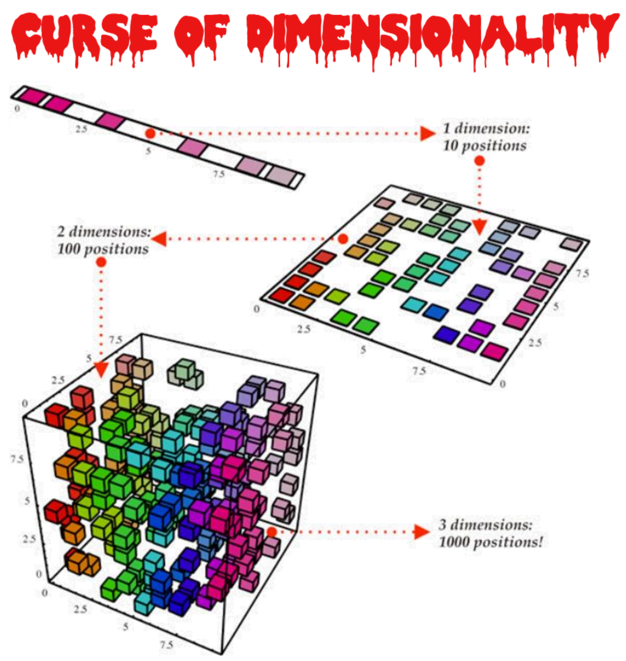
PREPARAÇÃO DOS DADOS: SELEÇÃO E EXTRAÇÃO DE CARACTERÍSTICAS

André Gustavo Adami
Daniel Luis Notari

INTRODUÇÃO

O desempenho do sistema depende do relacionamento entre o número de exemplos, número de características e a complexidade do algoritmo de aprendizagem

- O número de parâmetros do algoritmo de aprendizagem cresce de forma exponencial com o número de características ("Maldição da Dimensionalidade")
- Em um espaço de características com muitas dimensões, as amostras se tornam esparsas e pouco similares (muito distantes)



Conjunto **reduzido** de características **relevantes** também facilita a interpretação dos resultados do modelo, permite uma maior generalização e reduz o custo computacional (acelerando o tempo de treinamento)

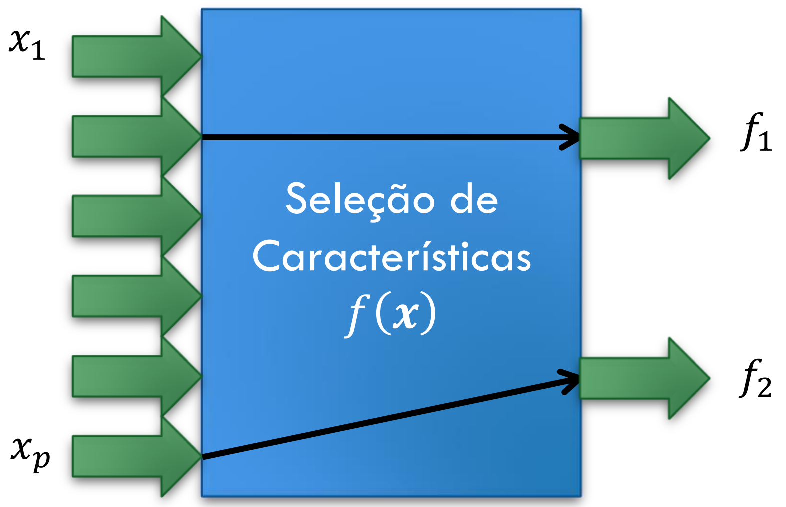
INTRODUÇÃO

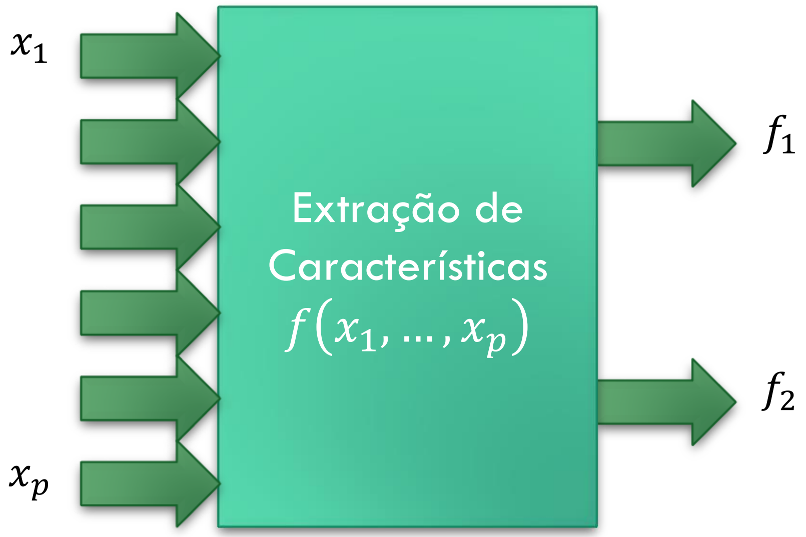
Um conjunto reduzido de características relevantes

- Facilita a interpretação dos resultados do modelo
- Permite uma maior generalização (sem ruídos ou detalhes irrelevantes)
- Reduz o custo computacional, resultando na aceleração do tempo de treinamento)
- Permite o uso de modelos mais simples que demandam menos poder computacional e tempos de predição mais rápidos

INTRODUÇÃO

A redução da dimensionalidade pode ser obtida por meio da remoção de informações irrelevantes/redundantes ou uma representação compacta e informativa dos dados originais

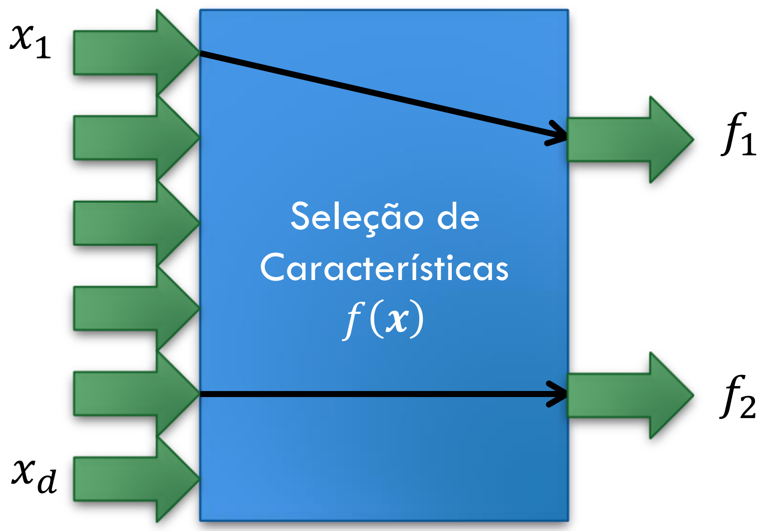




SELEÇÃO DE CARACTERÍSTICAS

Para um determinado conjunto de d variáveis, i.e., $\chi = \mathbb{R}^d$, o objetivo da seleção de características é produzir subconjunto de variáveis, i.e. $k \ll d$, para o projeto do algoritmo de aprendizagem

- Modelos mais simples
- Menor pegada de memória
- Treinamento mais rápido



Categorias

- Filtros
- Wrappers
- Embarcados/Intrínsecos (combina as categorias anteriores)

SELEÇÃO DE CARACTERÍSTICAS: WRAPPERS

Realiza uma busca de um subconjunto de variáveis no espaço de todos os possíveis subconjuntos de variáveis, avaliando cada subconjunto com base no desempenho de uma dado algoritmo de aprendizagem

Considerados algoritmos gulosos, podem ser computacionalmente custosos e frequentemente impraticáveis no caso de buscas exaustivas

Vantagens

- Detectam interações entre variáveis
- Eles encontram o subconjunto ótimo de variáveis para um determinado algoritmo de aprendizado de máquina

SELEÇÃO DE CARACTERÍSTICAS: WRAPPERS

Processo

1. Seleciona um subconjunto de variáveis
2. Treina um algoritmo de aprendizado
3. Avalia o desempenho do modelo
4. Vá para o Passo 1

O subconjunto pode ser selecionado

- Começando sem nenhuma variável e adicionando uma a uma (*Forward Feature Selection*)
- Começando todas as variáveis e removendo uma a uma (*Backward Feature Selection*)
- Tenta todas as possíveis combinações (*Exhaustive Feature Selection*)

Critério de Parada

- Desempenho do modelo diminui
- Um determinado número de variáveis é obtido

SELEÇÃO DE CARACTERÍSTICAS: FILTROS

Seleciona variáveis de um conjunto de dados sem o uso de um algoritmo de aprendizagem, mas avaliando o poder preditivo de cada variável

- Correlacionar a variável ao alvo (o que queremos reconhecer/predizer)
- Determinar o valor preditivo (ou informação) da variável

Vantagens

- Seleciona características/variáveis que podem ser usadas em qualquer algoritmo de aprendizagem (não depende do algoritmo de aprendizagem)
- Geralmente não demandam grande capacidade de processamento

Os métodos geralmente não produzem o melhor subconjunto de variáveis, mas são métodos básicos (e essenciais)

SELEÇÃO DE CARACTERÍSTICAS: FILTROS

Métodos

- **Básicos**: remover constantes ou quase constantes (variação abaixo de um limiar ou que ocorrem na maioria das amostras) e variáveis duplicadas

- **Correlação**: remover as variáveis que dependem uma da outra. Se podemos estimar uma variável a partir de outra, quer dizer que a predita não traz informação adicional alguma sobre os dados (informação redundante)

- **Ranqueamento** e **Estatísticos**: as variáveis são avaliadas com base em quão importantes elas são para discriminar as classes (mútua informação, distância probabilística, dependência probabilística, distância entre classes)

SELEÇÃO DE CARACTERÍSTICAS: CORRELAÇÃO

A correlação é a medida padronizada da relação entre duas variáveis e indica a força e direção do relacionamento **linear** entre duas variáveis aleatórias

Uma alta correlação entre variáveis é uma propriedade muito útil pois podemos predizer uma variável a partir da outra (redundância)

As medidas de correlação mais comuns são

- Coeficiente de correlação de Pearson (correlação linear)
- Coeficiente de correlação de Spearman (correlação não linear)

SELEÇÃO DE CARACTERÍSTICAS: CORRELAÇÃO

Coeficiente de correlação de Pearson é o mais utilizado

- Assume que ambas as variáveis possuem uma distribuição normal
- As variáveis possuem entre elas um relacionamento definido por uma linha reta
- Dados são igualmente distribuídos ao redor da linha de regressão

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

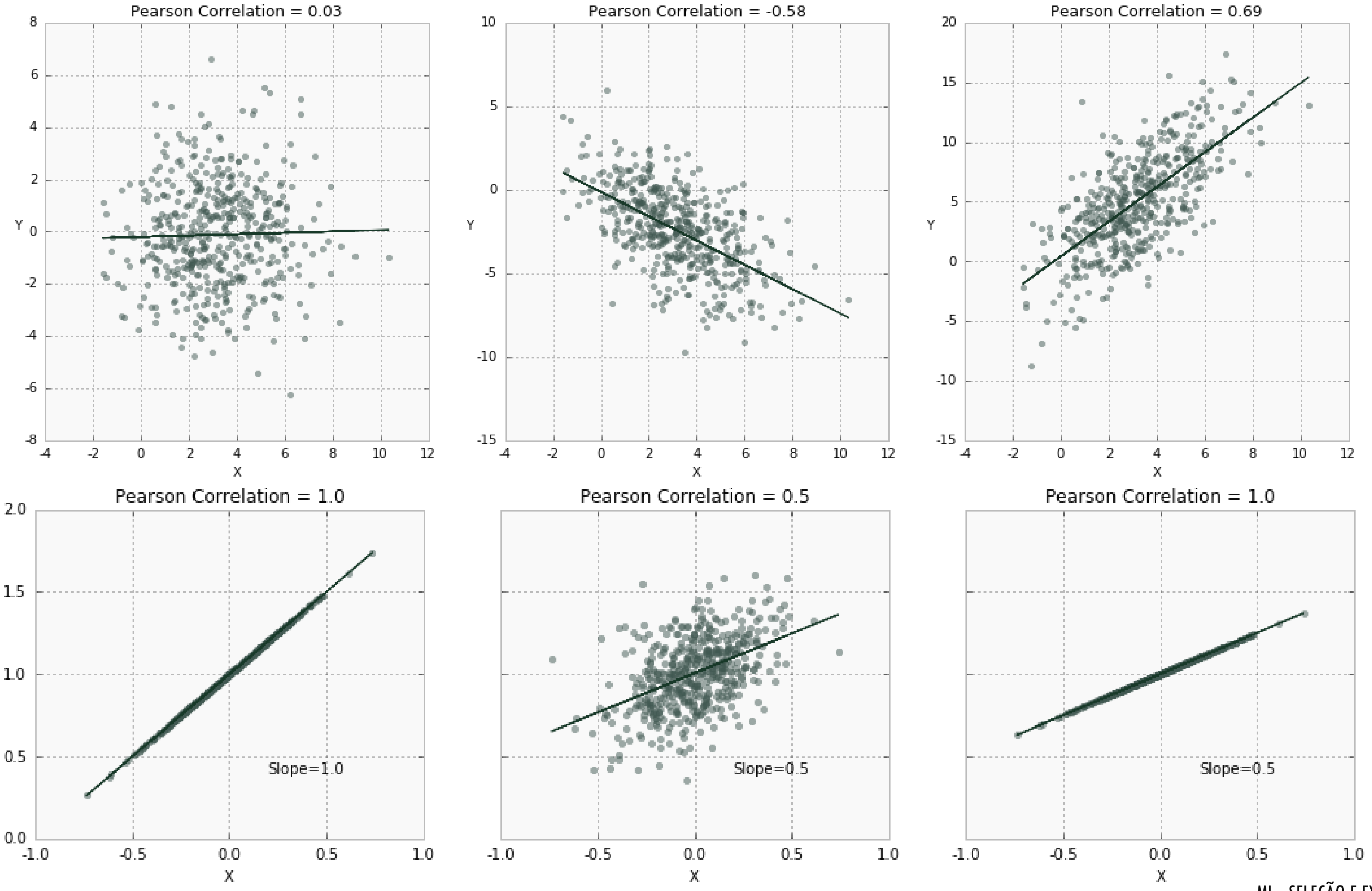
ou

$$r_{xy} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

A força do relacionamento entre as duas variáveis pode variar entre -1 e 1

- 1 significa correlação positiva: o valor de uma cresce quando o valor da outra cresce
- -1 significa correlação negativa: o valor de uma diminui quando o valor da outra cresce
- 0 significa nenhuma correlação linear entre as variáveis

SELEÇÃO DE CARACTERÍSTICAS: CORRELAÇÃO



SELEÇÃO DE CARACTERÍSTICAS: CORRELAÇÃO - PRÁTICA

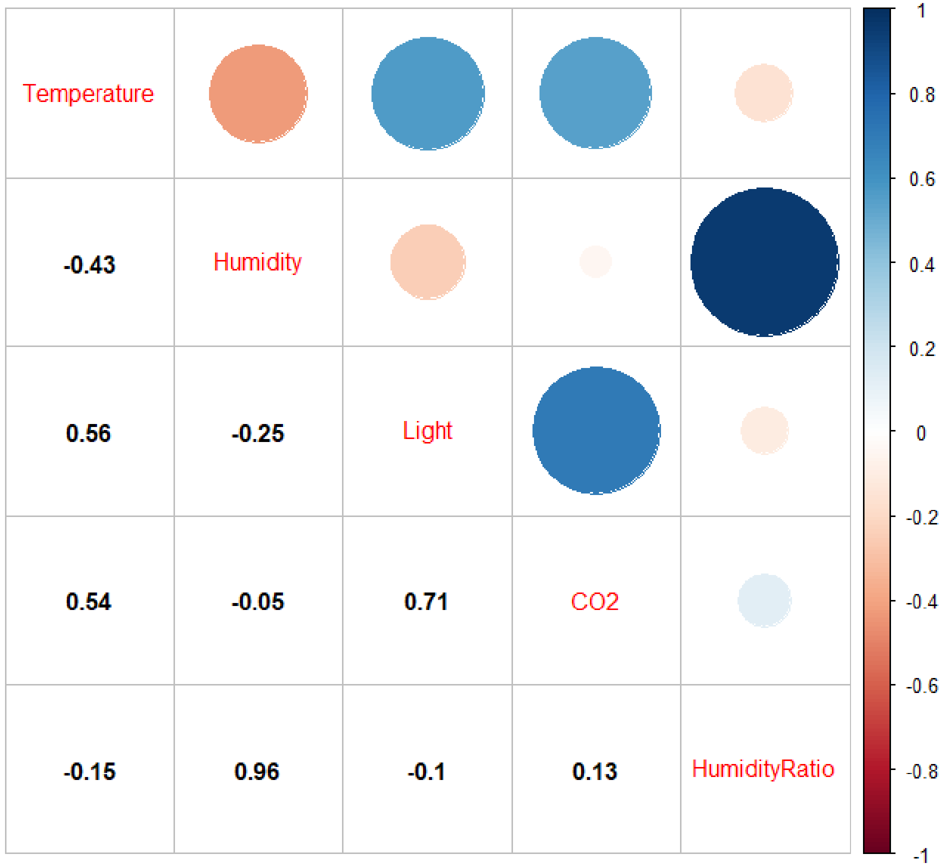
Para esta prática, assume-se que o conjunto de dados (prep-dados.txt) possua a seguinte estrutura (dados faltantes removidos, outliers removidos, rótulo transformado em *factor* e variáveis normalizadas)

```
> dados = read.csv("prep-dados.txt",header=T, na.strings="?")

> summary(dados)
  Temperature          Humidity             Light              CO2            HumidityRatio           Occupancy
 Min.   :-1.5262   Min.   :-1.63007   Min.   :-0.5014   Min.   :-0.7188   Min.   :-1.48233   Nao.Ocupado:5991
 1st Qu.:-0.9030   1st Qu.:-1.00585   1st Qu.:-0.5014   1st Qu.:-0.5299   1st Qu.:-0.94430   Ocupado    : 884
 Median :-0.1864   Median : 0.08487   Median :-0.5014   Median :-0.4439   Median : 0.05797
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.: 0.6548   3rd Qu.: 0.64279   3rd Qu.:-0.3222   3rd Qu.:-0.2045   3rd Qu.: 0.77962
 Max.   : 2.7839   Max.   : 2.24118   Max.   : 3.2344   Max.   : 3.5175   Max.   : 2.79306
```

SELEÇÃO DE CARACTERÍSTICAS: CORRELAÇÃO - PRÁTICA

Cálculo da matriz de correlação (dependência entre as todas as variáveis)

```
library(corrplot)

dadosCorrelacao = cor(dados[,1:5])

corrplot.mixed(dadosCorrelacao,
               lower.col="black")
```



Seleção das colunas altamente correlacionadas

```
correlacaoAlta = findCorrelation(dadosCorrelacao, cutoff=0.95)
```

SELEÇÃO DE CARACTERÍSTICAS: CORRELAÇÃO - PRÁTICA

É importante verificar se o coeficiente de correlação é significativo

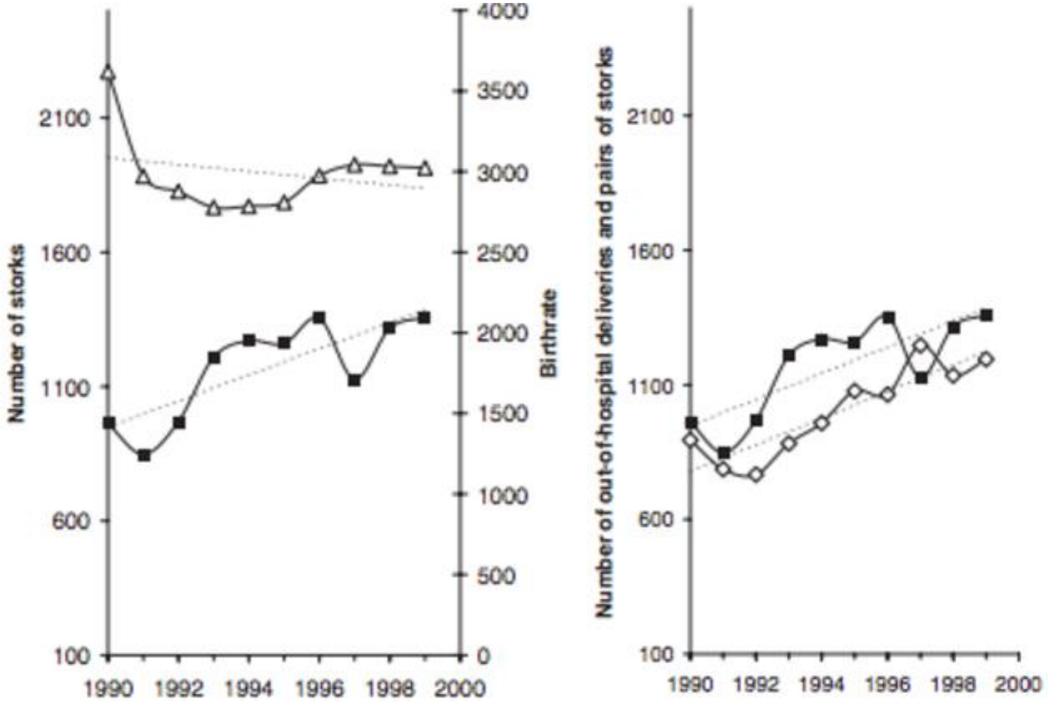
```
cor.test(dados$Humidity,dados$HumidityRatio)

    Pearson's product-moment correlation

data:  dadosNormalizadosZ$Humidity and dadosNormalizadosZ$HumidityRatio
t = 275.48, df = 6873, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9555688 0.9594971
sample estimates:
      cor
0.9575774
```

O resultado acima mostra que o *p-value* do teste é 2.2e-16, o que é menos do que o nível de significância α=0.05. Pode-se concluir *Humidity* e *HumidityRatio* são significantemente correlacionados com um coeficiente de correlação de 0.*9575774* e um *p-value* de 2.2e-16

SELEÇÃO DE CARACTERÍSTICAS: CORRELAÇÃO

Correlação não implica causalidade!!!!

Em 2004, os pesquisadores alemães Thomas Höfera, Hildegard Przyrembelb e Silvia Verlegerc mostraram pela correlação que **bebês eram trazidos por cegonhas**



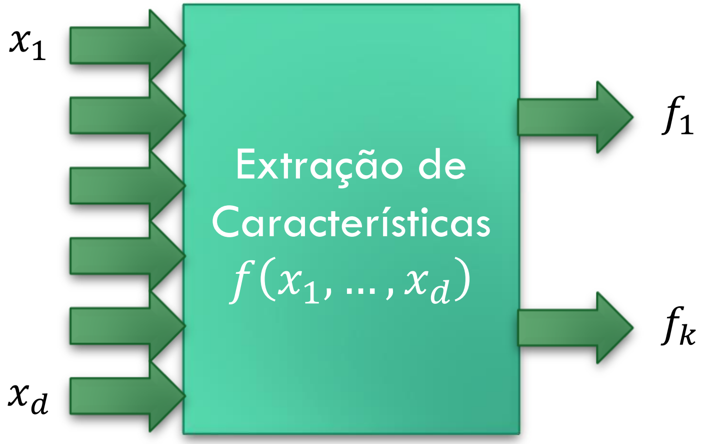
Höfer, T. , Przyrembel, H. and Verleger, S. (2004), New evidence for the Theory of the Stork. Paediatric and Perinatal Epidemiology, 18: 88-92. doi:10.1111/j.1365-3016.2003.00534.x

EXTRAÇÃO DE CARACTERÍSTICAS

Transformação linear ou não-linear das variáveis originais para obter um conjunto menor (projeção das características em um espaço dimensional menor)

- O efeito é o mesmo que o da seleção (redução da dimensionalidade), mas a extração realiza uma transformação em vez de somente selecionar um subconjunto

A transformação busca encontrar um novo conjunto de k dimensões que são uma combinação das d dimensões originais do conjunto de variáveis



EXTRAÇÃO DE CARACTERÍSTICAS

Transformação linear ou não-linear das variáveis originais para obter um conjunto menor (projeção das características em um espaço dimensional menor)

- O efeito é o mesmo que o da seleção (redução da dimensionalidade), mas a extração realiza uma transformação em vez de somente selecionar um subconjunto

A transformação busca encontrar um novo conjunto de k dimensões que são uma combinação das d dimensões originais do conjunto de variáveis

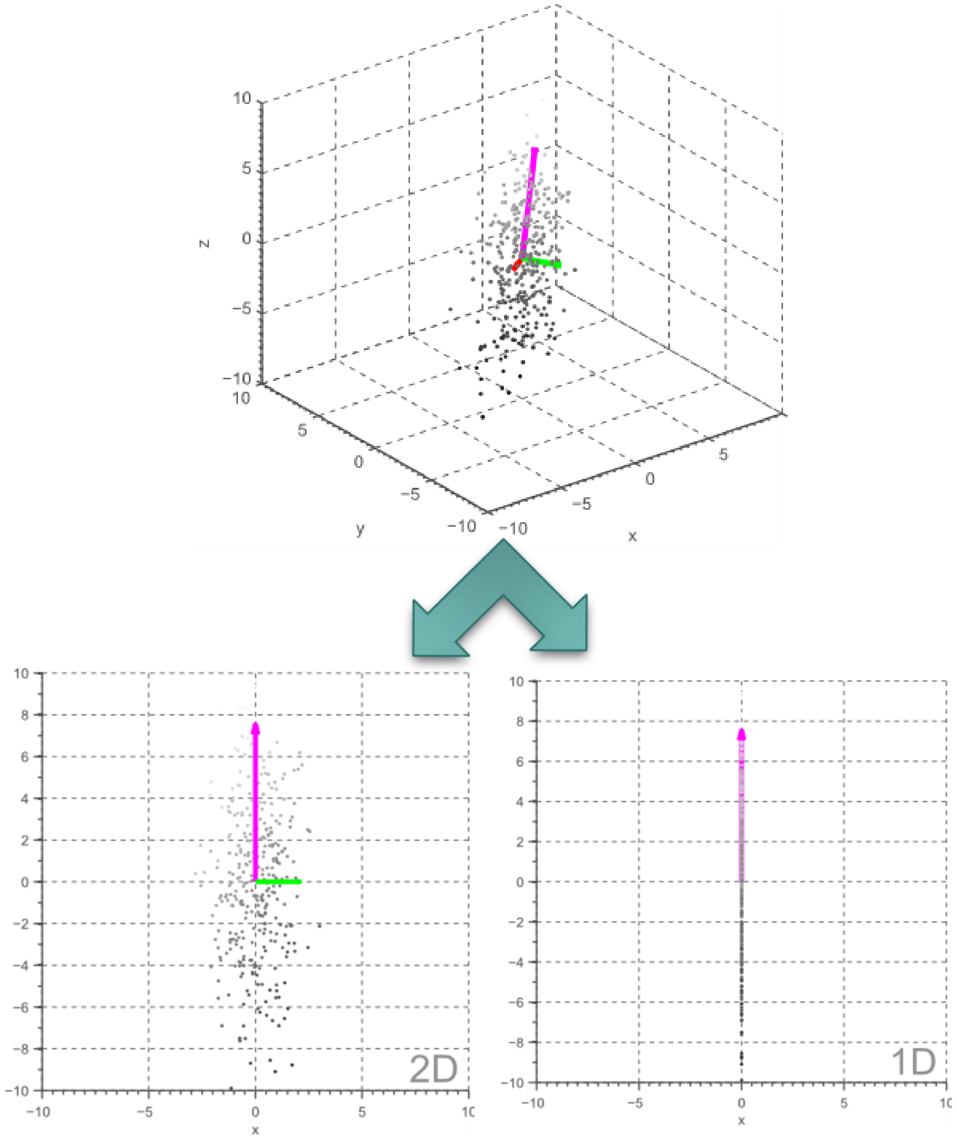
Um dos métodos mais comuns para transformação linear de um conjunto de características em um espaço menor de características é a Análise de Componentes Principais (*Principal Components Analysis* – PCA)

EXTRAÇÃO DE CARACTERÍSTICAS

O PCA deriva as novas variáveis (em ordem decrescente de importância) que são combinações lineares das variáveis originais ($Y = A^T X$) e são não correlacionadas ($corr(Y) = 0$)

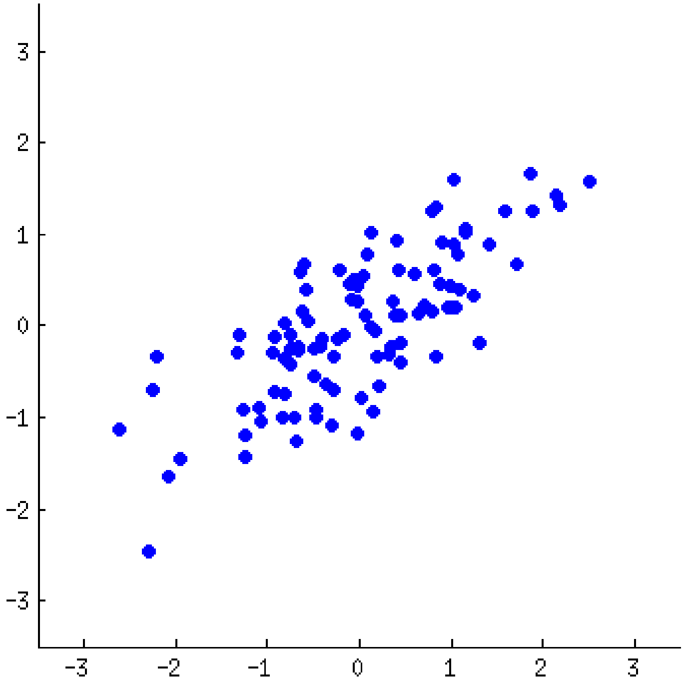
O PCA realiza a rotação dos eixos do sistema de coordenadas original para um novo conjunto de eixos ortogonais em termos da quantidade de variação que descreve os dados originais

É também chamada de Transformada Discreta de Karhunen-Loève (KLT) ou ainda Transformada Hotelling, em homenagem a Kari Karhunen, Michel Loève [1907-1979] e Harold Hotelling

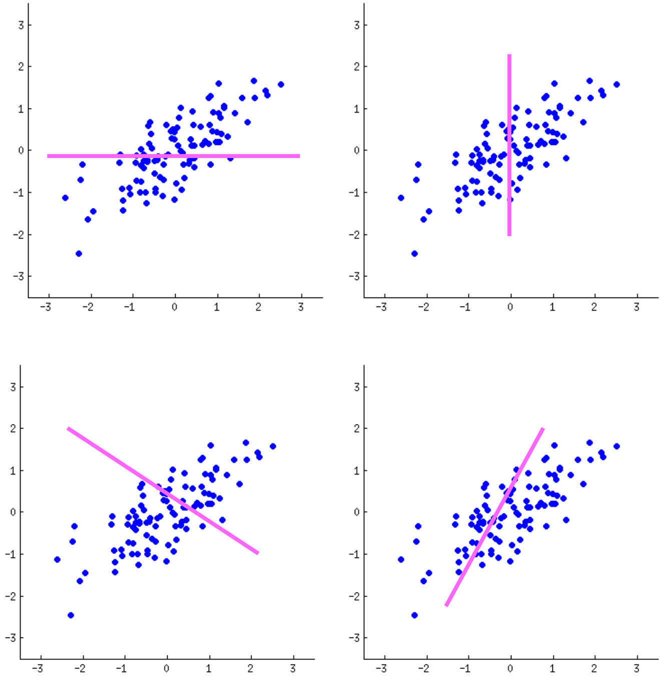


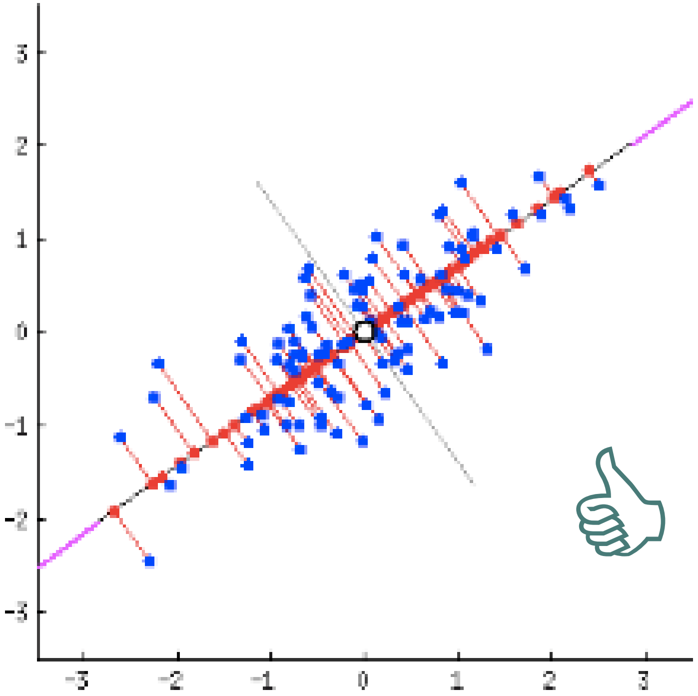
EXTRAÇÃO DE CARACTERÍSTICAS: PCA

Em um problema de regressão linear, buscamos posicionar uma reta com o objetivo de reduzir o erro da diferença dos dados em relação a esta reta



Qual é a melhor regressão?





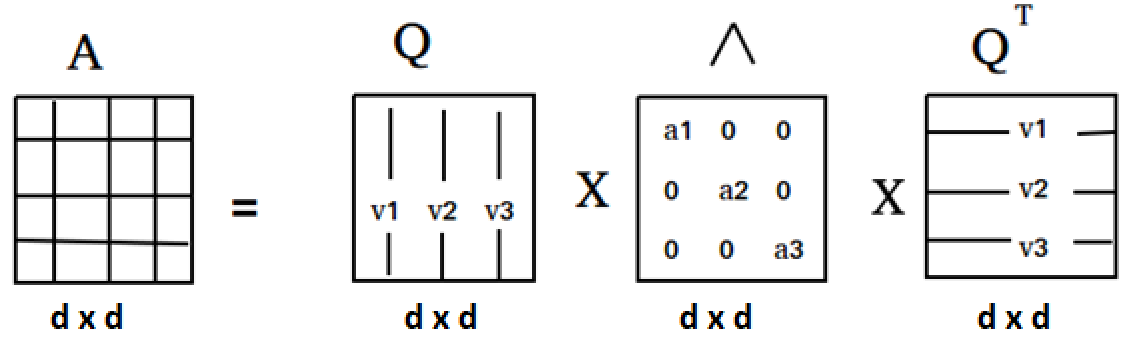
EXTRAÇÃO DE CARACTERÍSTICAS: PCA

A direção dos novos eixos devem capturar o máximo de variação dos dados (isto é, a direção onde a variância é máxima)

- O espalhamento dos dados é descrito pela matriz de Covariância Σ, matriz quadrada que descreve a variância dos dados e a covariância entre as variáveis

As direções e a sua respectiva magnitude da variabilidade dos dados podem ser definidas pelos auto-vetores (Q) e auto-valores (Λ), respectivamente

- As variâncias são os valores principais!
- Dado que $\chi = \mathbb{R}^d$, serão d auto-vetores e d auto-valores

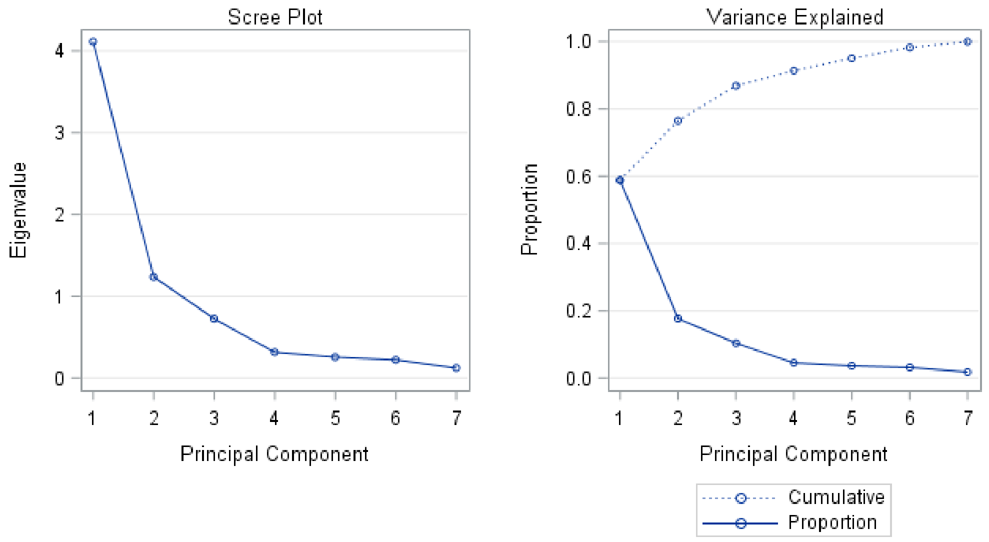


EXTRAÇÃO DE CARACTERÍSTICAS: PCA

Os componentes principais (auto-vetores) não são correlacionados (isto é, são independentes)

A magnitude dos auto-valores é utilizada para ordenar de forma decrescente os auto-vetores estimados a partir da matriz de covariância

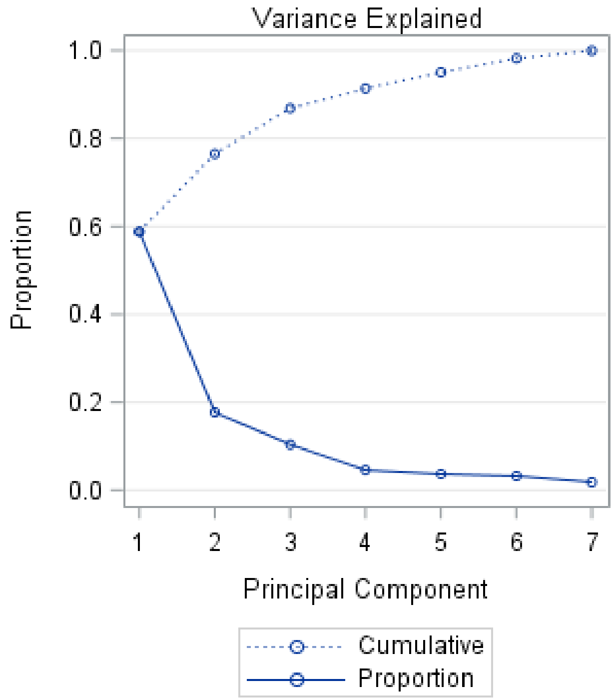
- O 1° auto-vetor (que possui o maior auto-valor), componente principal, explica a maior parte da variância
- O 2° auto-vetor (que possui o 2° maior auto-valor) explica a 2ª maior parte da variância
- O 3° auto-vetor (que possui o 3° maior auto-valor) explica a 3ª maior parte da variância
- …



EXTRAÇÃO DE CARACTERÍSTICAS: PCA

No caso da redução de dimensionalidade, o mapeamento das entradas em um espaço original de d dimensões é realizado para um novo espaço com dimensões k (onde $k \ll d$), com uma perda mínima de informações

Selecionam-se os auto-vetores com base na variância acumulada que explica os dados originais (95%) para produzir uma matriz de transformação



EXTRAÇÃO DE CARACTERÍSTICAS: PCA

Processo

1. Normalizar os dados: os componentes principais são dependentes das escalas utilizadas para medir as variáveis originais (mesmo se elas são medidas na mesma unidade – intervalos diferentes). Além disso, com o objetivo dos componentes principais terem média zero, deve-se remover a médias das variáveis
2. Estimar a matriz de covariância dos dados
3. Estimar os auto-valores e auto-vetores da matriz de covariância
4. Ordenar os auto-vetores com base nos auto-valores (variância explicada). A matriz de transformação/projeção A_{dxk} possui como colunas os auto-vetores

$$z = A^T x$$

Os auto-vetores devem ser armazenados para transformar futuros dados

EXTRAÇÃO DE CARACTERÍSTICAS: PCA

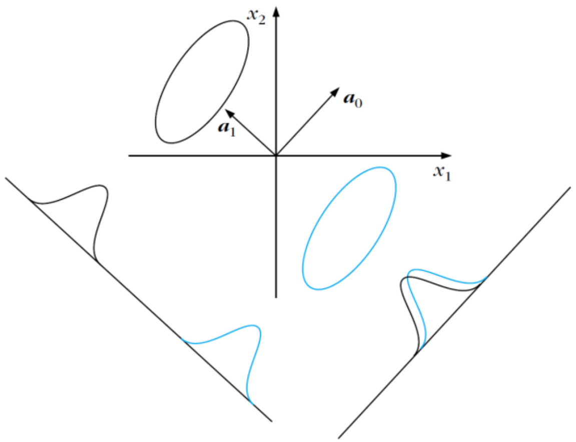
Desvantagens/Limitações

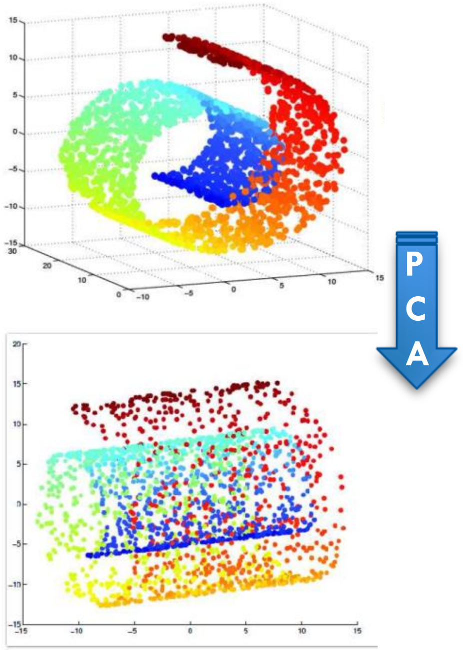
- Como o método realiza a rotação dos eixos do sistema de coordenadas original de forma não supervisionadas, o PCA não leva em conta a distribuição das classes em problemas de classificação

- Os dados projetados não são interpretáveis

- Assume que a distribuição do espaço original aproxima a normalidade (média e covariância não descreve algumas distribuições)

- Assume que as variáveis são linearmente correlacionadas

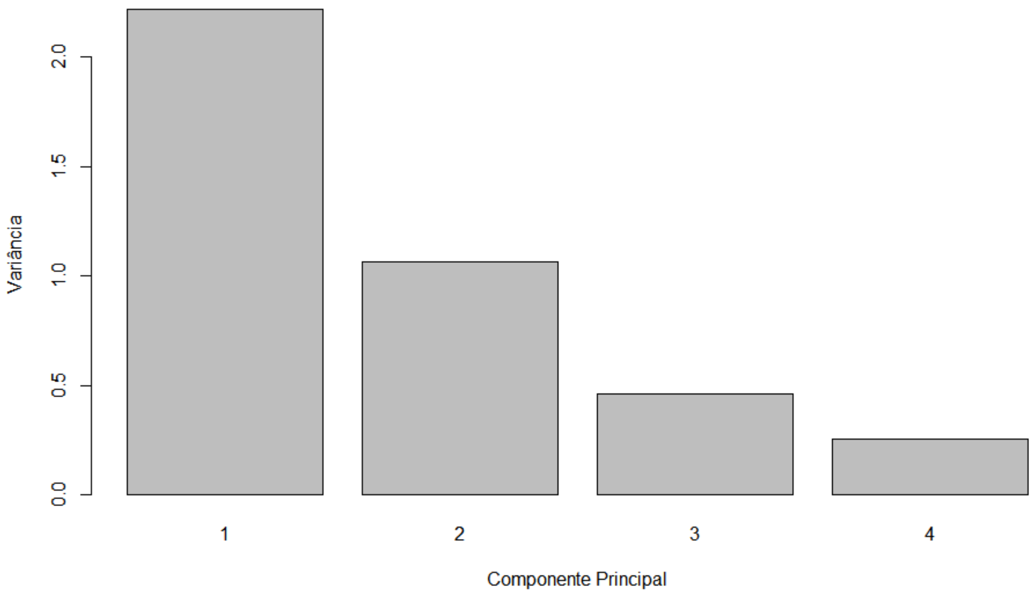




EXTRAÇÃO DE CARACTERÍSTICAS: PCA - PRÁTICA

Estimar algebricamente os auto-vetores (matriz de projeção)

```
sigma = cov(scale(dados[,1:4]))   # covariância dos dados normalizados
eig = eigen(sigma)  # eig$values eig$vectors

print(eig)
eigen() decomposition
$`values`
[1] 2.2159514 1.0645045 0.4637687 0.2557753

$vectors
            [,1]        [,2]       [,3]        [,4]
[1,]  0.54505948  0.16692958  0.8184914  0.07152949
[2,]  0.59764234  0.01796915 -0.4590769  0.65707618
[3,]  0.58520673 -0.26652564 -0.2728136 -0.71559058
[4,] -0.05715665 -0.94909205  0.2118794  0.22597463
```

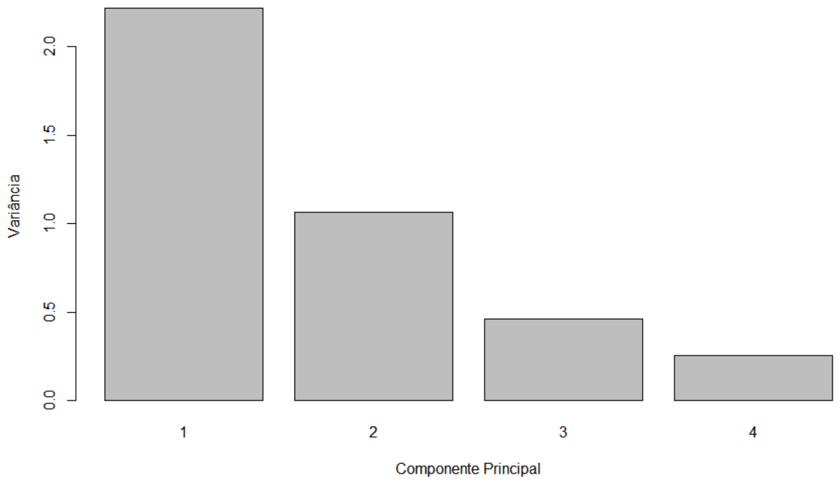


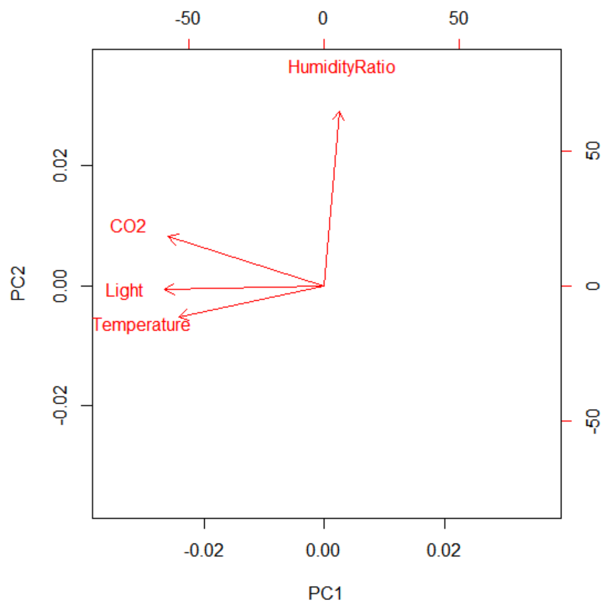

```
barplot(eig$values,ylab  = "Variância",xlab = "Componente Principal",names.arg =1:4)
```

EXTRAÇÃO DE CARACTERÍSTICAS: PCA - PRÁTICA

Estimar via função da base do R

```
pca = prcomp(dados[,2:5], center=TRUE, scale=TRUE)

print(pca)
Standard deviations (1, .., p=4):
[1] 1.4886072 1.0317483 0.6810057 0.5057424

Rotation (n x k) = (4 x 4):
                    PC1         PC2        PC3         PC4
Temperature   -0.54505948 -0.16692958 -0.8184914 -0.07152949
Light         -0.59764234 -0.01796915  0.4590769 -0.65707618
CO2           -0.58520673  0.26652564  0.2728136  0.71559058
HumidityRatio  0.05715665  0.94909205 -0.2118794 -0.22597463

barplot(pca$sdev^2,ylab  = "Variância",
              xlab = "Componente Principal",names.arg =1:4)

biplot(pca,xlabs = rep("", nrow(dados[,1:4])))
```



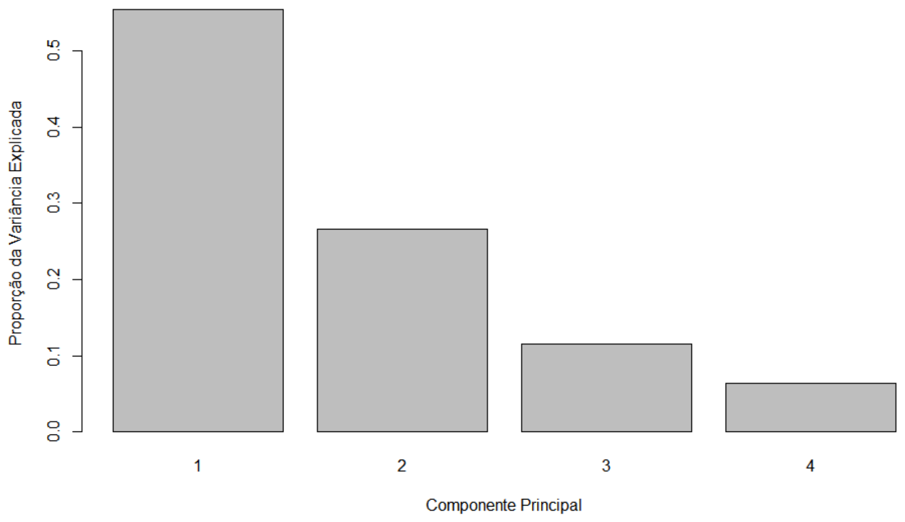


EXTRAÇÃO DE CARACTERÍSTICAS: PCA - PRÁTICA

O quanto cada autovetor explica a variância dos dados?

```
summary(pca)
Importance of components:
                         PC1    PC2    PC3     PC4
Standard deviation     1.489 1.0317 0.6810 0.50574
Proportion of Variance 0.554 0.2661 0.1159 0.06394
Cumulative Proportion  0.554 0.8201 0.9361 1.00000
```

```
pca_var = pca$sdev^2  # ou pca_var = eig$values
prop = pca_var/sum(pca_var)
barplot(prop,
    ylab  = "Proporção da Variância Explicada",
    xlab = "Componente Principal",
    names.arg =1:4)
```



EXTRAÇÃO DE CARACTERÍSTICAS: PCA - PRÁTICA

Como realizar a transformação?

```
dadosTransformados = predict(pca, dados)
```

```
dadosTransformados = as.matrix(dadosProcessados[,2:5]) %*% eig$vectors
```

```
summary(dadosTransformados)
      PC1                PC2                PC3                PC4
 Min.   :-4.4152   Min.   :-1.6804   Min.   :-2.3457   Min.   :-2.16507
 1st Qu.:-0.3265   1st Qu.:-0.9799   1st Qu.:-0.4778   1st Qu.:-0.19901
 Median : 0.6136   Median : 0.1092   Median : 0.1407   Median : 0.02519
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.: 1.0438   3rd Qu.: 0.7809   3rd Qu.: 0.5207   3rd Qu.: 0.23688
 Max.   : 1.5148   Max.   : 3.1242   Max.   : 1.7426   Max.   : 2.35976
```

```
cor(dadosTransformados)
              PC1           PC2           PC3           PC4
PC1  1.000000e+00 -6.025500e-16 -1.452276e-14  1.879051e-14
PC2 -6.025500e-16  1.000000e+00 -5.852002e-15 -2.278114e-15
PC3 -1.452276e-14 -5.852002e-15  1.000000e+00 -2.490354e-14
PC4  1.879051e-14 -2.278114e-15 -2.490354e-14  1.000000e+00
```

PRÁTICA: EXTRAÇÃO DE CARACTERÍSTICAS - PCA

Como realizar a redução da dimensionalidade?

```
nComp = 3
dadosReduzidos = predict(pca, dados)[,1:nComp]
summary(dadosReduzidos)
      PC1                PC2                PC3
 Min.   :-4.4152   Min.   :-1.6804   Min.   :-2.3457
 1st Qu.:-0.3265   1st Qu.:-0.9799   1st Qu.:-0.4778
 Median : 0.6136   Median : 0.1092   Median : 0.1407
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 1.0438   3rd Qu.: 0.7809   3rd Qu.: 0.5207
 Max.   : 1.5148   Max.   : 3.1242   Max.   : 1.7426
```

```
dadosReduzidos = as.matrix(dadosProcessados[,1:4]) %*% -eig$vectors[,1:nComp]
```

EXTRAÇÃO DE CARACTERÍSTICAS: PCA - PRÁTICA

PCA no pacote caret

```
pcaParametros = preProcess(dados, method = c("center", "scale", "pca"), thresh=1.0)
```

É possível definir o número máximo de componentes principais baseado na variância explicada com a opção **thresh** (default é 95%)

```
pcaParametros = preProcess(dados,  method = c("center", "scale", "pca"), thresh=0.95)
```

Ou pelo número de componentes (dimensões) com a opção **pcaComp**

```
pcaParametros = preProcess(dados,  method = c("center", "scale", "pca"), pcaComp=2)
```

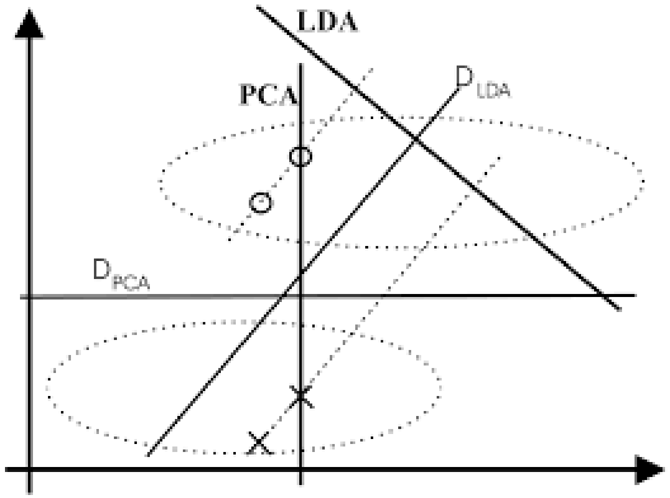
Para fazer a projeção, deve-se utilizar a função predict(), igual ao modo de que é feito para a função predict()

```
dadosProjetados = predict(pcaParametros,dados)
```

EXTRAÇÃO DE CARACTERÍSTICAS: PCA - PRÁTICA

Para as situações onde temos a informação das classes pode-se aplicar o método de Análise de Discriminantes Lineares (*Linear Discriminant Analysis* – LDA)

Com a informação das classes, esta transformação tem por objetivo de reduzir a dispersão dos dados da classe enquanto aumenta a separação entre as classes (o que é excelente para a tarefa de classificação)



Mas vamos deixar para depois....