Insper

Análise de Componentes Principais (PCA) e k-médias

Aula 3

Magno T F Severino
PADS - Aprendizagem Estatística de Máquina II

Objetivos de Aprendizagem

Ao final dessa aula você deverá ser capaz de

- compreender a técnica de análise de componentes principais (PCA);
- interpretar as componentes principais geradas pela PCA;
- aplicar PCA para redução de dimensionalidade dos dados e para modelagem usando `tidymodels`;
- compreender como funciona a análise de conglomerados (clusters);
- usar o algoritmo k-médias;
- interpretar os agrupamentos obtidos pelos métodos.

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Find hidden dependencies

Identify sequences

ASSOCIATION

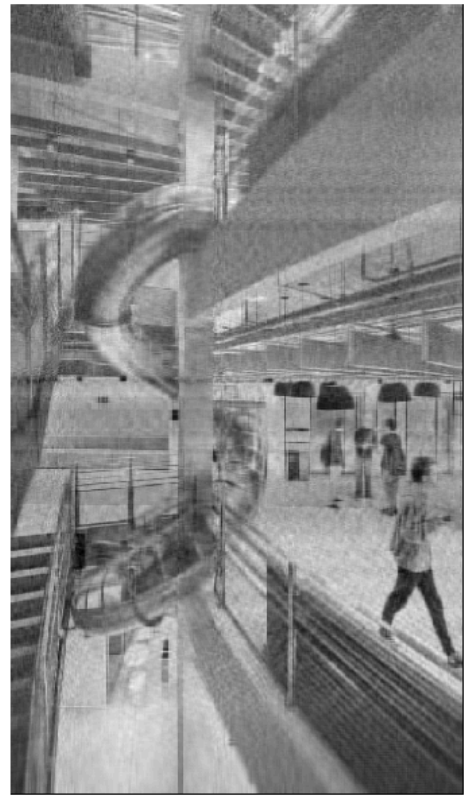
«Find what clothes I often wear together»

Desafio da Aprendizagem Não Supervisionada

- Nesse contexto, não existe uma variável preditora Y.
- A análise é mais subjetiva e depende do objetivo particular de cada aplicação.
- Geralmente, a aprendizagem não supervisionada é usada como parte da análise exploratória de dados.
- Dificuldade: não existe uma técnica definida para avaliar o desempenho de um modelo de aprendizagem não supervisionada, como as métricas usadas para avaliar modelos de regressão e classificação.

Uma Aplicação de PCA





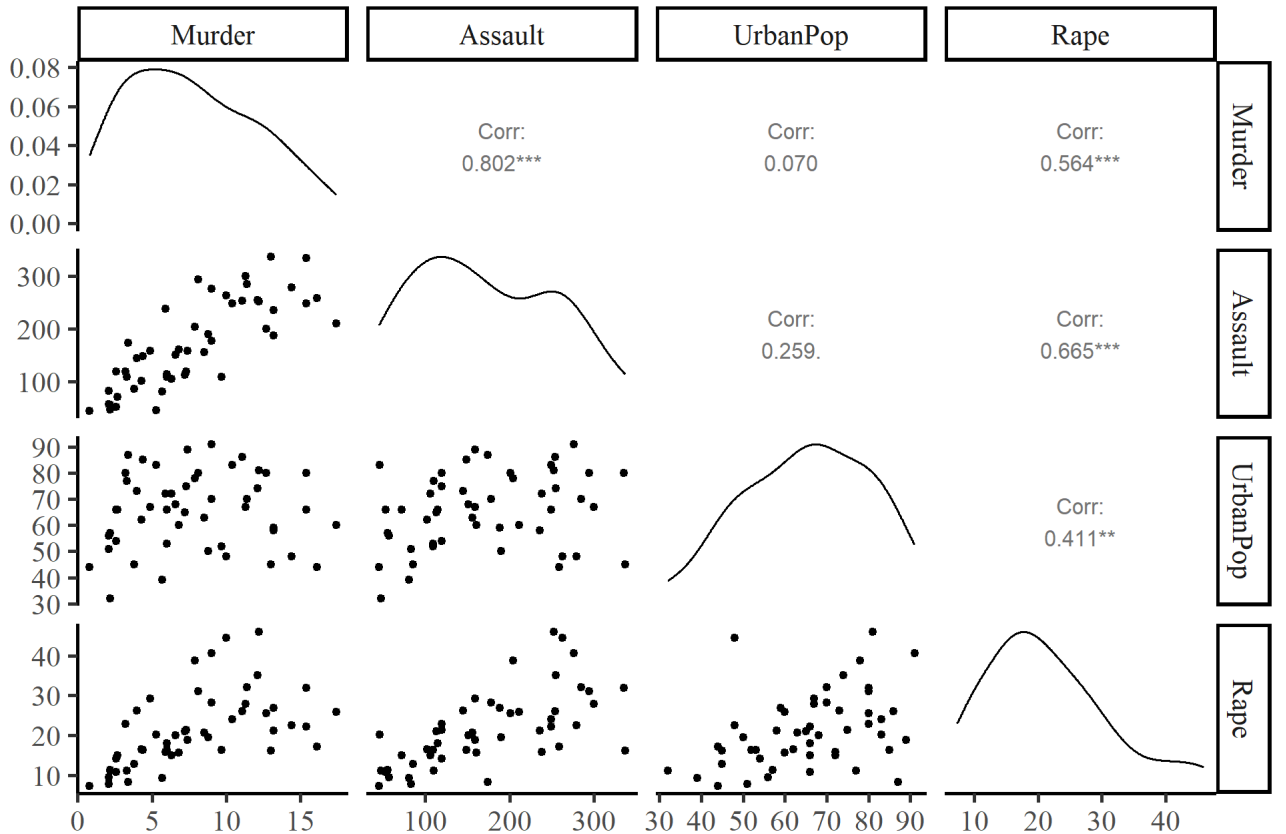
US Arrests Data

```
data("USArrests")
```

State	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0

US Arrests Data

Como visualizar a relação entre as variáveis?



Seria possível exibir essas informações apenas com um único gráfico? Quanta informação perderíamos ao fazer isso?

Análise de Componentes Principais

Considere que temos dados organizados numa estrutura como a seguinte:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

Em particular, para os dados `USArrest`, temos

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ \vdots & \vdots & \vdots & \vdots \\ x_{50,1} & x_{50,2} & x_{50,3} & x_{50,4} \end{pmatrix} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{50}^T \end{pmatrix}.$$

em que cada coluna em X, $(x_{\cdot 1}, x_{\cdot 2}, x_{\cdot 3}, x_{\cdot 4})$, representa `Murder`, `Assault`, `UrbanPop` e `Rape`, respectivamente.

Análise de Componentes Principais

O objetivo do PCA é aplicar uma transformação nos dados de forma a representá-los em um novo sistema de coordenadas.

Para os dados `USArrest`, a primeira componente principal de X é a variável Z_1 que

- é a combinação linear das variáveis de X, ou seja

$$z_{i1} = \phi_{11} x_{Murder_i} + \phi_{21} x_{Assault_i} + \phi_{31} x_{UrbanPop_i} + \phi_{41} x_{Rape_i},$$

- tem a maior variância possível com a restrição de que $\sum_{k=1}^{p} \phi_{k1}^2 = 1$.

Os elementos ϕ_{ij} são chamados de **cargas/loadings** e os valores z_{i1} são as **projeções/scores** do estado americano i na componente principal 1.

Da mesma maneira, a segunda componente principal de X é a variável Z_2 que

- é a combinação linear das variáveis de X, ou seja

$$z_{i2} = \phi_{12} x_{Murder_i} + \phi_{22} x_{Assault_i} + \phi_{32} x_{UrbanPop_i} + \phi_{42} x_{Rape_i},$$

- tem a maior variância possível com a restrição de que $\sum_{k=1}^{p} \phi_{k2}^2 = 1$,

- tem correlação zero com Z_1.

Este procedimento é repetido até obter a quarta componente principal.

Análise de Componentes Principais

Generalizando para uma matriz X com p colunas (variáveis):

O primeiro componente principal de X é a variável Z_1 que

- é a combinação linear das variáveis de X, ou seja

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p,$$

- tem a maior variância possível com a restrição de que $\sum_{k=1}^{p} \phi_{k1}^2 = 1$.

Os elementos ϕ_{ij} são chamados de **cargas/loadings** e os valores Z_{1i} são as **projeções/scores** do estado i na componente principal 1.

Da mesma maneira, a segunda componente principal de X é a variável Z_2 que

- é a combinação linear das variáveis de X, ou seja

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \cdots + \phi_{p2}X_p,$$

- tem a maior variância possível com a restrição de que $\sum_{k=1}^{p} \phi_{k2}^2 = 1$,

- tem correlação zero com Z_1.

Este procedimento é repetido até a p-ésima componente principal.

Um pouco de notação

Um estimador para a variância de uma variável aleatória X é o obtido por

$$\mathrm{Var}(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{N}.$$

Um estimador para a covariância entre as variáveis aleatórias X e Y

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y})}{N}.$$

Como ficam essas medidas se as variáveis X são padronizadas ($\overline{x} = 0$ e $\overline{y} = 0$)?

Neste caso, temos que

$$\mathrm{Var}(X) = \frac{\sum_{i=1}^{n}(x_i)^2}{N} \quad \text{e} \quad \mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i}{N}.$$

Análise de Componentes Principais

Como escolher $\phi_1 = (\phi_{11}, \cdots, \phi_{1p})^\top$?

Considere que os dados foram centralizados para ter médias iguais a zero. Dessa forma, queremos a combinação linear

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots + \phi_{p1} x_{ip},$$

que apresente a maior variância com a restrição de que $\sum \phi_{ji}^2 = 1$.

Note que

$$\bar{z}_{\cdot i} = \frac{1}{n} \sum_{i=1}^{n} z_{i1} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots + \phi_{p1} x_{ip} \right\} = 0.$$

Assim,

$$\text{Var}(Z_{\cdot 1}) = \frac{1}{n} \sum_{i=1}^{n} (z_{i1} - \bar{z}_{\cdot i})^2 = \frac{1}{n} \sum_{i=1}^{n} z_{i1}^2.$$

Análise de Componentes Principais

Lembre-se que queremos que a primeira componente apresente máxima variância e que, do slide anterior,

$$\text{Var}(z_{i1}) = \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2.$$

Então, temos o seguinte problema de otimização:

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 \right\}, \qquad \text{sujeito a } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

A expressão acima pode ser reescrita usando a definição das projeções/scores $z_{\cdot i}$:

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(\phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip} \right)^2 \right\}, \qquad \text{sujeito a } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

Por sua vez, essa expressão pode ser resumida usando a notação de somatório:

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1}x_{ij} \right)^2 \right\}, \qquad \text{sujeito a } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

Análise de Componentes Principais

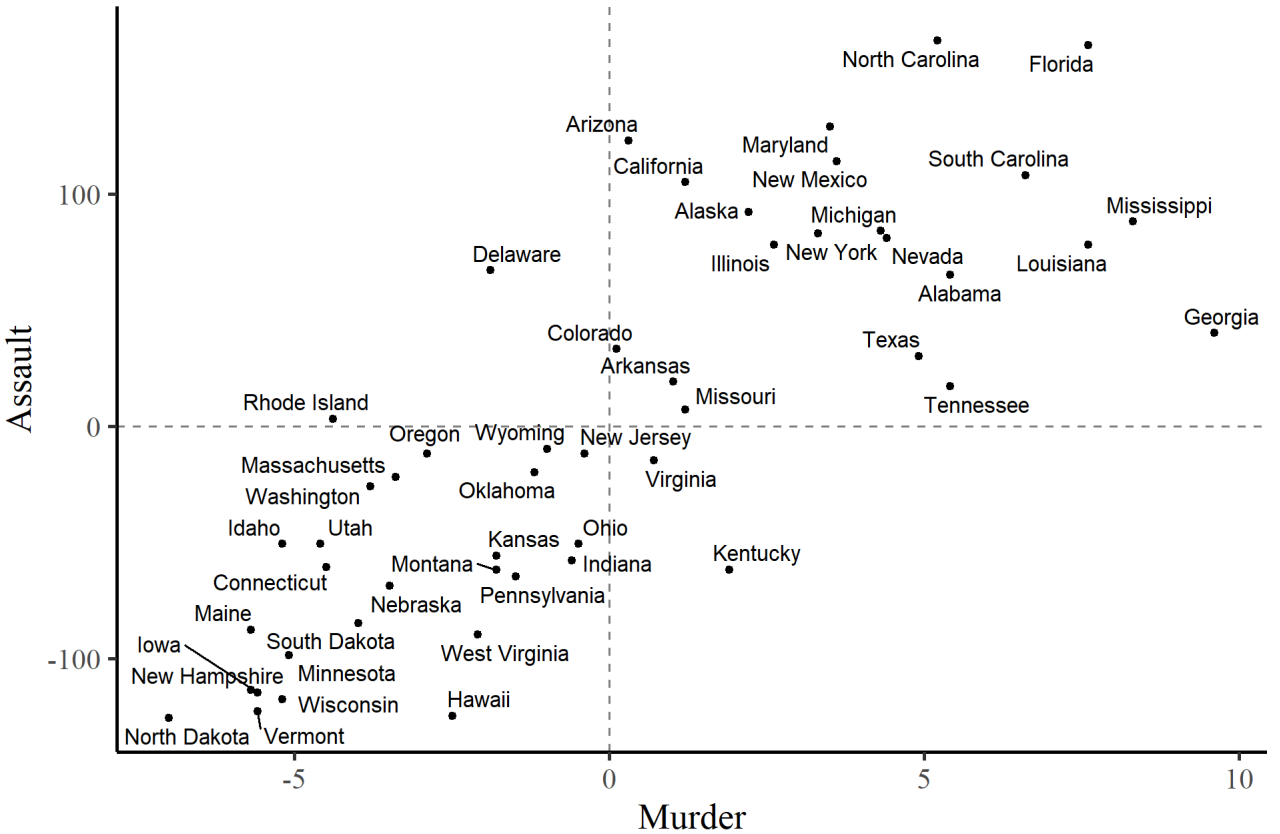
É possível obter as componentes principais a partir da matriz de covariâncias com a decomposição

$$\frac{1}{n} X^\top X = U \Sigma U^\top,$$

em que U é uma matriz $p \times p$ cujas colunas contém os autovetores da matriz de covariâncias $\frac{1}{n} X^\top X$ e Σ é uma matriz diagonal com seus autovalores. Os autovalores de U são ordenados de modo que seus respectivos autovalores sejam decrescentes.

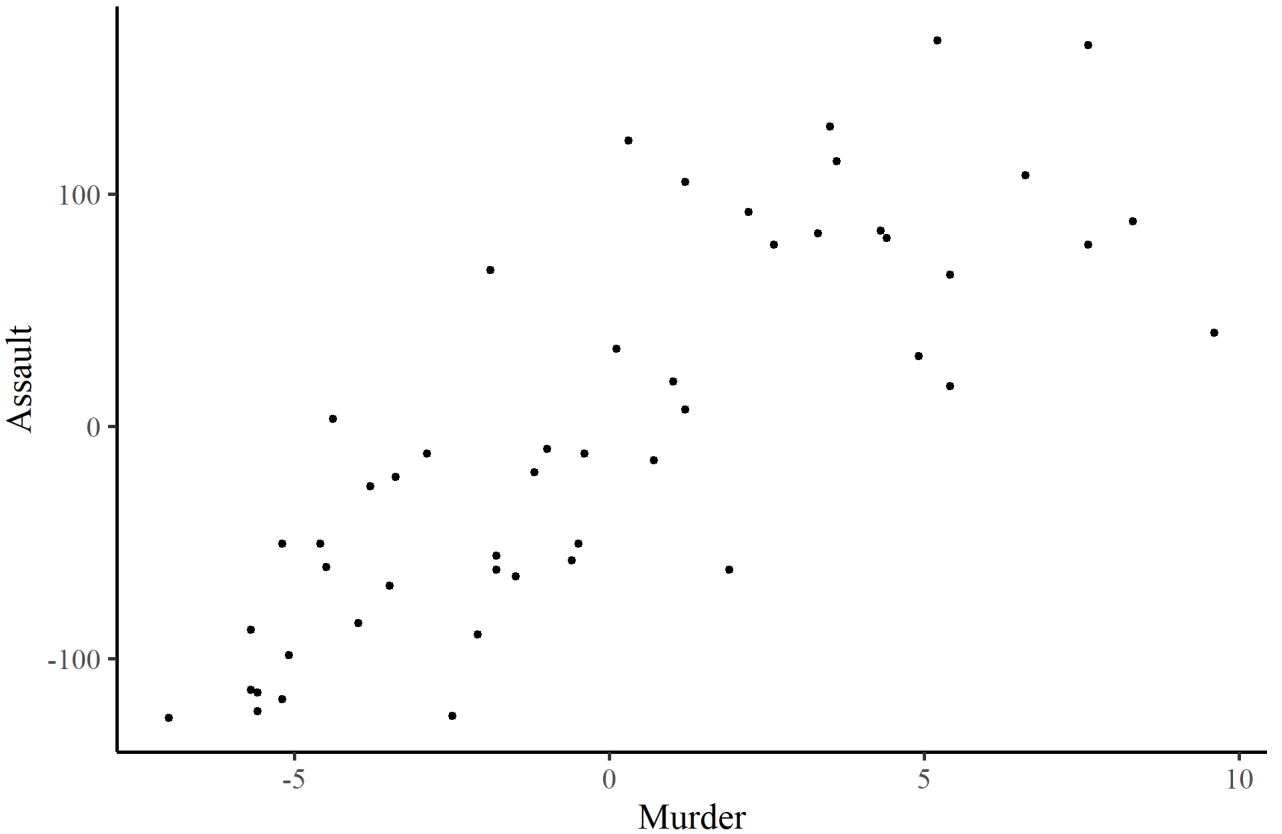
US Arrests Data

Vamos considerar inicialmente apenas as colunas `Murder` e `Assault`. Note que os dados abaixo estão centralizados (tem média zero).



US Arrests Data

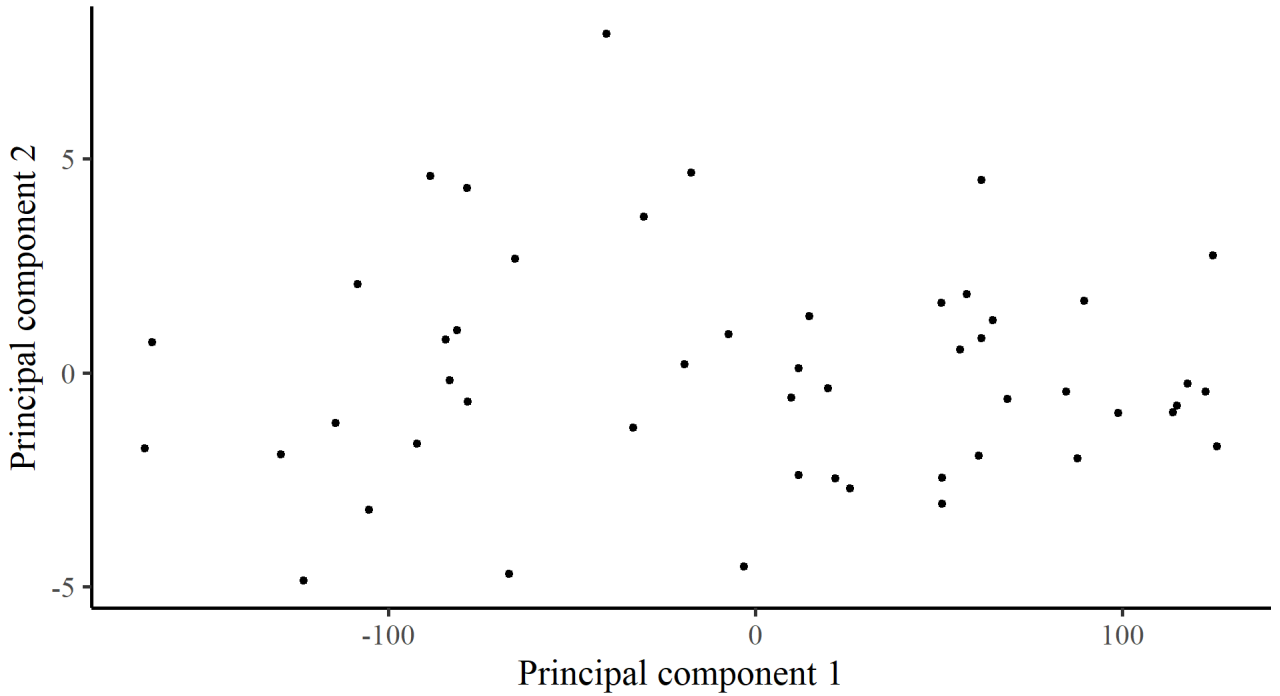
Vamos considerar inicialmente apenas as colunas `Murder` e `Assault`. Note que os dados abaixo estão centralizados (tem média zero).



US Arrests Data

O trecho de código abaixo aplica o método de PCA aos dados.

```
X <- USArrests %>%
  select(Murder, Assault) %>%
  scale(center=TRUE, scale=FALSE)
pca <- prcomp(X)
Z <- pca$x
```



Proportion of Variance Explained (PVE)

É importante saber o quanto perdemos de informação ao projetar os dados em uma dimensão reduzida. Qual é a PVE para cada componente?

A variância total dos dados centralizados é dada por

$$\sum_{j=1}^{p} \text{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2,$$

e a variância explicada pela m-ésima componente é

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2.$$

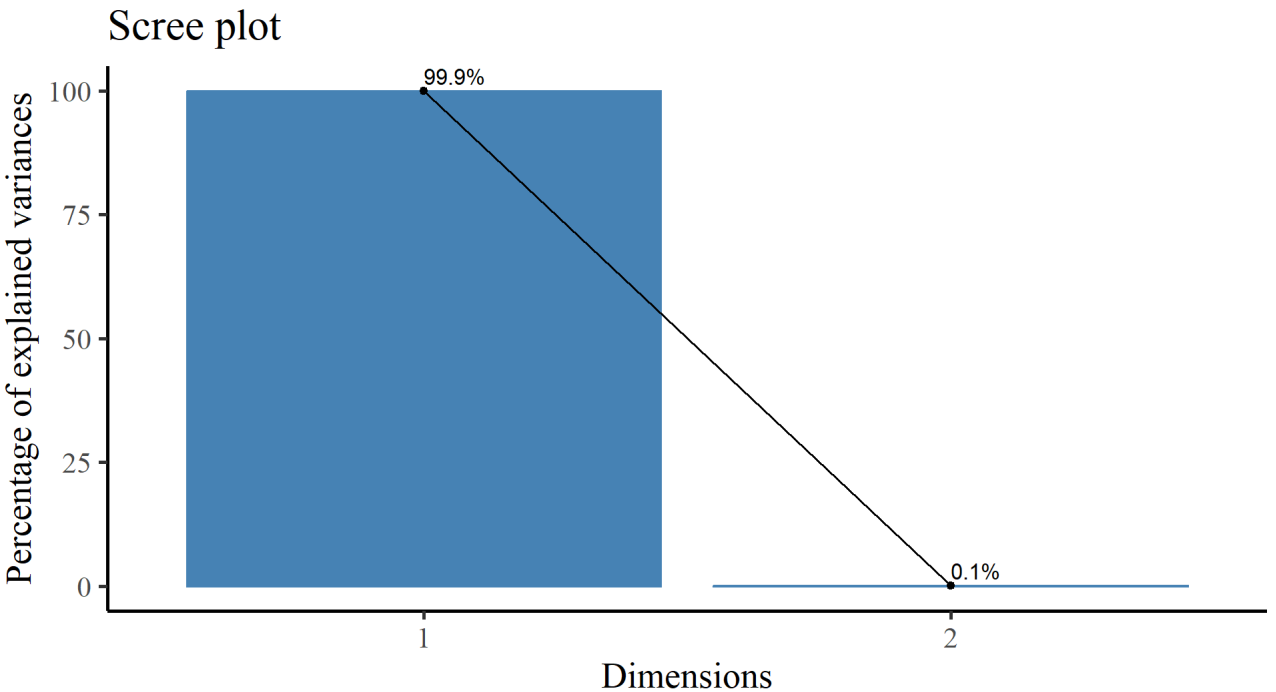
Portanto, a proporção de variância explicada pela m-ésima componente é dada por

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}.$$

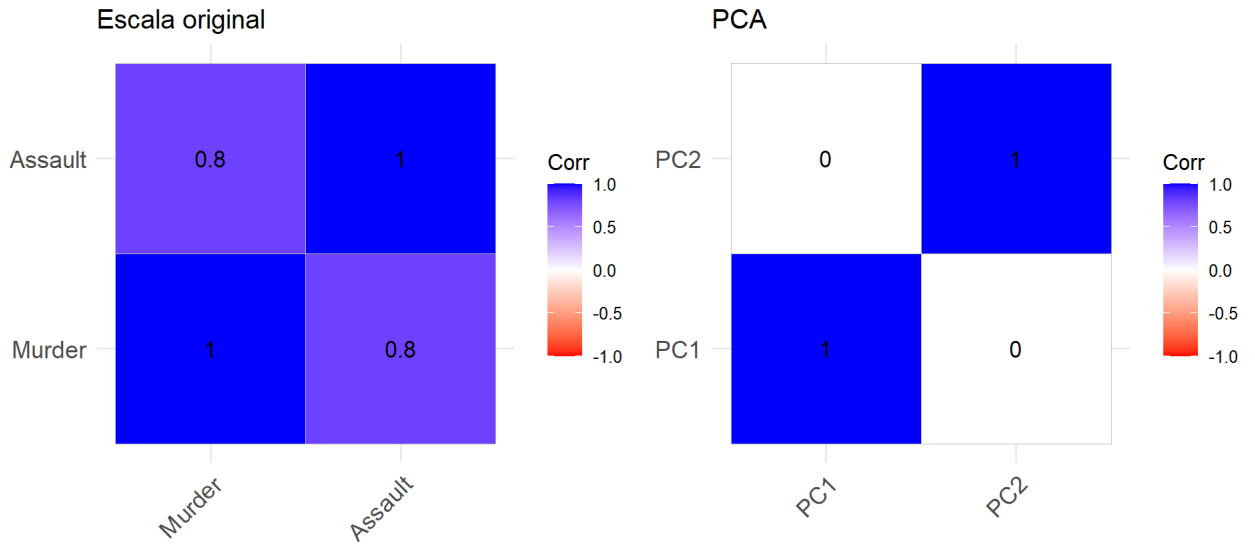
Gráfico de Cotovelo (scree plot)

A proporção de variância explicada é utilizada para determinar o número de componentes que serão utilizados. O gráfico de cotovelo pode ajudar nessa determiniação.

```
library(factoextra)

fviz_eig(pca, addlabels = TRUE)
```



Correlações

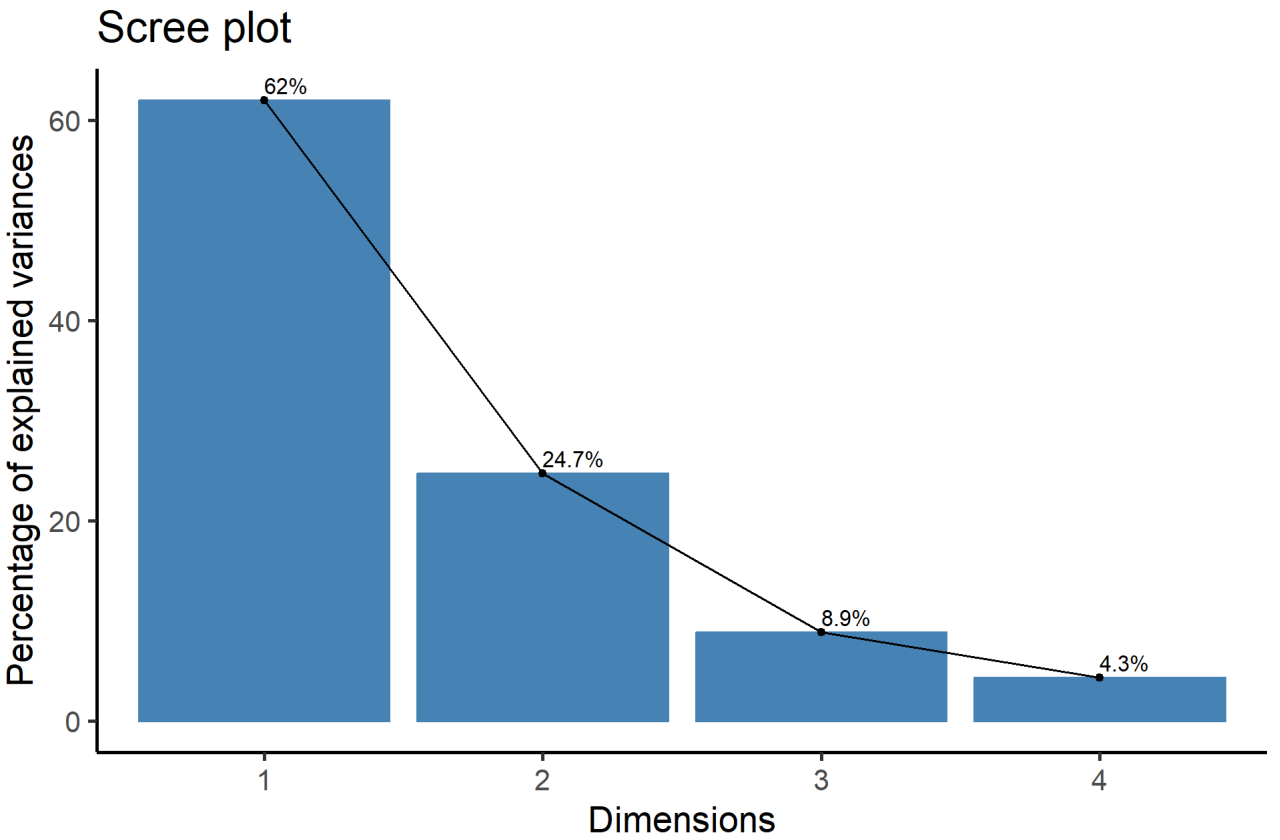


US Arrests Data

Agora, vamos considerar todas as quatro variáveis disponíveis na base de dados: `Murder`, `Assault`, `UrbanPop` e `Rape`.

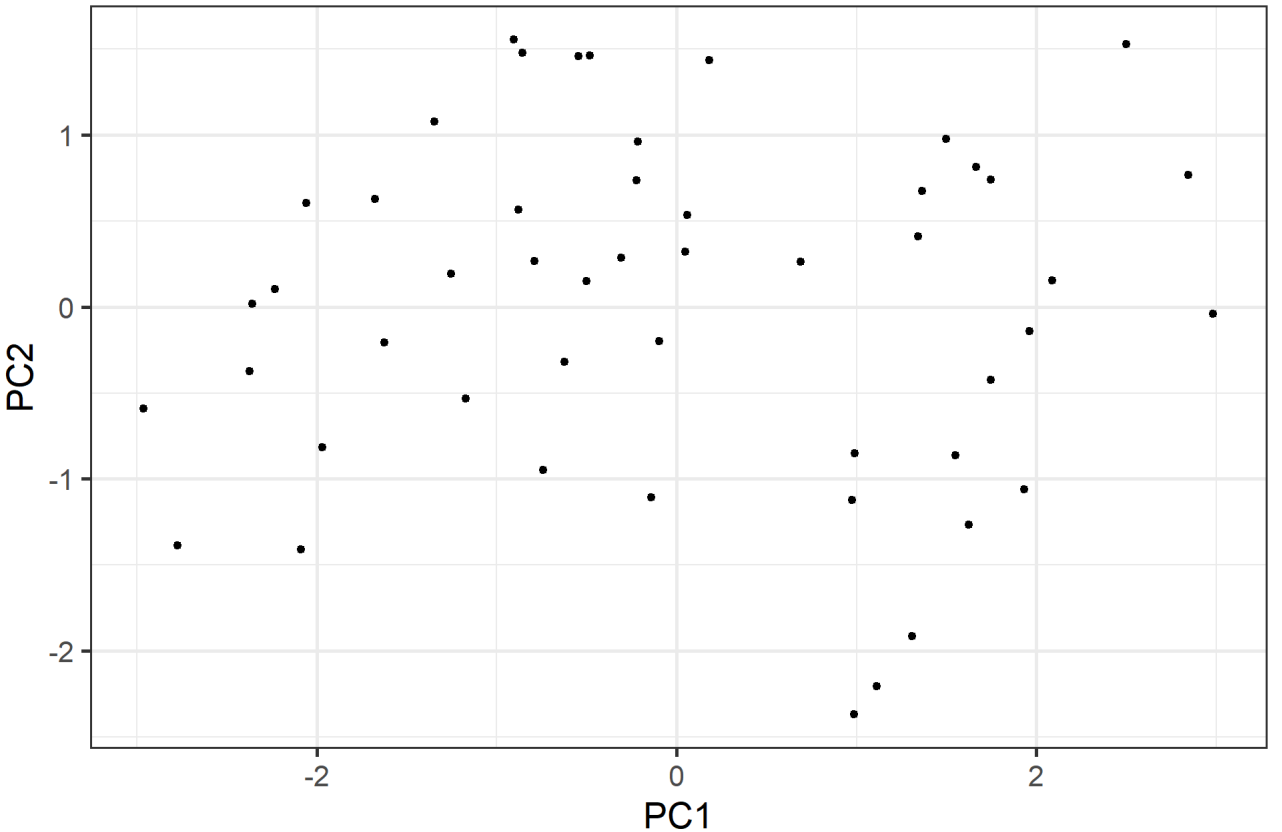
```
set.seed(1234)

X <- scale(USArrests, center = TRUE, scale = TRUE)

pca <- prcomp(X)

pca$rotation <- -pca$rotation
pca$x <- -pca$x

Phi <- pca$rotation
Z <- pca$x

fviz_eig(pca, addlabels = TRUE)
```

US Arrests Data



Note que, com apenas duas componentes principais, é possível explicar 86.8% da variabilidade dos dados originais.

Como interpretar as componentes principais?



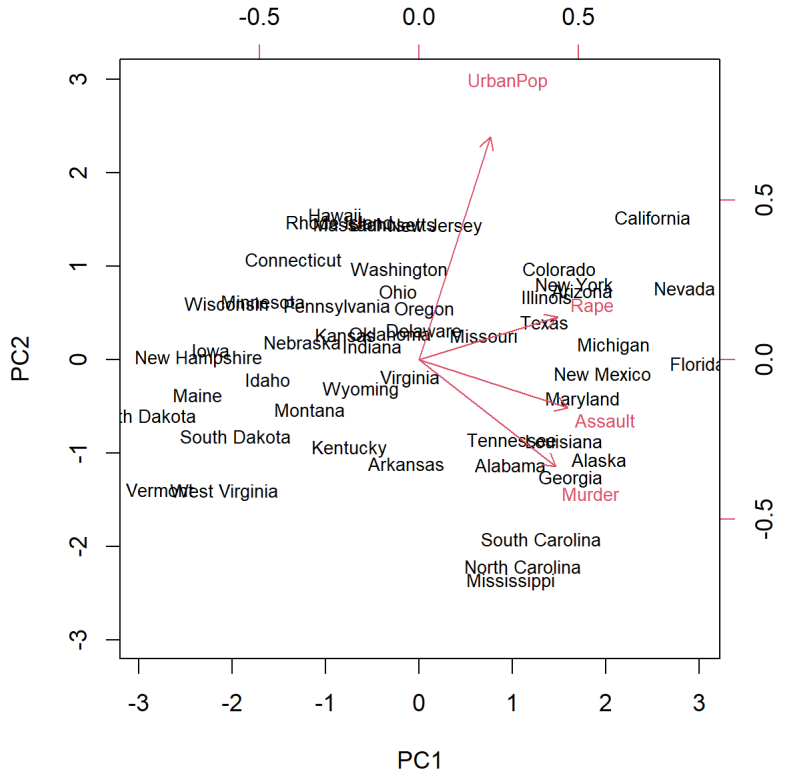
Como interpretar as cargas das componentes principais?

```
Phi %>%
  round(2)
```

	PC1	PC2	PC3	PC4
Murder	0.54	-0.42	0.34	-0.65
Assault	0.58	-0.19	0.27	0.74
UrbanPop	0.28	0.87	0.38	-0.13
Rape	0.54	0.17	-0.82	-0.09

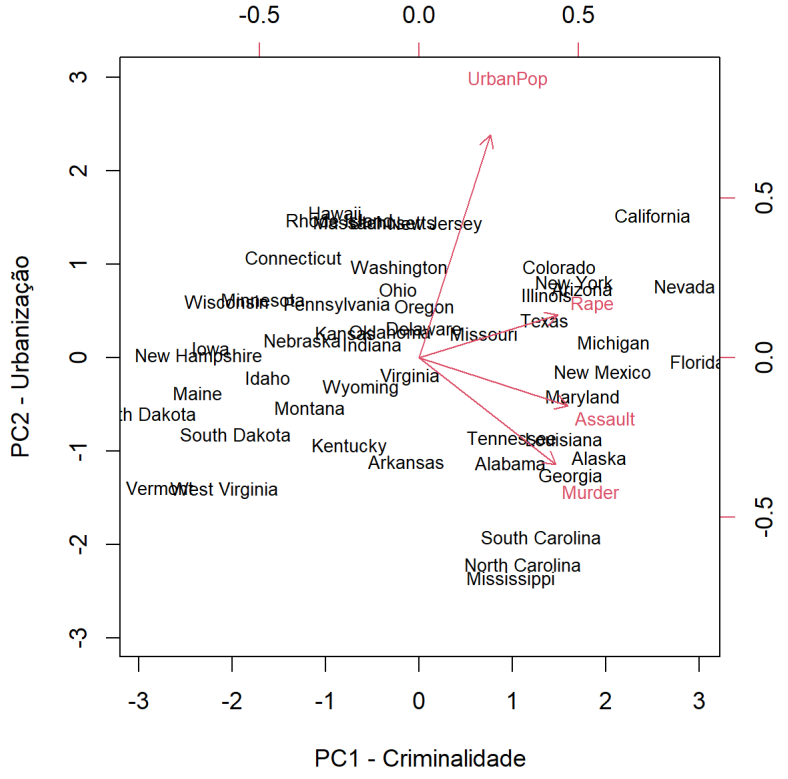
Como interpretar as componentes principais? - Biplot

```
biplot(pca, scale = 0, cex = 0.75,xlab = "PC1", ylab = "PC2")
```



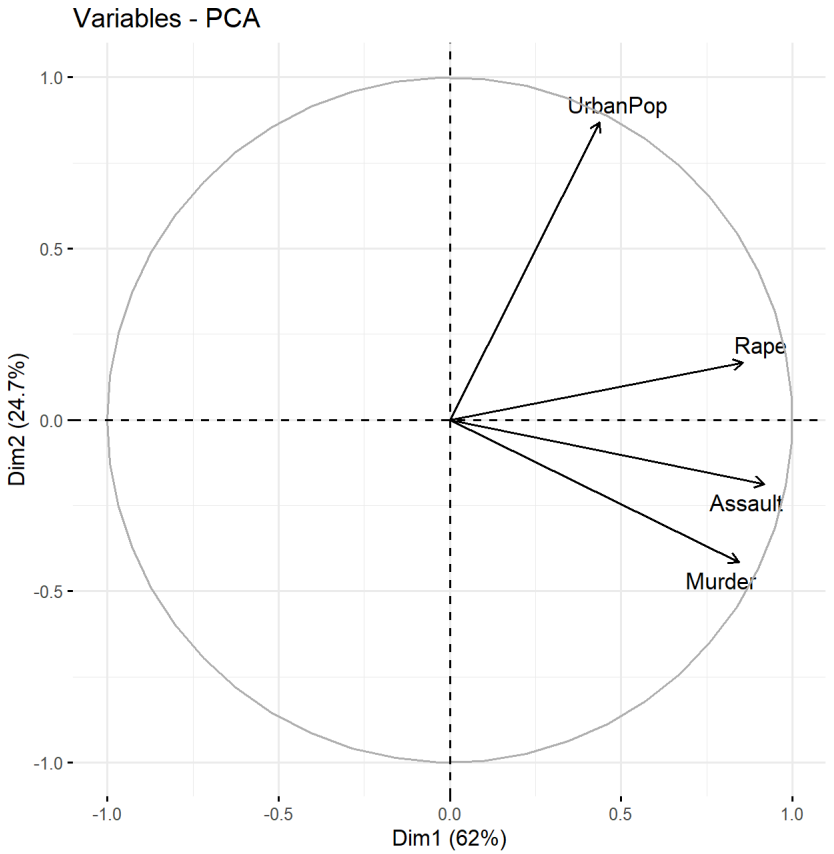
Como interpretar as componentes principais? - Biplot

```
biplot(pca, scale = 0, cex = 0.75,
       xlab = "PC1 - Criminalidade", ylab = "PC2 - Urbanização")
```



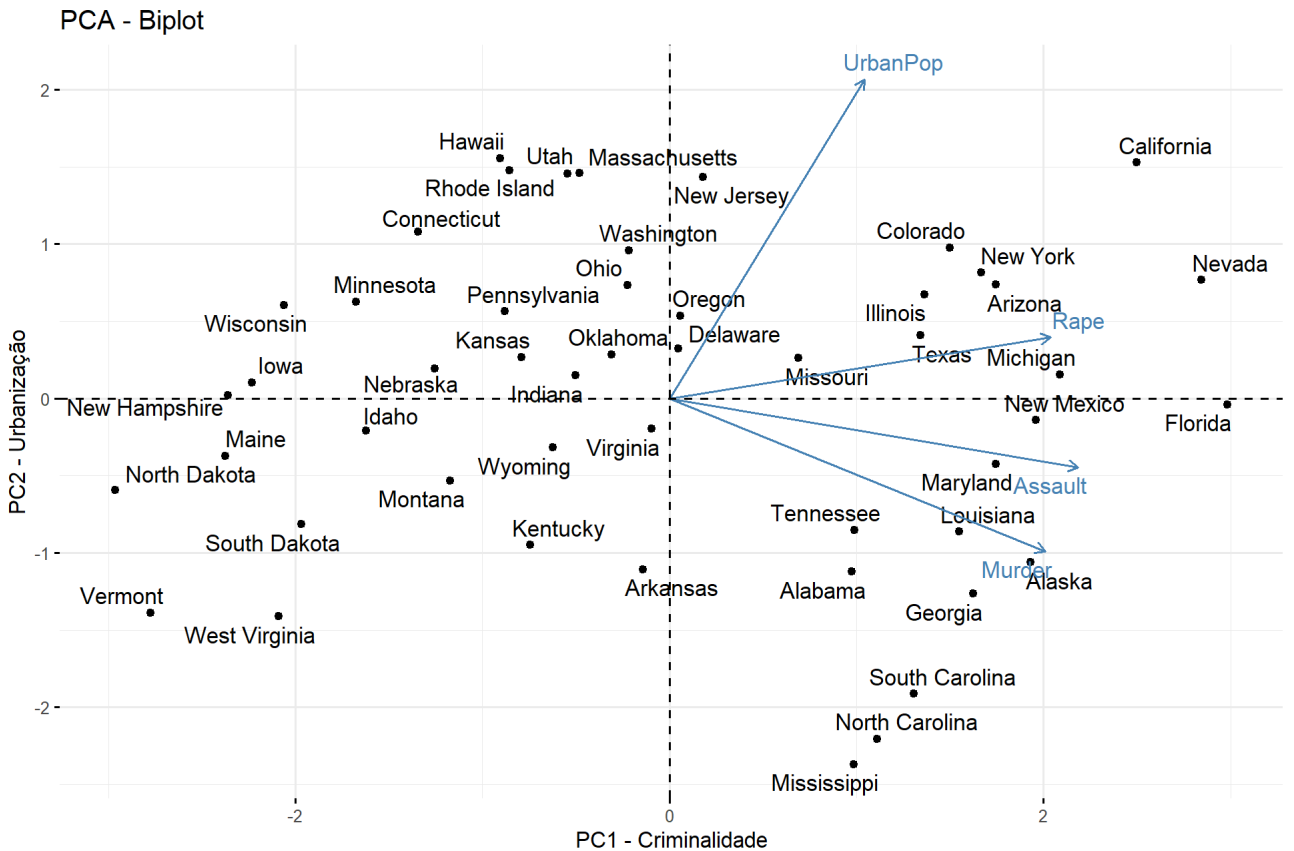
Como interpretar as componentes principais? - Alternativa

```
fviz_pca_var(pca, repel = TRUE, geom = c("arrow", "text"))
```



Como interpretar as componentes principais? - Alternativa

```
fviz_pca_biplot(pca, repel = TRUE, xlab = "PC1 - Criminalidade",
                ylab = "PC2 - Urbanização")
```



Estudo de Caso: Posicionamento de Marcas

Estudo de Caso: Posicionamento de Marcas

- Estes dados são apresentados no livro Data Science, Marketing & Business, dos professores Pedro J. Fernandez, Paulo C. Marques, Tiago Mendonça dos Santos e Hedibert F. Lopes.
- Nesta aplicação, 750 consumidores de alvejantes avaliaram um conjunto reduzido de marcas respondendo sobre preferências e avaliações de 29 atributos.

```
avaliacoes <- read_csv("Dados/avaliacoes.csv")
```

respondente	marca	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
1	RiUS	7	6	7	6	7	6	7	7	6
1	CanX	5	1	6	5	5	5	6	6	5
1	Candu	5	1	4	5	4	6	1	5	4
2	AbraxF	6	5	5	5	6	6	5	5	6
2	Mundo	5	5	6	3	3	5	5	5	4
3	Candu	5	4	4	5	3	3	5	6	5
3	RiUtil	7	6	6	6	6	6	6	6	6
3	Beach	6	6	2	5	3	6	6	5	6

Posicionamento de Marcas

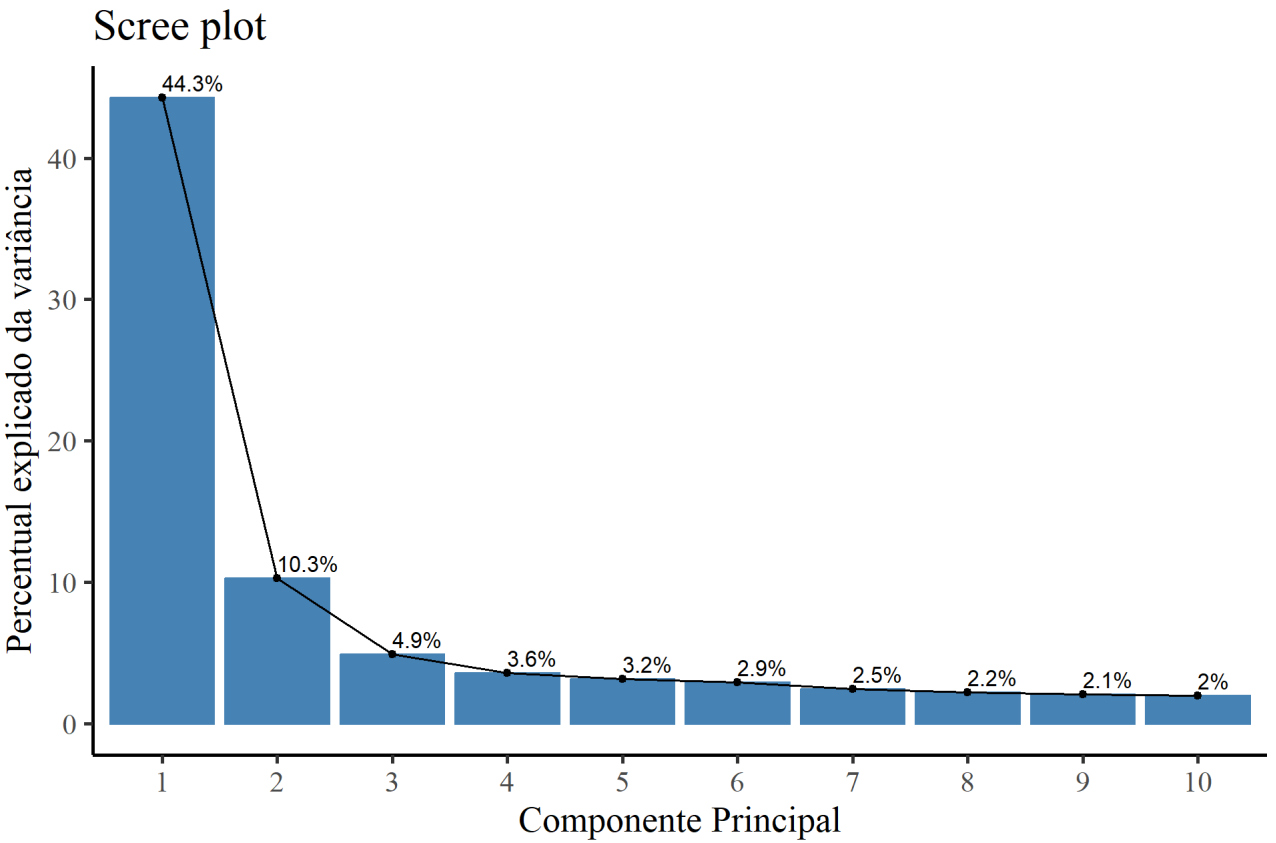
```
questoes <- read_csv("Dados/questoes.csv")
```

ID	Pergunta
Q1	Deixa as roupas mais brancas
Q2	É adequada para roupas coloridas
Q3	É a melhor para a remoção de manchas de gordura
Q4	Desinfeta melhor
Q5	É boa para limpar gordura
Q6	Deixa um aroma agradável na casa
Q7	É suave para as mãos
Q8	É adequada para a limpeza pesada

Posicionamento de Marcas - PCA

```
pca <- avaliacoes %>%
  select(starts_with("Q")) %>%
  prcomp(scale = TRUE)


fviz_eig(pca, addlabels = TRUE) +
  labs(x = "Componente Principal",
       y = "Percentual explicado da variância")

(cumsum(pca$sdev^2) / sum(pca$sdev^2))[1:3]

get_eigenvalue(pca)[1:10, 1:2]
```

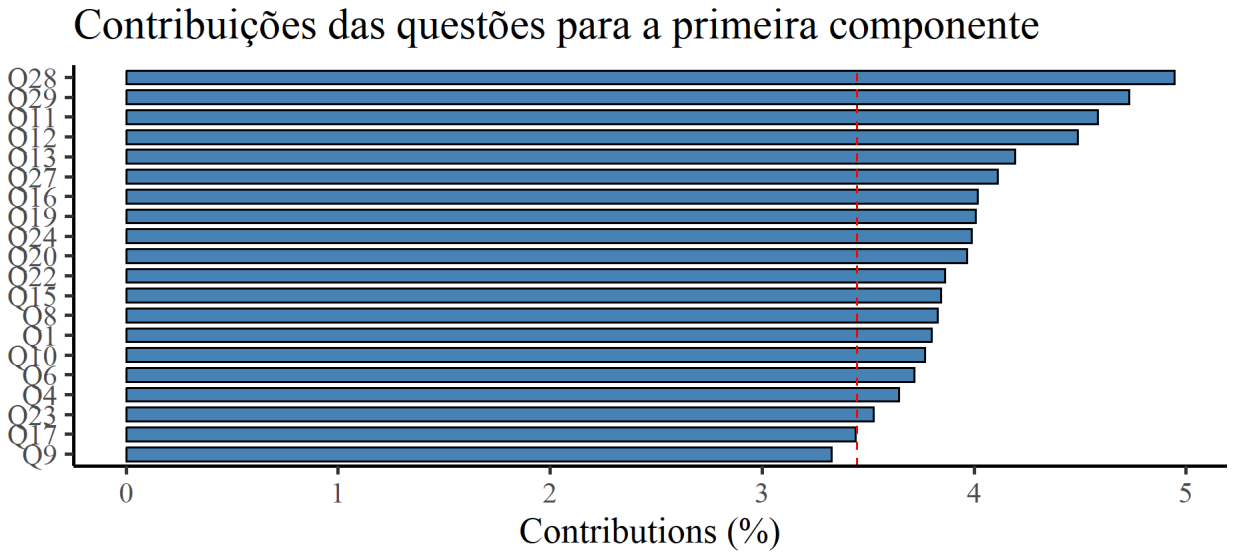
Posicionamento de Marcas - PCA



Note que as três primeiras componentes já explicam 59.4% da variabilidade dos dados.

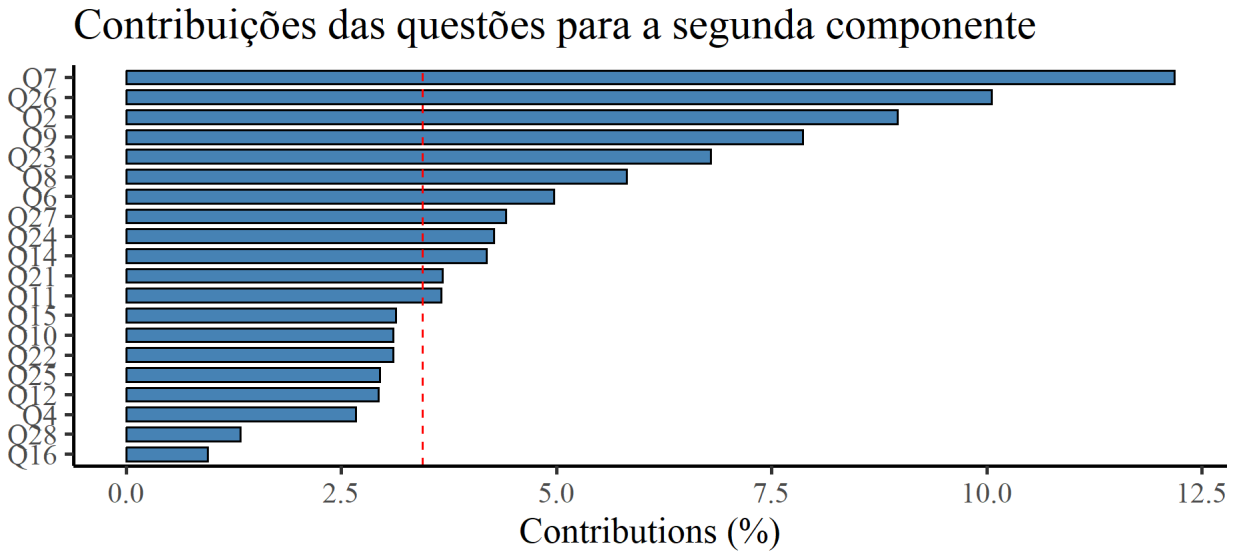
Interpretando as componentes principais

```
pca %>%
  fviz_contrib(choice = "var", axes = 1, sort.val = "asc", top = 20,
               fill = "steelblue", color = "black") +
  labs(x = "", title = "Contribuições das questões para a primeira compon
  coord_flip()+
  theme_set(theme_classic(base_size = 18))+
  theme(text = element_text(family = "serif"))
```



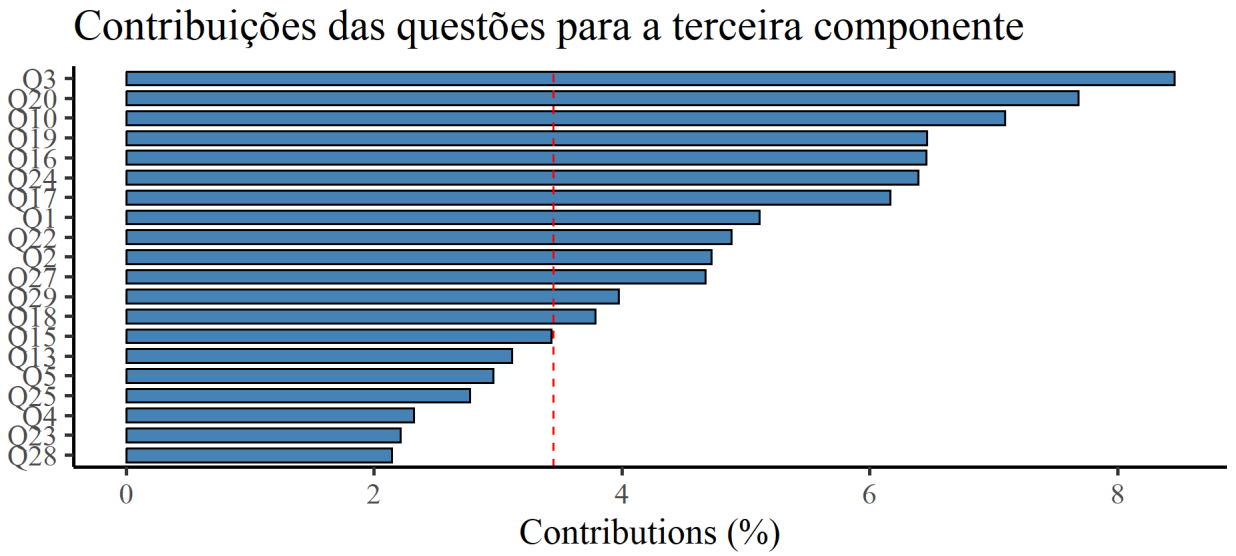
Posicionamento de Marcas - interpretando as PCs

```
pca %>%
  fviz_contrib(choice = "var", axes = 2, sort.val = "asc", top = 20,
               fill = "steelblue", color = "black") +
  labs(x = "", title = "Contribuições das questões para a segunda compone
  coord_flip()+
  theme_set(theme_classic(base_size = 18))+
  theme(text = element_text(family = "serif"))
```



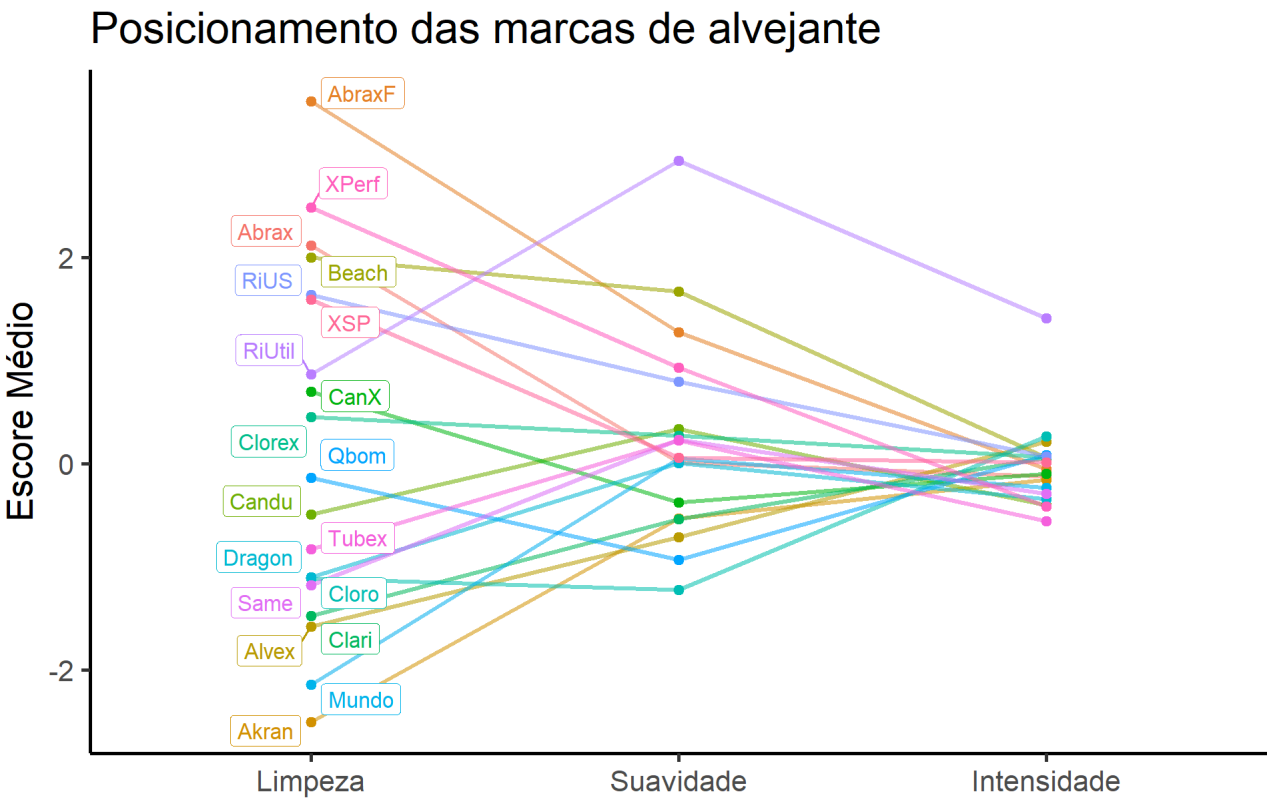
Posicionamento de Marcas - interpretando as PCs

```
pca %>%
  fviz_contrib(choice = "var", axes = 3, sort.val = "asc", top = 20,
               fill = "steelblue", color = "black") +
  labs(x = "", title = "Contribuições das questões para a terceira compon
  coord_flip()+
  theme_set(theme_classic(base_size = 18))+
  theme(text = element_text(family = "serif"))
```



Posicionamento de Marcas

O posicionamento das marcas é determinado pelo valor médio destas sobre os drivers (componentes principais).



PCA em Modelagem

Boston Data

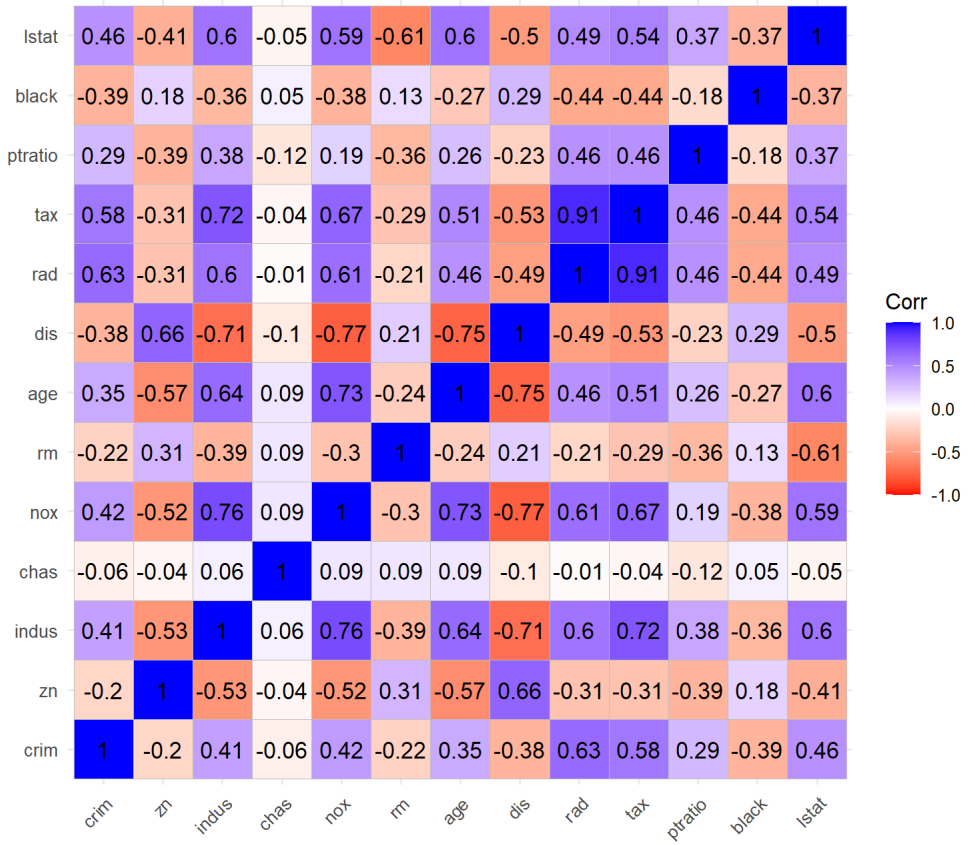
A seguir trabalharemos com os dados `Boston`.

- `crim`: per capita crime rate by town.
- `zn`: proportion of residential land zoned for lots over 25,000 sq.ft.
- `indus`: proportion of non-retail business acres per town.
- `chas`: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- `nox`: nitrogen oxides concentration (parts per 10 million).
- `rm`: average number of rooms per dwelling.
- `age`: proportion of owner-occupied units built prior to 1940.
- `dis`: weighted mean of distances to five Boston employment centres.
- `rad`: index of accessibility to radial highways.
- `tax`: full-value property-tax rate per $10,000.
- `ptratio`: pupil-teacher ratio by town.
- `black`: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
- `lstat`: lower status of the population (percent).
- `medv`: median value of owner-occupied homes in $1000s.

```
library(MASS)

Boston
```

Boston Data

Veja a correlação entre as preditoras no gráfico abaixo.



Boston Data

O gráfico anterior foi feito com o código abaixo.

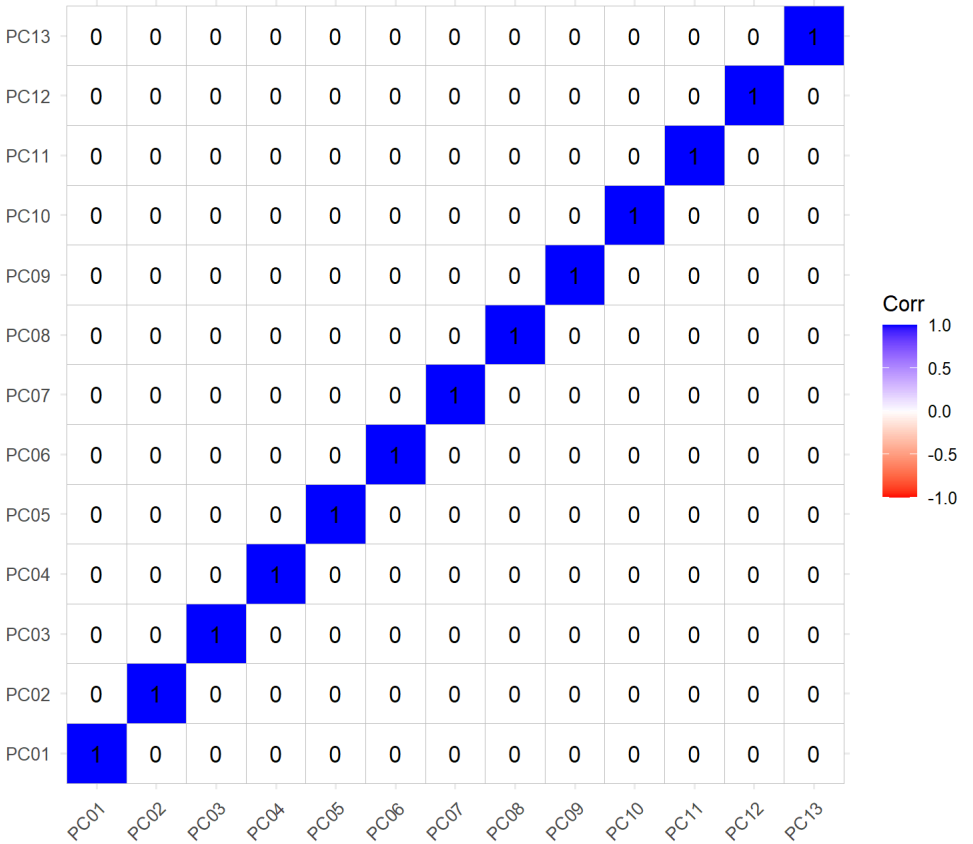
```
Boston %>%
  dplyr::select(-medv) %>%
  cor() %>%
  ggcorrplot(lab = TRUE)
```

Caso queira considerar uma PCA com 5 componentes, utilize:

```
receita <- recipe(medv ~ ., Boston) %>%
  step_normalize(all_predictors()) %>%
  step_pca(all_predictors(), num_comp = 5)
```

Como considerar um número de componentes principais que represente uma porcentagem mínima de explicação da variância? Consulte a documentação aqui.

Boston Data - PCA



Resumo da PCA

- A PCA permite sumarizar um conjunto de dados com um número grande de variáveis em um pequeno número de componentes representativos que coletivamente explicam o máximo da variabilidade dos dados originais.
- A redução de dimensionalidade é muito útil para a visualização dos dados e também para modelos preditivos supervisionados.
- Éste modelo é sensível à escala dos dados, é importante que os dados estejam normalizados.
- O número de componentes principais que devem ser usadas depende de uma análise subjetiva (baseada no gráfico de cotovelo).

O método k-médias para agrupamento

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Find hidden dependencies

ASSOCIATION

«Find what clothes I often wear together»

Identify sequences

Análise de agrupamento (clustering)

- O objetivo é definir grupos tais que os dados dentro de um mesmo grupo sejam **similares** segundo alguma perspectiva;
- Intuitivamente, queremos que dentro de cada cluster os objetos sejam **muito similares**; e que objetos de dois clusters distintos sejam **muito diferentes**;
- Existem duas formas principais de fazer análise de conglomerados: uma em que o número de grupos é definido **previamente** e outra em que o número de grupos é definido **posteriormente**.

Aplicações de métodos de agrupamento

- análise de dados;
- segmentação de clientes;
- detecção de *outliers*;
- segmentação de imagens.
- redução de dimensionalidade;

Análise de clusters

- Para cada observação $i = 1, \ldots, n$, conhecemos apenas o vetor $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$.

- Formalmente, um cluster C_r é o conjunto dos índices das observações que pertencem a ele.

- Queremos construir $k \geq 1$ clusters $C_1, \ldots, C_k$ tais que

$$\cup_{r=1}^{k} C_i = \{1, \ldots, n\}$$

e $C_i \cap C_j = \emptyset$ (*hard clustering*), para $i \neq j$ de modo a minimizar a dispersão intra-clusters

$$W = \frac{1}{2} \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i \in C_r} \sum_{j \in C_r} ||x_i - x_j||^2,$$

em que n_r é o número de observações no cluster C_r e $||x_i - x_j||^2 = \sum_{\ell=1}^{p} (x_{i\ell} - x_{j\ell})^2$ é o quadrado da distância Euclidiana entre x_i e x_j.

- Note que estamos considerando que k é um valor pré-definido. Na sequência vamos discutir sobre sua escolha.

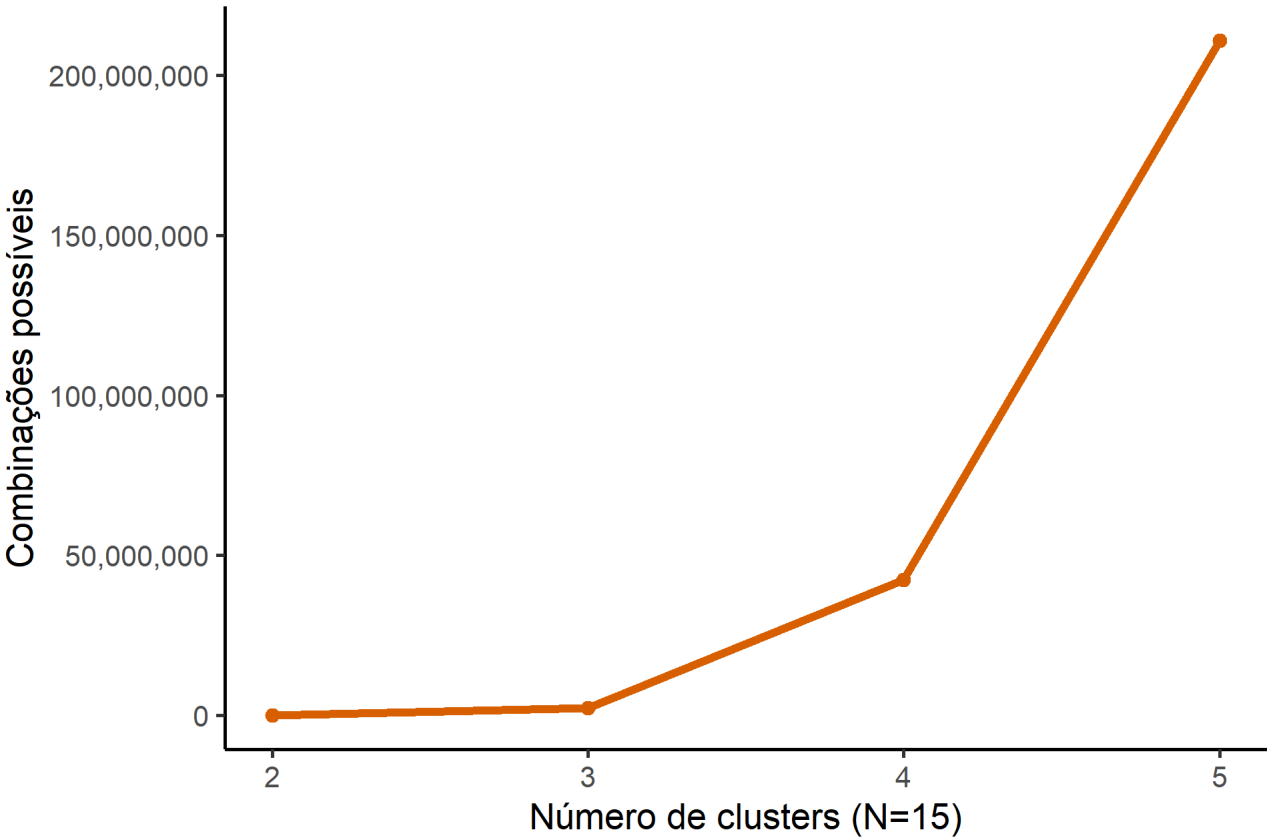
Função objetivo

Portanto nosso objetivo é encontrar $C_1, \ldots, C_k$ que minimizam o valor de W, ou seja,

$$\arg \min_{C_1,\ldots,C_k} \frac{1}{2} \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i \in C_r} \sum_{j \in C_r} ||x_i - x_j||^2.$$

Número de agrupamentos possíveis

Considere que temos apenas $N = 15$ observações.



Curiosidade matemática

O número de maneiras $A(n, k)$ de formar k clusters a partir de n observações é definido pela relação de recorrência abaixo:

$$A(n, k) = 1 \cdot A(n-1, k-1) + k \cdot A(n-1, k),$$

com as condições $A(n, 1) = A(n, n) = 1$.

```
A <- function(n, k){
  if(k == 1 ||  k ==n) return (1)
  return(A(n-1, k-1) + k * A(n-1, k))
}
```

Como interpretar a relação de recorrência acima?

Suponha que você seja a n-ésima observação. Então você tem duas escolhas, mutualmente exclusivas:

- você pode decidir criar um cluster apenas para você e as demais $n-1$ pessoas se agruparão em $k-1$ clusters de $A(n-1, k-1)$ maneiras;
- ou você pode escolher entrar em um dos k clusters e as demais $n-1$ pessoas se agruparão nestes n clusters de $A(n-1, k)$ maneiras.

Como encontrar os *clusters* então?

- Vimos que mesmo para números pequenos de N e k é computacionalmente inviável fazer uma busca exaustiva sobre todas as possíveis configurações de k clusters para N observações.
- Precisamos de uma alternativa!
- É possível mostrar que a seguinte igualdade vale:

$$\sum_{i \in C_r} \sum_{j \in C_r} ||x_i - x_j||^2 = 2n_r \sum_{i \in C_r} ||x_i - \bar{x}_r||^2,$$

- Aqui, $\bar{x}_r = \frac{1}{n_r} \sum_{i \in C_r} x_i$ é a média das observações pertencentes ao cluster C_r.
- A igualdade acima nos diz que a soma da distância entre elementos de um mesmo cluster é igual à soma das distâncias dos elementos de um cluster até o seu centróide.

Como encontrar os *clusters* então?

Então, nosso problema de otimização que era

$$\arg \min_{C_1,\ldots,C_k} \frac{1}{2} \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i \in C_r} \sum_{j \in C_r} ||x_i - x_j||^2.$$

pode ser reescrito como

$$\arg \min_{C_1,\ldots,C_k} \sum_{r=1}^{k} \sum_{i \in C_r} ||x_i - \bar{x}_r||^2.$$

Como encontrar os *clusters* então?

Outro fato que pode ser demonstrado matematicamente é que para os vetores $u_1, \ldots, u_m \in \mathbb{R}^p$, a quantidade

$$\sum_{i=1}^{m} ||u_i - c||^2$$

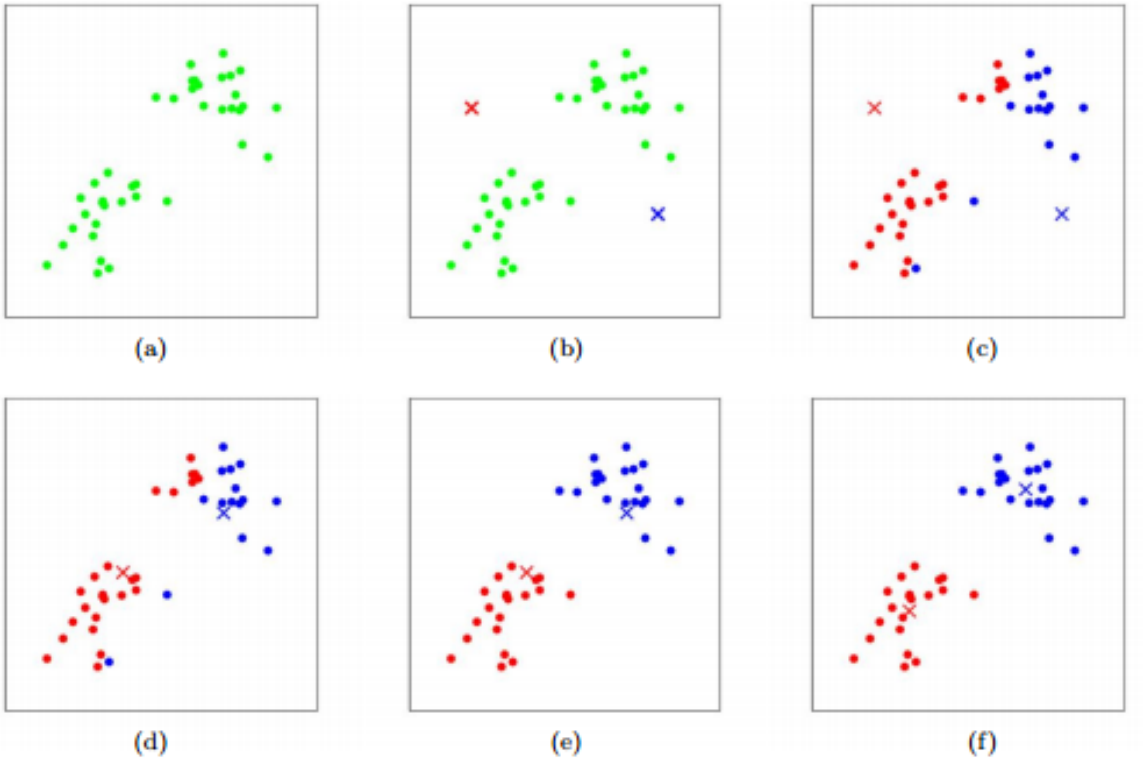
é minimizada escolhendo-se o vetor $c = \bar{u} = \frac{1}{m} \sum_{i=1}^{m} u_i$.

Com o resultado acima, resolver o problema original é equivalente à minimizar o custo

$$\arg \min_{\substack{C_1,\ldots,C_k \\ m_1,\ldots,m_k}} \sum_{r=1}^{k} \sum_{i \in C_r} ||x_i - m_r||^2.$$

Esta representação do problema sugere uma solução iterativa em que primeiramente fixamos os m_r's e minimizamos o custo escolhendo os C_r's adequadamente, e posteriormente fixamos os C_r's e minimizamos o custo escolhendo os m_r's como sendo as médias das observações nos respectivos clusters.

Algoritmo k-means



Algoritmo k-means (Lloyd, 1957)

(1). Inicializamos arbitratiramente $m_1, \ldots, m_k$.

(2). Alocamos a observação x_i no cluster C_r tal que

$$r = \arg \min_{1 \leq r \leq k} ||x_i - m_r||,$$

para $i = 1, \ldots, n$. Deste modo, determinamos $C_1, \ldots, C_k$.

(3). Fazemos $m_r = \bar{x}_r$, para $r = 1, \ldots, k$.

(4). Iteramos os dois passos anteriores até que o valor de W fique inalterado.

Observações

- O algoritmo converge, uma vez que o conjunto envolvido na iteração é finito;
- Não há nenhuma garantia de que encontraremos um mínimo global de W;
- Por esse motivo, recomenda-se executar o algoritmo diversas vezes com inicializações distindas para os m_r's.

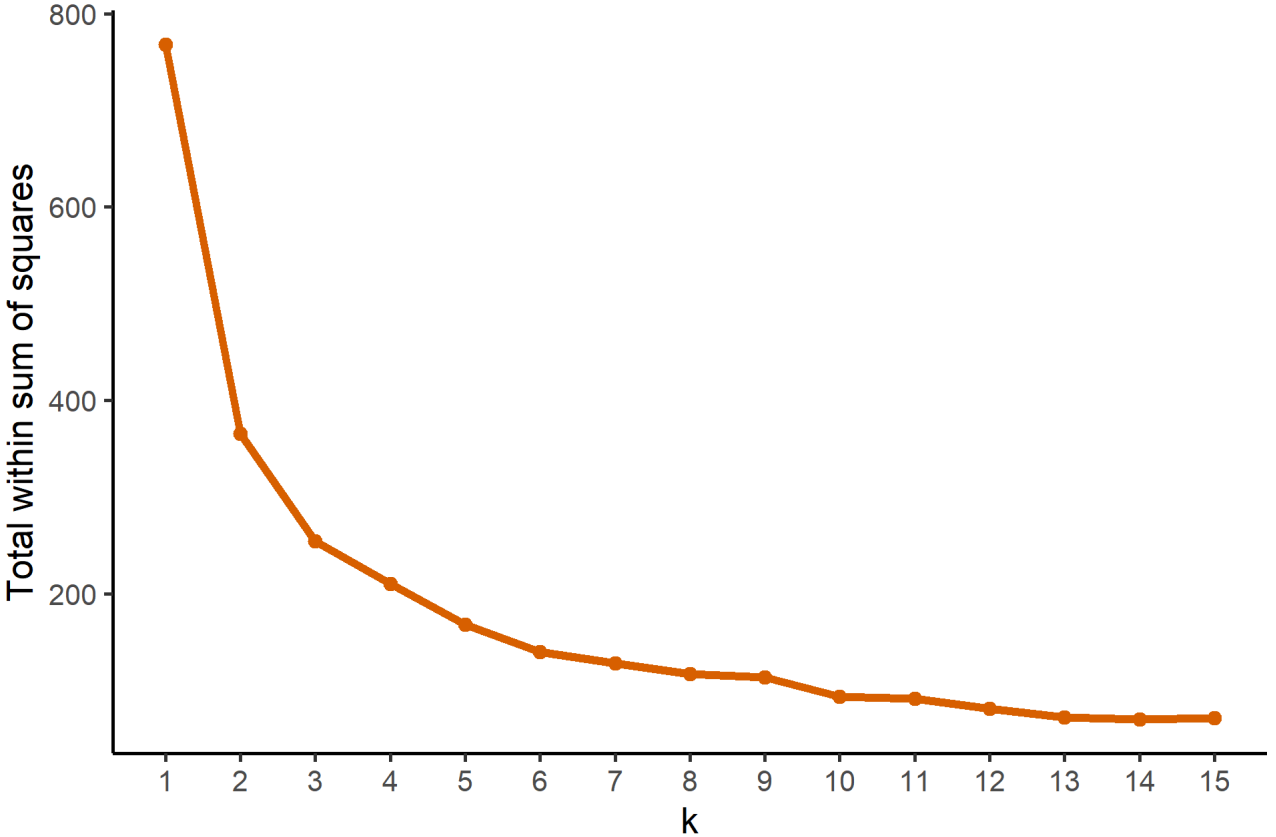
Definição do Número de clusters

Definição de k

Em muitas situações o número de clusters é definido previamente. Para as situações em que esse número não é pré fixado, podemos utilizar alguns métodos para isso:

- método do cotovelo;
- coeficiente de silhouette.

Método do cotovelo



Método silhouette

O coeficiente silhouette é dado por

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

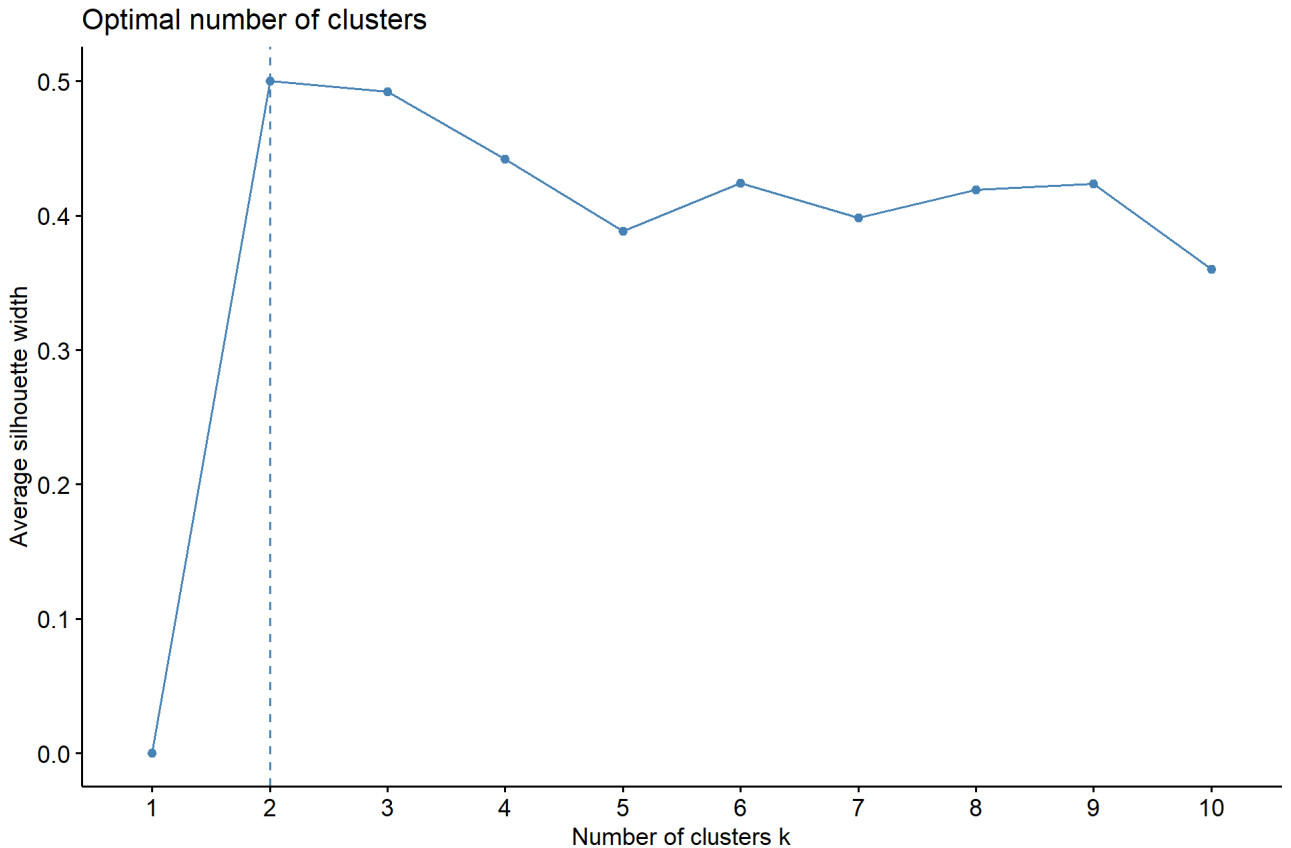
em que

- $a(i)$ é a distância média entre a observação i e as demais observações do mesmo cluster;
- $b(i)$ é a menor distância média da observação i a todos os pontos dos clusters aos quais i não pertence.

Note que $-1 \leq s(i) \leq 1$ e

- $s(i) \approx +1$ está "dentro" do próprio cluster e afastada dos demais;
- $s(i) \approx 0$ está próxima da fronteira entre clusters;
- $s(i) \approx -1$ pode ter sido agrupada de forma equivocada.

Método silhouette



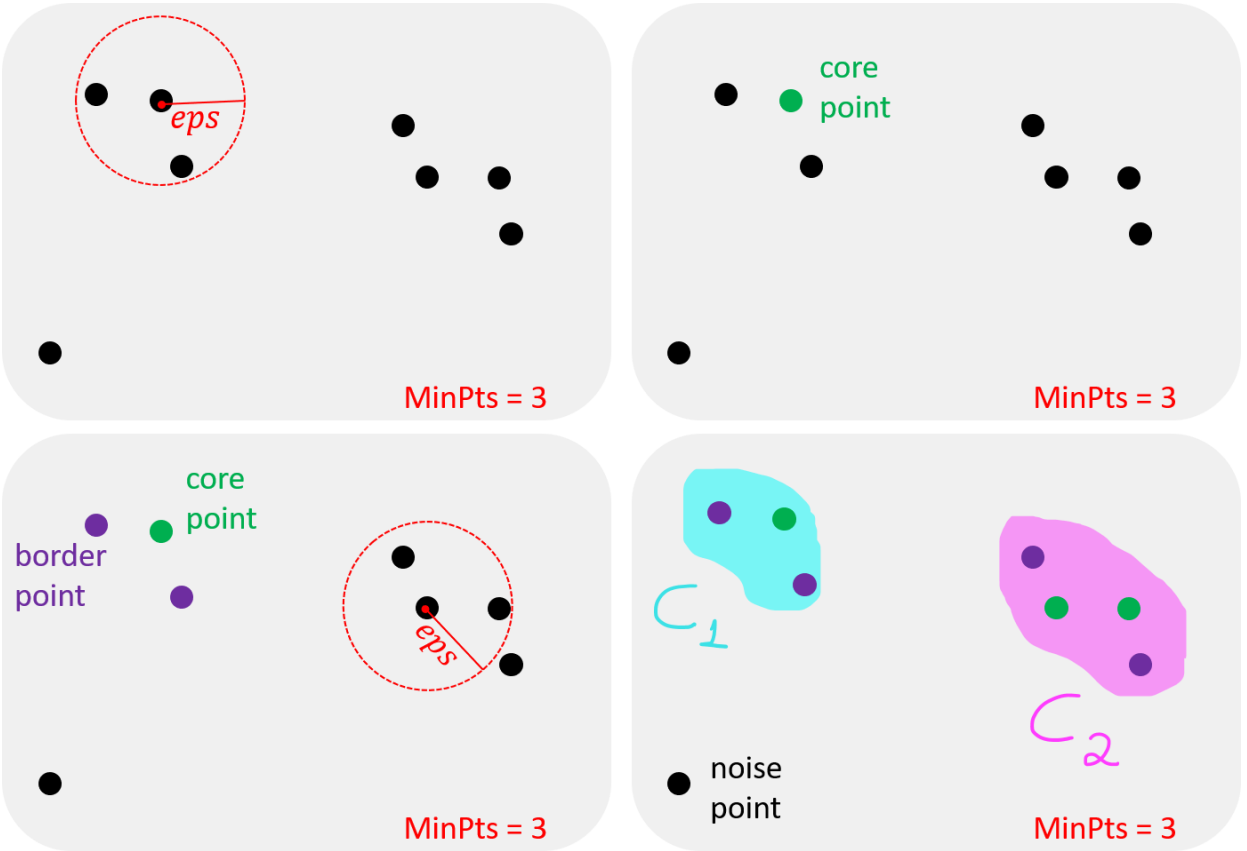
Métodos alternativos

A seguir está uma lista com alguns métodos que são alternativos ao k-médias:

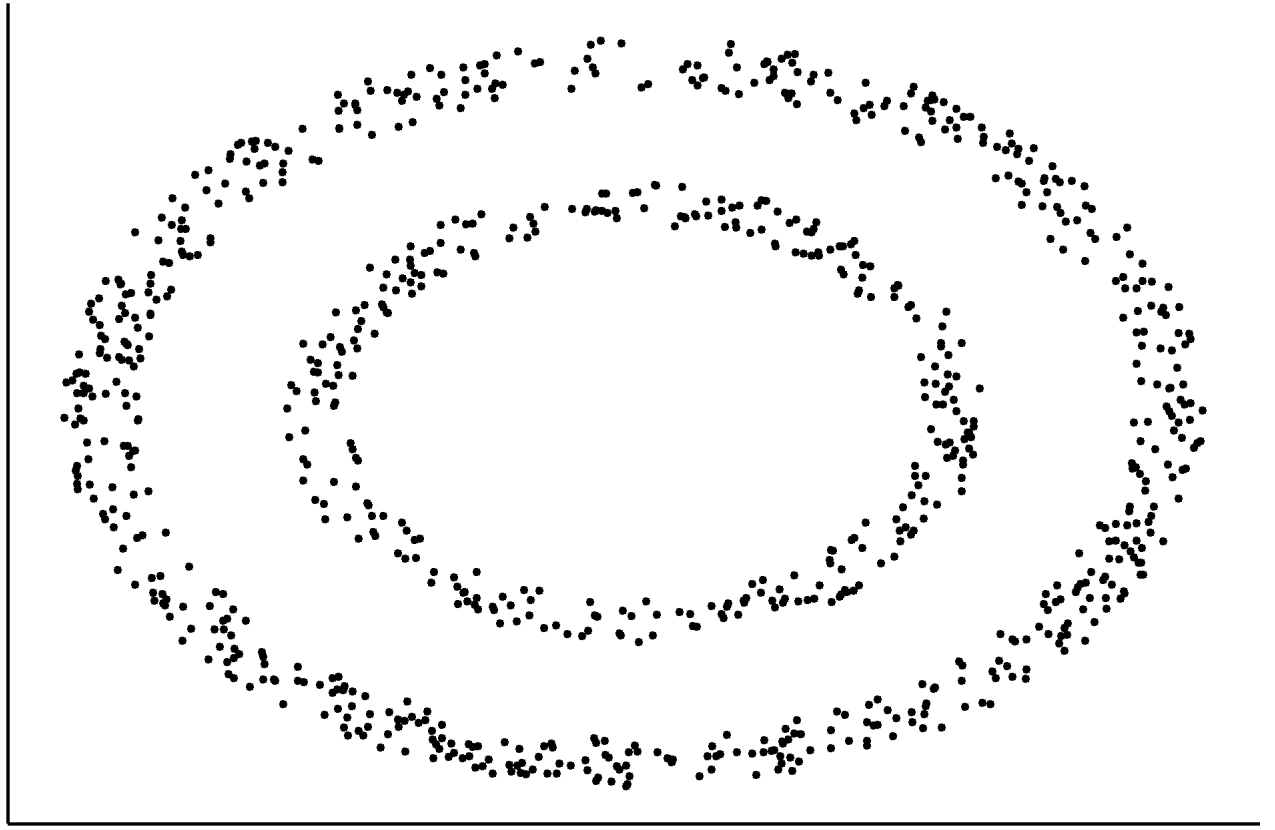
- DBSCan (Density-based spatial clustering of applications with noise);
- Spectral clustering;
- Mistura de Gaussianas;
- Self-Organizing Maps;
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies).

DBSCAN

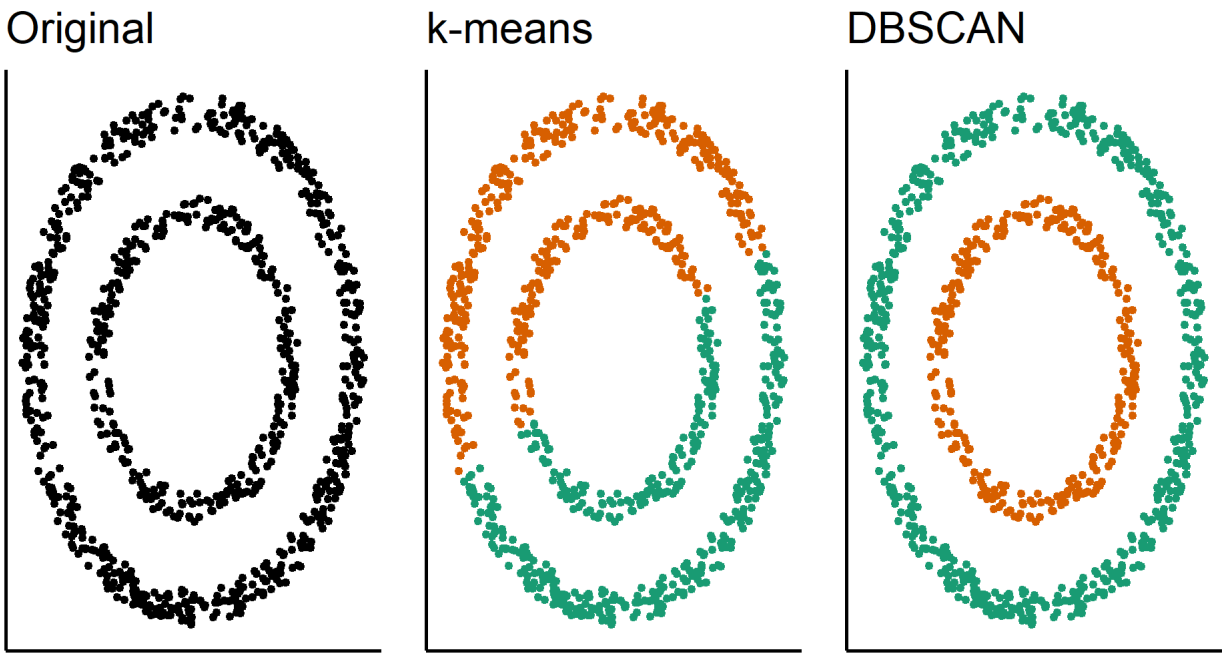
Density-based spatial clustering of applications with noise (DBSCAN)



Exemplo com dados sintéticos



Comparativo k-médias e DBSCAN



Resumindo k-médias

- Análise de conglomerados tem diversas aplicações práticas;
- O algoritmo k-means é uma técnica não supervisionada para clusterização.
- O algoritmo divide o conjunto de dados em k grupos baseado em sua similaridade.
- A média de cada cluster é usada como centróide e cada observação é associada ao centroide mais próximo.
- Os centróides são atualizados de forma iterativa até obter a convergência, resultado em k grupos e cada observação associada à um grupo.
- Limitação: o número de clusters deve ser definido previamente;
- Métodos como o gráfico de cotovelo e o coeficiente de silhouette nos ajudam a definir o número de clusters posteriormente.
- Aplicações:
    - Segmentação de imagens;
    - Segmentação de clientes;
    - Detecção de anomalias.

Obrigado!

magnotfs@insper.edu.br