Análise de Componentes Principais

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte

Introdução

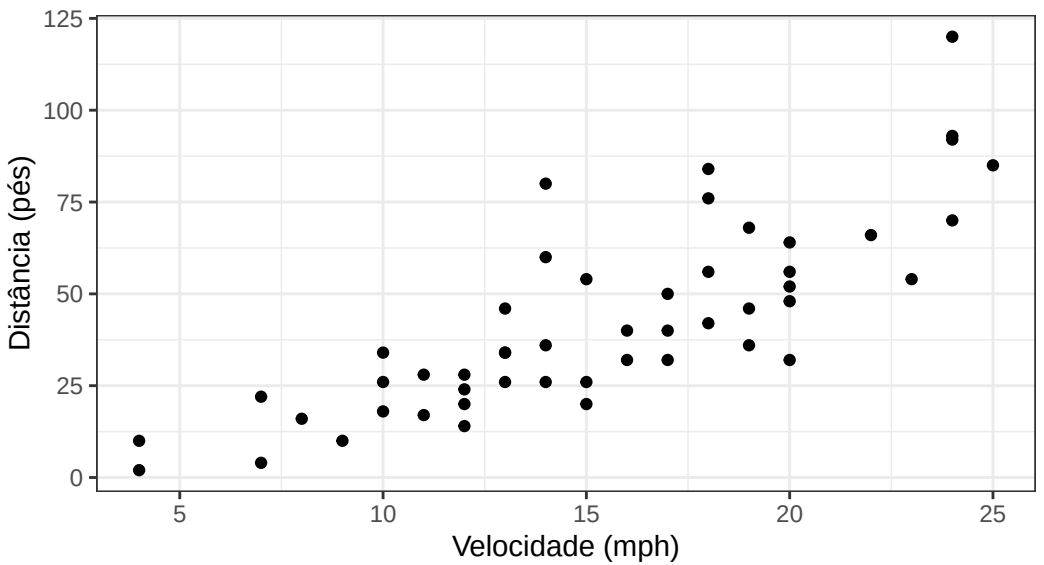
Visualização dos Dados

- Já sabemos como visualizar dados no R
- Os comandos `plot` e `ggplot` tem sido satisfatórios para nós
- Por exemplo, é fácil verificar se existe alguma relação entre a velocidade e a distância que os carros presentes no dataset `cars` levaram para parar

Visualização dos Dados

```
> ggplot(cars, aes(x = speed, y = dist)) +
+   geom_point() +
+   labs(x = "Velocidade (mph)", y = "Distância (pés)")
```

Visualização dos Dados



Visualização dos Dados

- Motivação: conjunto de dados `iris`
- Conjunto de dados muito conhecido
- Carregue-o na memória do R utilizando o comando `data(iris)`
- 150 observações de 3 espécies de plantas
- 4 medidas de cada planta: comprimento e largura da pétala, comprimento e largura da sépala

Visualização dos Dados

Iris setosa

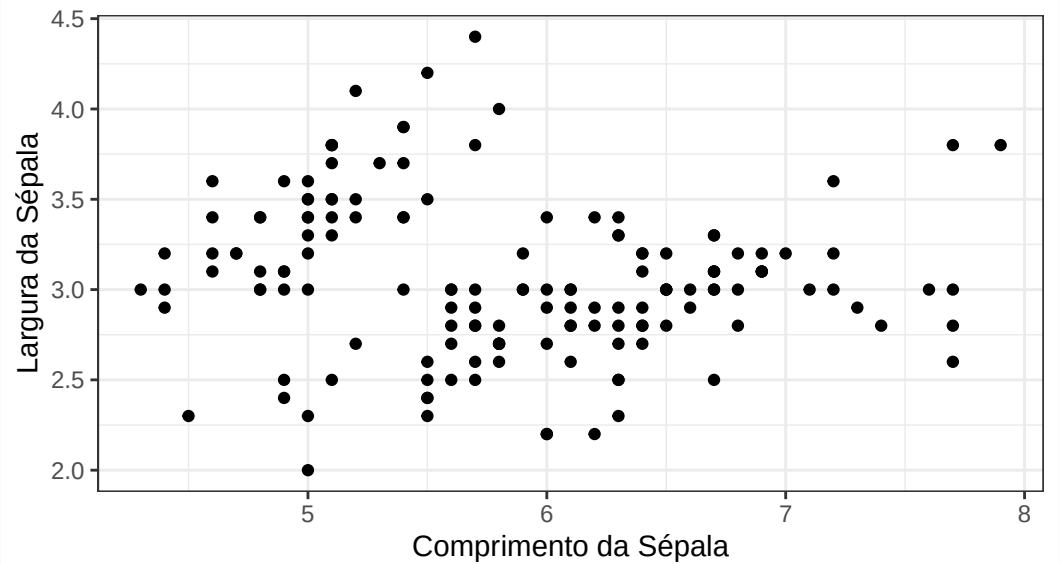
Iris versicolor

Iris virginica

Visualização dos Dados

```
> ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
+   geom_point() +
+   labs(x = "Comprimento da Sépala", y = "Largura da Sépala")
```

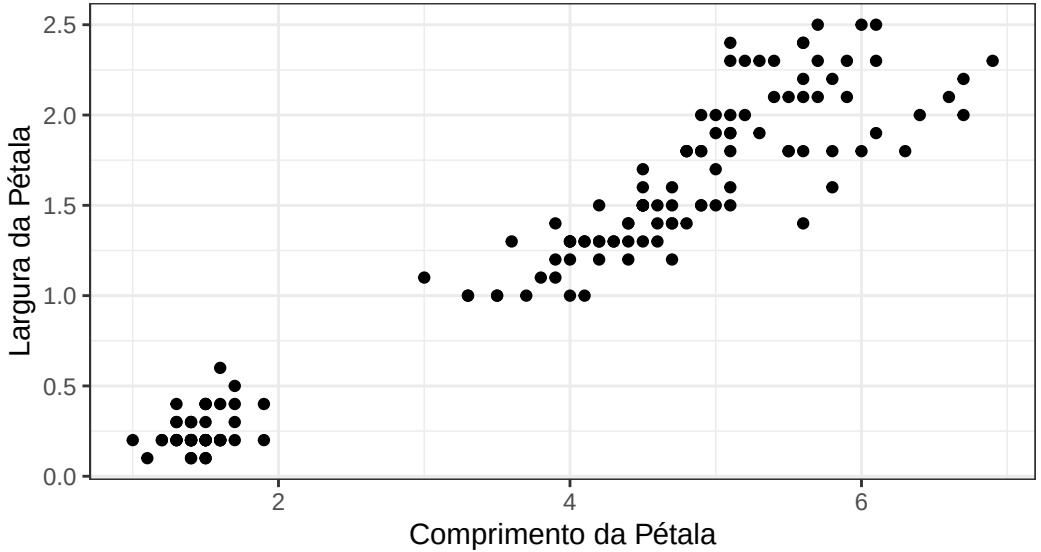
Visualização dos Dados



Visualização dos Dados

```
> ggplot(iris, aes(x = Petal.Length, y = Petal.Width)) +
+   geom_point() +
+   labs(x = "Comprimento da Pétala", y = "Largura da Pétala")
```

Visualização dos Dados



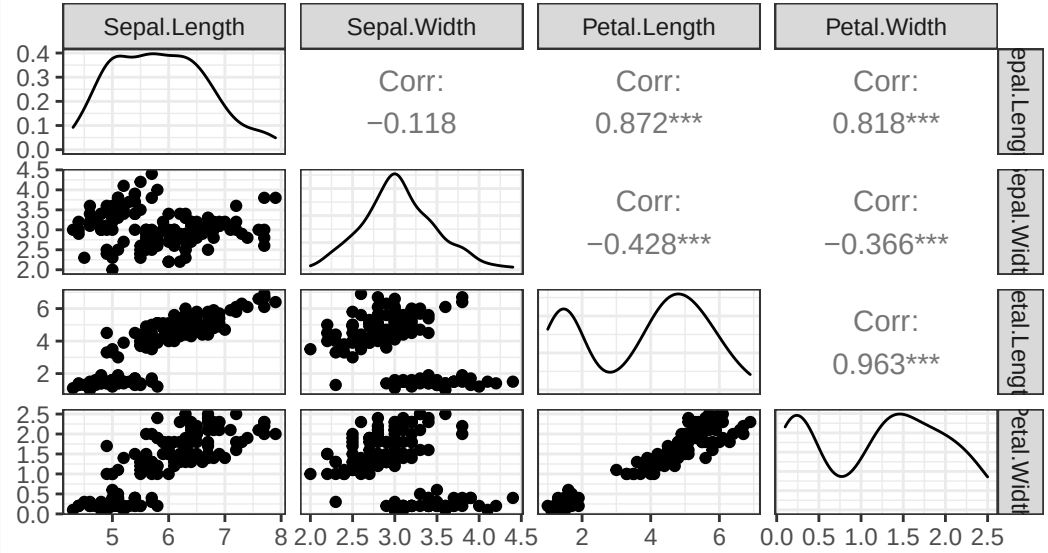
Visualização dos Dados

- Mas são quatro variáveis a serem plotados
- Se formos plotá-las duas a duas, devemos fazer $\binom{4}{2} = 6$ gráficos
- Há uma maneira mais direta de fazermos isto

Visualização dos Dados

```
> library(GGally)
> ggpairs(iris[, -5])
```

Visualização dos Dados



Visualização dos Dados

- As soluções que conhecemos funcionam razoavelmente bem para alguns conjuntos de dados
- Mas e se tivéssemos mais de 4 variáveis para analisar?
- E se fossem centenas de variáveis?

Visualização dos Dados

- Não é possível plotar todas as variáveis simultaneamente no mesmo gráfico
- Uma possível solução é escolher algumas variáveis para plotar
- Outra é projetar os dados em outras direções e visualizá-los a partir daí

Análise de Componentes Principais

Análise de Componentes Principais

- Diversas áreas do conhecimento apresentam problemas onde muitas variáveis são consideradas simultaneamente
- O conjunto de dados iris é um destes casos
- Mas aplicações assim aparecem em dados do mercado financeiro, em genética, em meteorologia

Análise de Componentes Principais

- Considerando o conjunto de dados iris, suponha que possamos representar as medidas de um dos espécimes i de plantas como um vetor de quatro posições:

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$$

ou

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{i4} \end{pmatrix}$$

- Em nossa notação, $\boldsymbol{x}_i'$ é o transposto do vetor $\boldsymbol{x}_i$

Análise de Componentes Principais

- De modo geral, um sujeito i com k medidas pode ser representado como um vetor no espaço $\mathbb{R}^k$:

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$$

ou

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}$$

Análise de Componentes Principais

- Assim, todos os sujeitos e todas as suas condições podem ser representados como

$$\boldsymbol{X}_{nk} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}$$

Análise de Componentes Principais

- Uma matriz é um vetor de vetores e também tem uma transposta

$$X'_{nk} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}' = (x_1, x_2, \cdots, x_n) = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}'$$

Análise de Componentes Principais

- Os produtos entre duas matrizes ou entre uma matriz e um vetor precisam ser compatíveis:

$$\boldsymbol{x}'\boldsymbol{y} = (x_1, x_2, \cdots, x_k) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix} = \sum_{i=1}^{k} x_i y_i$$

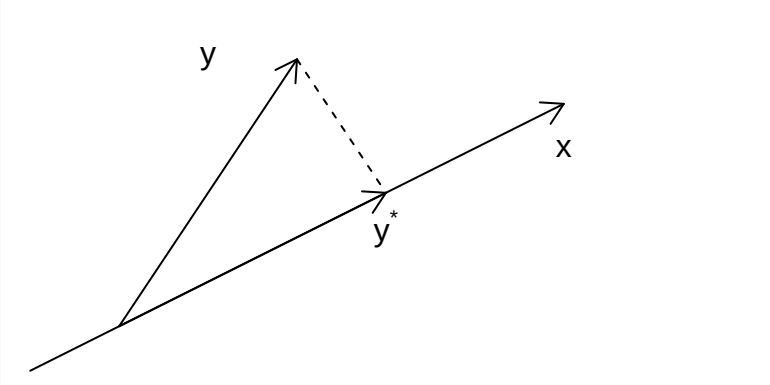
$$\boldsymbol{x}\boldsymbol{y}' = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} (y_1, y_2, \cdots, y_l) = \begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_l \\ x_2y_1 & x_2y_2 & \cdots & x_2y_l \\ \vdots & \vdots & \ddots & \vdots \\ x_ky_1 & x_ky_2 & \cdots & x_ky_l \end{pmatrix}$$

Análise de Componentes Principais

- Cada sujeito é um vetor de k valores, pois consideramos k condições
- Imagine cada sujeito como um ponto no espaço euclidiano $k-$dimensional
- Distância entre os vetores $\boldsymbol{x}$ e $\boldsymbol{y}$: $|\boldsymbol{x} - \boldsymbol{y}| = \sqrt{(\boldsymbol{x} - \boldsymbol{y})'(\boldsymbol{x} - \boldsymbol{y})}$
- Comprimento de um vetor $\boldsymbol{x}$: $|\boldsymbol{x}| = \sqrt{\boldsymbol{x}'\boldsymbol{x}}$
- Espaço linear gerado por $\boldsymbol{x}$: pontos de $\alpha\boldsymbol{x}$ para todo real α
- Ângulo entre os vetores $\boldsymbol{x}$ e $\boldsymbol{y}$: $\cos(\theta) = \frac{\boldsymbol{x}'\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}$ ($\boldsymbol{x}$ e $\boldsymbol{y}$ são ortogonais se $\boldsymbol{x}'\boldsymbol{y} = 0$)

Análise de Componentes Principais

- Projetar o vetor ***y*** em ***x*** (ou, mais precisamente, projetar ***y*** no espaço linear *R*(***x***) gerado por ***x***) é encontrar o ponto ***y**** que possui a menor distância para ***y***



Análise de Componentes Principais

- De acordo com a definição do slide anterior,

$$
\begin{aligned}
y^* &= |y| \cos(\theta) \frac{x}{|x|} \\
&= |y| \frac{x'y}{|x||y|} \frac{x}{|x|} \\
&= \left(\frac{x'y}{x'x} \right) x
\end{aligned}
$$

- Agora estamos preparados para entender a Análise de Componentes Principais

Análise de Componentes Principais

- A coleção de sujeitos e condições é uma matriz, onde cada linha é um sujeito e cada coluna é uma condição

$$X_{nk} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}$$

- Há n sujeitos e k condições
- Imagine que cada sujeito é um ponto num espaço $k-$dimensional
- Assim, temos n pontos (sujeitos) neste espaço

Análise de Componentes Principais

- Para a ACP funcionar satisfatoriamente, devemos realocar a origem das coordenadas para o centro dos dados
- Determinamos um conjunto de k direções ortogonais, ordenadas em termos da variabilidade dos dados
- É feita uma rotação dos dados originais
- Se procura a melhor maneira de visualizar os dados de acordo com a variabilidade das variáveis

Análise de Componentes Principais

- Assim, a ACP serve para encontrar as direções com maior variabilidade
- Os dados são projetados nessas direções, nos dando uma reconstrução em menor dimensão dos perfis dos sujeitos

Análise de Componentes Principais

- Como encontrar essas direções?

i) Encontre $v_1 = \arg\max_{|v_1|=1} \text{Var}(v_1' x)$
ii) Calcule $x_{-1} = x - v_1 v_1' x$
iii) Encontre $v_2 = \arg\max_{|v_2|=1} \text{Var}(v_2' x_{-1})$
iv) Calcule $x_{-1-2} = x_{-1} - v_2 v_2' x_{-1}$
v) Repita estes passos para encontrar outras direções

Análise de Componentes Principais

- Podemos mostrar que este método é equivalente a encontrar a Decomposição em Valores Singulares da matriz X
- Isto é, $X = U\Sigma V$, em que $V = (v_1, v_2, \cdots, v_k)$ são as direções de variabilidade
- É importante notar que $V = (v_1, v_2, \cdots, v_k)$ é uma matriz unitária, *i.e.*, $v_i' v_i = 1$ e $v_i' v_j = 0$, se $i \neq j$ (ou seja, as colunas de V são padronizadas e ortogonais)

Análise de Componentes Principais

- Pela Decomposição em Valores Singulares, $X = U\Sigma V$
- $V_{k\times k}$ é unitária e $V'V = I_{k\times k}$
- $\Sigma_{k\times k} = \begin{pmatrix} \sqrt{\lambda_1(n-1)} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2(n-1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_k(n-1)} \end{pmatrix}$,
  em que λ_i mede a escala (ou variância) das direções correspondentes
- $U_{n\times k}$ são os "novos dados" reescalados e projetados nas novas coordenadas, tal que $U'U = I_{k\times k}$

Análise de Componentes Principais

- Dado $X = U\Sigma V$, temos que

$$X'X = (V\Sigma'U')(U\Sigma V') = V\Sigma^2 V'$$

- Se X for centrado, defina

$$\text{Cov}(X) = \frac{1}{n-1} X'X,$$

  chamada matriz de variância-covariância
- Essa matriz mede a variabilidade e a covariabilidade dos dados

Análise de Componentes Principais

- Essa matriz é dada por

$$\text{Cov}(X) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k2} & \sigma_{k2} & \cdots & \sigma_k^2 \end{pmatrix}$$

$$= V'\Lambda V = V' \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{pmatrix} V,$$

em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ mede a variância dos dados em cada direção de V

Análise de Componentes Principais

- Para obtermos os dados nas novas coordenadas, usamos a transformação linear

$$X^* = XV$$

- Ou seja, para cada sujeito i sob a condição j, $x^*_{ij} = \sum_{l=1}^{k} x_{il} v_{jl}$

$$\text{Cov}(X^*) = \frac{1}{n-1}(X^* X^*) = \frac{1}{n} V' X' X V = V V' \mathbf{\Lambda} V' V \, \mathbf{\Lambda}$$

- $V = (v_1, v_2, \cdots, v_k)$ são os autovetores da matriz $\text{Cov}(X)$
- $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_k)$ são os autovalores desta matriz

Análise de Componentes Principais

- Depois de obtidas as direções principais da variabilidade dos dados, temos as seguintes opções:
  - Reduzir a dimensionalidade dos dados
  - Capturar padrões básicos na amostra
  - Limpar o ruído dos dados
  - Compressão de informação

Análise de Componentes Principais

- Uma boa redução dos dados ocorre quando a variabilidade “útil” dos dados é capturada dentre as componentes selecionadas
- Existem diversas maneiras de escolhermos as componentes importantes

i) Manter uma certa proporção (digamos 80%) da variância nos dados
ii) Manter as componentes cujo λ está acima de algum valor (a média, por exemplo)
iii) Criar um *cutoff* a partir da área plana do gráfico
iv) Testar a significância de uma direção
v) Métodos de reamostragem para atingir a estabilidade das direções

Análise de Componentes Principais

i) Considere a proporção da variabilidade explicada e mantenha tantas direções quanto necessárias para explicar uma certa proporção

- Esta é boa maneira de reduzir a dimensionalidade ocorre quando a variabilidade "útil" dos dados é capturada entre as componentes selecionadas
- Seja λ_i a variabilidade da no $i-$ésimo componente $\boldsymbol{v}_i$ e que os $\boldsymbol{v}_i$ são ortogonais, temos

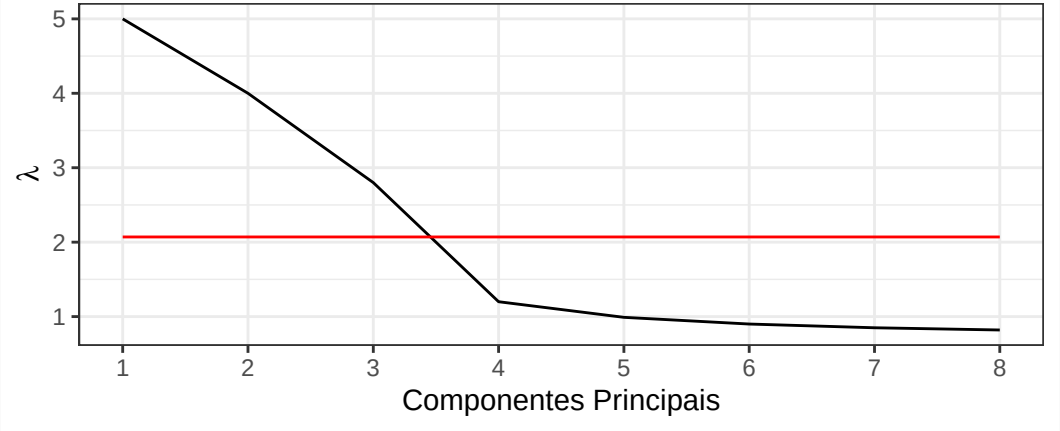
$$\lambda_1 + \lambda_2 + \cdots + \lambda_k = \text{variância total}$$

- Selecione direções $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r$ tais que

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{k} \lambda_i} \geq 0{,}80$$

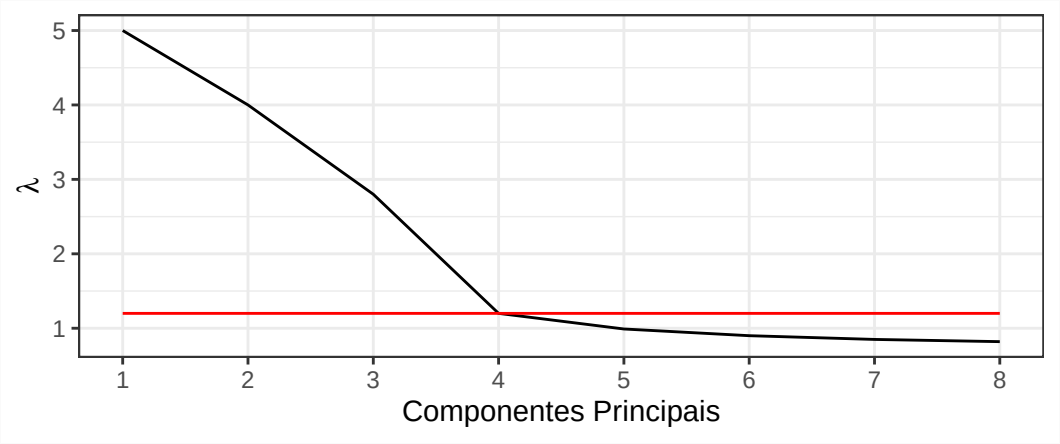
Análise de Componentes Principais

ii) Manter as componentes cujo λ está acima de algum valor (a média, por exemplo)



Análise de Componentes Principais

3. Criar um *cutoff* a partir da área plana do gráfico



Análise de Componentes Principais

iv) Para testar as componentes importantes, é possível realizar uma sequência de testes que determina quantos dos r últimos autovalores são estatisticamente iguais, desde que os dados sejam normais
Este teste é baseado numa estatística χ^2, desde que n seja grande o suficiente

$$H_0 : \lambda_{k-r+1} = \lambda_{k-r+2} = \cdots = \lambda_k$$
$$H_1 : \text{algum } \lambda_i \text{ listado acima é diferente dos demais}$$

Análise de Componentes Principais

Sob H_0,

$$\left(n - \frac{2k-11}{6}\right)\left(r\log(\overline{\lambda}) - \sum_{i=k-r+1}^{k}\log(\lambda_i)\right) \sim \chi^2\left(\frac{1}{2}(r-1)(r+2)\right)$$

Análise de Componentes Principais

v) Alternativamente, é possível realizar reamostragens ou permutações e checar o quão persistentes são as componentes principais

Reamostragem

- Bootstrap em um subconjunto de sujeitos
- Deletar aleatoriamente alguns sujeitos

Permutação

- Permutar as colunas da matriz
- Permutar os sujeitos dentro das colunas
- Permutar a tabela inteira

Análise de Componentes Principais

- Selecionar componentes principais leva, inevitavelmente, à perda de informação
- Usando o método i), uma porcentagem fixa da variabilidade é perdida
- Usando os métodos ii), iii) e iv), uma parte desconhecida da variabilidade é perdida
- Usando os métodos i) e ii), variabilidade informativa pode ser perdida
- Mesmo que a variabilidade de cada direção seja pequena, a soma de todas elas pode ser grande
- Uma perda de 2% não gera preocupações; uma perda de 20% deve ser tratada com cuidado

Análise de Componentes Principais

- Esta característica da PCA permite utilizá-la como método de seleção de variáveis
- Podemos manter apenas $k < p$ componentes principais e fazer nossas análises a partir deles

Análise de Componentes Principais

- É possível retornar os dados transformados para recuperar os dados originais, desde que nada seja jogado fora
- Para transformar os dados de volta para as coordenadas originais, usamos

$$X^0 = (X, 0)V'$$

- O número de componentes principais não significa nada além de "padrões básicos"
- A ACP não se importa com a ordem natural dos dados

Exemplo

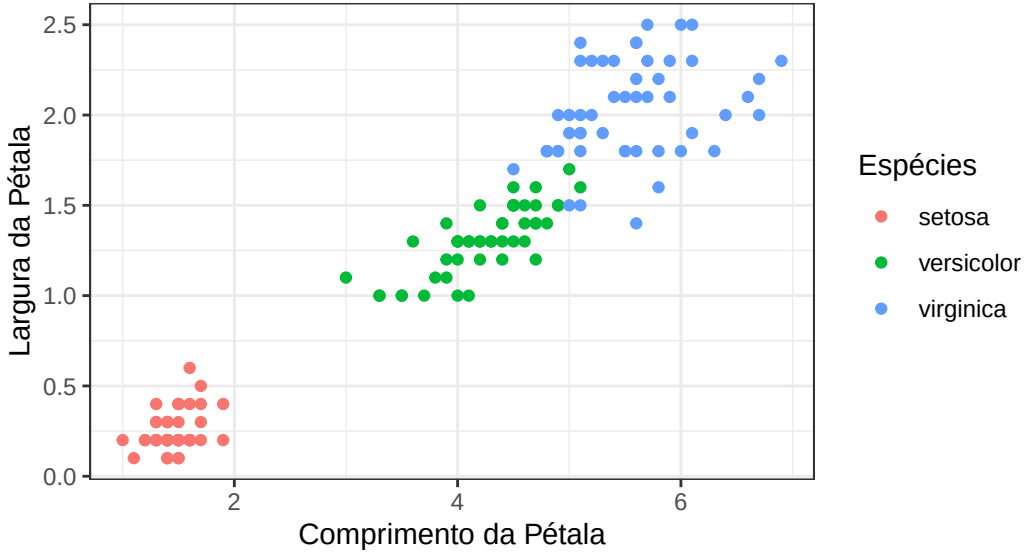
Exemplo - Iris

```
> iris.pca <- prcomp(iris[, -5], center = TRUE,
+   scale. = TRUE)
> names(iris.pca)

## [1] "sdev"     "rotation" "center"   "scale"    "x"

> head(iris.pca$x)

##            PC1        PC2         PC3          PC4
## [1,] -2.257141 -0.4784238  0.12727962  0.024087508
## [2,] -2.074013  0.6718827  0.23382552  0.102662845
## [3,] -2.356335  0.3407664 -0.04405390  0.028282305
## [4,] -2.291707  0.5953999 -0.09098530 -0.065735340
## [5,] -2.381863 -0.6446757 -0.01568565 -0.035802870
## [6,] -2.068701 -1.4842053 -0.02687825  0.006586116
```

Exemplo - Iris

```
> ggplot(iris, aes(x = Petal.Length, y = Petal.Width)) +
+   geom_point(aes(colour = Species)) +
+   labs(x = "Comprimento da Pétala", y = "Largura da Pétala",
+   colour = "Espécies")
```

Exemplo - Iris

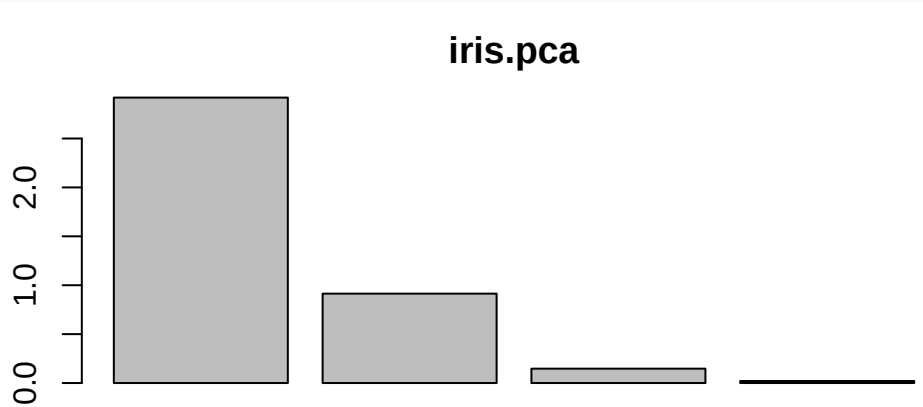


Exemplo - Iris

```
> summary(iris.pca)

## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7084 0.9560 0.38309 0.14393
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518
## Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

Exemplo - Iris

```
> plot(iris.pca)
```



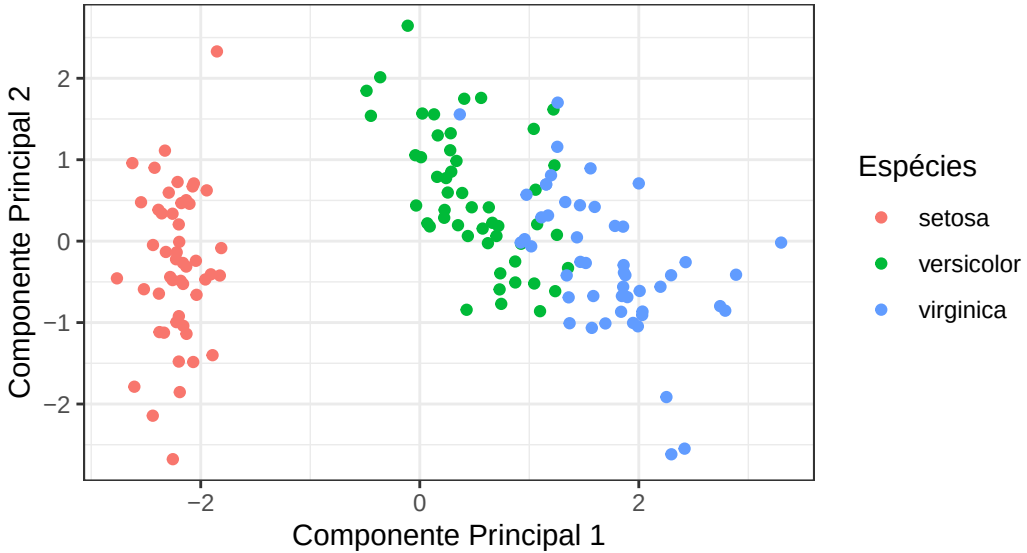
Exemplo - Iris

```
> iris.transformado <- data.frame(iris.pca$x,
+   iris$Species)
> head(iris.transformado)

##         PC1        PC2         PC3          PC4
## 1 -2.257141 -0.4784238  0.12727962  0.024087508
## 2 -2.074013  0.6718827  0.23382552  0.102662845
## 3 -2.356335  0.3407664 -0.04405390  0.028282305
## 4 -2.291707  0.5953999 -0.09098530 -0.065735340
## 5 -2.381863 -0.6446757 -0.01568565 -0.035802870
## 6 -2.068701 -1.4842053 -0.02687825  0.006586116
##   iris.Species
## 1       setosa
## 2       setosa
## 3       setosa
```

Exemplo - Iris

```
> ggplot(iris.transformado, aes(x = PC1, y = PC2)) +
+   geom_point(aes(colour = iris.Species)) +
+   labs(x = "Componente Principal 1", y = "Componente Principal 2",
+   colour = "Espécies")
```

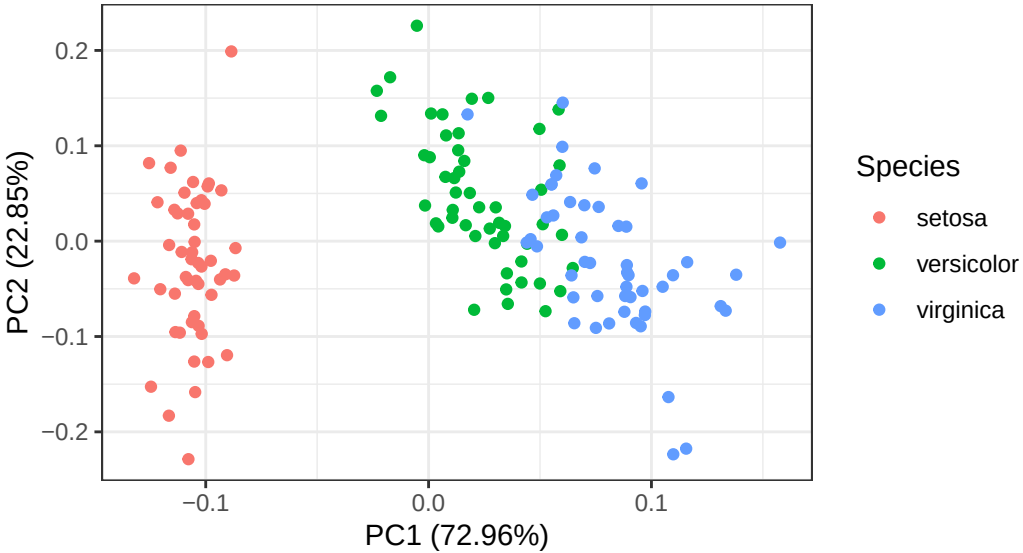
Exemplo - Iris



Análise de Componentes Principais

```
> library(ggfortify)
> autoplot(iris.pca, data = iris, colour = "Species")
```

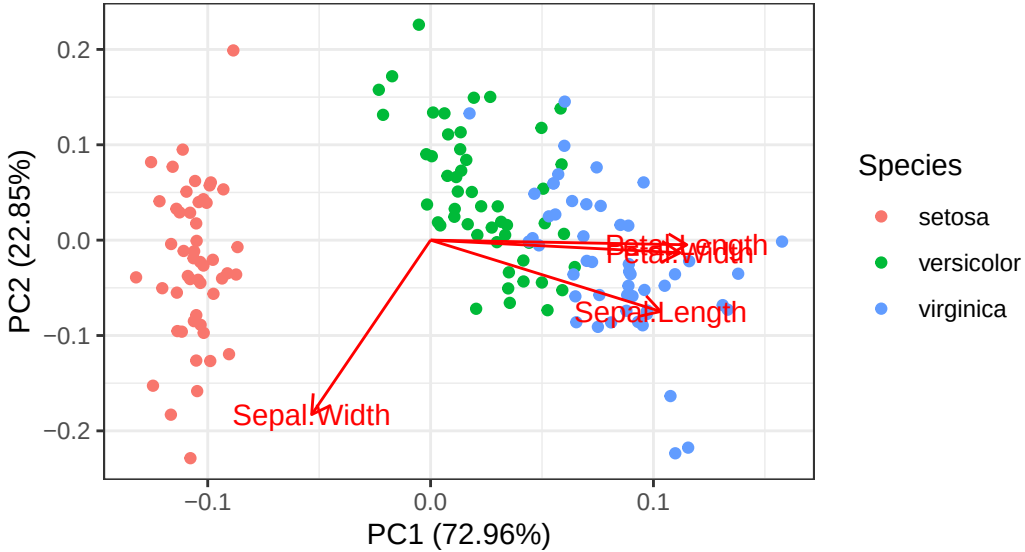
Análise de Componentes Principais



Análise de Componentes Principais

```
> autoplot(iris.pca, data = iris, colour = "Species",
+   loadings = TRUE, loadings.label = TRUE)
```

Análise de Componentes Principais



Aplicação

Aplicação - Heptatlo

- Heptatlo é uma competição de atletismo composta por sete provas
- É disputado apenas por mulheres; homens disputam o decatlo
- São dois dias de competição e a vencedora é determinada por uma pontuação específica

Aplicação - Heptatlo

- As provas do heptatlo são
  1. 100 metros com barreiras
  2. Salto em altura
  3. Arremesso de peso
  4. 200 metros rasos
  5. Salto em distância
  6. Lançamento de dardo
  7. 800 metros rasos

Aplicação - Heptatlo

- Vamos analisar os resultados do heptatlo da Olimpíada de 1988, em Seul
- A Análise de Componentes Principais será realizada em cima dos resultados de 25 melhores colocadas na prova
- Queremos procurar padrões e relações interessantes nestes dados

Aplicação - Heptatlo

```
> library(dplyr)
> library(ggplot2)
> library(GGally)
> library(corrplot)
>
> heptatlo <- read.csv(file = "heptatlo.csv")
>
> names(heptatlo)

## [1] "Nome"            "Barreiras100m"   "SaltoAltura"
## [4] "ArremessoPeso"   "Corrida200m"     "SaltoDistancia"
## [7] "LancamentoDardo" "Corrida800m"     "Pontos"
```

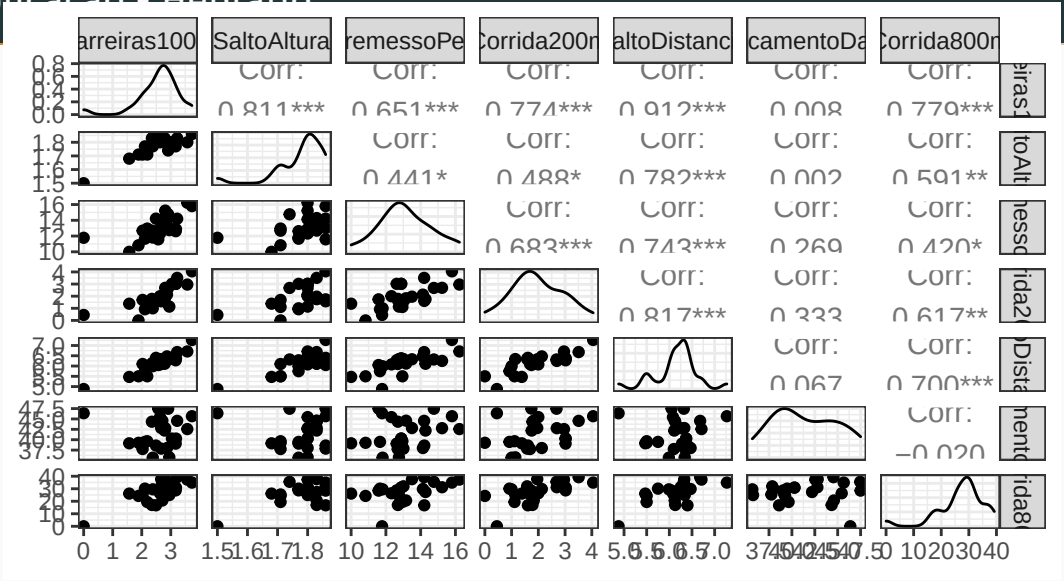
Aplicação - Heptatlo

```
> transformacao <- function(x){
+   return(max(x) - x)
+ }
>
> heptatlo <- heptatlo %>%
+   mutate(Barreiras100m = transformacao(Barreiras100m)) %>%
+   mutate(Corrida200m   = transformacao(Corrida200m)) %>%
+   mutate(Corrida800m   = transformacao(Corrida800m))
>
> heptatlo.novo <- heptatlo %>%
+   select(-c(Nome, Pontos))
```

Aplicação - Heptatlo

```
> ggpairs(heptatlo.novo)
```

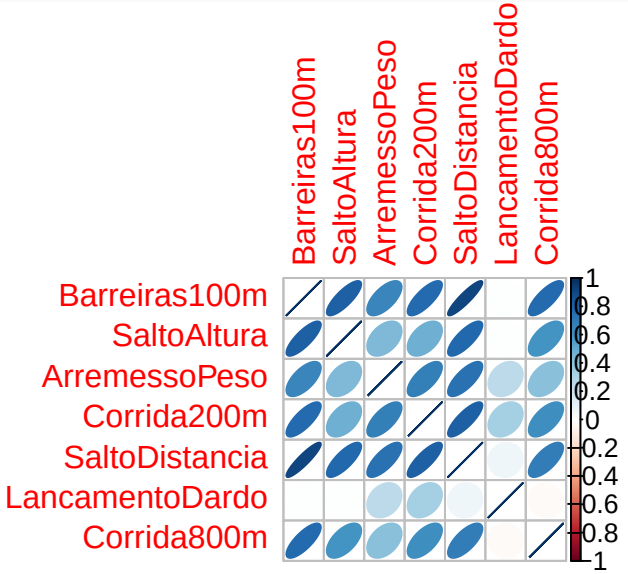
Aplicação - Heptatlo



Aplicação - Heptatlo

```
> corrplot(cor(heptatlo.novo), method = "ellipse")
```

Aplicação - Heptatlo

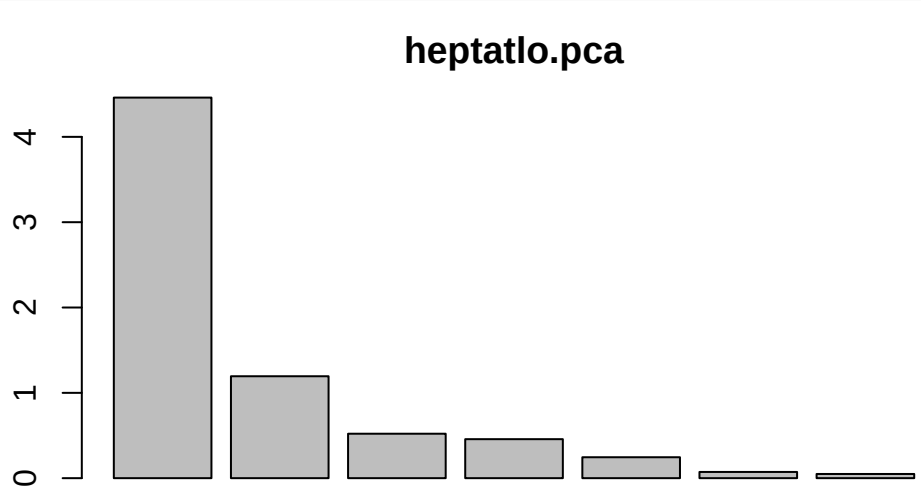


Aplicação - Heptatlo

```
> heptatlo.pca <- prcomp(heptatlo.novo,
+   center = TRUE, scale. = TRUE)
> summary(heptatlo.pca)

## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     2.1119 1.0928 0.72181 0.67614
## Proportion of Variance 0.6372 0.1706 0.07443 0.06531
## Cumulative Proportion  0.6372 0.8078 0.88223 0.94754
##                           PC5     PC6    PC7
## Standard deviation     0.49524 0.27010 0.2214
## Proportion of Variance 0.03504 0.01042 0.0070
## Cumulative Proportion  0.98258 0.99300 1.0000
```

Aplicação - Heptatlo

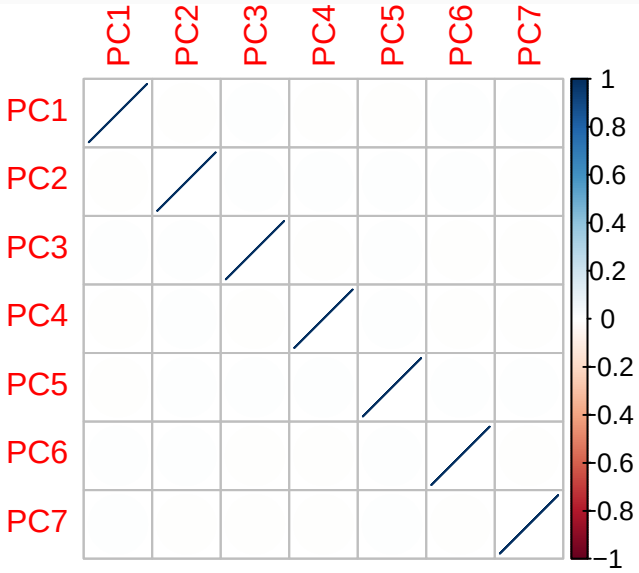
```
> plot(heptatlo.pca)
```



Aplicação - Heptatlo

```
> corrplot(cor(heptatlo.pca$x), method = "ellipse")
```

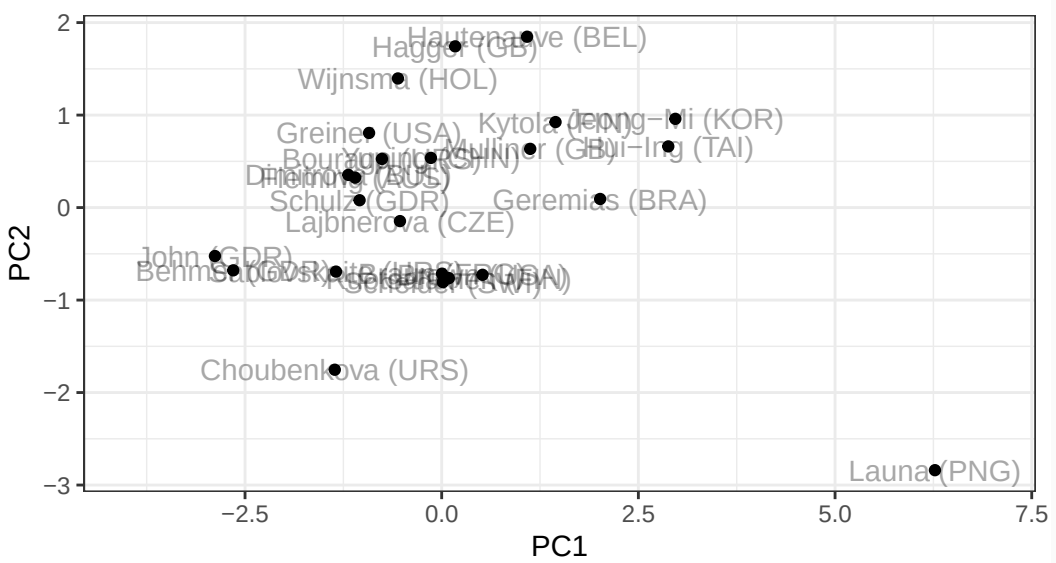
Aplicação - Heptatlo



Aplicação - Heptatlo

```
> heptatlo.transformado <- data.frame(heptatlo.pca$x,
+  Nome = heptatlo$Nome)
> ggplot(heptatlo.transformado, aes(x = PC1, y = PC2)) +
+   geom_point() +
+   geom_text(aes(label = Nome), alpha = 0.35) +
+   xlim(-4, 7)
```

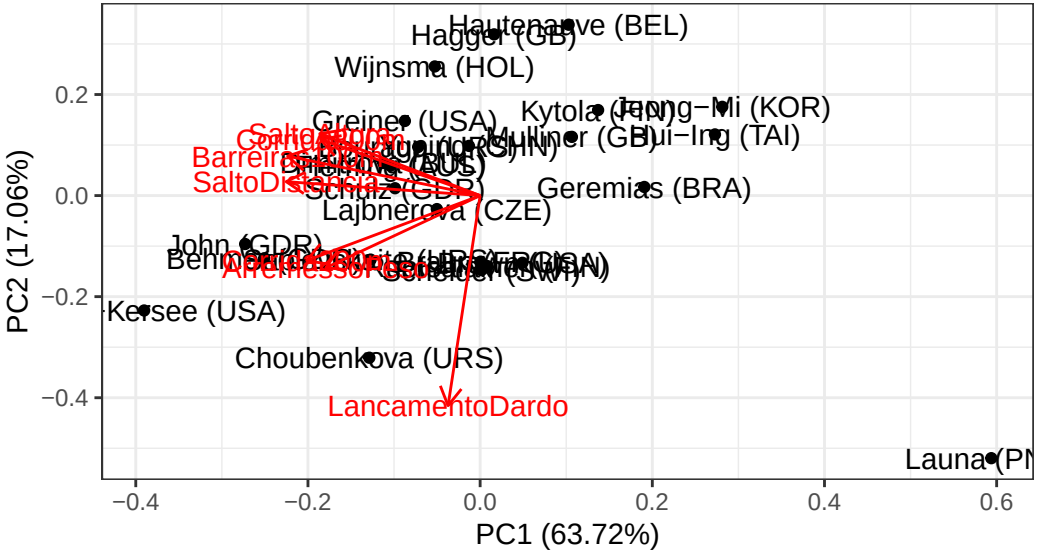
Aplicação - Heptatlo



Aplicação - Heptatlo

```
> rownames(heptatlo.novo) <- heptatlo$Nome
>
> autoplot(heptatlo.pca, data = heptatlo.novo,
+   label = TRUE, loadings = TRUE, loadings.label = TRUE)
```

Aplicação - Heptatlo



Exercícios

Exercícios

1. Refaça a PCA do conjunto de dados iris, mas escolha três combinações de componentes principais diferentes de PC1 e PC2 para construir seu gráfico. O que é possível perceber?
2. Aplique a função `autoplot` do pacote `ggfortify` na PCA do conjunto iris e interprete o resultado
3. Utilize o nome das espécies para identificar cada ponto no gráfico de dispersão das duas primeiras componentes principais

Exercícios

4. O arquivo `AlimentacaoReinoUnido.txt` mostra o consumo de diversos alimentos no Reino Unido em 1997. Importe este conjunto de dados para o R.
5. Faça a PCA deste conjunto de dados.
6. Quantas componentes principais são necessárias para que 95% da variância seja explicada?
7. Faça o gráfico de barras das contribuições das variâncias para confirmar sua resposta para o item anterior.

Exercícios

8. Faça um gráfico de dispersão com as duas primeiras componentes principais, identificando cada alimento por seu nome. O que é possível perceber?
9. Explique o que ocorre quando utilizamos o comando `autoplot` do pacote `ggfortify` para analisar a PCA.
10. Refaça a análise para o conjunto de dados presente no arquivo `AlimentacaoReinoUnido.txt`, mas utilize a transposta da matriz original de dados. O que mudou? O que é possível perceber com esta nova análise?

Análise de Componentes Principais

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte