Parte 6. Análise de Agrupamentos

Introdução

A Análise de Agrupamentos (em inglês: “cluster analysis”) é uma técnica estatística destinada a identificar e classificar dados, ou variáveis, em grupos com características similares. Na engenharia de processos, essa classificação é um auxiliar importante na identificação de diferenças e semelhanças em condições operacionais de uma unidade industrial, possibilitando a seleção de dados para análises mais detalhadas, ajuste de modelos etc.

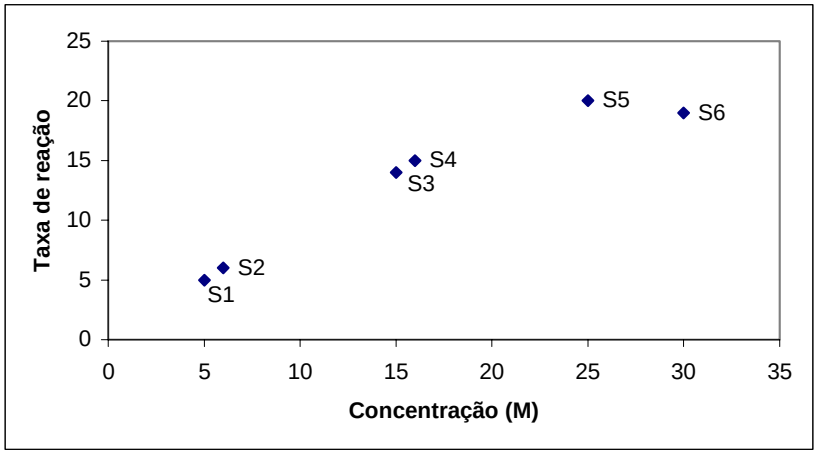
Para um caso geral, em que seja considerada uma amostra com n dados de um conjunto de p variáveis de processo, X_1, X_2, ..., X_p, a Análise de Agrupamentos objetiva:

1) Classificar os dados em g grupos, de acordo com critérios de semelhança (similaridade), ou de diferença entre os dados, ou entre os grupos de dados.
2) Possibilitar a interpretação dos resultados (conjunto dos g grupos de dados obtidos).

O resultado da Análise de Agrupamentos depende dos critérios adotados para agrupar os dados. Um exemplo que pode ilustrar a aplicação da técnica é o seguinte. Consideremos o seguinte conjunto de dados de laboratório, referentes a medidas da taxa de uma reação:

Tabela 1: Dados experimentais referentes à taxa de uma reação.

Item	Concentração (M)	Taxa de reação
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19



É possível verificar a existência de dados próximos nas condições de baixa concentração (S1 e S2), região intermediária (S3 e S4) e para excesso do reagente (S5 e S6). As técnicas de análise de agrupamentos consistem em agrupar observações, ou variáveis, com base em critérios de similaridade. A seguir, serão utilizados esses dados para ilustrar a aplicação das técnicas, nas várias etapas de execução.

Seqüência de Etapas

A análise de agrupamentos envolve normalmente as seguintes atividades, as quais serão detalhadas no texto.

1. Selecionar um critério de similaridade entre os dados.
2. Definir a técnica de agrupamento a ser adotada: hierárquica, ou não hierárquica.
3. Selecionar um critério de similaridade entre grupos.

4. Definir o número de grupos a ser obtido.
5. Interpretação dos resultados.

Medidas de similaridade entre dados

Há vários critérios para cálculo da distância entre dados. As mais comuns derivam da expressão geral, denominada “distância de Minkowski”:

$$D_{ij} = \left[\sum_{k=1}^{p} \left(\left| X_{ik} - X_{jk} \right| \right)^{n} \right]^{1/n},$$

em que n = 1, 2,..., ∞. Para n = 2, obtem-se a “distância euclideana”, que é a mais comumente adotada em medidas de distância entre dados:

$$D_{ij} = \left[\sum_{k=1}^{p} \left(\left| X_{ik} - X_{jk} \right| \right)^{2} \right]^{1/2}.$$

Quando n = 1, obtem-se a distância baseada em quadras (“city block”), ou “distância de Manhattan”:

$$D_{ij} = \sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|.$$

Outras medidas de distância:

Distância estatística (ou de Mahalanobis):

$$D_{ij} = \left(\mathbf{X}_i - \mathbf{X}_k \right)^T \mathbf{\Sigma}^{-1} \left(\mathbf{X}_i - \mathbf{X}_k \right)$$

em que **X** é o vetor p x 1 das variáveis e **Σ** é a matriz de covariância (pxp). Essa matriz é diagonal, para variáveis não correlacionadas. Se as variáveis forem padronizadas, obtem-se a matriz de correlação **R** (pxp), a qual se torna a matriz identidade **I** para variáveis não correlacionadas.

Matriz de correlação:

A matriz de correlação **R** (pxp) pode ser usada como medida de distância, principalmente no caso de agrupamentos de variáveis.

Medidas de distância para variáveis não quantitativas:

Há uma série de critérios adotados para variáveis não quantitativas, como, por exemplo, variáveis binárias. Por exemplo, seja o caso de medidas do vetor **X** *p* x 1 (*p* = 10), referentes a 10 critérios de comparação de duas unidades industriais:

Unidade	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
A	1	1	0	1	0	1	1	0	0	1
B	1	0	0	1	0	1	1	1	0	0

Em casos como esses, a medida baseia-se nos "coeficientes de associação", ou de "concordância". Monta-se uma tabela com as freqüências associadas a cada unidade industrial A e B:

		Unidade A		
		1	0	Total
Unidade B	1	a = 4	b = 1	a + b = 5
	0	c = 2	d = 3	c + d = 5
Total		a +c =6	b + d = 4	*p* = 10

Pode-se calcular, nesse exemplo, a distância euclideana média:

$$D_{ij} = \left[\frac{1}{p} \sum_{k=1}^{p} \left(\left| X_{ik} - X_{jk} \right| \right)^2 \right]^{1/2}$$

em que *i* e *j* referem-se, respectivamente, às unidades A e B. O resultado obtido é uma medida da dissimilaridade entre as unidades:

$$D_{ij} = \left[\frac{b+c}{p = a+b+c+d} \right]^{1/2} = 0{,}55$$

Essa medida varia entre 0 e 1. Quanto mais próxima de 0, mais similares são as duas unidades (observações, ou os dados).

Outros critérios utilizados:

"coeficiente de dissimilaridade simples":

$$D_{ij} = \frac{b+c}{p = a+b+c+d} = 0{,}3$$

“coeficiente de concordância simples”:

$$D_{ij} = \frac{a+d}{p=a+b+c+d} = 0,7$$

entre outros.

O cálculo das distâncias entre variáveis resulta na “matriz de similaridade”, **D** (simétrica, *p* x *p*, com diagonal principal igual a zero). Para o exemplo da taxa de reação, a matriz de similaridade obtida pelo cálculo da distância euclideana entre dados é a seguinte:

Matriz de Similaridade: Distância euclideana:

Dados	S1	S2	S3	S4	S5	S6
S1	0.00	1.41	13.45	14.87	25.00	28.65
S2	1.41	0.00	12.04	13.45	23.60	27.30
S3	13.45	12.04	0.00	1.41	11.66	15.81
S4	14.87	13.45	1.41	0.00	10.30	14.56
S5	25.00	23.60	11.66	10.30	0.00	5.10
S6	28.65	27.30	15.81	14.56	5.10	0.00

Matriz de Similaridade: Distância euclideana ao quadrado:

Dados	S1	S2	S3	S4	S5	S6
S1	0	2	181	221	625	821
S2	2	0	145	181	557	745
S3	181	145	0	2	136	250
S4	221	181	2	0	106	212
S5	625	557	136	106	0	26
S6	821	745	250	212	26	0

A matriz de similaridade é a base de todos os métodos de agrupamento.

Técnicas de agrupamento

As técnicas de construção dos grupos de dados são divididas em “hierárquicas” e “não hierárquicas”. As técnicas hierárquicas baseiam-se na ordenação de distâncias, sendo iniciadas a partir da menor distância detectada na matriz **D**. Os dados são agrupados seqüencialmente e, em cada etapa, forma-se um novo grupo que envolve dois ou mais grupos precedentes, segundo uma escala de distância entre grupos. Assim, parte-se de um número inicial de grupos igual ao número de objetos a serem agrupados. Os objetos vão sendo agrupados sucessivamente, até se chegar a um

único grupo que contém todos os objetos. Fica a critério do usuário definir quantos grupos representam melhor as características do sistema em estudo. Há, também, técnicas hierárquicas divisivas para formação de grupos: partem de um único grupo e fazem divisões sucessivas segundo critérios de distância.

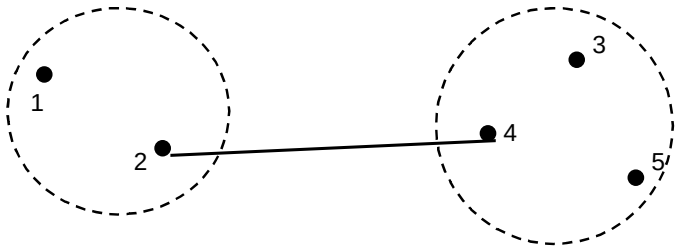
As técnicas não hierárquicas partem de um número pré-fixado de grupos (denominados sementes) e agrupam os dados segundo critérios de homogeneidade e de heterogeneidade entre eles. Diferentemente das técnicas hierárquicas, neste caso o número de grupos obtidos é definido *a priori* pelo usuário.

Técnicas hierárquicas

O agrupamento sucessivo de objetos (que podem ser observações ou variáveis, mas também grupos) é feito segundo critérios de similaridade entre grupos. Os diferentes critérios adotados são apresentados a seguir.

Ligação simples, ou vizinho mais próximo ("single linkage, nearest neighbor")

Baseia-se na menor distância entre quaisquer dois objetos dos dois grupos, o que equivale à distância entre os objetos mais próximos dos dois grupos:



Procedimento:

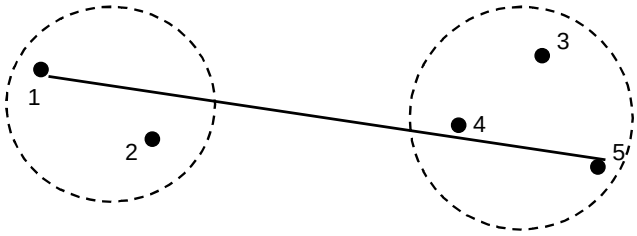
1) Percorrer a matriz de similaridade e detectar a menor distância d_{ij}; supondo que essa distância corresponda aos objetos *U* e *V*;
2) Juntar os dois objetos, formando o grupo (*UV*);
3) Atualizar a matriz de similaridade, com os novos objetos formados, tal que, para quaisquer dois novos objetos, *U* e *V*:

$$d_{UV} = \min\{d_{ij}\} \qquad i = 1, ..., N_U, \quad j = 1, ..., N_V$$

4) Repetir os passos 1 a 3 acima, até que o número remanescente de grupos seja 1.

Ligação completa, ou vizinho mais distante (“complete linkage, farthest neighbor”)

Baseia-se na maior distância entre quaisquer dois objetos dos dois grupos, o que equivale à distância entre os objetos mais distantes dos dois grupos:



Procedimento:

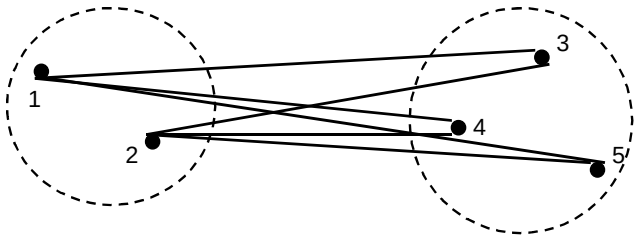
1) Percorrer a matriz de similaridade e detectar a menor distância d_{ij}; supondo que essa distância corresponda aos objetos *U* e *V*;
2) Juntar os dois objetos, formando o grupo (*UV*);
3) Atualizar a matriz de similaridade, com os novos objetos formados, tal que, para quaisquer dois novos objetos, *U* e *V*:

$$d_{UV} = \max\{d_{ij}\} \qquad i = 1, ..., N_U, \quad j = 1, ..., N_V$$

4) Repetir os passos 1 a 3 acima, até que o número remanescente de grupos seja 1.

Ligação média (“average linkage”)

Este método considera a distância entre objetos como sendo a média das distâncias entre pares de todos os componentes de cada objeto:



Procedimento:

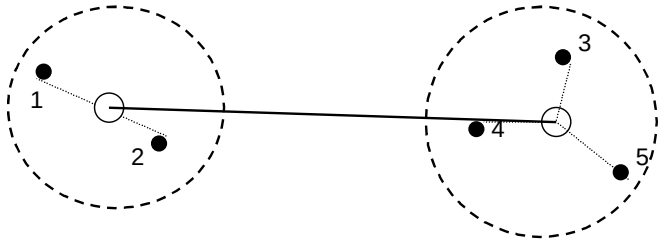
1) Percorrer a matriz de similaridade e detectar a menor distância d_{ij}; supondo que essa distância corresponda aos objetos *U* e *V*;
2) Juntar os dois objetos, formando o grupo (*UV*);
3) Atualizar a matriz de similaridade, com os novos objetos formados, tal que, para quaisquer dois novos objetos, *U* e *V*:

$$d_{UV} = \frac{\sum_{i=1}^{N_U} \sum_{j=1}^{N_V} d_{ij}}{N_U N_V}$$

4) Repetir os passos 1 a 3 acima, até que o número remanescente de grupos seja 1.

Centróide

Baseia-se nas distâncias entre valores médios dos objetos em cada grupo (centróides). A cada combinação de dois grupos, um novo grupo é formado e seu centróide é calculado novamente.



Procedimento:

1) Percorrer a matriz de similaridade e detectar a menor distância d_{ij}; supondo que essa distância corresponda aos objetos *U* e *V*;
2) Juntar os dois objetos, formando o grupo (*UV*);
3) Atualizar a matriz de similaridade, com os novos objetos formados, tal que, para quaisquer dois novos objetos, *U* e *V*, d_{UV} é a distância entre as médias das coordenadas dos objetos contidos em *U* e *V*.
4) Repetir os passos 1 a 3 acima, até que o número remanescente de grupos seja 1.

Exemplo de aplicação, com os dados de taxa de reação:

Agrupamento das observações utilizando: distância euclideana entre os dados, método do centróide. Resultados gerados pelo aplicativo Minitab:

Euclidean Distance, Centroid Linkage

Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	5	95.06	1.414	3	4	3	2
2	4	95.06	1.414	1	2	1	2
3	3	82.20	5.099	5	6	5	2
4	2	60.03	11.454	3	5	3	4
5	1	44.98	15.764	1	3	1	6

Cada linha da Tabela corresponde a um passo do processo de agrupamento e apresenta as seguintes informações: número de grupos, nível de similaridade, distância mínima detectada entre elementos da matriz de similaridade, objetos agrupados, número de ordem do novo grupo formado e número de observações no novo grupo. No caso, a coluna de distância corresponde à distância euclideana. O nível de similaridade é definido como:

$$s = 100\left(1 - \frac{d_{ij}}{\max\left(d_{ij}\right)}\right)$$

em que d_{ij} é a distância euclideana entre pares de objetos na matriz de similaridade. As tabelas a seguir detalham o que acontece a cada passo.

No primeiro passo, as observações S3 e S4 foram unidas, formando o novo "cluster 3". A nova matriz de similaridade fica então, da seguinte forma:

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Cluster1	0.0000	1.4142	14.1598	25.0000	28.6531
Cluster2	1.4142	0.0000	12.7475	23.6008	27.2947
Cluster3	14.1598	12.7475	0.0000	10.9772	15.1822
Cluster4	25.0000	23.6008	10.9772	0.0000	5.0990
Cluster5	28.6531	27.2947	15.1822	5.0990	0.0000

Passo 2: observações S1 e S2 unidas, formando o novo “cluster 1”. Matriz de similaridade atualizada:

```
Distances Between Cluster Centroids

               Cluster1     Cluster2     Cluster3     Cluster4
Cluster1         0.0000      13.4536      24.3002      27.9732
Cluster2        13.4536       0.0000      10.9772      15.1822
Cluster3        24.3002      10.9772       0.0000       5.0990
Cluster4        27.9732      15.1822       5.0990       0.0000
```

Passo 3: observações S5 e S6 unidas, formando o novo “cluster 3”. Matriz de similaridade atualizada:

```
Distances Between Cluster Centroids

               Cluster1     Cluster2     Cluster3
Cluster1         0.0000      13.4536      26.0768
Cluster2        13.4536       0.0000      13.0000
Cluster3        26.0768      13.0000       0.0000
```

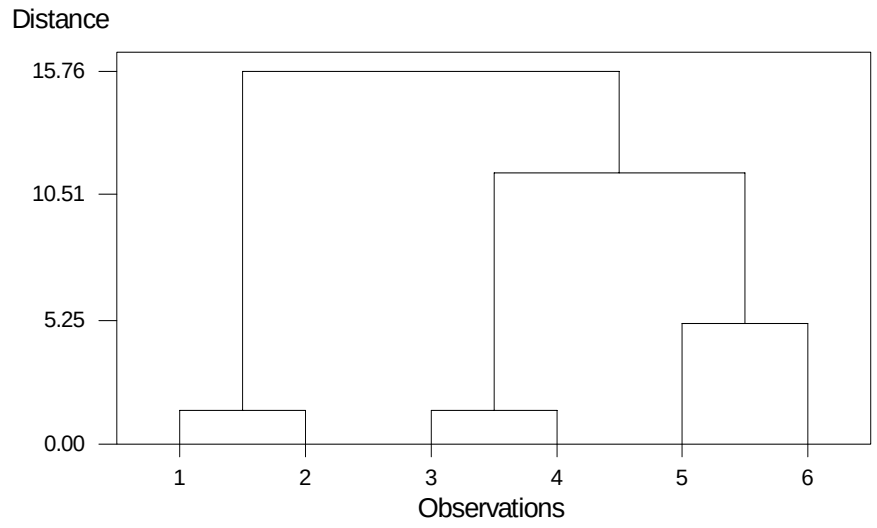
Passo 4: objetos 2 e 3 unidos, formando o novo “cluster 2”. Matriz de similaridade atualizada:

```
Distances Between Cluster Centroids

               Cluster1     Cluster2
Cluster1         0.0000      19.7041
Cluster2        19.7041       0.0000
```

No Passo 5, forma-se um único grupo,contendo todos os objetos.

O processo de formação de grupos por métodos hierárquicos é representado graficamente como uma árvore hierárquica, ou “dendrograma”:



Método de Ward

Neste método, os grupos são formados minimizando-se os quadrados dos desvios dos componentes de cada grupo, em relação ao valor médio de cada grupo (centróide do grupo). Define-se, para um grupo *k, ESS_k* como:

$$ESS_K = \sum_{j=1}^{N_k} \left(\mathbf{X}_j - \overline{\mathbf{X}}\right)^T \left(\mathbf{X}_j - \overline{\mathbf{X}}\right)$$

em que N_k é o número de componentes do grupo *k*, $\mathbf{X}_j$ é um vetor de observações (dados multivariados) contido no grupo *k* e $\overline{\mathbf{X}}$ é o centróide do grupo *k*. Assim, o total da soma dos quadrados dos desvios dos grupos é:

$$ESS = \sum_{J=1}^{k} ESS_j$$

O processo de agrupamento inicia-se com *n* grupos (igual ao número de observações). A cada passo do processo todos os pares de grupos, *i*, *j*, são considerados e é selecionado para compor o novo grupo o par que representar o menor incremento em *ESS*. Ou seja, por este método, a matriz de similaridade é composta pelos valores de *ESS* correspondentes a cada par *i, j.*

Critérios para seleção do número de grupos a ser utilizado, ao final de uma análise pelo método hierárquico

Além do bom senso na interpretação dos resultados, existem vários critérios estatísticos para definição do número de grupos a ser retido no final de uma análise. Entre essas técnicas estão:

a) Nível de similaridade:

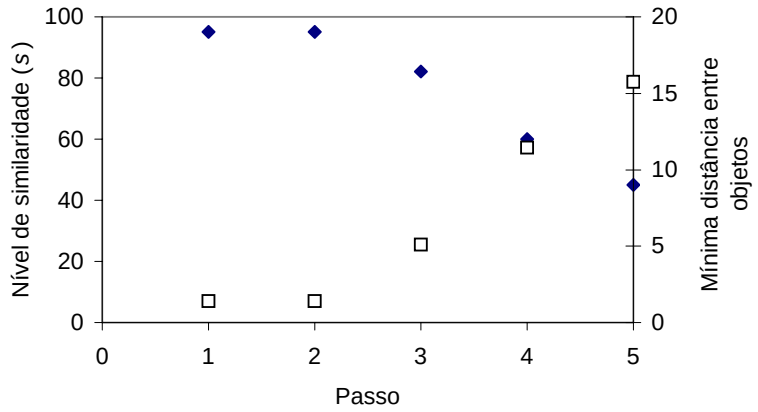
$$s = 100\left(1 - d_{ij}\right) / \max\left\{d_{ij}\right\};$$

b) Valor da distância mínima entre objetos, a cada atualização da matriz de similaridade. Esta informação pode ser vista, também, graficamente, no dendrograma, na escala de distâncias a cada passo do processo.

Examinando-se o gráfico desses parâmetros a cada passo do processo é possível visualizar mudanças abruptas, que indicam pontos de parada no processo.

Exemplo:

A Figura a seguir apresenta a variação do nível de similaridade e da distância mínima entre grupos em função dos passos de agrupamento para o caso apresentado, da taxa de reação, com as observações agrupadas pelo método da centróide. Observa-se que há uma variação clara de inclinação das curvas a partir do passo 3, que corresponde à formação de 3 grupos, não havendo alterações posteriores. Por esse critério, deveriam ser considerados 2 grupos para agrupamento dos dados.



Em geral recomenda-se que seja refeita a análise fixando-se o número de grupos definido na primeira análise, para obtenção da configuração mais adequada.

Técnicas não hierárquicas

Nas técnicas não hierárquicas, em geral o número de grupos a serem formados deve ser definido previamente.

Uma vez que as técnicas não hierárquicas não envolvem o cálculo de uma matriz de similaridades, nem a retenção de todos os dados na memória de computadores, são mais adequados para análise de grandes quantidades de dados.

Os algoritmos seguem a seguinte seqüência geral:

1) Selecionar de *k* centróides para os grupos, ou partição dos dados em *k* segmentos;
2) Alocar cada observação ao centróide mais próximo (com base, por exemplo, na distância euclideana);
3) Recalcular os centróides;
4) Realocar cada observação aos novos centróides;

5) Repetir os passos 3 e 4 até que não haja mais variação nos valores dos centróides, ou até que outro critério estabelecido pelo usuário seja atendido.

Os vários algoritmos desenvolvidos para análise de agrupamentos diferem entre si segundo os seguintes critérios:

a) método de estimativa inicial dos *k* segmentos, ou centróides.

Alguns métodos geram aleatoriamente as sementes iniciais, enquanto outros dividem a faixa de variação das variáveis em *k* segmentos, ou ainda utilizam sementes fornecidas pelo usuário.

b) regras de realocação de observações em relação aos centróides.

Nesses casos, são adotados diferentes critérios estatísticos:

1) Minimização do traço (soma das diagonais) da matriz de somas dos desvios quadráticos de cada grupo, **SSCP** ("sum of squares and cross products"), o que equivale a minimizar o desvio quadrático *ESS*.
2) Minimização do determinante da matriz **SSCP** de cada grupo;
3) Minimização do traço da matriz $\mathbf{W}^{-1}\mathbf{B}$, em que **W** e **B** são, respectivamente, as partes interna e entre grupos da matriz **SSCP**;
4) Minimização do maior autovalor da matriz $\mathbf{W}^{-1}\mathbf{B}$.

A matriz **SSCP** indica a dispersão de uma população em relação ao valor médio. É calculada da seguinte forma:

Dadas *g* populações (grupos de observações multivariadas):

População 1: $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n1}$

População 2: $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n2}$

.
.
.

População *g*: $\mathbf{X}_{g1}, \mathbf{X}_{g2}, \ldots, \mathbf{X}_{gng}$

Sejam os valores $\overline{\mathbf{X}}_k$ os valores médios da população *k* e $\overline{\mathbf{X}}$ o valor médio de todas as populações. Então, as seguintes matrizes são válidas para comparações de dispersões internas de cada população e entre populações (ou grupos):

Tipo de dispersão	Matriz de dispersão	Graus de liberdade
Entre grupos	$\mathbf{B} = \sum_{k=1}^{g} n_k \left(\mathbf{X}_k - \overline{\mathbf{X}}\right)^T \left(\mathbf{X}_k - \overline{\mathbf{X}}\right)$	g - 1
Interna a cada grupo	$\mathbf{W} = \sum_{k=1}^{g} \sum_{j=1}^{n_k} \left(\mathbf{X}_{kj} - \overline{\mathbf{X}_k}\right)^T \left(\mathbf{X}_{kj} - \overline{\mathbf{X}_k}\right)$	$\sum_{k=1}^{g} n_k - g$
Totais	**SSCP** = **B** + **W**	$\sum_{k=1}^{g} n_k - 1$

Entre os algoritmos utilizados para análise não hierárquica de agrupamentos, o algoritmo denominado “K-means” (ou “método das k médias”) é o mais utilizado. Segue a seqüência apresentada, alocando cada observação à semente com centróide mais próximo (distância euclideana), atualizando os centróides dos grupos receptor e doador a cada observação alocada. O processo é interrompido quando não há mais variações nos valores dos centróides.

EXEMPLO DE AGRUPAMENTO PELO MÉTODO DAS K MÉDIAS

Supondo o caso seguinte, em que 4 observações de duas variáveis devem ser agrupadas em K = 2 grupos:

Obs.	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Supondo que seja decidido arbitrariamente que os valores representativos dos dois grupos sejam os centróides das observações (AB) e (CD).

Assim, o Passo 1 do processo é o cálculo dos centróides dos dois grupos:

Grupo	$\bar{X}_1$	$\bar{X}_2$
(AB)	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(CD)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

No Passo 2, calcula-se a distância Euclideana entre cada observação e o centróide dos grupos e aloca-se cada item ao grupo com centróide mais próximo. Se qualquer observação mudar de grupo, os centróides devem ser recalculados antes de seguir com o procedimento. Cálculo das distâncias:

$$d^2(A,(AB)) = (5-2)^2 + (3-2)^2 = 10$$

$$d^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61$$

Como A está mais próximo de (AB) do que de (CD), ele não é realocado. Para a observação B:

$$d^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10$$

$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9$$

Portanto, B é realocado no grupo (CD), resultando no novo grupo (BCD). Os novos centróides são:

Grupo	$\bar{X}_1$	$\bar{X}_2$
A	5	3
BCD	-1	-1

Novamente, a distância de cada observação em relação aos centróides dos grupos é calculada:

Grupo	A	B	C	D
A	0	40	41	89
BCD	52	4	5	5

Como todas as observações estão devidamente alocadas nos grupos com centróides mais próximos, o processo termina.