os resultados da análise possam ser interpretados apropriadamente (por exemplo, se são padronizados por observação).

Medidas de densidade

Com exceção da distância de Mahalanobis (que leva em consideração a estrutura de covariância dos dados), todas as medidas que discutimos não levam em conta o contexto. Em outras palavras, as medidas de proximidade entre os objetos i e j são as mesmas, independente da posição de i e j em relação a outros objetos no conjunto de dados. Isso ocasionalmente torna-se um problema com algumas rotinas de agrupamento, as quais são frequentemente míopes por natureza. Demonstremos como. A Figura 8.9 representa um diagrama de dispersão de pontos em que uma estrutura de dois agrupamentos é bem evidente pelo exame visual. Dois pares de pontos estão destacados na figura: par A e par B. Observe que cada par de pontos está separado exatamente pela mesma distância (em termos

euclidianos). Entretanto, com respeito à nossa certeza sobre qual par pertence ao mesmo agrupamento, há diferenças importantes. Parece claro que os objetos no par A pertencem ao mesmo conjunto não somente porque estão próximos entre si, mas também porque estão na vizinhança de pontos que estão também todos próximos uns dos outros.

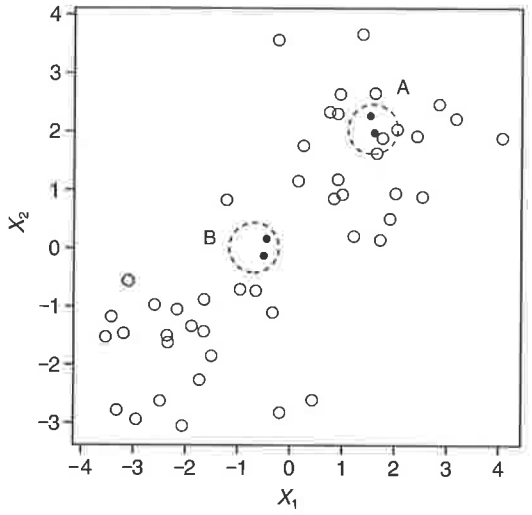


Figura 8.9 Dois pares diferentes de pontos: mesma distância, densidades diferentes.

Em outras palavras, os pontos no par A estão em uma região de alta densidade. Embora os pontos no par B estejam também próximos uns dos outros, não há outros objetos nos arredores. Esses dois pontos estão em uma região de baixa densidade.

Todas as rotinas de agrupamentos hierárquicos aglomerativos operam sequencialmente. Elas começam colocando objetos que são "próximos" entre si no mesmo agrupamento e prosseguem até que todos os objetos tenham sido reunidos em um único agrupamento. Essas rotinas são míopes no sentido de que não há retorno: uma vez que os objetos tenham sido reunidos em um único agrupamento, não podem mais ser separados. Nesse exemplo, em que estamos muito mais confiantes de que os pontos no par A pertencem ao mesmo agrupamento, necessitamos de uma medida que reflita as diferenças na densidade local da região que o cerca.

Uma possibilidade é o uso de uma medida baseada na k-ésima densidade de vizinhança mais próxima. A abordagem, detalhada abaixo, pode ser utilizada com avaliações diretas da distância (ou dissimilaridade) ou com medidas construídas indiretamente (por exemplo, a métrica euclidiana). Usamos a notação d^* para diferençar a medida de densidade de outros tipos de medida de distância descritos nesta seção.

ETAPAS PARA SE CALCULAR A K-ÉSIMA DENSIDADE DE VIZINHANÇA MAIS PRÓXIMA

1. Para cada objeto i, calcular a distância ao k-ésimo vizinho mais próximo. Represente essa distância por $d_i(k)$. Observe que $d_i(k)$ está inversamente relacionado à densidade relativa da região que cerca o objeto i. Se a distância $d_i(k)$ for pequena, então a densidade é alta.
2. Conecte todos os objetos i e j em que i está na k-ésima vizinhança mais próxima de j ou j está na k-ésima vizinhança mais próxima de i (ou ambos). Essa conexão estabelece quais pares de pontos são os melhores candidatos para inclusão no mesmo agrupamento. Note que, em vizinhanças de alta densidade, dois objetos serão conectados somente se estiverem especialmente próximos um do outro em termos da distância euclidiana. Em regiões de baixa densidade, dois objetos podem ser conectados mesmo se estiverem separados (em termos euclidianos).

3. Estabeleça $d_{ij}* = [d_i(k) + d_j(k)]/2$ para todos os pontos conectados i e j. Observe que d_{ij}^* é uma medida do tipo distância (quanto menor seu valor, mais próximos os dois objetos) que reflete a vizinhança local dos objetos i e j. Os primeiros objetos a serem agrupados serão aqueles em regiões de alta densidade que estejam próximos uns dos outros.

Alguns pontos permanecerão "desconectados" pela medida de densidade k-ésima de vizinhança mais próxima. Efetivamente, tratamos esses pares de pontos como se fossem separados por uma distância infinita (isto é, $d_{ij}^* = \infty$). Se algum subconjunto de pontos estiver inteiramente desconectado dos pontos restantes na amostra, ele formará um agrupamento isolado. Isso tende a ocorrer quando um conjunto de pontos está separado de outro por uma região de densidade baixa (isto é, quando a distribuição subjacente de dados é multimodal).

Há um quê de arte na escolha correta do k. Se for muito grande (por exemplo, k se aproximando de n), todas as estimativas de densidade local tendem a se aproximar da densidade média da amostra inteira. No limite, todos os objetos estão conectados a todos os outros objetos no mesmo nível de proximidade. Se k for pequeno demais, as estimativas da densidade local não serão confiáveis e o resultado será muitos agrupamentos pequenos e isolados. Uma regra de ouro é tentar vários valores de k na vizinhança de $n^{1/2}$ (tanto acima quanto abaixo).

8.4 AGRUPAMENTO AGLOMERATIVO: SEU FUNCIONAMENTO

8.4.1 Intuição

A ideia básica por trás do agrupamento aglomerativo é simples. Comece com cada objeto em seu próprio agrupamento isolado (isto é, n agrupamentos de tamanho 1). Em cada etapa do processo, encontre dois agrupamentos "mais próximos" (em um sentido bem definido) e junte-os. Continue até que reste um agrupamento de tamanho n. O algoritmo é simples e notavelmente eficiente.

AGRUPAMENTO DE LIGAÇÃO SIMPLES

Etapas do processo iterativo

Etapa 0. Comece com todos os objetos em agrupamentos separados (isto é, n agrupamentos com um objeto cada). Represente esses agrupamentos por $C_1, C_2, C_3, \ldots, C_n$. Nesse passo inicial, a distância entre dois agrupamentos é definida como a distância entre dois objetos neles contidos; isto é,

$$d_{C_iC_j} = d_{ij}$$

Sendo $t = 1$ um índice do processo iterativo.

Etapa 1. Encontre a menor distância entre dois agrupamentos quaisquer. Represente esses dois agrupamentos mais próximos por C_i e C_j.

Etapa 2. Combine os agrupamentos Ci e C_j para formar um novo agrupamento denominado C_{n+t}.

Etapa 3. Defina a distância entre o novo agrupamento C_{n+t} e todos os agrupamentos C_k da seguinte maneira:

$$d_{C_{n+t}C_k} = \min\{d_{C_iC_k}, d_{C_jC_k}\}.$$

Etapa 4. Adicione o agrupamento C_{n+t} como um novo agrupamento e remova os agrupamentos C_i e C_j. Considere $t = t + 1$.

Etapa 5. Volte à etapa 1 e continue até que reste um agrupamento.

Resultado do agrupamento

Ao final, todos os mecanismos da ligação simples terminam no mesmo ponto: um grande agrupamento de todas as observações. De fato, não é o ponto final da análise que é particularmente útil, mas a sequência de etapas que descreve quais objetos são reunidos em qual estágio da análise. A representação gráfica dessas etapas é conhecida como *dendrograma*; corresponde a uma estrutura hierárquica em árvore, gerada pela sequência iterativa descrita anteriormente. É essencial saber como se lê um dendrograma para se extrair *insight* de uma análise de agrupamento hierárquico.

Para fins de ilustração, introduzimos um exemplo simples com quatro objetos A, B, C e D localizados ao longo de um segmento de linha, como apresentado na Figura 8.10. É fácil verificar que, aplicando-se uma ligação simples a esses objetos, produzir-se-ia a seguinte sequência de soluções de agrupamentos:

Iteração 0: {A}, {B}, {C}, {D}
{A} se junta a {B} em distância $d = 2$
Iteração 1: {A, B}, {C}, {D}
{C} se junta a {D} em distância $d = 3$
Iteração 2: {A, B}, {C, D}
{A, B} se junta a {C, D} em distância $d = 6$
Iteração 3: {A, B, C, D}

O dendrograma que representa a solução de agrupamento de ligação simples é apresentado na Figura 8.11. Cada vez que dois objetos ou dois agrupamentos são reunidos, são representados por uma linha vertical conectando as duas linhas horizontais em uma distância em que os dois são agrupados juntos. Deste modo, há uma linha vertical a uma distância $d = 2$ juntando os objetos A e B. À direita da junção vertical, as duas linhas que representam os objetos A e B são substituídas por uma única linha representando o agrupamento {A, B}.

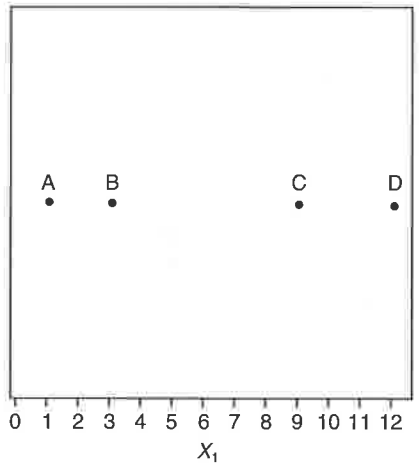


Figura 8.10 Ilustração: quatro pontos sobre um segmento de linha.

Quantos agrupamentos?

O agrupamento aglomerativo não fornece uma resposta definitiva à questão: quantos agrupamentos há? De fato, o dendrograma é uma representação gráfica de uma hierarquia de solução de agrupamentos aninhados: uma solução de um agrupamento, solução de dois agrupamentos e assim por diante, até uma solução de n agrupamentos. Desenhando-se uma linha vertical sobre o dendrograma (que corresponda a um valor específico de distância d), revela-se a solução de agrupamento no nível da distância e de pertencimento a diferentes agrupamentos. Por exemplo, uma linha vertical em $d = 4$ define a solução de dois agrupamentos, com agrupamentos {A, B} e {C, D}.

Então, como alguém pode dizer, ao analisar o dendrograma, se uma dessas soluções de agrupamentos aninhados fornece uma representação "melhor" dos dados? Uma coisa a se buscar é uma gama de

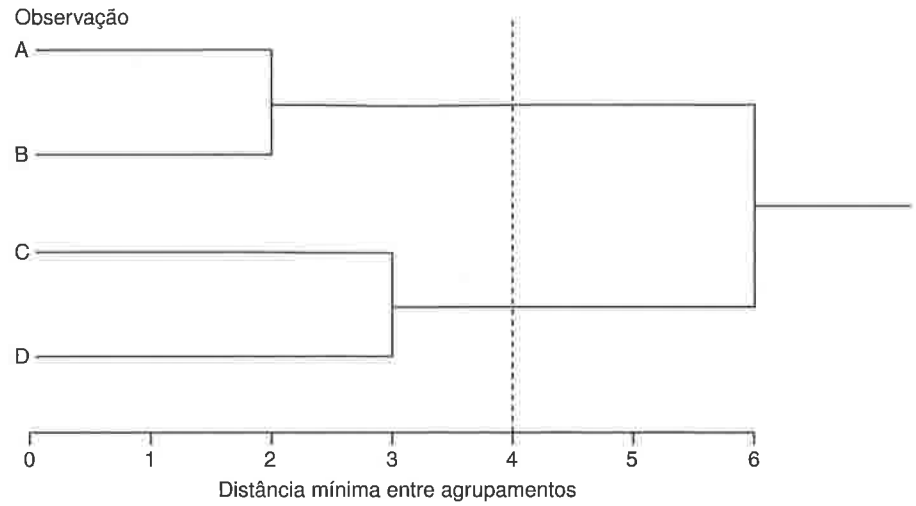


Figura 8.11 Dendrograma para solução de agrupamento de ligação simples com os dados da Figura 8.10.

distâncias relativamente ampla em relação às quais o número de agrupamentos na solução não muda. Nesse exemplo simples, a estrutura de dois agrupamentos é estável no intervalo de distâncias (3, 6). Não há dúvida de que a leitura do número de agrupamentos de um dendrograma (assim como a leitura do número de fatores de um gráfico scree) envolve um montante considerável de subjetividade e exige discernimento da parte do analista.

Propriedades da ligação simples

Os resultados de um agrupamento de ligação simples (e de todas as abordagens aglomerativas discutidas posteriormente nesta seção) são hierárquicos por natureza. Isso significa que uma solução de agrupamento perto do topo da árvore pode sempre ser obtida combinando-se os agrupamentos de qualquer solução mais próxima da base da árvore. Essa propriedade é uma consequência natural do algoritmo.

O agrupamento de ligação simples é eficiente, do ponto de vista computacional. À medida que o número de objetos n aumenta, a quantidade de esforço computacional requerida para o pior caso aumenta na ordem de n^2. O algoritmo é ainda mais eficiente para dados escassos (por exemplo, para estruturas de rede, em que cada objeto é conectado a uma fração dos outros objetos no conjunto apenas). Aqui o esforço computacional está na ordem de nA, em que A é o número médio de conexões para cada objeto no conjunto. Além do mais, o agrupamento de ligação simples não exige dados métricos. A construção do algoritmo descrito anteriormente também funciona bem para medidas ordinais de dissimilaridade.

Uma desvantagem da ligação simples é que ela tende a ser extremamente míope. Um objeto será adicionado a um agrupamento desde que esteja próximo a qualquer um dos outros objetos do agrupamento, mesmo que, por outro lado, esteja relativamente longe de todos os outros. Assim, a ligação simples possui uma tendência a produzir agrupamentos longos e encadeados, com formatos não convexos. Se os verdadeiros agrupamentos subjacentes forem não convexos, essa propriedade não será necessariamente uma coisa ruim; no entanto, na maioria dos casos, as modas que ocorrem naturalmente em nossos dados tenderão a ser convexas e compactas – e um reflexo melhor da homogeneidade interna. Como resultado direto, a abordagem não teve um bom desempenho nos estudos de Monte Carlo (veja, por exemplo, Milligan, 1980).

Alternativas para a ligação simples

Muitas abordagens diferentes têm sido desenvolvidas para se tratar da fraqueza inerente à ligação simples. Algumas dessas abordagens são descritas brevemente a seguir. Observe que todas essas abordagens são aglomerativas por natureza e produzem soluções de agrupamentos hierárquicos.

Ligação completa. Em lugar de definir a distância (ou dissimilaridade) entre os agrupamentos como a distância entre o par de objetos mais próximo (como na ligação simples), usamos a distância entre o par de objetos mais afastado. Isso assegura que cada objeto adicionado ao agrupamento esteja próximo de todos os objetos no agrupamento, e não somente de um. A única mudança exigida para se ir de um agrupamento de ligação simples para um de ligação completa é reescrever a etapa 3 da seguinte maneira:

$$d_{C_{n+1}C_k} = \text{máx}\{d_{C_iC_k}, d_{C_jC_k}\} \tag{8.6}$$

Em comparação à ligação simples, é muito mais provável que a ligação completa produza agrupamentos convexos que tendem a ser de diâmetro comparável. Embora a ligação completa tenha esta tendência de produzir agrupamentos convenientes e homogêneos, estes não são necessariamente levados pela modalidade natural de dados. Milligan (1980) descobriu que a ligação completa pode ser altamente sensível a discrepâncias nos dados. Quando uma ligação ocorre na etapa 3 (isto é, mais do que dois agrupamentos podem ser reunidos), a escolha pode afetar o formato subsequente da solução de agrupamentos (um problema que não ocorre no caso da ligação simples).

Ligação média. Esta abordagem pode ser considerada uma espécie de meio-termo entre a ligação simples e a ligação completa. Alguns autores preferem este método porque, com ele, chega-se mais perto de um ajuste de árvore que satisfaça o critério de minimização dos mínimos quadrados. Em lugar de usar a mínima (ligação simples) ou a máxima (ligação completa), a nova distância é definida como a distância média entre o agrupamento C_k e o novo agrupamento C_{n+t} (formado pela junção dos agrupamentos C_i e C_j). Assim, reescrevemos a etapa 3 da seguinte maneira:

$$d_{C_{n+t}C_k} = \frac{n_i d_{C_iC_k} + n_j d_{C_jC_k}}{n_i + n_j} \tag{8.7}$$

onde $n_i + n_j$ é o número de objetos no agrupamento recém-formado C_{n+t}. Note que se os dados forem não métricos, a média pode ser substituída pela mediana (nesse caso, o método é chamado de *ligação mediana*).

Método centroide. Em vez de definir a distância entre dois agrupamentos como a distância média entre todos os pares de objetos, é possível primeiro tirar a "média" dos objetos em cada agrupamento (ou seja, calcular os centroides do agrupamento) e então definir a distância entre os dois centroides. Esse método simplifica as coisas se trabalharmos com distâncias ao quadrado. Considere d_{ij}^2 a distância euclidiana ao quadrado entre os objetos i e j. Se o agrupamento $C = \{i, j\}$, então a distância ao quadrado entre o objeto k e o centroide do agrupamento C pode ser representada por

$$d_{kC}^2 = \frac{d_{ik}^2 + d_{jk}^2}{2} - \frac{d_{ij}^2}{4} \tag{8.8}$$

Em geral, a distância ao quadrado entre qualquer agrupamento C_k e um novo agrupamento C_{n+t} criado pela junção dos agrupamentos C_i e C_j pode ser representada por

$$d^2(C_k, C_i \cup C_j) = \frac{n_{C_i} d_{C_k,C_i}^2 + n_{C_j} d_{C_k,C_j}^2}{n_{C_i} + n_{C_j}} - \frac{n_{C_i} n_{C_j} d_{C_i,C_j}^2}{(n_{C_i} + n_{C_j})^2} \tag{8.9}$$

Determinando-se a regra na etapa 3 como uma função da distância euclidiana ao quadrado (em lugar de medidas de atributo X), o método centroide pode ser usado diretamente com medidas de proximidade avaliadas e também com medidas de distância derivadas (por exemplo, distâncias ao quadrado calculadas a partir de dados dos atributos). De acordo com Milligan (1980), o método centroide é robusto para as discrepâncias, mas pode ser superado pela ligação média.

Método de Ward. Os três métodos descritos anteriormente (ligação completa, ligação média e método centroide) são variações de uma abordagem aglomerativa geral chamada *método de grupo de pares* (*pair group method*), que difere somente em termos da relação de distância especificada na etapa 3. Por contraste, o método de Ward (às vezes chamado de *método de variância mínima*) adota uma estratégia

ligeiramente diferente na etapa 1. Em lugar de juntar os dois agrupamentos mais próximos, o método de Ward busca juntar os dois agrupamentos cuja fusão dá origem à menor soma de quadrados dentro do agrupamento (isto é, a variância mínima dentro do grupo).

O método de Ward tem uma tendência a produzir agrupamentos de tamanhos iguais (isto é, agrupamentos com aproximadamente o mesmo número de observações em cada) convexos e compactos. Como a abordagem é baseada na minimização das distâncias dentro dos agrupamentos, ela, com frequência, produz uma solução de agrupamento – se a árvore for "cortada" no lugar certo – que é similar aos métodos de partição descritos na Seção 8.5 a seguir (que também enfoca a minimização da soma de quadrados dentro do grupo).

Ligação por densidade. Outro modo de lidar com as tendências um tanto míopes da ligação simples se dá pela mudança das medidas de distância para explicar o contexto. A medida de densidade d^* do k-ésimo vizinho mais próximo (descrito na Seção 8.3) reflete não somente a relação direta entre dois objetos, mas também sua relação com os objetos circundantes. Ao lançarmos mão da ligação simples, usando d^* em vez de d, podemos tomar decisões melhores sobre a junção de grupos apropriados de objetos nos primeiros estágios da análise.

De acordo com Wong e Lane (1983), a ligação por densidade reúne a maioria dos pontos fortes da abordagem de ligação simples (principalmente sua eficiência) e tende a desempenhar um pouco melhor no fornecimento de agrupamentos compactos (dependendo, obviamente, do valor de k usado). Um benefício mais impressionante, que será demonstrado a seguir, é que o contexto fornecido pela medida de densidade dá a essa abordagem uma vantagem distinta na identificação da modalidade natural dos dados.

8.4.2 Exemplo

Para fins de ilustração, aplicamos agora três dos métodos descritos anteriormente (ligação simples, método de Ward e ligação por densidade) e comparamos os resultados. Com agrupamentos, é frequentemente mais fácil avaliar o resultado quando conhecemos as características subjacentes dos dados em primeiro lugar. Portanto, iniciamos usando os dados descritos na Seção 8.2: uma amostra de 50 observações extraídas de uma população bimodal e outra amostra de 50 observações retiradas de uma distribuição unimodal. Utilizamos a distância euclidiana nas rotinas de agrupamento.

Os resultados da análise de ligação simples são apresentados na Figura 8.12. Para a amostra bimodal, desenhar uma linha vertical em $d = 0{,}36$ sobre o dendrograma divide os dados em quatro agrupamentos: dois agrupamentos relativamente grandes (de tamanho 26 e 20, respectivamente) e dois pequenos (cada um com duas observações). Assim, a modalidade natural desses dados fica pelo menos evidente a partir da solução de agrupamentos, mesmo que a estrutura de dois agrupamentos não seja totalmente clara. Na amostra unimodal, a situação é mais obscura. A divisão do dendrograma em $d = 0{,}40$ produz dois agrupamentos relativamente grandes (um com 35 objetos e outro com 14) e um agrupamento único. Embora menos convincente que o agrupamento da amostra bimodal, parece seguro afirmar que os resultados de agrupamento de ligação simples não demonstram a diferença na modalidade dessas duas amostras de forma dramática.

Os resultados do método de Ward são mostrados na Figura 8.13. Em contraste com os resultados da ligação simples, a estrutura de dois agrupamentos da amostra bimodal surge alta e clara. Os agrupamentos não são exatamente do mesmo tamanho (a divisão é 28/22 em vez de 25/25), o que sugere que houve algum erro de classificação. Mas a separação entre os dois grupos, como mostrada no dendrograma, é inequívoca. Entretanto, há uma separação igualmente impressionante da segunda amostra em dois agrupamentos, mesmo ela sendo originária de uma população unimodal. Assim, o método de Ward falha em demonstrar a diferença na modalidade das duas amostras.

Os resultados da análise de ligação por densidade (baseada na densidade k-ésima de vizinho mais próximo com $k = 9$) são apresentados na Figura 8.14. O que é interessante aqui é o total contraste entre os dois dendrogramas. O primeiro (da amostra bimodal) mostra dois agrupamentos separados começando a se formar, o que é indicado pelas áreas circuladas no dendrograma na figura.

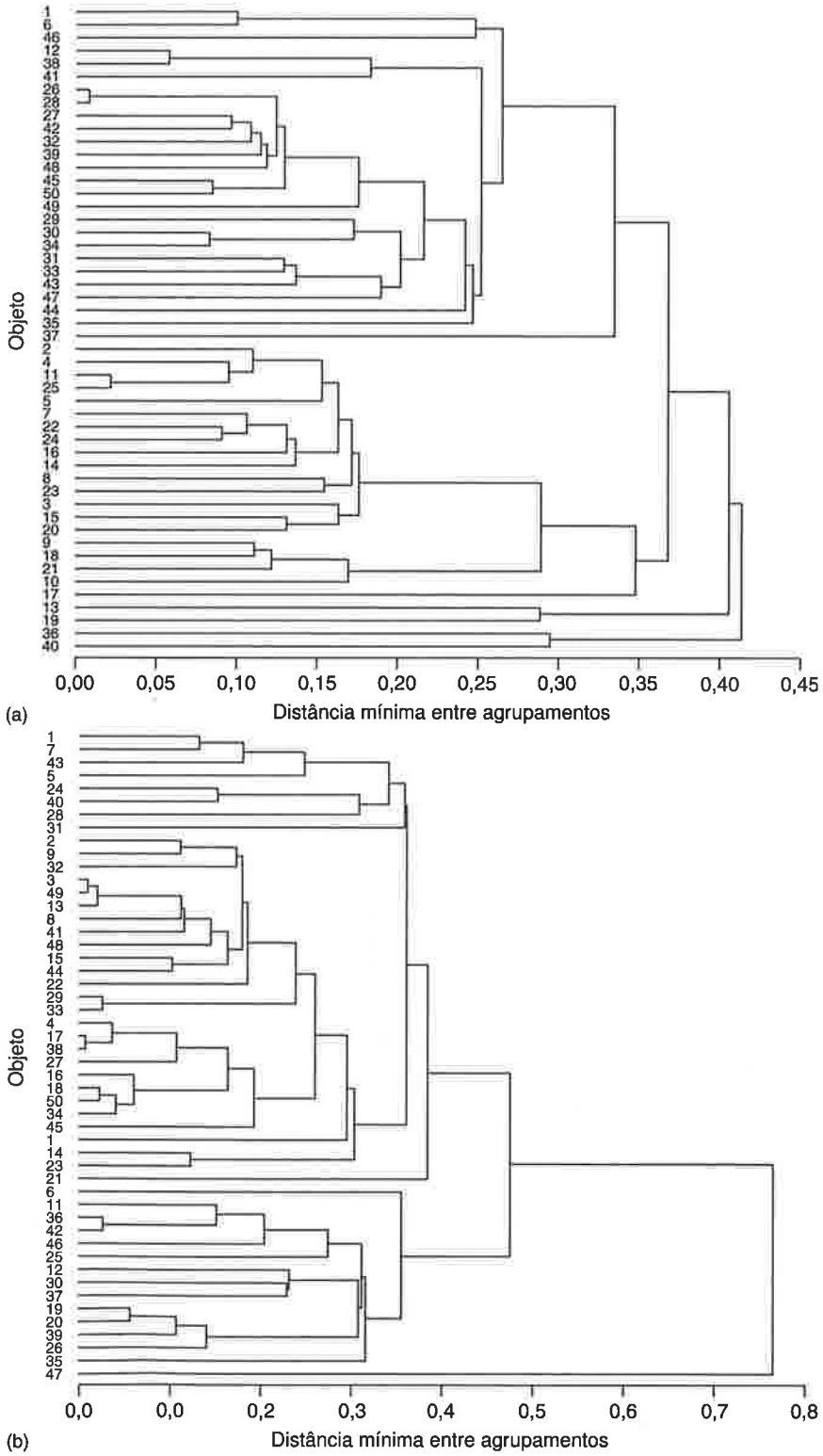


Figura 8.12 Dendrograma da análise de ligação simples dos dados da amostra (a) bimodal e (b) unimodal.

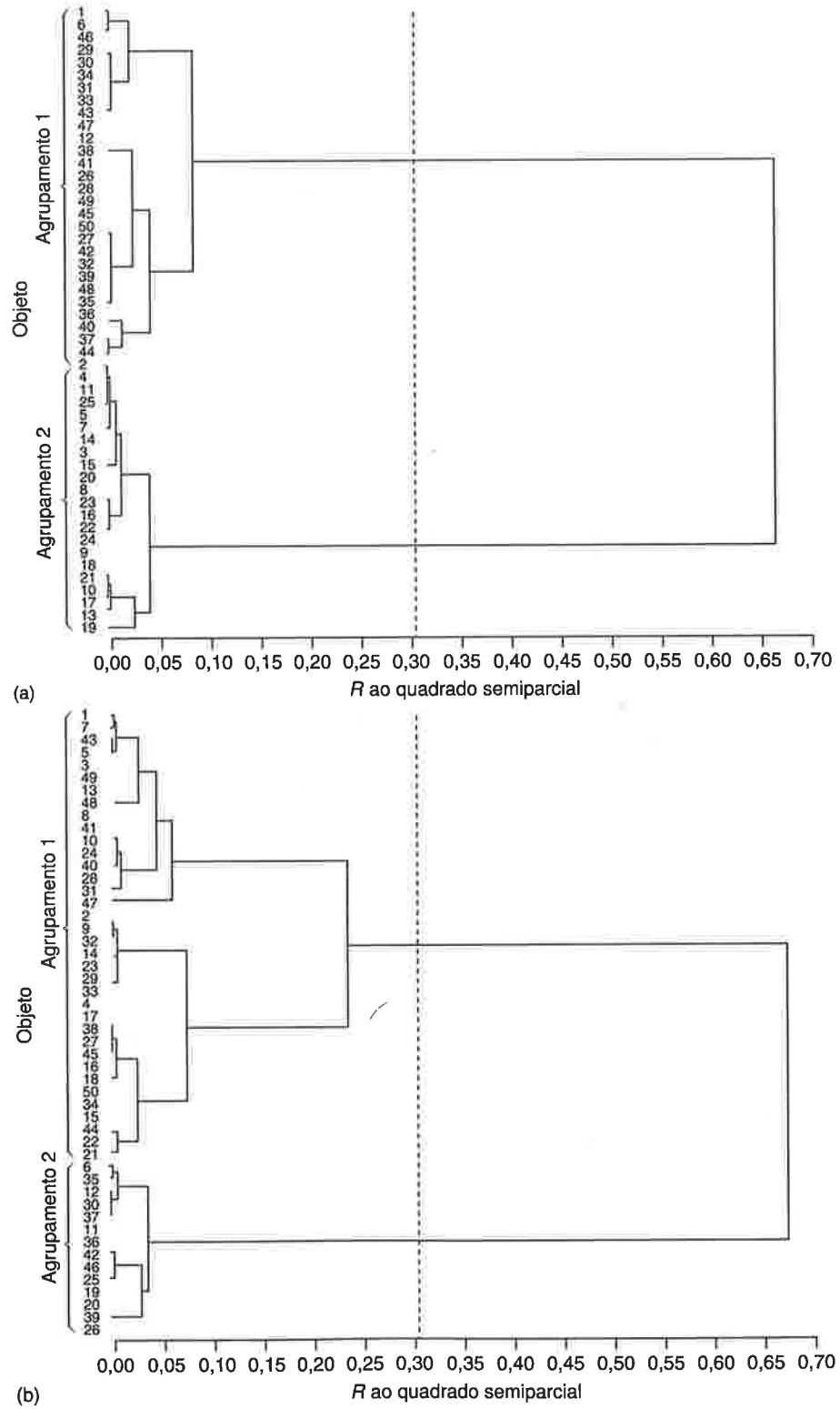


Figura 8.13 Dendrogramas da análise de Ward dos dados da amostra (a) bimodal e (b) unimodal.

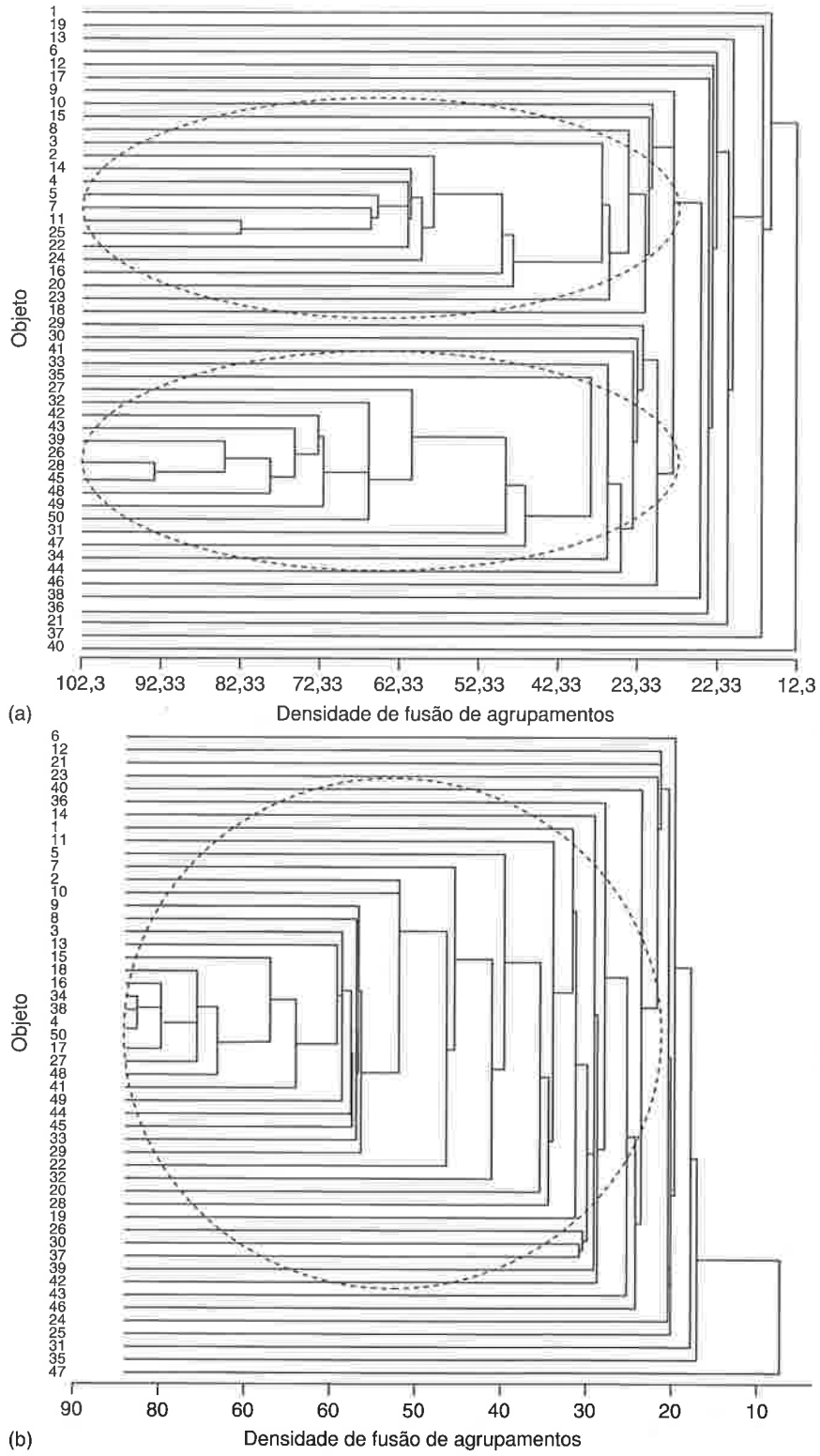


Figura 8.14 Dendrogramas da análise de densidade de agrupamento do *k*-ésimo vizinho mais próximo para dados da amostra (a) bimodal e (b) unimodal.

Cada grupo de objetos é chamado de *agrupamento modal*. Para valores diferentes de k (por exemplo, $k = 7, 8, 9, 10, 11$), observamos dois agrupamentos modais no dendrograma da amostra bimodal. Embora esses dois agrupamentos modais, ao final, se reúnam em níveis mais baixos de densidade, note que eles são bem distintos em níveis mais altos de densidade. Em contraste, para os mesmos valores de k observamos somente um único agrupamento modal no dendrograma da amostra unimodal; isso é verdade independente do nível de densidade. Assim, diagramas de agrupamentos estruturalmente diferentes de densidade de agrupamento do k-ésimo vizinho mais próximo revelam as diferenças na modalidade das duas amostras.

Por que o método de Ward encontra uma solução de dois agrupamentos para a amostra unimodal, ao passo que a ligação por densidade sugere um único agrupamento modal? A resposta tem a ver com as diferenças de objetivos. O método de Ward realiza um trabalho melhor de redução da heterogeneidade (isto é, da variância de partição). A ligação por densidade, porque explica o contexto, está em consonância com a descoberta de soluções de agrupamentos em que as regiões de alta densidade dos objetos são separadas por regiões de densidade relativamente mais baixa.

8.5 PARTIÇÃO: COMO FUNCIONA

8.5.1 Intuição

A partição é diferente do agrupamento aglomerativo. Com a partição, nossa meta é dividir a amostra em um número determinado K de grupos não superpostos, de maneira que os objetos dentro de cada grupo sejam relativamente similares e os objetos entre grupos sejam relativamente dissimilares. Para fazer isso, necessitamos achar um modo de medir a similaridade dentro do grupo e a diferença entre os grupos (para que possamos comparar duas partições e dizer qual é a melhor). Também precisamos encontrar a melhor dessas partições (uma que seja pelo menos localmente ótima, se não globalmente ótima) de acordo com a nossa medida escolhida.

A abordagem de partição descrita aqui é conhecida como *agrupamento de K-means* (Hartigan, 1975). O algoritmo, como a ligação simples, é simples e eficiente do ponto de vista computacional. No entanto, ele é propenso a encontrar soluções localmente ótimas apenas (porque é baseado em uma heurística que realiza somente melhorias locais para uma partição inicial, até que melhorias posteriores não sejam possíveis). Portanto, é necessário executar o agrupamento de K-means um grande número de vezes, com diferentes pontos iniciais (o que não é diferente de certos métodos de escalonamento multidimensional), para assegurar uma boa solução.

As etapas do algoritmo geral de agrupamento de K-means são descritas a seguir:

AGRUPAMENTO DE *K*-MEANS

Etapas do processo iterativo

Etapa 1. Selecione uma partição inicial dos dados nos agrupamentos K. Diversas abordagens à escolha dessa partição inicial são descritas no texto. Uma variante importante de K-means começa com um conjunto inicial de centroides de "semente" K que define a partição inicial designando cada objeto para o ponto de semente mais próximo. Em muitos casos, os pontos de semente são simplesmente os pontos reais de K, normalmente escolhidos de forma que sejam amplamente dispersos.

Etapa 2. Calcule o centroide para cada agrupamento C, $\bar{\mathbf{x}}_C$. Observe que é possível realizar esse cálculo utilizando-se dados do tipo atributo ou dados de distância (como no método de agrupamento centroide, descrito na Seção 8.4 anterior; veja DeSarbo, Carroll, Clark e Green, 1984).

Etapa 3. Calcule a soma das distâncias ao quadrado de cada objeto do seu centroide do agrupamento (isto é, o quadrado da soma de erros da partição, representado por ESS):

$$\text{ESS} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_{C(i)})'(\mathbf{x}_i - \bar{\mathbf{x}}_{C(i)})$$

onde $C(i)$ é o agrupamento para o objeto i. Como o ESS reflete as distâncias entre os objetos internos ao grupo, esse é o termo que queremos tornar tão pequeno quanto possível.

Etapa 4. Torne a relacionar cada objeto i ao agrupamento cujo centroide é mais próximo. Se ao final da etapa 4 os elementos do agrupamento permanecerem sem alterações, o processo convergiu para pelo menos um mínimo local. Se pelo menos um objeto do agrupamento modificar-se, volte à etapa 2 com a nova partição.

Observemos uma iteração do algoritmo K-means usando um conjunto simples de dados para analisar como ele funciona. A Figura 8.15 mostra um gráfico de sete observações medidas em duas dimensões (X_1 e X_2). Inicialmente, os objetos foram divididos em dois agrupamentos, de modo que os objetos 1 a 4 formam o primeiro agrupamento e os objetos 5 a 7 formam o segundo. Os centroides para os dois agrupamentos iniciais são (−0,75, 0,0) e (2,33, 0), respectivamente. O erro da partição inicial é dado por ESS = 7,836.

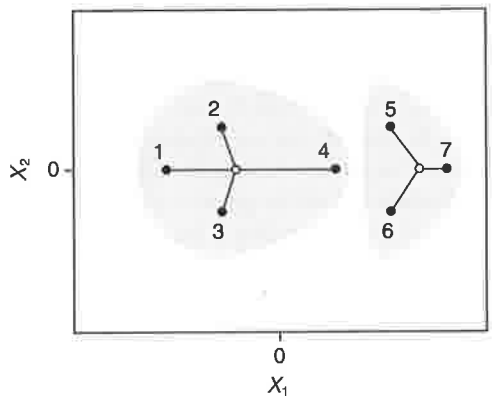


Figura 8.15 Partição inicial em dois agrupamentos de sete observações sobre duas variáveis.

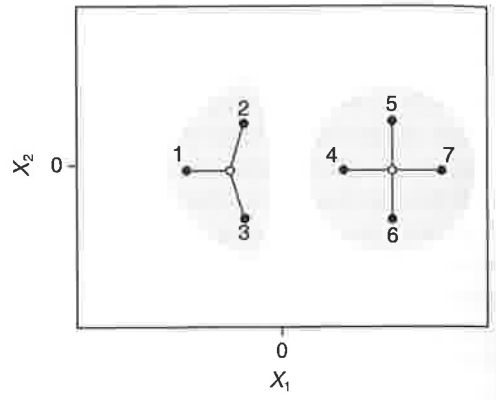


Figura 8.16 Partição final de dois agrupamentos após nova atribuição do objeto 4.

Após uma iteração do processo de agrupamento K-means delineado anteriormente, um objeto é novamente atribuído: o objeto 4 é mudado do agrupamento 1 para o 2 porque sua localização em (1, 0) é mais próxima do centroide do agrupamento 2 (2,33, 0) do que do centroide do agrupamento 1 (−0,75, 0). A Figura 8.16 mostra a configuração do agrupamento após a nova atribuição do objeto 4. Depois dessa nova atribuição, o erro de partição diminui para ESS = 6,775. Essa é a solução final; nenhuma nova atribuição é capaz de implicar em maior redução do ESS.

Selecionando a partição inicial

O procedimento K-means produz um resultado que é localmente apenas ótimo. É, portanto, importante tecer algumas considerações à escolha da partição inicial de agrupamentos. Em geral, quanto melhor o ponto inicial, melhor a solução final. Isso se torna ainda mais importante quando o conjunto de dados é relativamente pequeno.

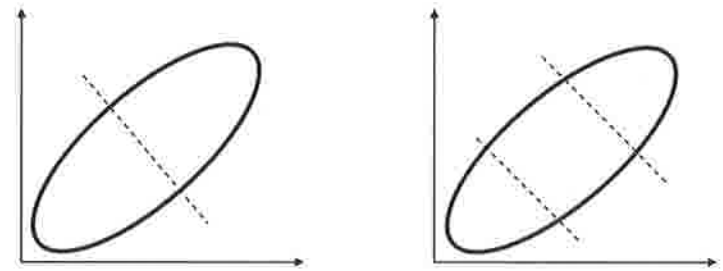
A maioria dos pacotes de software disponíveis oferece alguma heurística para selecionar-se a partição inicial. Uma abordagem é usar K observações dos dados como "sementes" e atribuir cada observação restante à "semente" mais próxima. As sementes podem ser escolhidas aleatoriamente ou sequencialmente para que cada uma esteja a uma certa distância mínima das outras.

Outra possibilidade é o analista especificar seu próprio conjunto de sementes de agrupamentos. Um bom ponto de partida é usar centroides de uma solução de agrupamento aglomerativo, como o método de Ward ou a ligação média. Observe que a abordagem de Ward é amplamente consistente com o objetivo de partição de variância (de fato, o critério usado no método de Ward é bem similar ao critério ESS usado por K-means). Além do mais, o método de Ward é facilmente aplicado a problemas relativamente pequenos (nos quais a importância de se determinar um bom ponto de partida é maior).

A conclusão final é: provavelmente faz sentido usar todos esses métodos (e, talvez, usar alguns inícios aleatórios também) e depois escolher a melhor solução de agrupamento.

Propriedades da solução

Diferente do que ocorre nos procedimentos aglomerativos discutidos na Seção 8.4 anterior, os resultados da partição não são hierárquicos por natureza. Com um método hierárquico, qualquer solução de agrupamentos perto do topo da árvore pode ser formada combinando-se agrupamentos em um ponto mais baixo (sem que se mude o pertencimento de qualquer objeto ao agrupamento). Tal não é o caso com um método de partição. Cada partição chega separadamente; isto é, a partição de três agrupamentos não toma a partição de dois agrupamentos como ponto de partida. A Figura 8.17 mostra um exemplo de soluções para K-means de dois e três agrupamentos para um conjunto hipotético de dados. A figura deixa claro que não há meio de se ir da solução de três agrupamentos para a solução de dois agrupamentos sem atribuir novamente os objetos.



(a) Solução de agrupamento $K = 2$ (b) Solução de agrupamento $K = 3$

Figura 8.17 Soluções de agrupamentos de métodos de partição não são necessariamente hierárquicas.

Como o K-means é focado na redução do ESS, ele é bem consistente com o objetivo de redução da heterogeneidade. Assim, como o procedimento de Ward, tende a produzir agrupamentos compactos e convexos.

Quantos agrupamentos?

O algoritmo K-means encontra uma solução de agrupamentos para um dado valor de K; compete ao analista decidir qual valor de K resulta na "melhor" solução de agrupamento. Então, como se decide o valor apropriado? A chave é conduzir análises com vários valores diferentes de K e depois escolher a solução que melhor corresponda aos objetivos da análise. Isso normalmente envolve uma análise de custo-benefício entre a simplicidade da solução (quando um número menor de agrupamentos é melhor) e sua adequação (caso a redução da heterogeneidade dentro do grupo seja o objetivo, mais agrupamentos é melhor). Com frequência, a decisão requer discernimento subjetivo por parte do analista.

Observe que alguns critérios não são particularmente úteis na avaliação dessa escolha. A medida ESS é problemática, porque vai diminuir necessariamente à medida que o número de partições aumenta (se isso não ocorrer, uma das soluções tem de ser meramente uma ótima local). Uma medida que capta a análise de custo-benefício entre a simplicidade e a adequação é chamada de *estatística pseudo-F*. Esse critério, usado por Calinski e Harabasz (1974), é reminiscência de um teste-F porque é efetivamente uma razão da soma média de quadrados entre os grupos pela soma média de quadrados dentro do grupo, e assim explicam-se os graus de liberdade, que são funções de K. A fórmula é dada por

$$\text{pseudo-}F = \frac{\text{tr}[\mathbf{B}/(K-1)]}{\text{tr}[\mathbf{W}/(n-K)]} \tag{8.10}$$

onde $\mathbf{B}$ é a matriz da soma de quadrados entre agrupamentos, $\mathbf{W}$ é a matriz da soma de quadrados dentro dos agrupamentos, K é o número de agrupamentos e n é o número de objetos.

Em geral, quanto maior o pseudo-F, mais "eficiente" é a partição na redução da heterogeneidade no interior do grupo (refletida no denominador). Embora tr($\mathbf{B}$) geralmente vá aumentar com o número de

agrupamentos, tr[**B**/(K − 1)] não aumenta. Em geral, a medida do pseudo-F não aumenta monotonicamente, mas atinge um máximo para determinado valor específico de K (dependendo, naturalmente, dos dados). É essa explicação do número de graus de liberdade exigido pelos agrupamentos adicionais que dá a esse critério sua capacidade de diagnóstico.

Vamos supor que temos um conjunto de dados com modalidade natural M. Seria de se esperar que uma estatística pseudo-F relativamente grande fosse associada com a partição K-means para $K = M$. Se fizermos a análise para $K = M + 1$, deveria haver mudanças relativamente pequenas em **B** e **W** (porque não podemos melhorar de maneira significativa a solução afastando-nos da modalidade natural). Como resultado, o aumento em K pode causar um aumento no denominador e um decréscimo no numerador, resultando em um valor mais baixo do pseudo-F para $K = M + 1$. Quando um aumento no número de agrupamentos resulta em um decréscimo no pseudo-F, há indicação de que a complexidade adicional da solução pode não valer a pena em termos da redução na heterogeneidade dentro do agrupamento. Milligan e Cooper (1985) conduziram uma grande simulação Monte Carlo na qual testaram 30 diferentes critérios para determinar o número de agrupamentos: o que funcionou melhor foi o pseudo-F de Calinski e Harabasz.

Interpretando a solução final

Ao final, é importante não somente selecionar o número de agrupamentos, mas também entender como interpretar os agrupamentos e como diferem uns dos outros. Provavelmente, a abordagem mais direta é examinar os centroides dos agrupamentos calculando-se o valor médio de cada variável dos objetos atribuídos a cada agrupamento. Isso revela quais agrupamentos têm relativamente muitas variáveis e quais têm poucas.

Entretanto, somente os centroides dos agrupamentos não dizem quantas superposições existem entre os agrupamentos de uma dada variável X. Por isso, também é importante considerar alguma medida da variância em X explicada pela solução de agrupamentos. Podemos decompor a soma total dos quadrados de X a uma soma de quadrados dentro do agrupamento (isto é, a soma dos desvios ao quadrado entre cada observação e a média de seu agrupamento) e a soma de quadrados entre os agrupamentos. Se a solução de agrupamentos possuir observações estreitamente agrupadas ao redor de suas respectivas médias, a soma de quadrados dentro dos agrupamentos será pequena e a soma dos quadrados entre os agrupamentos será grande (porque a sua soma é uma constante). Como medida resumida, consideramos a razão da soma de quadrados entre agrupamentos pela soma total de quadrados (que é análoga a uma medida R^2, que capta a quantidade de variação em X explicada pela média do agrupamento) ou podemos considerar a razão da soma de quadrados entre agrupamentos pela soma de quadrados dentro do agrupamento [que seria o mesmo que olhar para a razão $R^2/(1 - R^2)$, análoga ao pseudo-F para uma variável única)]. Comparando esses valores das variáveis, fica claro quais delas são mais importantes na definição das diferenças entre os agrupamentos.

8.5.2 Exemplo

Novamente, para fins de ilustração, retornamos aos dados da amostra (uma bimodal e outra unimodal) descritos na Seção 8.2. Os resultados da aplicação de agrupamentos K-means para os dados bimodais com $K = 2$ são apresentados na Tabela 8.6.

A solução corresponde bem proximamente ao nosso entendimento desses dados. Os centroides dos agrupamentos são localizados em (3,1, 2,8) e em (5,9, 5,5), que são razoavelmente próximos às modas da população em (3,0, 3,0) e (6,0, 6,0). A designação das observações aos agrupamentos é quase perfeita: somente um objeto está designado incorretamente. A solução de dois agrupamentos alcança boa separação em ambas as variáveis X_1 e X_2 (note que R^2 vale 0,67 e 0,70, respectivamente).

Como saber se a solução de dois agrupamentos é a melhor? Para fins de comparação, realizamos outra análise com $K = 3$. A solução de dois agrupamentos resultou em uma estatística pseudo-F de 103,7; esse valor cai para 82,33 quando $K = 3$. Embora não necessariamente conclusivo, o resultado sugere que a melhoria na homogeneidade dentro do agrupamento não compensou o custo de se acrescentar outro agrupamento à solução.

A Tabela 8.7 mostra os resultados da aplicação de agrupamentos K-means aos dados unimodais (novamente com $K = 2$). Em muitos aspectos, a solução de dois agrupamentos para dados unimodais é

bem similar à solução de dois agrupamentos para os dados bimodais. Os centroides dos agrupamentos são aproximadamente os mesmos: (3,4, 3,7) e (5,5, 5,9). A designação dos objetos não é igual (os tamanhos dos agrupamentos são 18 e 32); a variável X_2 parece desempenhar um papel relativamente mais importante do que X_1 na distinção entre os agrupamentos.

Então, como podemos dizer que a solução de dois agrupamentos é a apropriada? A estatística pseudo-F para a solução de dois agrupamentos é 51,6. Infelizmente, não há meio de se calcular a estatística pseudo-F para a solução de um agrupamento. Assim, não há meio para se observar se há um declínio no pseudo-F quando mudamos de um para dois agrupamentos.

Tabela 8.6 Resultados da partição K-means de dados da amostra bimodal para $K = 2$

Agrupamento	Frequência
1	26
2	24

Variável	STD Total	Dentro do STD	R^2	$R^2/(1 - R^2)$
X_1	1,708	0,993	0,669	2,022
X_2	1,624	0,900	0,700	2,328
Geral	1,666	0,947	0,684	2,160

	Média do agrupamento	
Variável	Agrupamento 1	Agrupamento 2
X_1	5,875	3,108
X_2	5,473	2,781

Estatística pseudo-F = 103,68

Tabela 8.7 Resultado da partição K-means de dados da amostra unimodal para $K = 2$

Agrupamento	Frequência
1	32
2	18

Variável	STD Total	Dentro do STD	R^2	$R^2/(1 - R^2)$
X_1	1,544	1,159	0,448	0,812
X_2	1,425	0,910	0,601	1,504
Geral	1,486	1,042	0,518	1,076

	Média do agrupamento	
Variável	Agrupamento 1	Agrupamento 2
X_1	3,367	5,499
X_2	3,665	5,943

Estatística pseudo-F = 51,65

Se aumentarmos para $K = 3$ agrupamentos, o pseudo-F sobe ligeiramente para 59,0. Note que ambos os valores são relativamente pequenos se comparados ao valor 103,7 da solução de dois agrupamentos dos dados bimodais. No final das contas, talvez faça sentido usar um método diferente (tal como o agrupamento de densidade do k-ésimo vizinho mais próximo) para que se obtenha alguma evidência convergente com relação à modalidade natural desses dados.

8.6 EXEMPLO DE PROBLEMA: SEGMENTAÇÃO DA PREFERÊNCIA

Apresentamos agora uma aplicação dos métodos de agrupamentos descritos anteriormente para a segmentação de preferência. Os dados que temos são de 303 estudantes de MBA que participaram de um estudo on-line indicando suas preferências para 10 carros com preços comparáveis: BMW 328i, Ford Explorer, Infiniti J30, Jeep Grand Cherokee, Lexus ES 300, Chrysler Town & Country, Mercedes C280, Saab 9000, Porsche Boxster e Volvo V90. Os estudantes indicaram suas preferências para cada carro em uma escala de 1 (baixa) a 10 (alta). Para fins de validação (a ser discutida mais tarde, na Seção 8.7), deixamos de lado aproximadamente metade dos dados (151 observações).

EXEMPLO *Segmentação da preferência de automóveis entre estudantes de MBA. MBA_CAR*

Nosso objetivo explícito nessa análise é a segmentação. Queremos dividir os estudantes em um número administrável de grupos com preferências relativamente homogêneas. Assim, encontrar a modalidade natural desses dados é menos importante do que a variância da partição. Por essa razão, focaremos o uso de métodos como o de Ward e o agrupamento de *K*-means.

Começamos conduzindo um agrupamento de Ward dos dados, e examinamos o dendrograma para obtermos uma ideia do número apropriado de agrupamentos. Com base no dendrograma da Figura 8.18, os resultados sugerem uma solução de dois agrupamentos (correspondente à linha traçada em 0,10) ou talvez uma solução de cinco agrupamentos (correspondente à linha em 0,05).

Para os fins desta ilustração, prosseguiremos com a solução de cinco agrupamentos. Neste caso, poderíamos nos preocupar com o fato de a solução de dois agrupamentos ser muito simples; ela pode

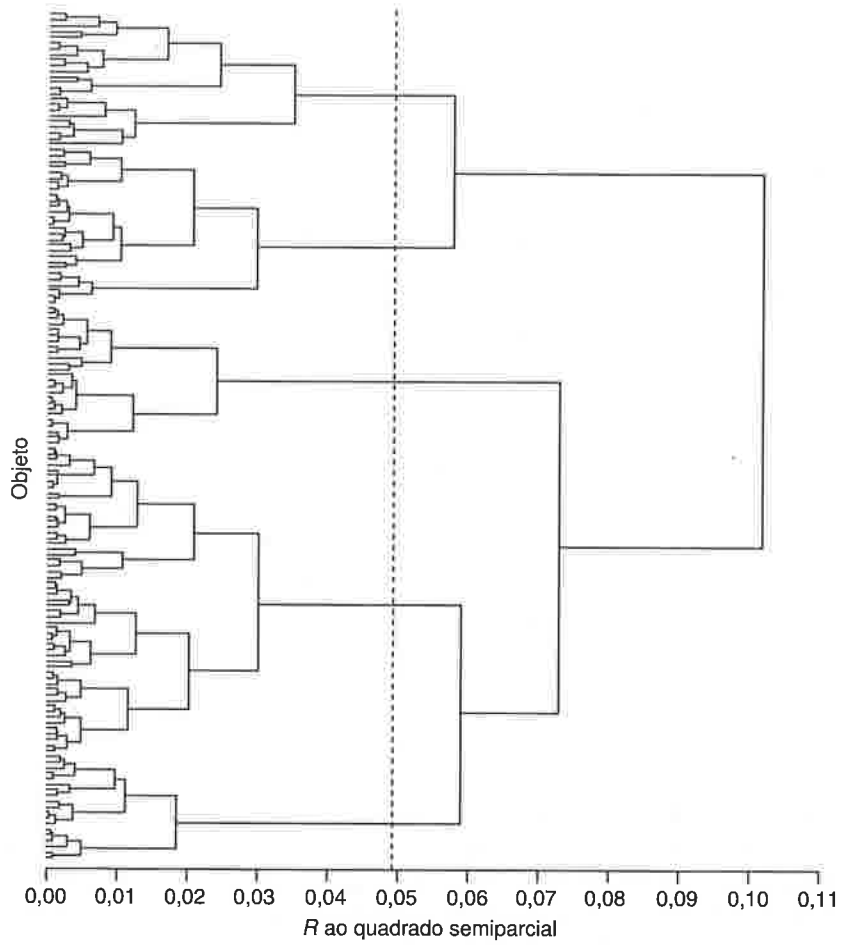


Figura 8.18 Dendrograma do método de Ward aplicado aos dados de preferência de carro.

não representar adequadamente as diferenças nas preferências dos estudantes. Uma solução de cinco agrupamentos é ainda administrável e possui o potencial de oferecer mais *insight* sobre os padrões diferentes de preferência presentes nos dados. Usando os centroides da solução de cinco agrupamentos do método de Ward como sementes, realizamos uma análise de agrupamento de *K*-means com $K = 5$. Verificamos que o pseudo-F da partição de cinco agrupamentos foi maior que aqueles das soluções com $K = 4$ e $K = 6$.

É importante lembrar que essa solução de cinco agrupamentos não corresponde necessariamente à modalidade natural desses dados. Para verificar, realizamos uma análise de agrupamentos hierárquica usando o método de densidade do k-ésimo vizinho mais próximo. Para qualquer valor de k maior que cinco, o resultado foi um agrupamento modal.

A Tabela 8.8 mostra o resumo dos agrupamentos – incluindo a filiação ao agrupamento, centroides dos agrupamentos e estatísticas para as medidas de preferência – para a solução de cinco agrupamentos.

Tabela 8.8 Resumo da análise de agrupamentos *K*-means de dados da preferência de carros para solução de agrupamentos $K = 5$

Agrupamento	Frequência
1	27
2	35
3	34
4	39
5	17

	Estatística para variáveis			
Variável	**STD Total**	**Dentro do STD**	R^2	$R^2/(1-R^2)$
BMW	1,880	1,577	0,314	0,459
Ford	2,172	1,898	0,257	0,346
Infiniti	1,795	1,692	0,136	0,157
Jeep	2,068	1,691	0,349	0,535
Lexus	1,917	1,747	0,192	0,237
Chrysler	1.348	1,268	0,139	0,161
Mercedes	1,818	1,645	0,203	0,255
Saab	2,150	1,989	0,167	0,201
Porsche	2,480	1,483	0,652	1,870
Volvo	2,237	1,590	0,508	1,032
Geral	2,008	1,669	0,327	0,487

	Média do agrupamento				
Variável	Agrupamento 1	Agrupamento 2	Agrupamento 3	Agrupamento 4	Agrupamento 5
BMW	6,93	6,74	7,24	5,90	3,71
Ford	2,93	6,37	4,59	4,74	4,88
Infiniti	4,37	4,31	4,50	2,97	3,18
Jeep	3,26	6,66	5,35	5,67	7,06
Lexus	5,74	5,66	5,44	3,67	4,82
Chrysler	1,11	2,46	1,41	1,54	2,24
Mercedes	7,11	6,17	6,06	4,87	4,88
Saab	5,85	5,94	4,24	3,95	4,41
Porsche	7,33	6,97	2,97	7,15	2,82
Volvo	2,93	5,37	2,74	2,08	6,53
Estatística pseudo-F = 17,89					

Os segmentos são relativamente próximos em tamanho (o menor é 17; o maior, 39). Verificando as estatísticas para cada medida de preferência, notamos que os agrupamentos separam-se mais fortemente com respeito à preferência pelo Porsche Boxster (o R^2 é 0,65, maior que qualquer outra variável de preferência). Examinando os centroides dos agrupamentos, percebemos que os segmentos 1, 2 e 4 expressam uma grande preferência pelo Porsche (aproximadamente 7 de 10), ao passo que os segmentos 3 e 5 expressam menor preferência (aproximadamente 3 em 10). Assim, a solução de agrupamentos capta o fato de as preferências dos estudantes pelo Porsche serem altamente polarizadas.

Em contraste, os agrupamentos não são particularmente distintos com respeito à preferência pelo Chrysler Town & Country (R^2 = 0,14). Isso se deve ao fato de os estudantes expressarem quase uniformemente pouca preferência por este carro, que é a única minivan incluída no estudo (não chega a ser um resultado surpreendente se considerarmos o fato de que os estudantes, nesta amostra em particular, são jovens, solteiros e sem filhos em sua maioria).

Também podemos comparar e contrastar os perfis dos segmentos. Os segmentos 1 e 2, por exemplo, exibem altas preferências pelo BMW e pelo Porsche. A diferença entre esses dois segmentos é que o segmento 1 expressa uma baixa preferência pela perua Volvo e pelos dois utilitários esportivos, Ford Explorer e Jeep Grand Cherokee (aproximadamente 3 de 10), enquanto o segmento 2 expressa preferências maiores (5 a 6 de 10). Os segmentos 2 e 5 dão as maiores classificações aos SUVs; a diferença é que o segmento 5 expressa preferências bem baixas para os carros esportivos (BMW e Porsche). Utilizando essas abordagens de perfil, é possível rotularmos os agrupamentos da mesma maneira de quando usamos as cargas fatoriais para rotular os fatores na análise fatorial.

8.7 QUESTÕES RELATIVAS À APLICAÇÃO DA ANÁLISE DE AGRUPAMENTOS

8.7.1 Como os dados devem ser escalonados?

Uma vez que o agrupamento envolve agrupar objetos que são similares ou próximos, é preciso saber como os dados são escalonados. Na construção de uma medida de distância, é importante que dimensões diferentes sejam comparavelmente escalonadas; do contrário, a variável com maior variância figurará com mais destaque na solução. É por isso que, como nos componentes principais, as variáveis são normalmente padronizadas. Esta geralmente deve ser a opção padrão, a menos que alguma outra base para o escalonamento de variáveis seja fornecida.

Como exemplo deste último, as variáveis na segmentação de preferência da Seção 8.6 são escalonadas comparavelmente: todas as preferências são medidas em uma escala de 10 pontos, de forma a tornar altamente plausível que esses números sejam comparáveis de carro a carro. Isso não implica que todas as medidas tenham a mesma variância; de fato, observamos uma variância muito maior para o Porsche (com média 5,72 e desvio-padrão 2,48) do que para o Chrysler Town & Country (com média 1,72 e desvio-padrão 1,35). Se dividirmos o desvio-padrão para padronizar essas medidas, estamos dizendo que uma diferença de um ponto na preferência pelo Chrysler (por exemplo, de 1 a 2) é efetivamente a mesma para a diferença em 4 pontos na preferência pelo Porsche (por exemplo, de um 3 para um 7). Isso pode ou não ser apropriado. Algumas vezes, podemos optar por conferir menos peso a uma variável que carrega menos informações (isto é, variância mais baixa) sobre as diferenças nas preferências dos estudantes.

Se nos interessarmos mais pelas classificações relativas das preferências do que por seus níveis absolutos (por exemplo, se uma classificação de preferência 4 de um estudante não for diretamente comparável ao 4 de outro estudante), podemos considerar a possibilidade de centrar na média ou padronizar os dados para cada respondente (isto é, por linha em vez de coluna). Note que isso tem basicamente o mesmo efeito de se usar o coeficiente de correlação como medida de similaridade, e é um procedimento apropriado somente quando o padrão da resposta é mais relevante do que a proximidade em nível. Não padronizamos os dados de preferência dos carros porque havia diferenças reais entre os estudantes em relação a quanto eles gostavam do conjunto de carros escolhidos para o nosso estudo. Isso foi sustentado por dados adicionais coletados sobre a intenção de aquisição de algum dos 10 carros do conjunto: os estudantes com baixos níveis de preferência por todos os carros relataram uma baixa

probabilidade de comprar algum deles, ao passo que os estudantes com altos níveis de preferência por alguns carros relataram uma grande probabilidade de eventualmente comprá-los.

8.7.2 Validação

Quando falamos sobre validação, estamos normalmente interessados em verificar a capacidade de generalização da solução. Em outras palavras, as descobertas de uma amostra específica podem ser estendidas para a população da qual a amostra foi retirada? Por exemplo, se tivéssemos calculado a relação entre atitude e comportamento para uma amostra de indivíduos, poderíamos testar e verificar se a relação estimada da amostra é compatível ao comportamento de outro grupo de indivíduos retirado da população. No contexto do agrupamento, no entanto, a solução é específica para um conjunto particular de objetos ou indivíduos. Não podemos usar a solução de agrupamento (isto é, os indicadores de quais objetos pertencem a quais agrupamentos) para dizer qualquer coisa sobre um objeto pertencente a um agrupamento de fora da amostra.

Se, entretanto, tivermos as medidas dos objetos em alguns atributos subjacentes – por exemplo, se os dados usados para a análise de agrupamentos são medidas de distância construídas e não avaliações diretas (possivelmente não métricas) da proximidade ou similaridade do objeto –, podemos nos concentrar nos centroides dos agrupamentos. A questão passa a ser: se aplicarmos a mesma metodologia de agrupamentos a duas amostras diferentes da mesma população, terminamos com os mesmos centroides de agrupamentos? E esses conjuntos diferentes de centroides resultam em uma atribuição similar de objetos aos agrupamentos? Essa abordagem não é diferente da abordagem de confiabilidade teste-reteste para se determinar a validade da solução de agrupamentos.

Uma abordagem possível a esse tipo de validação é delineada a seguir. Começamos com duas amostras da mesma população (neste caso, dividimos aleatoriamente uma amostra em duas). A primeira amostra é a de calibração e a segunda é a de validação. Usando o método escolhido de agrupamentos, realizamos então uma análise de agrupamentos sobre os dados de calibração, determinamos o número de agrupamentos e calculamos os centroides dos agrupamentos. Esses centroides, se a solução for válida, devem refletir as tendências centrais dos agrupamentos que existem na população como um todo. Assim, se tirarmos outra amostra da população, devemos ser capazes de determinar a que grupo pertence cada objeto e designá-lo para o centroide mais próximo. Então, é o que fazemos: atribuímos cada objeto na amostra de validação para o centroide mais próximo com base na solução dos dados de calibração.

PASSOS PARA A VALIDAÇÃO DE AGRUPAMENTO

1. Divida os dados em duas amostras aleatórias: calibração e validação.
2. Utilize o método de agrupamento para agrupar os dados de calibração. Determine o número apropriado de agrupamentos e calcule os centroides.
3. Use os centroides de agrupamentos dos dados de calibração, atribua cada observação da amostra de validação ao centroide mais próximo. Represente essa solução de agrupamento como S_1.
4. Use o mesmo método de agrupamento do passo 2 para agrupar os dados de validação. Escolha a solução com o mesmo número de agrupamentos, como determinado no passo 2. Represente essa solução de agrupamento como S_2.
5. As soluções de agrupamentos S_1 e S_2 representam diferentes atribuições do mesmo conjunto de observações para os agrupamentos. Para avaliar a concordância entre as duas soluções, faça uma tabulação cruzada S_1 *versus* S_2.

Em seguida, utilizamos o mesmo método de agrupamentos diretamente sobre os dados de validação para determinar uma solução de agrupamento com o mesmo número de agrupamentos. Isso nos dá um segundo conjunto de afiliações aos agrupamentos para os objetos na amostra de validação. A diferença é que o primeiro conjunto de atribuições é baseado na solução de agrupamento de uma amostra diferente de dados (isto é, a amostra de calibração). Se as amostras compartilharem as mesmas características da

população subjacente (e o sinal da população é forte em relação ao ruído da variação idiossincrática da amostra), devemos ter então atribuições altamente similares de agrupamentos. Uma maneira de verificar se isso ocorre é a tabulação cruzada de duas soluções. Se houver K agrupamentos, teremos uma tabela $K \times K$, na qual as linhas são atribuições de agrupamento baseadas nos centroides da amostra de calibração e as colunas são as atribuições de agrupamento de um agrupamento direto da amostra de validação. Se houver um alto grau de convergência, devemos esperar que a maioria das observações caia nas células K (não necessariamente ao longo da diagonal, porque o número de agrupamentos é arbitrário). Se as observações da amostra estiverem espalhadas por toda a tabela, isso sugere que a solução de agrupamento é específica da amostra e não necessariamente generalizável à população como um todo.

Ilustramos a abordagem usando os dados de preferência por carros descritos na seção precedente. Nossa análise anterior sugere que uma solução de cinco agrupamentos é apropriada.

Tabela 8.9 Solução de cinco agrupamentos (usando Ward e *K*-means) para validação dos dados da amostra

Agrupamento	Frequência
1	49
2	35
3	36
4	21
5	10

	Estatística para as variáveis			
Variável	**STD Total**	**Dentro do STD**	R^2	$R^2/(1-R^2)$
BMW	1,723	1,517	0,246	0,326
Ford	2,185	1,829	0,318	0,467
Infiniti	1,972	1,636	0,330	0,493
Jeep	2,100	1,787	0,296	0,420
Lexus	2,057	1,601	0,410	0,696
Chrysler	1,776	1,507	0,299	0,426
Mercedes	1,912	1,728	0,205	0,258
Saab	2,269	1,989	0,252	0,337
Porsche	2,396	1,693	0,514	1,058
Volvo	2,372	1,796	0,442	0,793
Geral	2,088	1,714	0,344	0,524

	Média do agrupamento				
Variável	Agrupamento 1	Agrupamento 2	Agrupamento 3	Agrupamento 4	Agrupamento 5
BMW	7,20	6,83	7,11	6,48	3,70
Ford	3,24	5,74	5,78	6,24	5,60
Infiniti	5,18	2,60	4,81	2,86	4,90
Jeep	3,71	6,34	6,08	6,05	5,60
Lexus	6,27	3,41	6,40	3,81	4,80
Chrysler	1,82	1,40	3,06	1,43	4,90
Mercedes	6,82	5,86	7,03	4,95	4,40
Saab	4,47	3,83	6,31	6,85	4,50
Porsche	6,43	7,77	6,86	3,05	2,44
Volvo	2,86	2,14	5,58	2,76	7,00
Estatística pseudo-F = 19,13					

Tabela 8.10 Tabulação cruzada das soluções de dois agrupamentos diferentes para dados de validação mostrando uma concordância relativamente alta

Tabela de Solução S_1 pela Solução S_2						
	1	2	3	4	5	Total
1	**33**	1	3	2	0	39
2	0	3	**31**	6	1	41
3	12	0	1	**9**	0	22
4	4	**30**	0	2	0	36
5	0	1	1	2	**9**	13
Total	49	35	36	21	10	151

Utilizando os centroides desses cinco agrupamentos, atribuímos cada indivíduo da amostra de teste para o agrupamento com o centroide mais próximo. Os tamanhos dos agrupamentos são 39, 41, 22, 36 e 13, respectivamente. Depois, agrupamos as observações diretamente na amostra de validação. Utilizamos o mesmo procedimento que foi usado com o primeiro conjunto de dados: primeiro executamos o método de Ward, usamos essa solução para fornecer as sementes iniciais para o algoritmo de K-means, e depois executamos K-means com $K = 5$. Os resultados são apresentados na Tabela 8.9.

A tabulação cruzada de atribuições para dois agrupamentos diferentes para a validação é apresentada na Tabela 8.10. Observe que há uma concordância razoavelmente forte entre as duas soluções: 112 das 151 observações caem nas cinco células que refletem a correspondência mútua mais alta entre os dois grupos de agrupamentos (essas células são mostradas em negrito na tabela). É uma taxa de acerto de 74%, que parece razoavelmente alta. Se isso representa um nível aceitável de validade, depende do julgamento do analista.

Uma possível medida de resumo para a correspondência entre duas soluções de agrupamento é o Índice de Rand (Rand, 1971). Esse índice é a proporção de todos os pares possíveis de objetos em que as duas partições diferentes concordam; isto é, em que o par de objetos é atribuído ao mesmo agrupamento em ambas as soluções ou em que o par de objetos é atribuído a diferentes agrupamentos em ambas as soluções. No exemplo acima, com $n = 151$, há $n!/(n-2)!2! = 151 \times 150/2 = 11.325$ pares distintos de objetos atribuídos aos agrupamentos. O número de pares nos quais as duas soluções estão em concordância pode ser calculado através da seguinte fórmula:

$$A = \binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left[\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2\right] \tag{8.11}$$

em que n_{ij} é o número de objetos na célula (i, j) da classificação cruzada apresentada na Tabela 8.10 (isto é, o número de objetos em comum entre o agrupamento i da solução S_1 e o agrupamento j da solução S_2) e $n_{i\cdot}$ e $n_{\cdot j}$ são a linha e a coluna das frequências marginais, respectivamente. Para o problema anterior, $A = 9{,}280$, resultando em um Índice de Rand de 9,280/11,325 = 0,82, o que sugere uma forte correspondência. Hubert e Arabie (1985) discutem algumas modificações do Índice de Rand e oferecem ideias sobre o teste de significância.

8.7.3 Nota adicional sobre o número de agrupamentos

É importante perceber que o teste de validação descrito anteriormente não é projetado para se validar o número de agrupamentos (veja Green e Krieger, 1999). Em outras palavras, só porque obtemos uma correspondência relativamente alta entre as duas soluções de agrupamentos não significa que podemos concluir que a modalidade natural dos dados é cinco. Na verdade, se não estamos interessados na modalidade natural dos dados, devemos usar um método como a densidade do k-ésimo vizinho mais próximo. Embora não haja testes estatísticos para o número de agrupamentos, podemos usar o método da densidade do k-ésimo vizinho mais próximo para obter alguma noção de quão fortemente multimodais os dados são. Essa é a ideia subjacente à abordagem proposta por Wong e Schaack (1982).

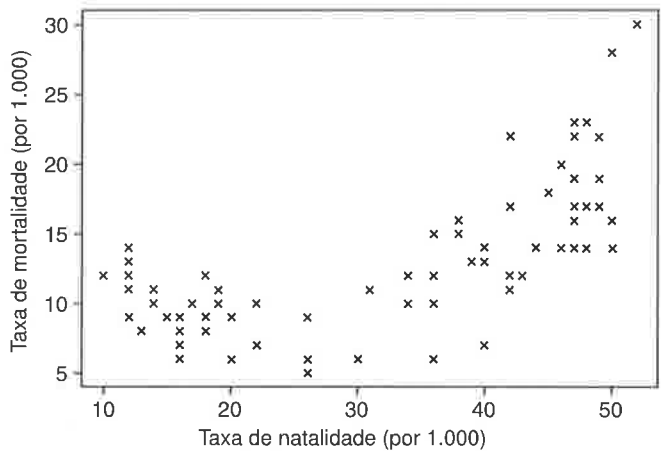


Figura 8.19 Gráfico da taxa de natalidade *versus* a taxa de mortalidade de 74 países.

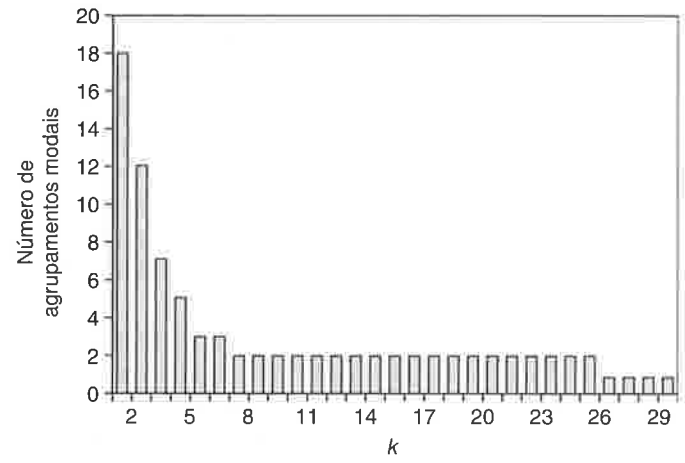


Figura 8.20 Número de agrupamentos modais nos dados sobre taxa de natalidade e taxa de mortalidade encontrados na análise de agrupamentos pela densidade do *k*-ésimo vizinho mais próximo para diferentes valores de *k*.

Lembre-se de que a abordagem utiliza uma estimativa da densidade na vizinhança de cada ponto para estabelecer uma medida de proximidade entre os pontos. Para valores pequenos de k, as estimativas da densidade local são baseadas em poucas observações. Por exemplo, quando $k = 1$, podemos observar muitos pares de pontos em que cada observação está sozinha na k-ésima vizinhança mais próxima de outra. Cada subconjunto constitui uma "moda" nos dados porque nenhuma observação cai na vizinhança de qualquer outro ponto dos dados. Para pequenos valores de k, tendemos a observar muitos desses subconjuntos desconexos nos dados (isto é, nenhuma observação em um grupo está conectada a outra observação em outro grupo). À medida que o valor de k aumenta, o tamanho da vizinhança usada para se estimar a densidade local também aumenta. Isso "suaviza" as estimativas da densidade local, e o número de modas nos dados decresce. Finalmente, para alguns valores de k, o número de modas cai para 1.

Pode-se, razoavelmente, esperar que dados fortemente multimodais levem a um valor maior de k antes que as estimativas da densidade local sejam suficientemente suavizadas para que observemos apenas um agrupamento modal. Por exemplo, considere os conjuntos de dados unimodal e bimodal introduzidos na Seção 8.2. Para os dados unimodais, um único agrupamento modal (isto é, um conjunto conectado) é encontrado para todos os valores de $k \geqslant 4$. Para os dados bimodais, um único agrupamento modal não ocorre antes de $k \geqslant 12$. Portanto, se virmos mais de um conjunto desconexo de observações para uma grande variedade de valores de k, teremos uma indicação mais forte da multimodalidade nos dados.

Para ilustrar, considere os dados das taxas de natalidade e mortalidade (por 1.000 habitantes) para 74 países em 1976. Esses dados são apresentados na Figura 8.19. Realizamos uma análise de agrupamento do k-ésimo vizinho mais próximo sobre esses dados para todos os valores de k de 3 a 50. O número de agrupamentos modais para cada valor de k é esquematizado no gráfico da Figura 8.20. A figura sugere fortemente que esses dados são pelo menos bimodais: dois agrupamentos modais são formados para todos os valores de k entre 8 e 26.

8.8 RESUMO DA APRENDIZAGEM

- A análise de agrupamento envolve categorização: dividir um grande número de objetos em grupos menores para que os objetos dentro de cada grupo sejam relativamente similares e os objetos em diferentes grupos sejam relativamente dissimilares.
  - A maioria dos algoritmos de agrupamento é heurística por natureza porque é computacionalmente inviável buscar uma solução ótima através de todos os resultados de agrupamento do tipo prescrito.
  - Normalmente, o objetivo da análise de agrupamentos é tratar a heterogeneidade em cada grupo de dados. No entanto, algumas vezes, a meta é procurar agrupamentos que ocorrem naturalmente. Observe que, mesmo quando os dados são unimodais (isto é, somente um agrupamento ocorre naturalmente), ainda é possível reduzir posteriormente a heterogeneidade, dividindo-se os dados em agrupamentos (ou segmentos) *ad hoc*.
- Neste capítulo, discutimos métodos de agrupamento hierárquico, aglomerativo e métodos de partição.
  - Os métodos aglomerativos começam com todos os objetos em agrupamentos separados e prosseguem reunindo-os. O resultado de uma análise de agrupamento aglomerativo é uma hierarquia de partições aninhadas, em que a solução de agrupamento k pode ser obtida reunindo-se dois grupos da solução de agrupamento $(k + 1)$.
  - Os métodos de partição dividem o conjunto de todos os objetos em um número determinado de grupos diferentes mutuamente exclusivos e coletivamente exaustivos. Esses métodos não são hierárquicos.
- Os métodos de agrupamento aglomerativo geralmente usam como entrada alguma medida de proximidade (direta ou derivada) entre os objetos. Essas medidas podem tomar formas diferentes:
  - *distância* (uma medida métrica; por exemplo, a distância euclidiana entre dois objetos de todo um conjunto de características medidas);
  - *dissimilaridade* (uma medida métrica ou não métrica; por exemplo, as classificações da escala intervalar ou da escala ordinal de pares de objetos do mais similar ao menos similar); ou
  - *densidade* (uma medida métrica que reflita não somente a proximidade de dois objetos, mas também o número de outros objetos na vizinhança que os cerca).
- Métodos de partição geralmente exigem dados multivariados (por exemplo, multiatributos).
- O agrupamento aglomerativo é baseado em um algoritmo heurístico que reúne os dois agrupamentos mais próximos, calcula uma medida de proximidade entre o agrupamento novo e os agrupamentos remanescentes e procede com a iteração até que todos os objetos tenham sido reunidos. Alguns dos modos diferentes de se definir a proximidade entre os agrupamentos são:
  - *Ligação simples.* Se definirmos a dissimilaridade entre dois agrupamentos como a menor dissimilaridade entre qualquer objeto em um agrupamento e qualquer objeto em outro, o método é conhecido como *agrupamento de ligação simples.* Essa abordagem é bem eficiente do ponto de vista computacional e pode ser usada com dados de proximidade métricos ou não métricos. Porém, ela tem a tendência de produzir agrupamentos como em cadeia.
  - *Ligação completa.* É similar à ligação simples, com exceção de a dissimilaridade entre dois agrupamentos ser definida como a maior dissimilaridade entre qualquer objeto em um agrupamento e qualquer objeto no outro. Essa abordagem tende a produzir agrupamentos convexos e de tamanhos iguais, e pode ser suscetível a discrepâncias.

- *Ligação média.* É similar à ligação simples, com exceção de a dissimilaridade entre dois agrupamentos ser definida como a dissimilaridade média entre todos os pares de objetos (i, j) quando o objeto i está no primeiro dos dois agrupamentos e o objeto j está no segundo.
- *Método de Ward.* Em vez de reunir os dois agrupamentos mais próximos, o método de Ward agrega dois agrupamentos que levam à menor soma de quadrados dentro do agrupamento em cada etapa. Essa abordagem possui a tendência de produzir agrupamentos convexos de tamanhos iguais.
- *Ligação por densidade.* Para evitar a miopia associada com a ligação simples, essa abordagem usa a medida da densidade do k-ésimo vizinho mais próximo em vez de uma medida simples de dissimilaridade. Esse método tem probabilidade maior de revelar a modalidade subjacente dos dados (porque tende a reunir objetos que estão em vizinhanças de alta densidade).

- Dispomos os resultados dos métodos de agrupamento aglomerativo usando um diagrama do tipo árvore chamado de *dendrograma*. O dendrograma é uma representação gráfica que exibe a hierarquia das soluções de agrupamentos aninhados de n agrupamentos a um agrupamento. As "quebras" no dendrograma sugerem o que pode ser um número apropriado de agrupamentos usados para resumir os dados.
- Discutimos uma abordagem amplamente usada para partição conhecida como agrupamento K-means. Trata-se de um procedimento de otimização interativo que começa com uma partição inicial, calcula os centroides de cada agrupamento na partição e depois atribui novamente os objetos ao agrupamento com o centroide mais próximo. O procedimento continua até que a partição permaneça sem mudança.
  - Como o K-means possui a tendência de chegar a soluções localmente ótimas (as quais podem ser relativamente afastadas do ótimo global), é importante realizar a análise várias vezes usando diferentes pontos iniciais e, então, escolher a melhor solução.
- Para decidir o número apropriado de agrupamentos do agrupamento K-means, é necessário realizar a análise de vários valores diferentes de K e então comparar as soluções. Uma medida métrica para comparar as soluções é chamada de *estatística pseudo-F*, que é dada por

$$\text{pseudo-}F = \frac{\text{tr}[\mathbf{B}/(K-1)]}{\text{tr}[\mathbf{W}/(n-K)]}$$

  onde **B** é a matriz da soma dos quadrados entre os agrupamentos e **W** é a matriz da soma dos quadrados dentro do agrupamento.
- Se um objetivo da análise de agrupamento for a generalização, é importante validar os resultados de algum modo. A análise de agrupamentos pode ser fortemente influenciada pela variação idiossincrática da amostragem das observações usadas para determinar a solução do agrupamento.

LEITURAS SELECIONADAS

Geral

ANDERBERG, M. *Cluster analysis for applications.* Nova York: Academic Press, 1973.

ARABIE, P.; HUBERT L.; DESOETE G. *Clustering and classification.* River Edge: World Scientific, 1996.

HARTIGAN, John A. *Clustering algorithms.* Nova York: John Wiley and Sons, 1975.

SNEATH, P. H. A.; SOKAL, R. R. *Numerical taxonomy.* San Francisco: Freeman Press, 1963.

Ligação simples, completa e média

CARROLL, J. D. "Hierarchical clustering". In GREEN; WIND, *Multiattribute Decisions in Marketing.*

HARTIGAN, John A. "Representations of similarity matrices by trees", *Journal of American Statistical Association*, v. 62, p. 1140-1158, 1967.

JOHNSON, S. C. "Hierarchical clustering schemes", *Psychometrika.* v. 32, p. 241-254, 1967.

Método de Ward

WARD, J. "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association*, v. 58, p. 236-244, 1963.

***K*-ésimo vizinho mais próximo**

WONG, M. Anthony; LANE, T. "A k^{th} nearest neighbor clustering procedure", *Journal of the Royal Statistical Society*, Series B, v. 45, p. 362-368, 1983.

***K*-means**

HARTIGAN, John; WONG, M. Anthony. "Algorithm AS136: a K-means clustering program", *Applied Statistics*, v. 28, p. 100-128, 1979.

Avaliação e Validação

GREEN, Paul E.; KRIEGER, Abba M. "A cautionary note on using internal cross validation to select the number of clusters", *Psychometrika*, v. 64, n. 3, p. 341-353, 1999.

HUBERT, L.; ARABIE Phipps. "Comparing partitions", *Journal of Classification*, v. 2, p. 193-218, 1985.

MCINTYRE, R. M.; BLASHFIELD, R. K. "A nearest centroid technique for evaluating the minimum variance clustering procedure", *Multivariate Behavioral Research*, v. 15, p. 225-238, 1980.

MILLIGAN, G. "A monte carlo study of thirty internal criterion measures for cluster analysis", *Psychometrika*, v. 46, p. 187-199, 1981.

MILLIGAN, Glenn W.; COOPER, Martha C. "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, v. 50, n. 2, p. 159-179, 1985.

EXERCÍCIOS

8.1 Como os agrupamentos obtidos utilizando-se o método de Ward podem diferir daqueles obtidos usando-se o agrupamento de ligação simples? O que explica essa diferença?

8.2 Responda às seguintes questões usando os dados sobre íris de Fischer (descritos em mais detalhes no problema 4.5 e disponíveis no arquivo *IRIS*). Use somente as medidas do comprimento da sépala e da largura e comprimento da pétala como dados para sua análise.

a. Qual é a modalidade natural desse conjunto de dados? Em outras palavras, parece haver mais de um agrupamento natural nesses dados?
b. Usando um método de agrupamento apropriado, encontre uma solução de três agrupamentos para os dados das íris. Como a sua solução de agrupamento corresponde às espécies reais identificadas no conjunto de dados?

8.3 Considere a matriz de julgamentos de similaridade descrita no problema 7.2 (e disponível no arquivo *SIMILARITY*). Faça uma análise de agrupamento desses dados e descreva seus resultados. Como as suas conclusões se comparam àquelas da análise de MDS dos mesmos dados?

8.4 Considere os dados sobre detergentes de lavar roupa que as famílias adquirem em lojas descritas no problema 7.9 (e disponíveis no arquivo *STORE_SHARE*). Faça uma análise de agrupamento para identificar grupos de consumidores com padrões similares de compra de detergente para lavar roupa. Descreva seus resultados.

8.5 Um pesquisador acredita que duas lojas competem quando os consumidores que compram em uma loja tendem também a comprar na outra. Quanto maior o potencial de mudança entre as lojas, maior a competição. Usando os dados do problema 8.4 (no arquivo *STORE_SHARE*), ele calcula a seguinte medida de força da competição entre a loja i e a loja j, representada por C_{ij}:

$$C_{ij} = \sum_{h=1}^{45} m_i^h m_j^h$$

onde h é um índice de famílias e m_j^h é a participação da escolha da loja pela família h pela loja j. A matriz **C** é apresentada na Tabela 8.11 (e está disponível no arquivo *SHOPPING*).

Tabela 8.11 Matriz de produto cruzado de compartilhamento de escolha de lojas

1	2	3	4	5	6	7	8	9	10	11	12	13
0,706	0,178	0,456	0,088	0,003	0,152	0,024	0,000	0,517	0,083	0,000	0,047	0,005
0,178	0,850	0,866	0,107	0,242	0,012	0,038	0,136	0,141	0,299	0,049	0,296	0,146
0,456	0,866	4,867	0,041	0,192	0,479	0,553	1,159	0,458	1,052	0,319	1,018	0,645
0,088	0,107	0,041	0,464	0,020	0,179	0,012	0,000	0,129	0,177	0,077	0,062	0,000
0,003	0,242	0,192	0,020	0,610	0,072	0,015	0,035	0,127	0,336	0,007	0,194	0,004
0,152	0,012	0,479	0,179	0,072	0,960	0,000	0,162	0,492	0,035	0,177	0,072	0,000
0,024	0,038	0,553	0,012	0,015	0,000	0,556	0,063	0,000	0,149	0,008	0,000	0,036
0,000	0,136	1,159	0,000	0,035	0,162	0,063	1,236	0,133	0,048	0,037	0,259	0,221
0,517	0,141	0,458	0,129	0,127	0,492	0,000	0,133	1,273	0,131	0,071	0,256	0,025
0,083	0,299	1,052	0,177	0,336	0,035	0,149	0,048	0,131	1,694	0,078	0,275	0,174
0,000	0,049	0,319	0,077	0,007	0,177	0,008	0,037	0,071	0,078	0,389	0,348	0,043
0,047	0,296	1,018	0,062	0,194	0,072	0,000	0,259	0,256	0,275	0,348	1,191	0,285
0,005	0,146	0,645	0,000	0,004	0,000	0,036	0,221	0,025	0,174	0,043	0,285	0,556

Usando a informação na matriz **C**, o pesquisador quer representar de algum modo o padrão de competição no mercado. Realize uma análise que permitirá ao pesquisador entender melhor a competição de lojas neste mercado. Certifique-se de justificar a medida que você usa em sua análise (pode ser necessária alguma manipulação de dados em **C**) e a abordagem que você tomou.

8.6 Considere os dados do estudo de Bucklin e Lattin sobre a competição entre lojas descrito no começo deste capítulo (e disponível no arquivo *STORE_SWITCH*). Baseado em sua análise, Bucklin e Lattin identificaram quatro agrupamentos de lojas concorrentes no mercado: (3, 26, 36), (4, 16, 43, 45), (7, 10, 24) e (18, 21, 19). Conduza sua própria análise de agrupamentos dos dados em *STORE_SWITCH*. O que você conclui sobre a estrutura de agrupamento de lojas em competição nesse mercado?

8.7 Considere a coincidência dos dados da cesta de mercado descrita no problema 7.7 (e disponível no arquivo *BASKET*).

a. Examine os dados você mesmo usando uma análise de agrupamentos. Qual método você considera ser o mais apropriado? Por quê? Como os dados precisam ser escalonados (se é que precisam)? Faça uma apresentação sucinta de suas descobertas.

8.8 Fader e Lodish coletaram dados no *IRI Marketing Factbook* sobre 10 variáveis para 331 categorias diferentes de mantimentos durante o ano calendário de 1986. As primeiras cinco variáveis captaram o que Fader e Lodish chamaram de "características estruturais" da categoria (isto é, os aspectos da categoria que provavelmente não seriam alterados substancialmente por uma atividade promocional de curto prazo); essas variáveis (disponíveis no arquivo *FACTBOOK*) estão descritas na Tabela 8.12.

Tabela 8.12

PENET	Porcentagem de famílias que fizeram pelo menos a aquisição de uma categoria.
PCYCLE	Tempo médio entre as aquisições.
PRICE	Média de dólares gastos na categoria por ocasião da aquisição.
PVTSH	Participação combinada no mercado para todos os produtos de rótulo privado e genéricos.
PUR/HH	Número médio de ocasiões de compra por família durante o ano.

Em sua análise das características estruturais das categorias de mantimentos, Fader e Lodish concluíram que "as configurações dos agrupamentos envolvendo [as variáveis] *PVTSH* e *PRICE* não eram muito estáveis ou significativas". Essa é a oportunidade para você mesmo examinar melhor as declarações. Faça uma análise de agrupamentos usando as variáveis estruturais *PENET*, *PURHH*, *PCYCLE*, *PRICE* e *PVTSH* e responda as seguintes questões:

a. Se você tivesse que separar essas categorias em agrupamentos, com base em suas características estruturais, quantos agrupamentos você escolheria? Explique as razões para sua recomendação.
b. Descreva as diferenças entre os agrupamentos da solução proposta por você.
c. Você concorda, como determinaram Fader e Lodish, que os resultados da análise de agrupamentos não são estáveis nem significativos? Por que sim ou por que não?

8.9 Considere os dados sobre a proporção de famílias em diferentes países com diferentes tipos de alimentos (descritos no problema 7.13 e disponíveis no arquivo *INTL_FOODS*).

a. Proponha e calcule uma medida apropriada de similaridade entre os países (baseada em padrões similares de consumo de alimento). Usando a medida proposta, faça uma análise de agrupamentos dos países e descreva seus resultados.
b. Proponha e calcule uma medida apropriada de similaridade entre alimentos (baseada em padrões similares de demanda dos países). Utilizando a medida proposta, realize uma análise de agrupamentos dos países e descreva seus resultados.
c. Como os *insights* em sua análises dos itens "a" e "b" anteriores comparam-se àqueles obtidos na análise de revelação dos dados no problema 7.13?

8.10 Grover e Srinivasan (1987) examinaram o comportamento de mudança de preferência de mais de 4.500 consumidores por 11 marcas e tipos diferentes de cafés instantâneos. A Tabela 8.13 (disponível no arquivo *COFFEE*) mostra a mudança de comportamento relatado por um subconjunto de 1.553 consumidores.

Note, por exemplo, que o número de vezes em que observamos a mudança do descafeinado normal High Point e do descafeinado normal Sanka é 43, mas o número de vezes em que observamos uma mudança entre o descafeinado normal High Point e o descafeinado liofilizado Sanka é somente 1. Assim, podemos concluir que há um grau de competição muito maior entre as diferentes marcas do mesmo formato do que entre formatos diferentes.

a. Qual medida você usaria para refletir o nível de competição entre as diferentes alternativas de café instantâneo?
b. Usando a medida proposta no item "a", realize uma análise de agrupamento dos dados de Grover e Srinivasan. Como você descreveria a estrutura competitiva do mercado de café instantâneo?

Tabela 8.13 Mudando a preferência por marcas e tipos de café instantâneo

	1	2	3	4	5	6	7	8	9	10	11
High Point Decaffeinated Regular	93	7	17	19	18	43	1	4	6	7	10
Tasters Choice Caffeinated Freeze Dried	9	80	12	11	24	7	4	2	6	3	3
Tasters Choice Decaffeinated Freeze Dried	9	14	46	3	7	7	4	2	2	0	9
Folgers Caffeinated Regular	19	18	4	82	29	14	0	4	9	2	6
Maxwell House Caffeinated Regular	26	11	6	35	184	24	3	11	18	6	6
Sanka Decaffeinated Regular	15	13	8	13	28	127	4	3	3	8	8
Sanka Decaffeinated Freeze Dried	2	0	3	2	1	7	17	3	0	1	4
Maxim Caffeinated Freeze Dried	4	3	4	3	6	5	2	27	1	0	4
Nescafe Caffeinated Regular	5	3	2	4	16	4	0	1	46	9	2
Nescafe Decaffeinated Regular	6	1	4	1	5	9	0	0	11	15	2
Brim Decaffeinated Freeze Dried	10	4	4	4	2	10	2	2	5	2	27

8.11 Considere os dados coletados por John Jones que descrevem as dissimilaridades entre 30 organizações (descritos no problema 7.10 e disponíveis no arquivo *ORG_DISSIM*).

a. Qual método de agrupamento você usaria para analisar esses dados e por quê?
b. Analise os dados e interprete seus resultados.

8.12 O arquivo *RANDOM_1* contém 100 observações retiradas aleatoriamente de uma distribuição uniforme de duas dimensões; isto é, do quadrado da unidade $U[0,1] \times U[0,1]$. O arquivo *RANDOM_2* contém uma amostra de teste de 100 observações retiradas do mesmo espaço.

a. Usando agrupamento de K-means, encontre uma solução de quatro agrupamentos para os dados em *RANDOM_1*.
b. Seguindo o procedimento descrito no texto, valide a solução de agrupamento do item "a", usando os dados em *RANDOM_2*. Descreva seus resultados. Você diria que a solução de quatro agrupamentos é válida?
c. Repita os itens "a" e "b" acima para uma solução de três agrupamentos. Quais são as diferenças entre a validação de três agrupamentos e a validação de quatro agrupamentos? Como você as explicaria?

8.13 Como no problema 8.12, os dados nos arquivos *ROUND_1* e *ROUND_2* são também retirados de uma distribuição aleatória. Cada arquivo contém 100 observações; os dados em *ROUND_2* servem como amostra de teste.

a. Usando agrupamentos de *K*-means, encontre uma solução de quatro agrupamentos para os dados em *ROUND_1*.
b. Seguindo o procedimento descrito no texto, valide a solução de agrupamento do item "a", utilizando os dados em *ROUND_2*. Descreva seus resultados. Qual a diferença entre os resultados da validação de quatro agrupamentos do problema 8.12? O que explica essas diferenças na sua opinião?