Curso
Ciência de Dados
aplicada à Saúde

Aprendizagem não supervisionada
PCA + K-means

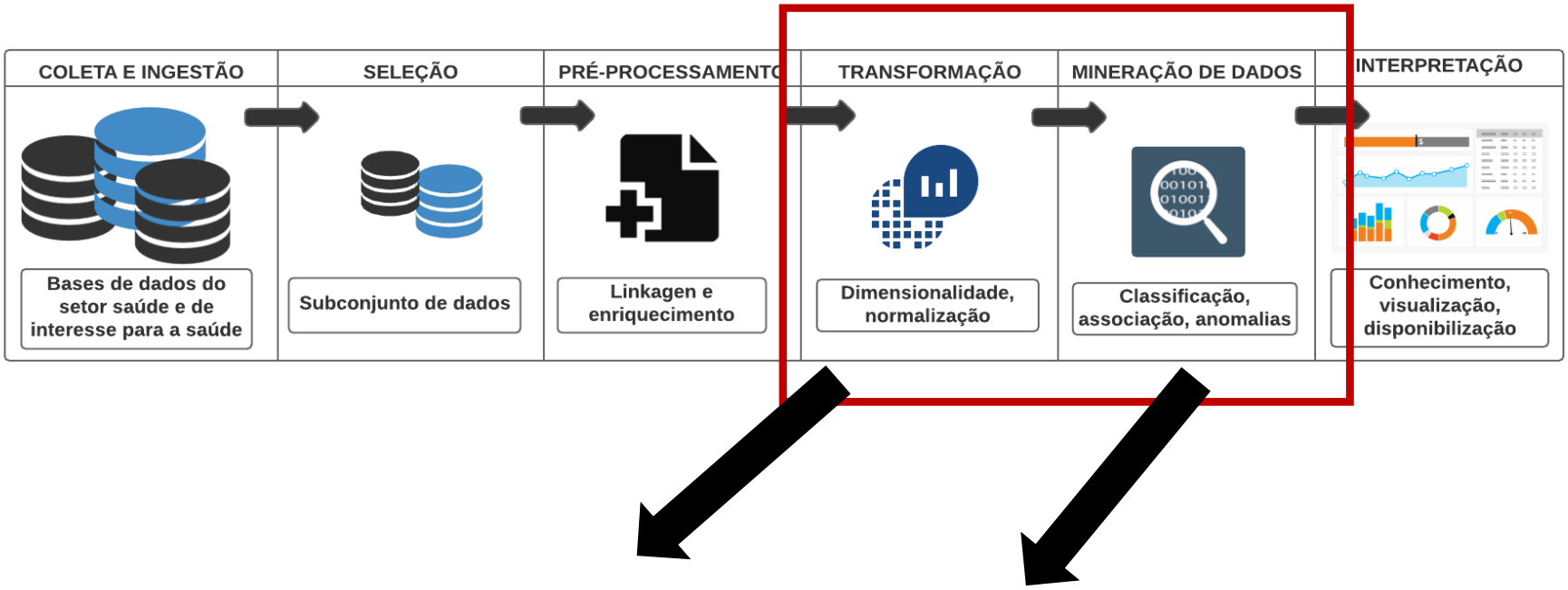
Prof. Dr. Marcel Pedroso
Pesquisador em Saúde Pública

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde
Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
SUS
Ministério da Saúde
GOVERNO FEDERAL
BRASIL
PÁTRIA EDUCADORA

PCA
Principal Component Analysis

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde
Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
SUS
Ministério da Saúde
GOVERNO FEDERAL
BRASIL
PÁTRIA EDUCADORA

Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases – KDD*)



PCA – Principal Component Analysis

K-means – Análise de agrupamentos

DEFINIÇÃO MINERAÇÃO DE DADOS

- **Etapa** do processo de **KDD**
- **Aplicação de algoritmos** capazes de **extrair conhecimento** a partir de dados pré-processados

- Análise Descritiva (medidas de distribuição, tendência central e variância)
- Análise Preditiva (classificação e regressão)
- **Análise de Agrupamento (segmentação de bases de dados)**
- Detecção de anomalias e associação

DEFINIÇÃO MACHINE LEARNING / APRENDIZADO DE MÁQUINA

PARADIGMAS DE APRENDIZADO

- **Aprendizado Supervisionado**

Baseado em um conjunto de objetos para os quais as saídas **desejadas são conhecidas** (exemplos Árvores de Decisão, Regressão linear e logística, k-NN, naïve Bayes, Redes Neurais Artificiais, SVM, Regras de Classificação)

- **Aprendizado Não Supervisionado**

Baseado em um conjunto de objetos para os quais as saídas desejadas **NÃO** são conhecidas ou a tarefa é de **categorização** (**K-means**, G-means, DBSCANC, Redes Neurais Artificiais)

PCA – PRINCIPAL COMPONENT ANALYSIS

Definição: Técnica de Redução de Dimensionalidade baseada em álgebra linear

- Dimensionalidade = número de atributos, features, variáveis de um dataset

Benefícios: Redução de Dimensionalidade

- Muitos algoritmos de mineração de dados funcionam **melhor** se a dimensionalidade for **pequena**
- Eliminar atributos irrelevantes e reduzir **ruídos** (erros)
- Tornar o modelo mais **compreensível**
- Facilitar a **visualizaçã**o dos dados
- Reduzir quantidade de tempo de **processamento** dos modelos
- Minimizar os efeitos da **"maldição da dimensionalidade"**

PCA – PRINCIPAL COMPONENT ANALYSIS

Maldição da dimensionalidade

- Análises de dados se tornam **mais difíceis** quando a dimensionalidade aumenta
- **Dados dispersos** no espaço que ocupam na matriz

➢ Impacto em **técnicas de classificação**: não há objetos de dados suficientes para a modelagem confiável das classes para todos os objetos possíveis

➢ Impacto em **técnicas de agrupamento**: problemas na definição de densidade e distância entre os pontos são críticas e os tornam menos significativos

Problemas: Exatidão da classificação reduzida e grupos de qualidade inferior

PCA – PRINCIPAL COMPONENT ANALYSIS

Objetivos PCA

- Criar um **novo conjunto de atributos** (dimensões)
- Manter a **variabilidade** dos dados

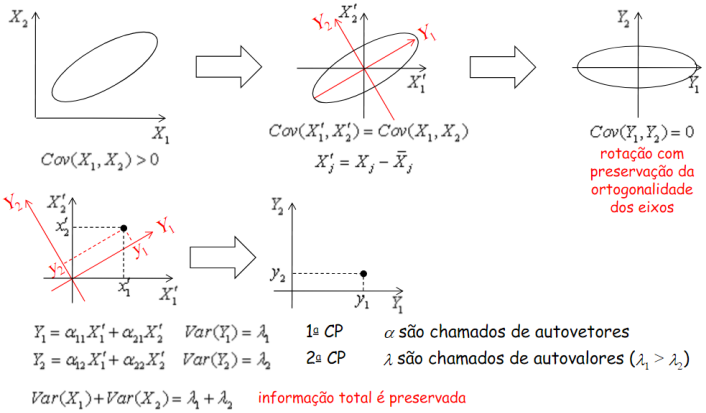
Características

- Tendência a **identificar padrões** nos dados
- Frequentemente uma **pequena fração dos atributos** é capaz de capturar a maior parte da variabilidade dos dados
- **Eliminação dos ruídos** (espera-se que os erros sejam mais fracos que os padrões)

PCA – PRINCIPAL COMPONENT ANALYSIS

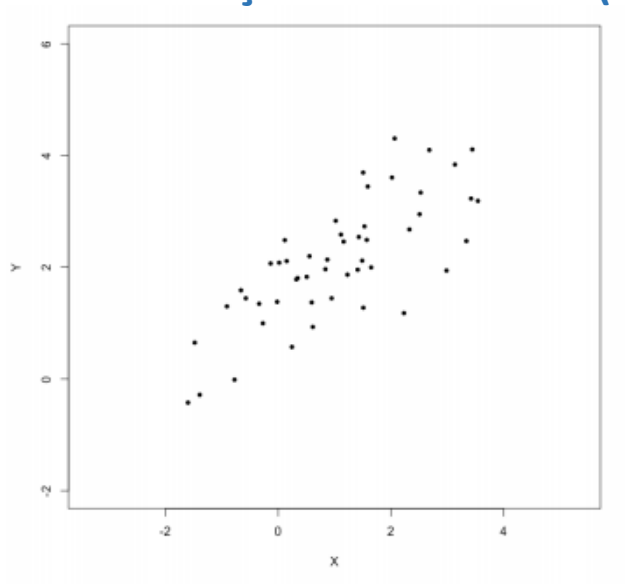
Processo básico do PCA

- **Normalização** dos dados e construção de uma **matriz de covariância**
- Cálculo dos **autovetores** e **autovalores** da matriz de covariância
- Ordenação dos novos atributos de acordo com a **variância capturada**
- O **componente principal** (a_vetor com maior a_valor) captura a maior variância
- Os **demais componentes** capturam o restante da variância (ortogonal)

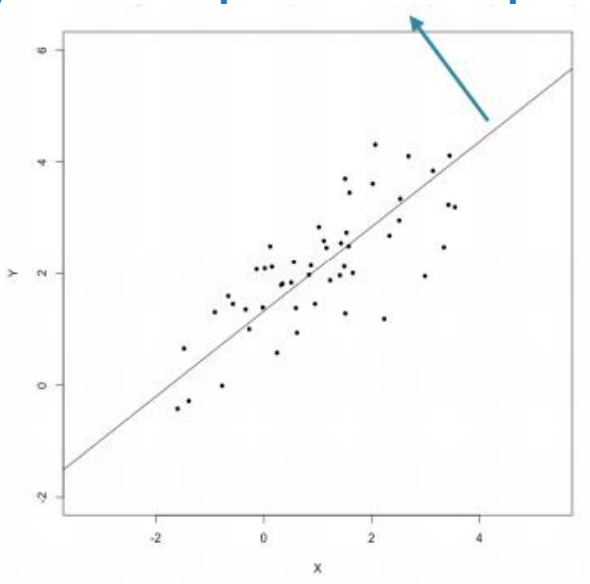


PCA – PRINCIPAL COMPONENT ANALYSIS

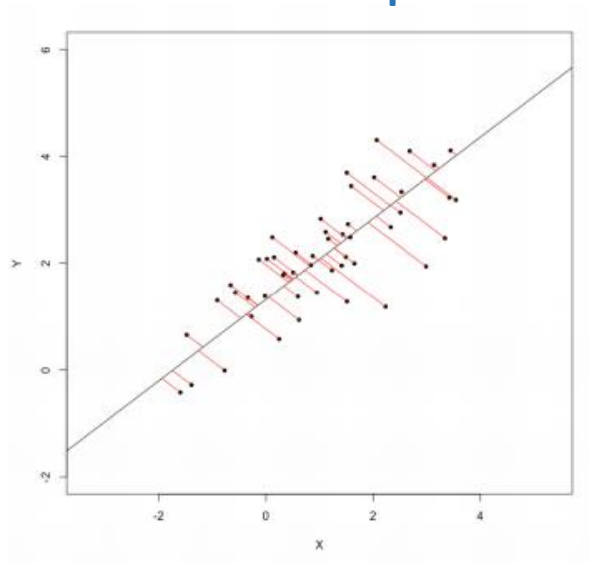
Observações 2 dimensões (x, y)



Componente Principal



Scores do Componente

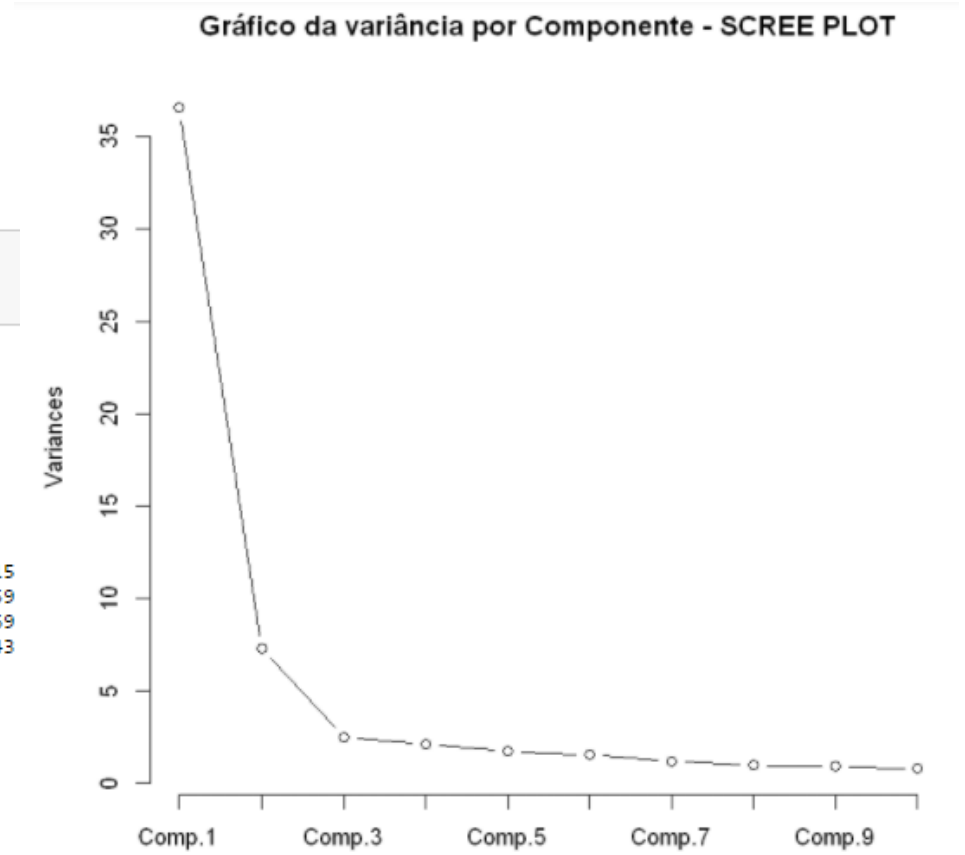


PCA – PRINCIPAL COMPONENT ANALYSIS

Principais métricas e gráficos

```
# PACOTE - PRINCIPAL COMPONENT ANALYSIS (PCA)
PCA <- princomp(pnud_PCA, scores=TRUE, cor=TRUE)
summary(PCA)
```

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     6.0492169 2.6980308 1.57866864 1.44939532 1.31040892
Proportion of Variance 0.5902101 0.1174092 0.04019669 0.03388301 0.02769632
Cumulative Proportion  0.5902101 0.7076193 0.74781598 0.78169899 0.80939531
                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.24023890 1.09220258 0.98852001 0.95403852 0.89445305
Proportion of Variance 0.02480956 0.01924043 0.01576084 0.01468048 0.01290397
Cumulative Proportion  0.83420487 0.85344529 0.86920613 0.88388660 0.89679058
                          Comp.11    Comp.12    Comp.13     Comp.14     Comp.15
Standard deviation     0.85377392 0.81706816 0.76940620 0.718610831 0.704640959
Proportion of Variance 0.01175693 0.01076775 0.00954816 0.008329057 0.008008369
Cumulative Proportion  0.90854751 0.91931526 0.92886342 0.937192474 0.945200843
```



K-MEANS CLUSTERING

ANÁLISE DE AGRUPAMENTOS – CLUSTERING

Aprendizado Não Supervisionado

- Baseado em um conjunto de objetos para os quais as saídas desejadas **NÃO** são conhecidas ou a tarefa é de **categorização**

Objetivos

- Dado um conjunto de objetos descritos por **múltiplos valores** (atributos)
- **Atribuir grupos** (clusters) aos objetos particionando os dados
  - Maximizar a similaridade **intra**-clusters
  - Minimizar a similaridade **inter**-clusters

ANÁLISE DE AGRUPAMENTOS – CLUSTERING

Métodos de Agrupamento

Os métodos de agrupamento podem ser divididos em **três principais** categorias

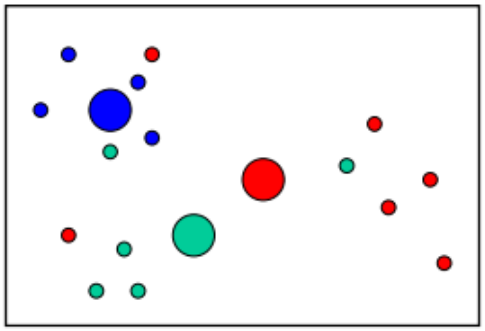
- **Métodos de particionamento**: Dados n objetos, esse método constrói k partições dos dados, onde cada partição representa um cluster (com $k \leq n$)
- **Métodos hierárquicos**: decomposição hierárquica de um dado conjunto de objetos, em que os clusters são aninhados e organizados em uma árvore de dendograma
- **Métodos baseados em grade e densidade**: constrói clusters com base na densidade das regiões, ou seja, no número de objetos ou pontos

K-MEANS CLUSTERING

- **Método de particionamento**: Dados n objetos, esse método constrói k partições dos dados, onde cada partição representa um cluster (com $k \leq n$)
- Busca **minimizar a distância** dos elementos a um conjunto de k centros **iterativamente**
- **O parâmetro K** é o número de clusters (grupos) e precisa ser definido *a priori*
- Dado um objeto, é **calculada a distância** (euclidiana por exemplo) desse objeto ao centro de cada cluster
- Para calcular o centro de cada grupo, **basta calcular a média (mean)** dos valores dos objetos que estão naquele grupo

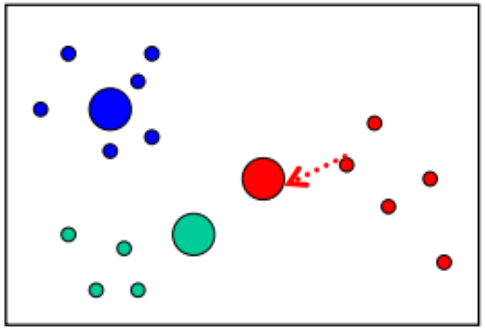
K-MEANS CLUSTERING

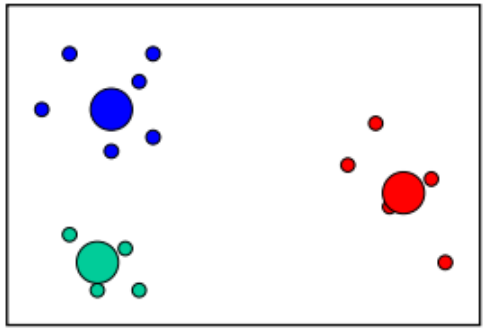
- **Exemplo K = 3**



Passo 1: Escolha aleatória de clusters e cálculo dos centróides (círculos maiores)

Passo 2: Atribua cada ponto ao centróide mais próximo





Passo 3: Recalcule centróides (neste exemplo, a solução é agora estável)

K-MEANS CLUSTERING

Métricas principais

- Inertia: é uma **medida de variância** intra-cluster
  (soma dos erros dentro do cluster ou “withinss”)

- Silhouette: é uma **medida relativa de similaridade** de um objeto para o seu próprio cluster em comparação com outros clusters
- Varia de **-1 a 1**, **quanto mais próximo de 1** indica que o objeto está bem adaptado ao seu próprio cluster e mal adaptado aos clusters vizinhos
- Reflete a **qualidade da alocação** dos objetos no grupos e auxilia a **escolha do número** ótimo de **K**

Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

www.facebook.com/fiocruz.icict

twitter.com/@Icict_fiocruz

www.youtube.com/videosaudefio

www.icict.fiocruz.br

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

SUS

Ministério da Saúde

GOVERNO FEDERAL
BRASIL
PÁTRIA EDUCADORA