Análise de Cluster

Introdução

Dos livros A.D.M (D.Peña) e A.D.M. (J.Lattin, et al)

[1] EACH-USP
Universidade de São Paulo

Outline

Outline

Introdução

É uma técnica que divide um grande grupo em grupos menores para que as observações dentro de cada novo grupo sejam relativamente similares.

Comparações com MDS

- 1. Ambas são metodologias de construção de representações de objetos baseados em similaridades ou dissimilaridades.
- 2. A principal diferença é que em MDS se contrõem representações espaciais e contínuas e em A.Agrupamentos, representações não espaciais e discretas, podendo ser conjuntos sobrepostos ou não sobrepostos.
- 3. MDS fornece variáveis de escala intervalar que definem as localizações de cada objeto no espaço, já A.A. fornece variáveis de escala nominal que indicam se cada objeto pertence ou não a um conjunto (em determinado tipo de agrupamentos).
- 4. A.A. é realizado com objetivo de tratar a heterogeneidade dos dados.

Tipos

Existem principalmente dois métodos para trabalhar com A.A:

- 1. Métodos hierárquicos (estrutura de árvore)
- 2. Métodos de partição (separam as observações em um número distinto de grupos)

Propriedades:

- A atribuição resultante dos objetos nos agrupamentos é mutuamente exclusiva e coletivamente exaustiva
- Há métodos que permitem agrupamentes sobrepostos e agrupamentos indistintos.

Métodos Hierárquicos

- 1. De baixo para cima (métodos aglomerativos, até obter um agrupamento determinado n)
- 2. De cima para baixo (médodos divisíveis, divide-se em dois a cada etapa até que restem n agrupamentos de tamanho 1.

Observações:

- 1. quase todos os problemas de agrupamentos de qualquer tamanho apreciável exigem uma solução heurística.
- 2. O número de modos diferentes de dividir n objetos em n agrupamentos de tamanho $n_1, n_2, ...n_m$

$$A = \frac{n!}{n_1!n_2!, ...n_m!}$$

Por exemplo, se n=20, m=4 e $n_i = 5 \forall i$ $A \geq 488$ m.
Como é difícil saber as subdivisões dos conjuntos, adota-se uma abordagem heurística.

Aplicações Potenciais

- 1. Taxonomia numérica: identificar e discriminar diferentes espécies e sub-espécies
- 2. Segmentação de Mercado: Dividir o mercado em grupos menores e mais homogêneos e oferecer os produtos. Exemplo: Dados de preferência por cerveja medidos em escala de 9 pontos.
- 3. Análise de estrutura de mercado: quando queremos saber quais consumidores estão comprando o produto. I,é, quais outros produtos estão no conjunto competitivo

Objetivos da AA

- 1. Tratar a heterogeneidade dos dados
- 2. Encontrar uma modalidade natural dos dados.

Medidas de distância

- 1. Distância euclidiana: quando as variáveis possuem propriedades métricas:

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

onde d_{ij} é a distância euclidiana entre objetos i e j (é aplicada para dados na mesma escala, normalmente padronizados).

- 2. Métrica p ou L_p de Minkowski: Pode ser mais apropriada em alguns casos:

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

onde $p = 2$ é a distância euclidiana, é um caso particular.

$p = 1$ é

$$d_{ij}(1) = \sum_k |x_{ik} - x_{jk}| \tag{3}$$

é chamada métrica de quarteirão que é como andar de um ponto A a um ponto B em uma cidade disosta em sistema de grades de ruas, formando ângulos retos entre si.
$p = \infty$ da a sup-métrica:

$$d_{ij}(\infty) = mx(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, ..., |x_{ip} - x_{jp}|) \tag{4}$$

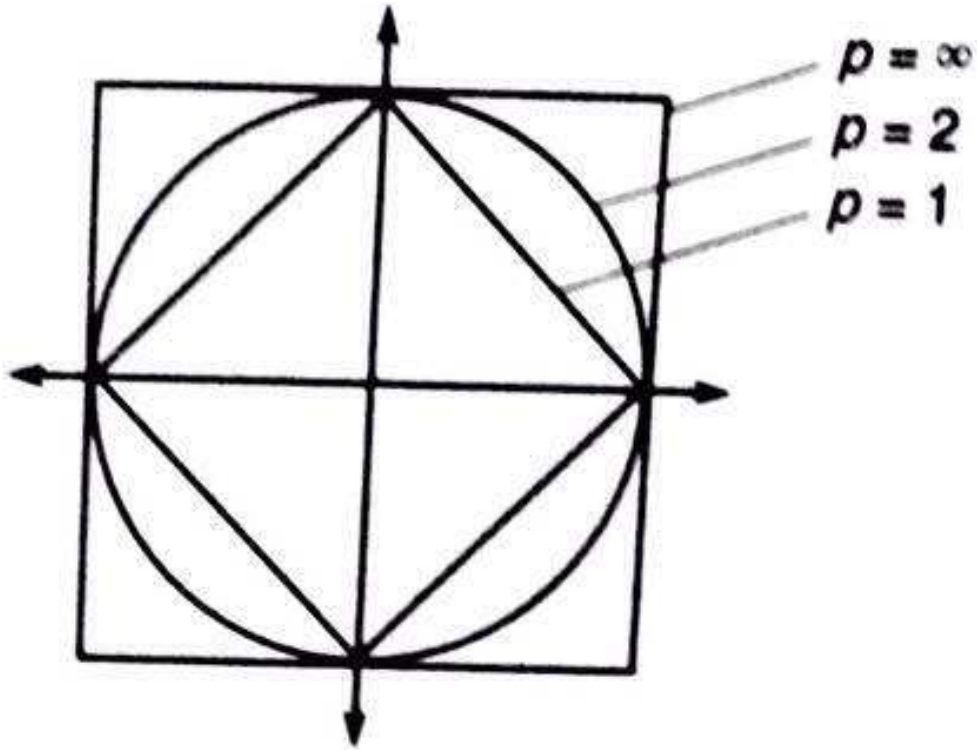


Figura 8.7 Círculos de unidade para três diferentes métricas de Mink

- 3. Distância de Mahalanobis: é uma medida ajustada para a covariância:

$$D_{ij}^2 = (\boldsymbol{x_i} - \boldsymbol{x_j})' \boldsymbol{\Sigma^{-1}} (\boldsymbol{x_i} - \boldsymbol{x_j}) \tag{5}$$

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

A medida D^2 pode ser mais bem entendida no contexto de dados distribuidos de acordo com uma normal. A figura abaixo mostra um exemplo de pontos retirados de uma distribuição normal bivariada (centrada na origem) com covariância positiva. Observe que A está na mesma distancia de mahalanobis a partir da origem que o ponto B. (tem a mesma probabilidade de terem sido extraídos de uma normal bivariada com centro em $(0, 0)$.

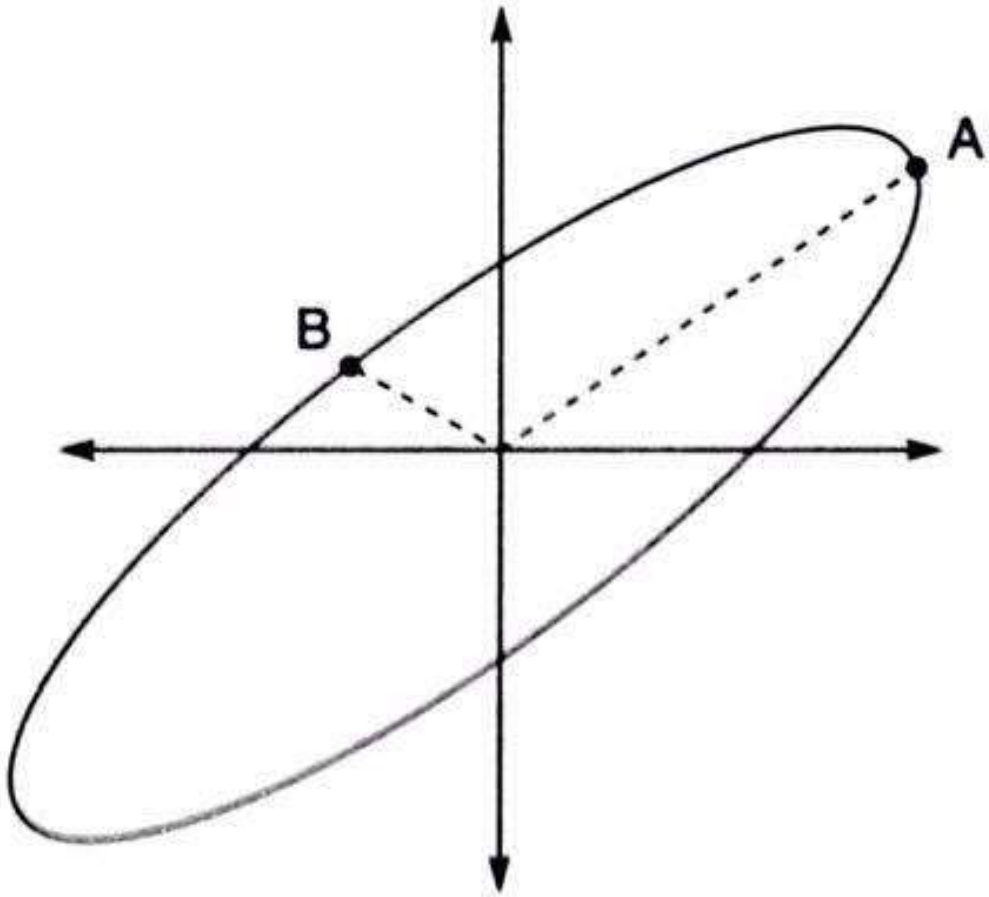


Figura 8.8 Diagrama demonstrando a distância de Mahalanobis: um da isodistância em duas dimensões.

Medidas de correspondência

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

cautela sobre correlação

o ρ nem sempre é uma medida apropriada de similaridade, pois é uma medida de covariância que é também um tipo de proximidade, as não necessáriamente uma medida de similaridade.

Medidas de densidade

As medidas anteriores (com exceção da distância de Mahalanobis), não levam em consideração a posição dos pontos em relação aos outros objetos. Por exemplo, na figura abaixo, os pontos A estão na vizinhança de um conjunto de dados (alta densidade) enquanto os pontos B estão em uma zona de baixa densidade, mesmo eles tendo a mesma distância.

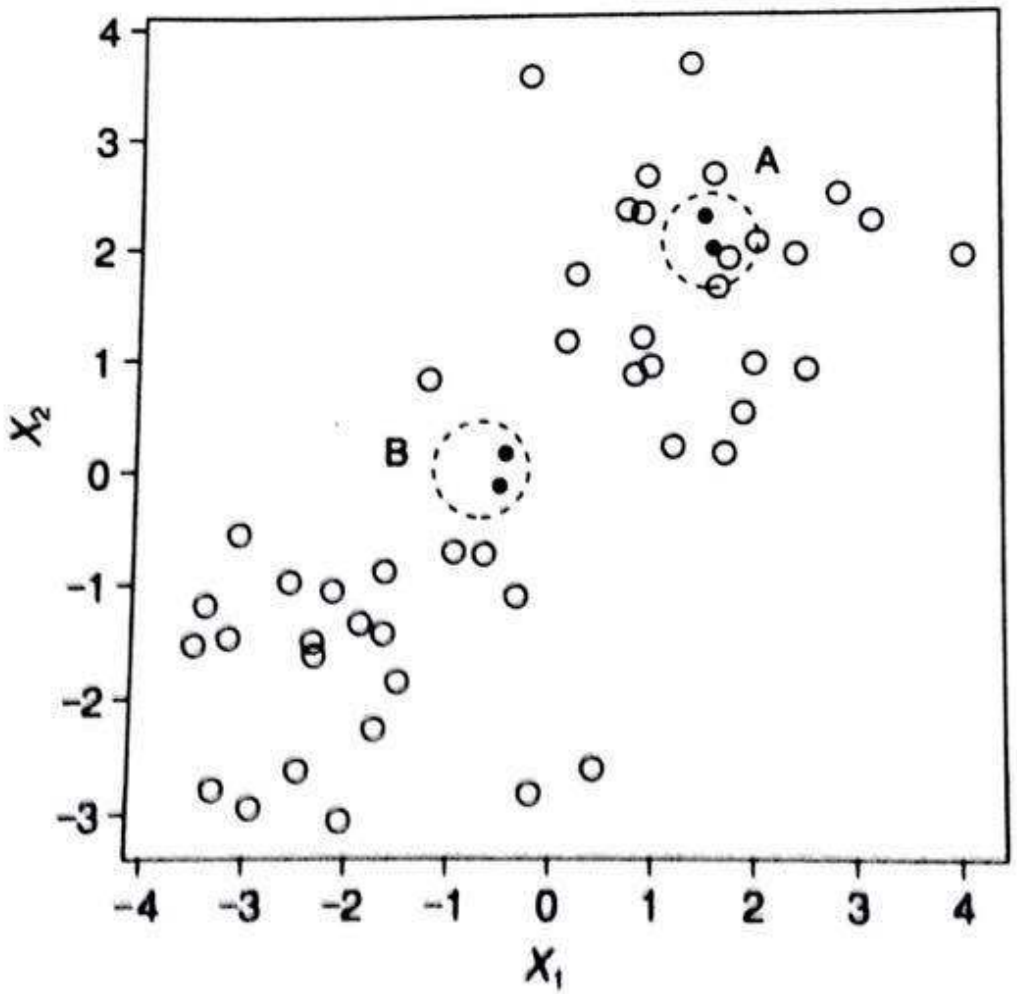


Figura 8.9 Dois pares diferentes de pontos: mesma distância, densidades diferentes.

Etapas para calcular a k-ésima densidade de vizinhança

- 1. para cada i calcular a distância ao k-ésimo vizinho mais próximo $(d_i(k))$. Observe que $(d_i(k))$ está inversamente relacionado à densidade relativa da região que cerca o objeto i. Se $(d_i(k))$ for pequena então a densidade é alta.
- 2. Conecte todos os i e j em que i está na k-ésima vizinhança mais próxima de j ou j está na k-ésima vizinhança de i (ou ambos). Aqui selecionamos os melhores candidatos para o mesmo grupo. Em vizinhanças de alta densidade, dois objetos serão conectados somente se estiverem especialmente próximos um do outro em termos de distância euclidiana, ao contrário em regiões de baixa densidade.
- 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. Os primeiros objetos a serem agrupados serão aqueles em regiões de alta densidade que estejam próximos uns dos outros.

Intuição

A ideia é: começe com cada objeto em seu próprio agrupamento isolado (n grupos de tamanho 1), em cada etapa do processo, encontre dois agrupamentos mais próximos e junte-os. Continue até que reste um agrupamento de tamanho n.

Etapas

- 0. n agrupamentos com um objeto: $C_1, C_2, ..., C_n$. A distância ente dois agrupamentos é definida como a distância entre dois objetos nele contidos

$$d_{C_i C_j} = d_{ij}$$

Seja $t = 1$ um índice do processo iterativo:

- 1. Encontre a menor distância entre dois agrupamentos quaisquer. Represente-os por C_i e C_j
- 2.Combine C_i e C_j para formar um novo agrupamento C_{n+1}
- 3. Defina a distância entre C_{n+1} e todos os C_k como

$$d_{C_{n+1} C_k} = min\{d_{C_i C_k}, d_{C_j C_k}\}$$

- 4. Adicione C_{n+1} e remova C_i e C_j. Considere $t = t + 1$
- 5. Volte à etapa 1 e continue até que reste um agrupamento

Outline

Dendrograma

é um gráfico que mostra a sequência de etapas que descreve quais objetos são reunidos em qual estágio da análise.
Como exemplo, considere o gráfico de quatro objetos A, B , C e D localizados ao longo de um segmento de linha: aplicando a ligação simples, se produz 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: $\{A, B, C, D\}$

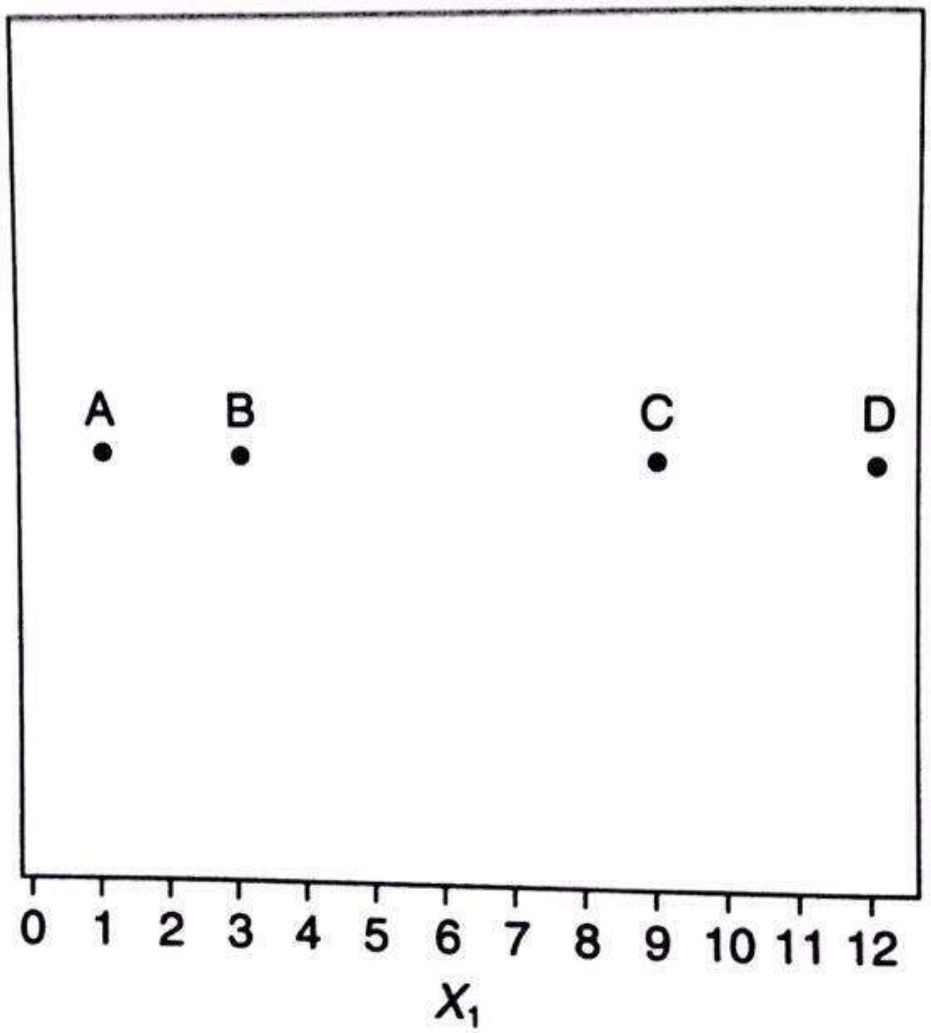


Figura 8.10 Ilustração: quatro pontos sobre um segmento de linha.

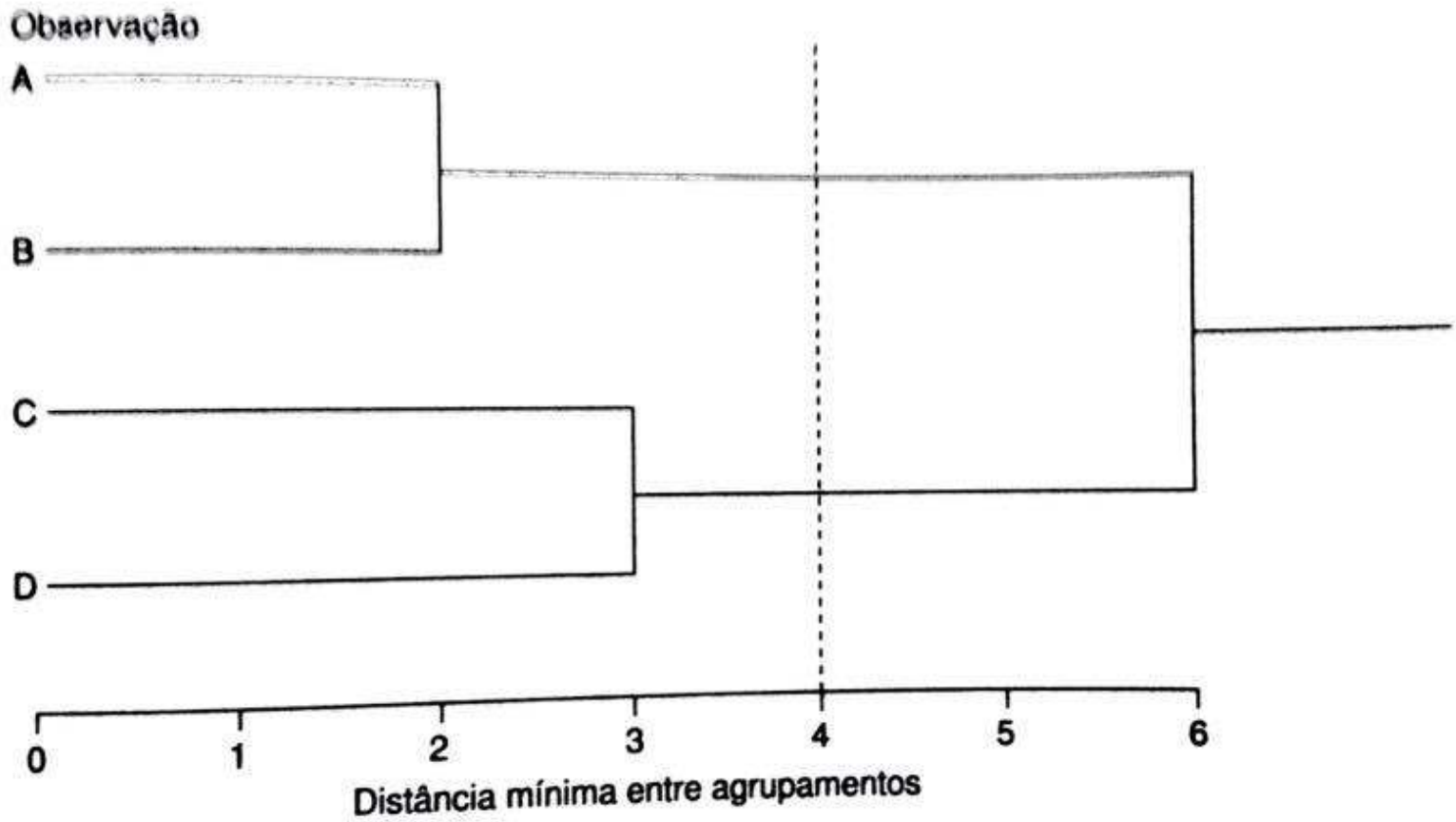


Figura 8.11 Dendrograma para solução de agrupamento de ligação simples com os dados da Figura 8.10.

Quantos Agrupamentos?

O Agrupamento aglomerativo não fornece uma resposta definitiva à questão.
Uma coisa a se buscar é uma gama de distâncias relativamente ampla em relação às quais, o número de agrupamentos na solução não muda.

Propriedades da ligação simples

- 1. São hierárquicos por natureza
- 2. Se n aumenta o esforço computacional para o pior caso é da ordem n^2
- 3. Para dados escassos o esforço computacional está na ordem nA, em que A é o número médio de conexões para cada objeto no conjunto.
- 4. Não exige dados métricos (podem ser medidas ordinais de dissimilaridade).
- 5. Uma desvantagem é que ela é míope: um objeto será adicionado a um grupo desde que esteja próximo a qualquer um dos outros objetos do agrupamento, mesmo que, por outro lado, esteja relativamente longe de todos os outros.

Alternativas à ligação simples

- 1. Ligação completa: Usamos neste caso a distância entre o par de objetos mais afastados, o que asegura que cada objeto adicionado ao grupo, esteja próximo de todos os objetos no agrupamento:

$$d_{C_{n+1}C_k} = max\{d_{C_iC_k}, d_{C_jC_k}\} \tag{6}$$

- 2. Ligação Média: meio-termo entre ligação simples e completa, com ele, chega-se perto de um ajuste de árvore que satisfaz o critério de minimização dos MQ. A nova distância é definida como a distância média entre o agrupamento C_k e o novo agrupamento C_{n+1} (formado pela junção dos agrupamentos C_i e C_j). Assim, reescrevemos a etapa 3 da seguinte maneira:

$$d_{C_{n+1}C_k} = \frac{n_i d_{C_iC_k} + n_j d_{C_iC_k} j}{n_i + n_j} \tag{7}$$

$n_i + n_j$ é o número de objetos em C_{n+1}. Se os dados forem não métricos,substituir a média pela mediana (ligaçã mediana)

Alternativas à ligação simples

- 3. Método Centroide: calcula a "média" dos objetos em cada agrupamento (centroides de cada agrupamento) e então define a distância entre dois centroides. Simplifica as coisas se usarmos distâncias ao quadrado. Se d_{ij}^2 é a distância ao quadrado entre i e j, e seja $C = \{i, j\}$, então a distância ao quadrado entre o objeto k e o centroide do agrupamento C é:

$$d_{kC}^2 = \frac{d_{ik}^2 + d_{jk}^2}{2} - \frac{d_{ij}^2}{4} \tag{8}$$

Em geral $d_{C_k C_{n+t}}^2$ com C_{n+t} agrupamento criado pela junção dos agrupamentos C_i e C_j pode ser representada por:

$$d^2(C_k, C_i \cup C_j) = \frac{n_{C_i} d_{C_k,C_i}^2 + n_{C_j} d_{C_k,C_j}^2}{n_{C_i} + n_{C_j}} - \frac{n_{C_i} n_{C_j} d_{C_i,C_j}^2}{(n_{C_i} + n_{C_j})^2} \tag{9}$$

Alternativas à ligação simples

determinando-se a regra na etapa 3 como uma função da distância euclidiana ao quadrado, o método centroide pode ser usado diretamente como medida de proximidade avaliadas também como medidas de distância derivadas.

- 4. Método de Ward: as 3 medidas anteriores são variações de uma abordagem aglomerativa (método de grupo de pares), o método de Ward (método de variância mínima) na etapa 1, em lugar de juntar dois agrupamentos mais próximos, busca juntar os dois agrupamentos cuja fusão da origem à menor soma de quadrados dentro do agrupamento (variância mínima dentro do grupo).
  Tende a produzir agrupamentos de tamanhos iguais, e com frequência produz uma solução de agrupamento (se a árvore for cortada no lugar certo)

Alternativas à ligação simples

- 5. Ligação por densidade: tenta explicar o contexto: d^* a densidade do k-ésimo vizinho mais próximo reflete não somente a relação direta entre dois objetos mas também sua relação com objetos circundantes.

Exemplo

Uma amostra de 50 observações extraídas de uma distribuição bimodal e outra de 50 extraída de uma distribuição unimodal. As figuras abaixo apresentam os resultados de ligação simples. Para o caso da amostra bimodal, uma linha vertical em d=0,35 divide os dados em quatro agrupamentos, dois deles de 26 e 20 e dois com duas observações. A modalidade natural fica pelo menos evidente embora não muito clara.
Para o caso unimodal, uma linha em d=0,40, produz dois agrupamentos grandes (35 e 14) e um único. Os resultados não demonstram a diferença na modalidade das duas amostras de forma dramática.

No caso do método de Ward, a estrutura de dois agrupamentos da amostra bimodal é alta e clara ainda que os agrupamentos não são do mesmo tamanho (erro de classificação?). No caso da amostra unimodal, ela separa em dois agrupamentos, indicando a falha do método em demonstrar a diferença na modalidade das duas amostras. Na análise de ligação por densidade (baseada na densidade k-ésima de vizinho mais próximo com k=9) é também apresentada. É possível ver o contraste entre os dois dendogramas, o bimodal mostra dois agrupamentos separados. Cada grupo de objetos é chamado de agrupamento modal. Para diferentes k's, observamos dois agrupamentos modais na amostra bimodal. Em contraste, para os mesmos valores de k observamos somente um único agrupamento modal.

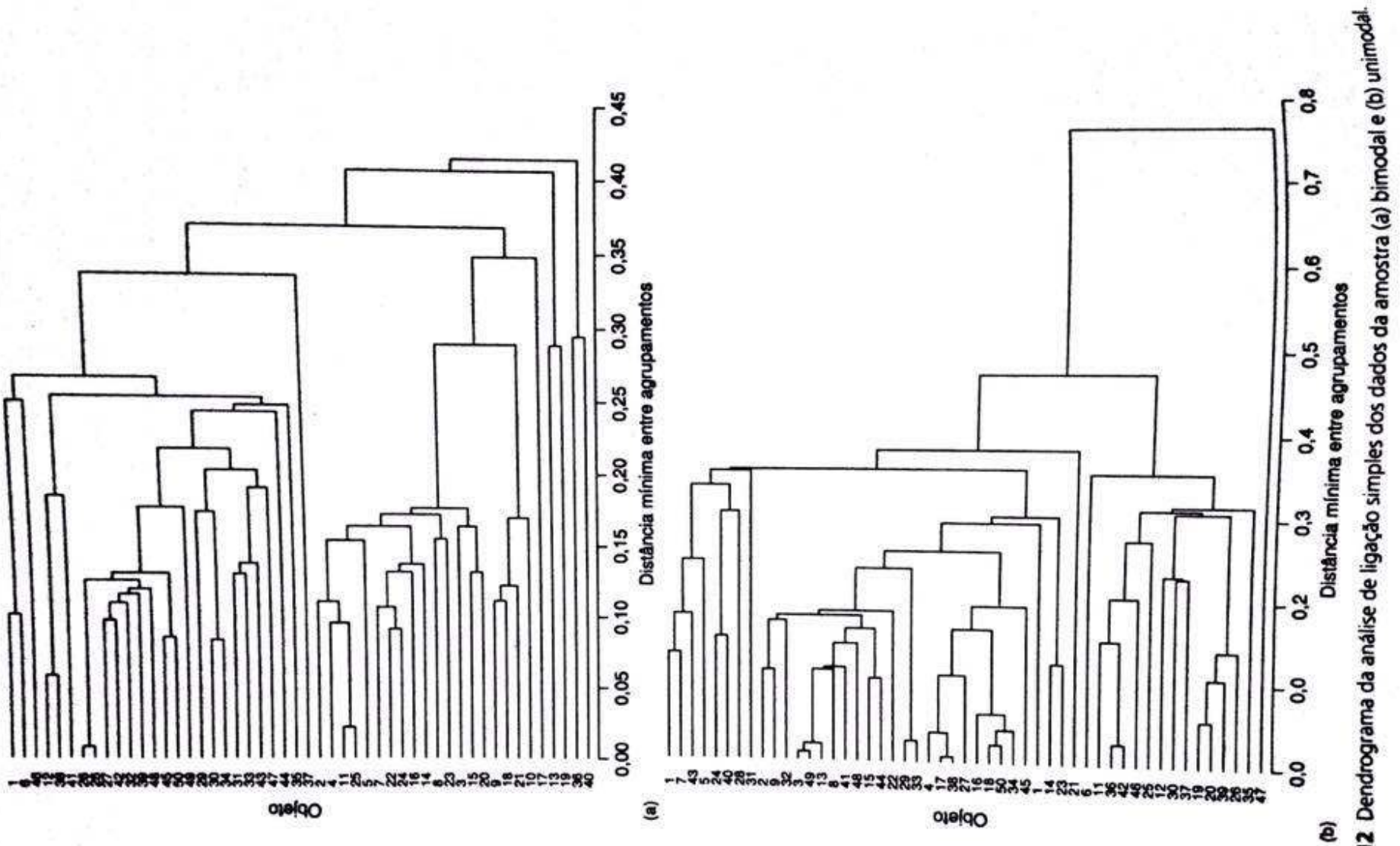
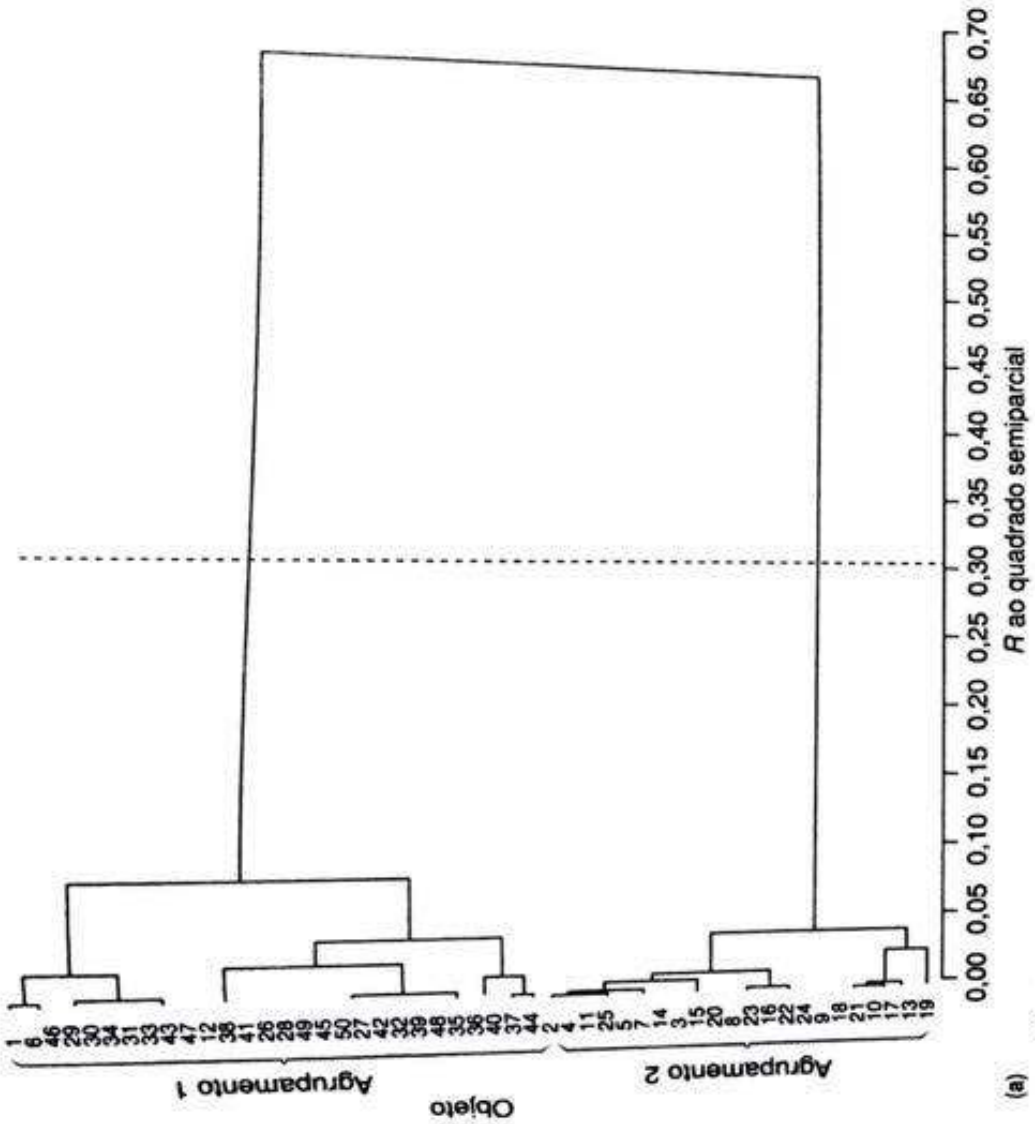
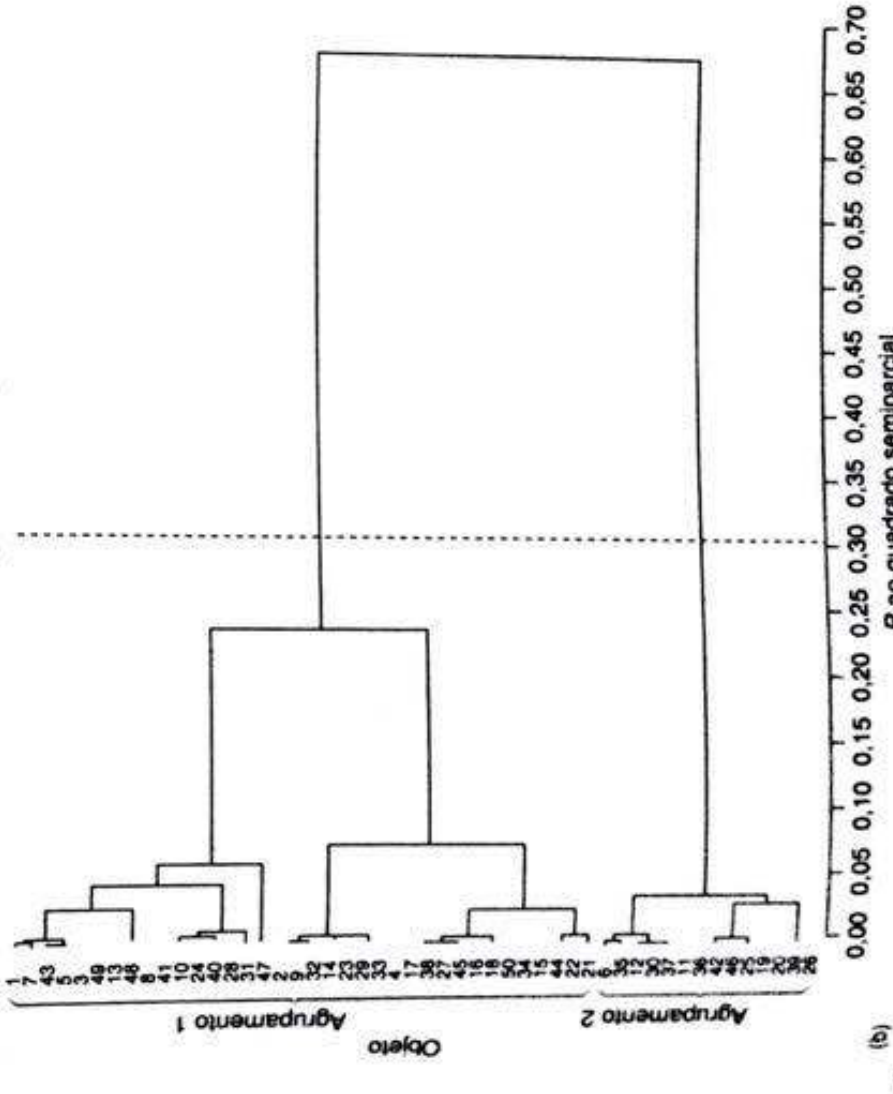
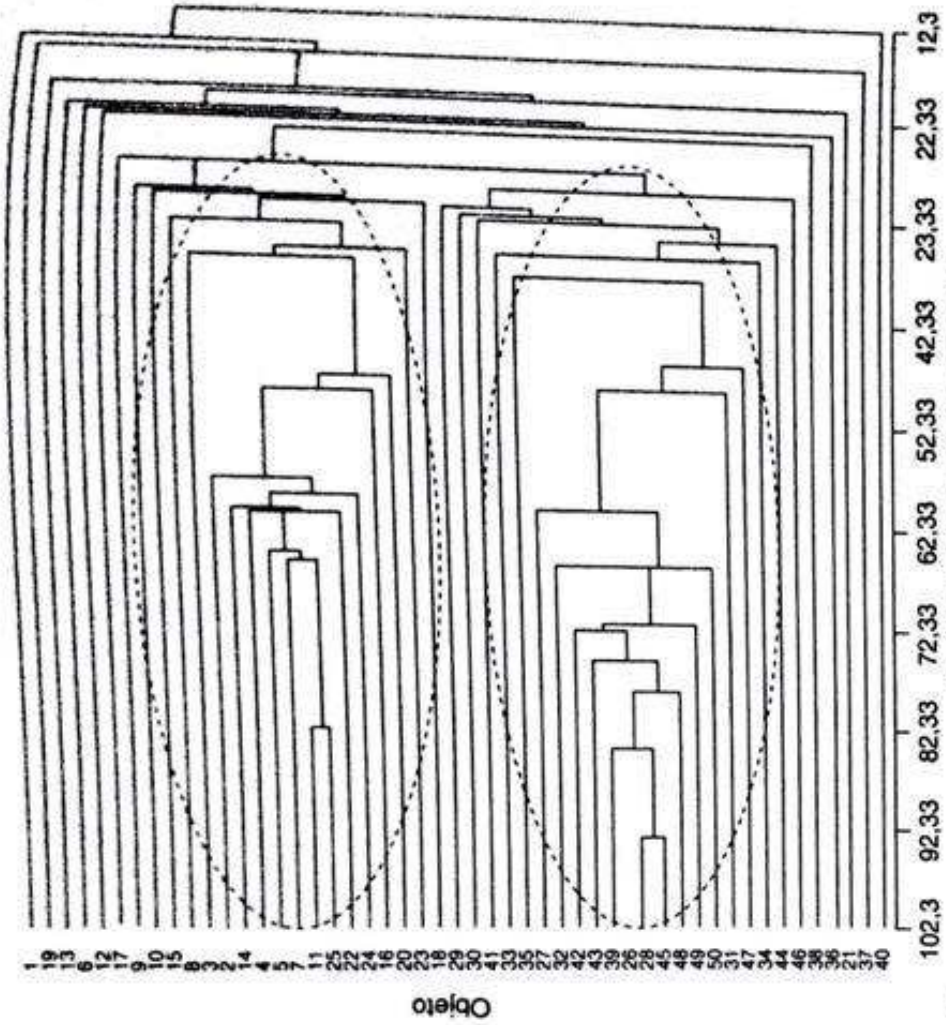
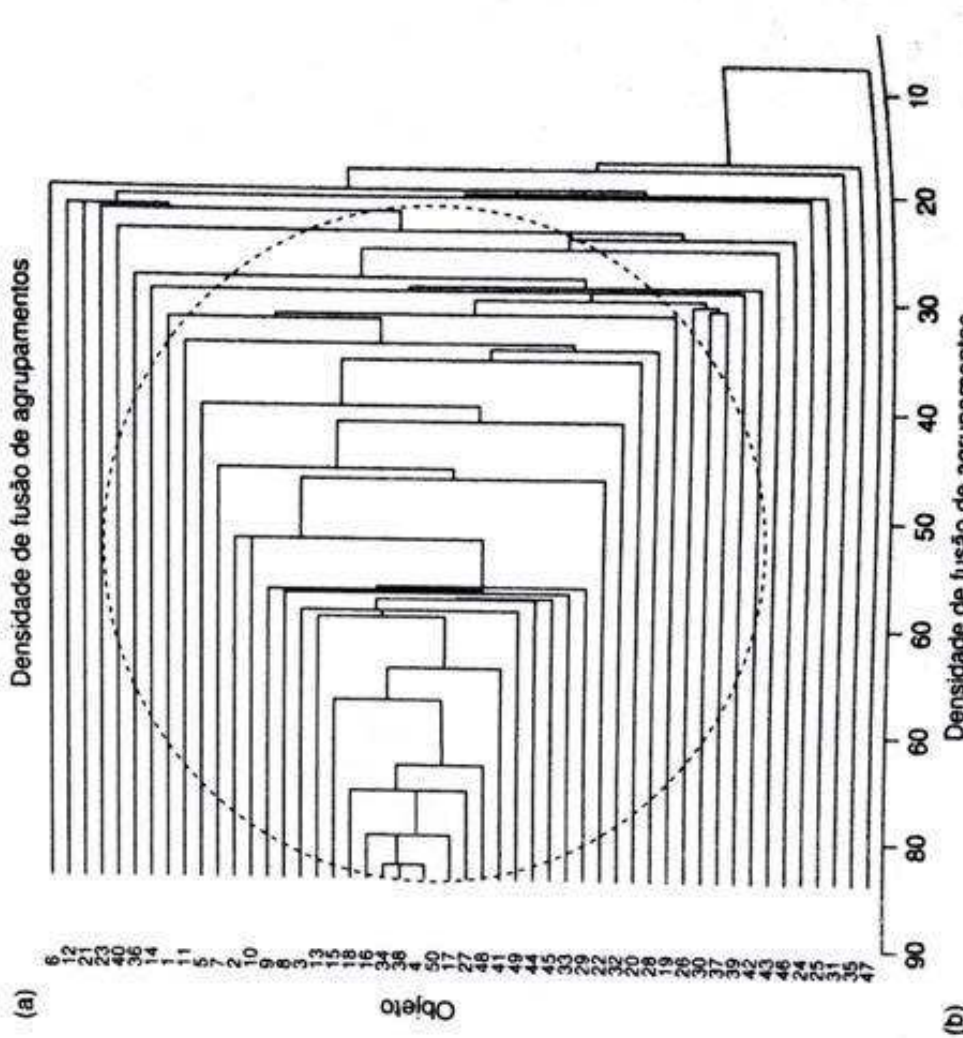


Figura 8.12 Dendrograma da análise de ligação simples dos dados da amostra (a) bimodal e (b) unimodal.


0,00 0,05 0,10 0,15 0,20 0,25 0,30 0,35 0,40 0,45 0,50 0,55 0,60 0,65 0,70
R ao quadrado semiparcial
Agrupamento 1
Agrupamento 2
Objeto
(a)


0,00 0,05 0,10 0,15 0,20 0,25 0,30 0,35 0,40 0,45 0,50 0,55 0,60 0,65 0,70
Agrupamento 1
Agrupamento 2
Objeto
(b)


Densidade de fusão de agrupamentos
12,3
22,33
23,33
42,33
52,33
62,33
72,33
82,33
92,33
102,3
Objeto
(a)


Densidade de fusão de agrupamentos
10
20
30
40
50
60
60
80
90
Objeto
(b)

Intuição

Nossa meta agora é dividir a amostra em um determinado K de grupos não superpostos, de maneira que os objetos dentro da cada grupo sejam relativamente similares e os objetos entre grupos sejam relativamente dissimilares.
A abordagem de partição é conhecida como agrupamento $K - means$ (Hartigan 1975). Como este algoritmo é localmente ótimo, é necessário executar um grande número de vezes, com diferentes pontos iniciais, para assegurar uma boa solução

Agrupamento K-means (etapas do algoritmo)

- 1. Selecione uma partida inicial dos dados nos agrupamentos K. Uma abordagem pede para utilizar um conjunto inicial de centroides de "semente" . Outra abordagem pede escolher pontos amplamente dispersos.
- 2. Calcule o centroide para cada agrupamento C, $\bar{X}_C$. (Pode ser realizado o cálculo para dados do tipo atributo ou distância)

Agrupamento K-means (etapas do algoritmo)

- 3. Calcule a soma das distâncias ao quadrado de cada objeto do seu centroide do agrupamento (soma de quadrados de erros da partição ESS):

$$ESS = \sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{C(i)})`(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{C(i)})$$

  Onde $C(i)$ é o agrupamento para i. Queremos tornar o ESS tão pequeno quanto possível
- 4. Torne a relacionar cada i ao agrupamento cuja centroide é mais próximo. Se ao final da etapa 4 os elementos do agrupamento permanecerem sem alterações, o processo convergiu para pelo menos o mínimo local. Se pelo menos um objeto do agrupamento modificar-se, volta à etapa 2 com a nova partição.

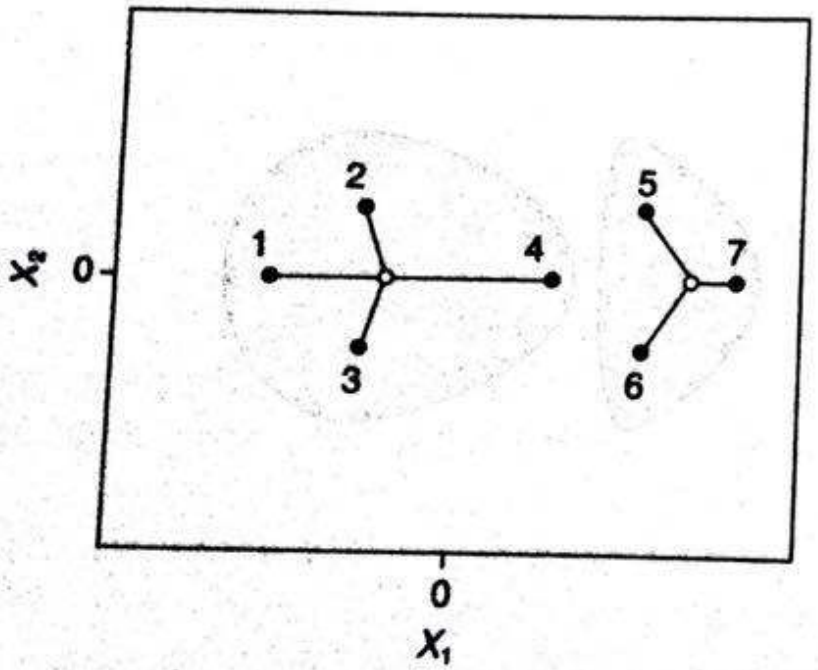


Figura 8.15 Partição inicial em dois agrupamentos de sete observações sobre duas variáveis.

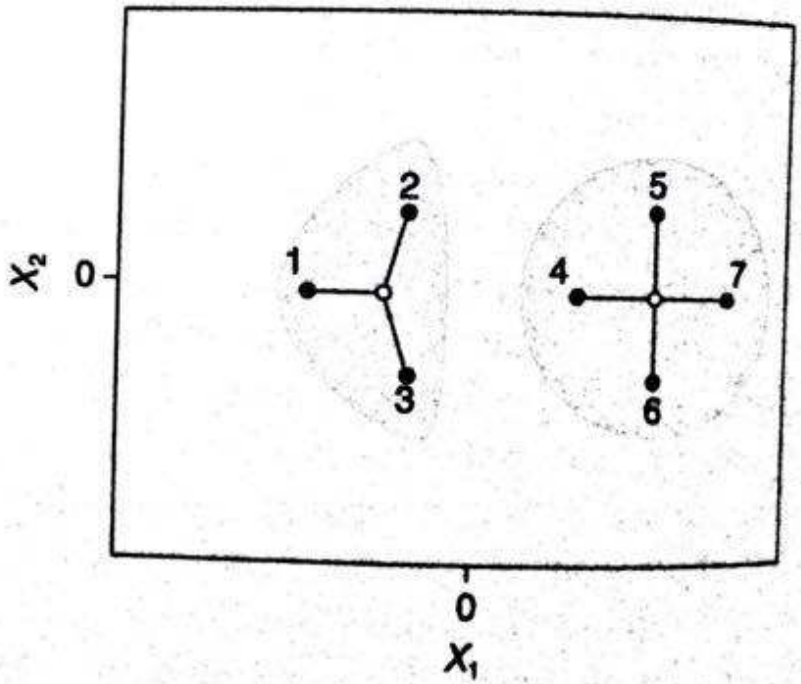


Figura 8.16 Partição final de dois agrupamentos após nova atribuição do objeto 4.

Quantos Agrupamentos?

Compete ao analista decidir qual valor de K resulta na melhor solução de agrupamento. A chave é conduzir análises com vários valores de K e depois escolher a solução que melhor corresponda aos objetos, isso involve uma análise de custo-benefício entre a simplicidade e a sua adequação. Como o ESS diminui quando k aumenta, usamos a estatística Pseudo-F (razão da soma média de quadrados entre os grupos pela soma de quadrados dentro do grupo):

$$pseudo - F = \frac{tr[\boldsymbol{B}/(K-1)]}{tr[\boldsymbol{W}/(n-K)]} \tag{10}$$

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

Questões relativas à aplicação da análise de agrupamentos

Como escalonar?: Na construção de uma medida de distância, é importante que dimensões diferentes sejam comparávelmente escalonadas, do contrário a variável com maior variância figurará com mais destaque na solução

validação: queremos verificar a capacidade de generalização da solução, com os seguintes passos:

- 1. Divida os dados em duas amostras aleatórias (calibração e validação)
- 2. Utilize o método de agrupamento para os dados de calibração, determine o número de grupos e calcule os centroides
- 3. Use os centroides dos agrupamentos dos dados de calibração, atribua cada observação da amostra de validação ao centroide mais próximo. (Essa solução será S_1).
- 4. Use o mesmo método do passo 2 para agrupar os dados de validação. Escolha a solução com o mesmo número de agrupamentos (S_2).
- 5. S_1 e S_2 são atribuições do mesmo conjunto de observações. Para avaliar a concordância, faça uma tabulação cruzada S_1 versus S_2.