

# Introdução a Aprendizagem De Máquina

Pós-graduação em Ciência de Dados e Machine Learning  
Módulo 3 - Data Mining e Machine Learning

Professor Msc. Ricardo José Menezes Maia

# Análise do Componente Principal (PCA)

# Análise do Componente Principal (PCA)

Seção 10.2 do Introduction to Statical Learning de Gareth James

# Análise do Componente Principal (PCA)

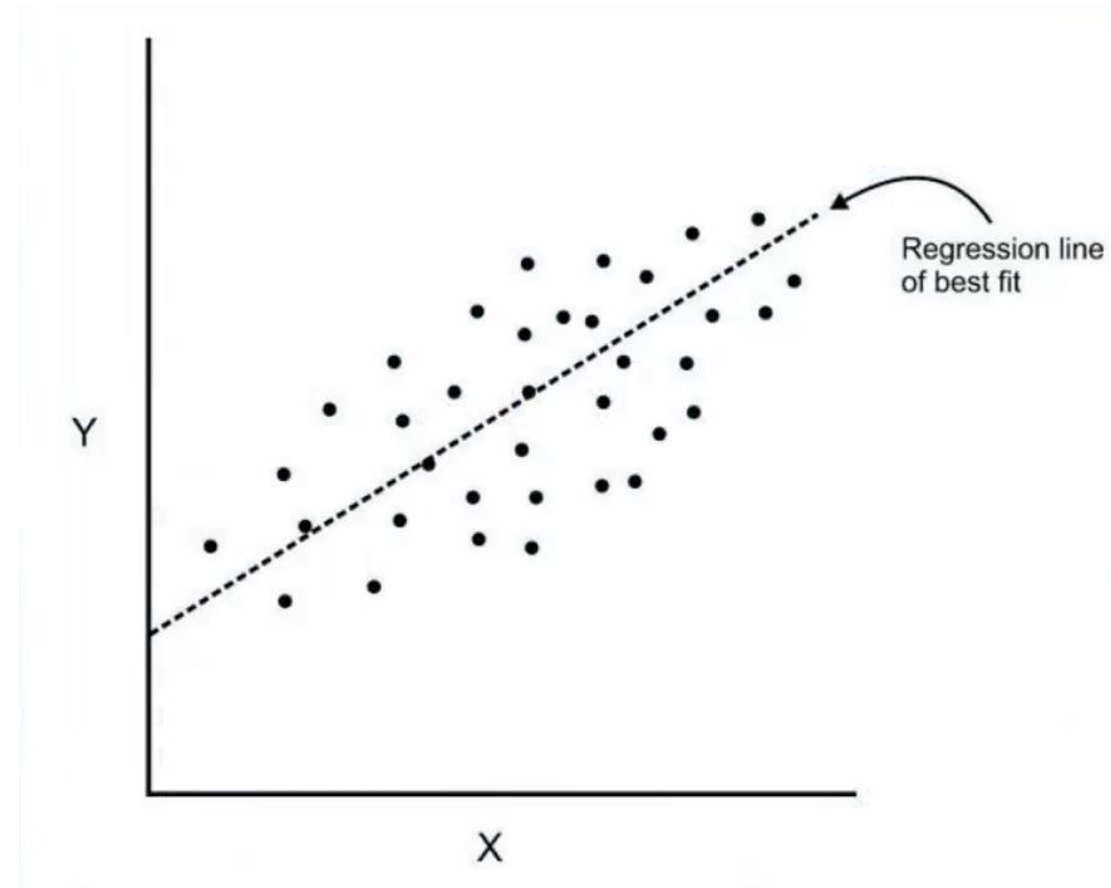
- É um método não supervisionado estatístico, usado para examinar relações entre um conjunto de variáveis, com objetivo de identificar uma estrutura básica por trás das mesmas.
- Também é conhecido como análise de fatores.

# Análise do Componente Principal (PCA)

- Enquanto uma regressão determina a linha que melhor se ajusta ao conjunto de dados, PCA determina o melhor conjunto ortogonal de linhas para ajustar nosso modelo.
- Ortogonal significa “perpendicular”. Estamos procurando então por linhas perpendiculares  $n$   $n$ -dimensional.
- Espaço  $n$ -dimensional é o espaço vetorial do grupo de parâmetros.
  - Existirão tantas dimensões neste espaço vetorial quanto parâmetros no seu conjunto de dados. Assim, um conjunto com 4 variáveis terá 4 dimensões.

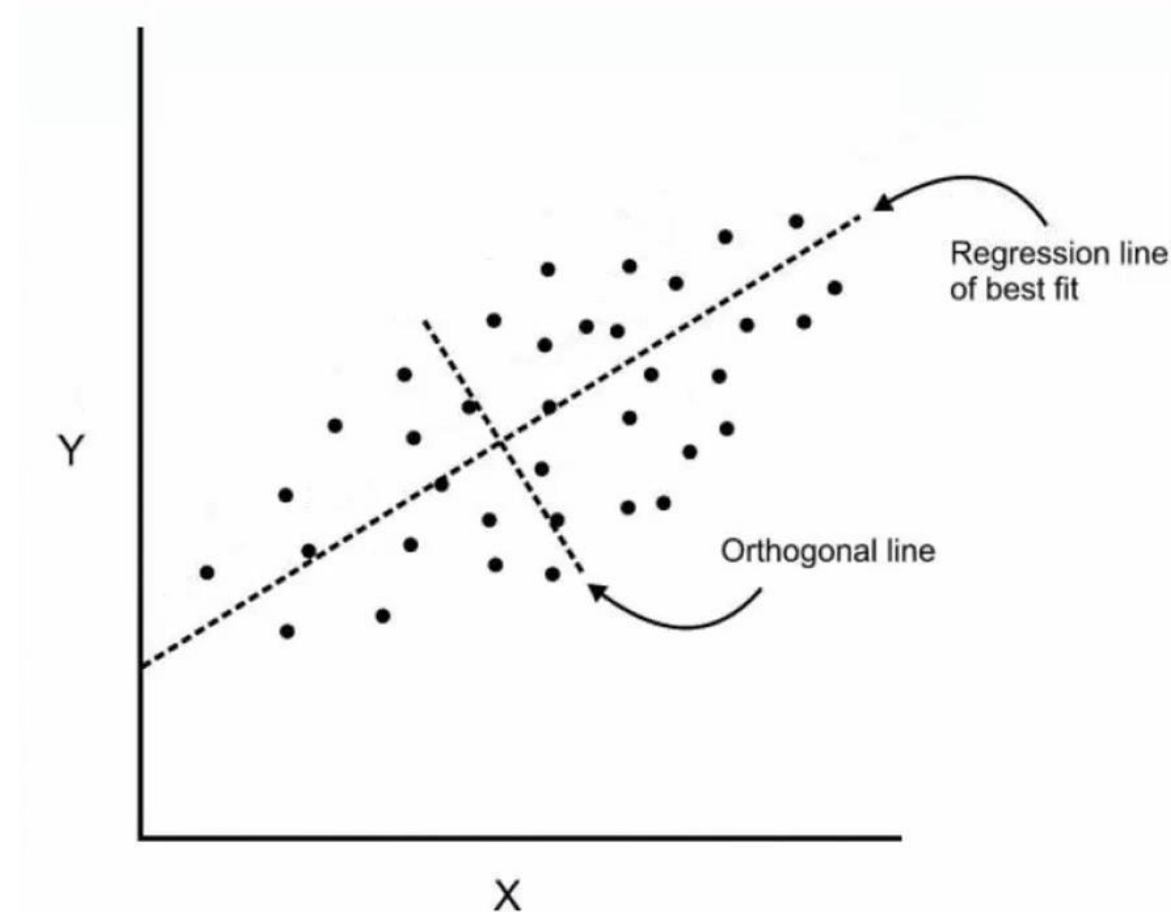
# Análise do Componente Principal (PCA)

- Temos alguns dados plotados ao longo do eixo x e y.



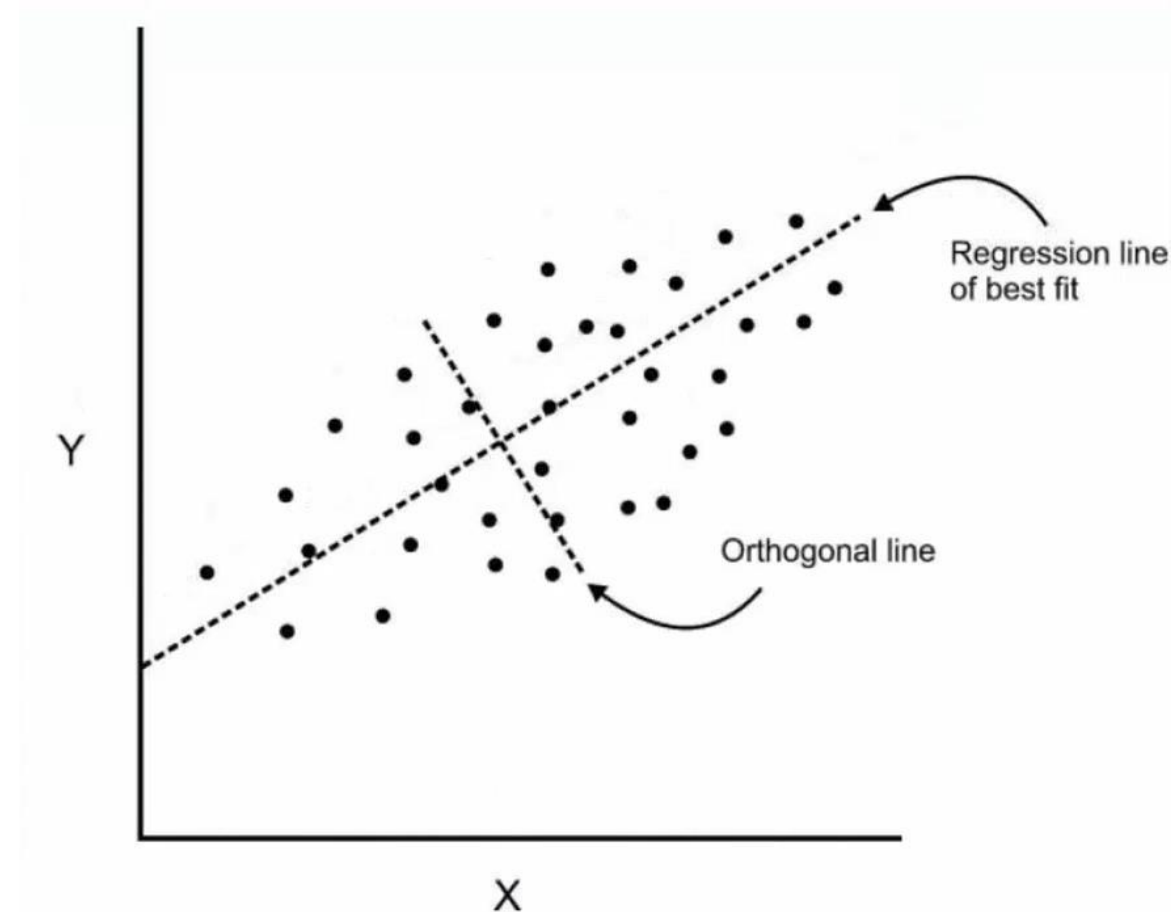
# Análise do Componente Principal (PCA)

- Podemos adicionar uma linha ortogonal
- Agora começaremos a entender os componentes



# Análise do Componente Principal (PCA)

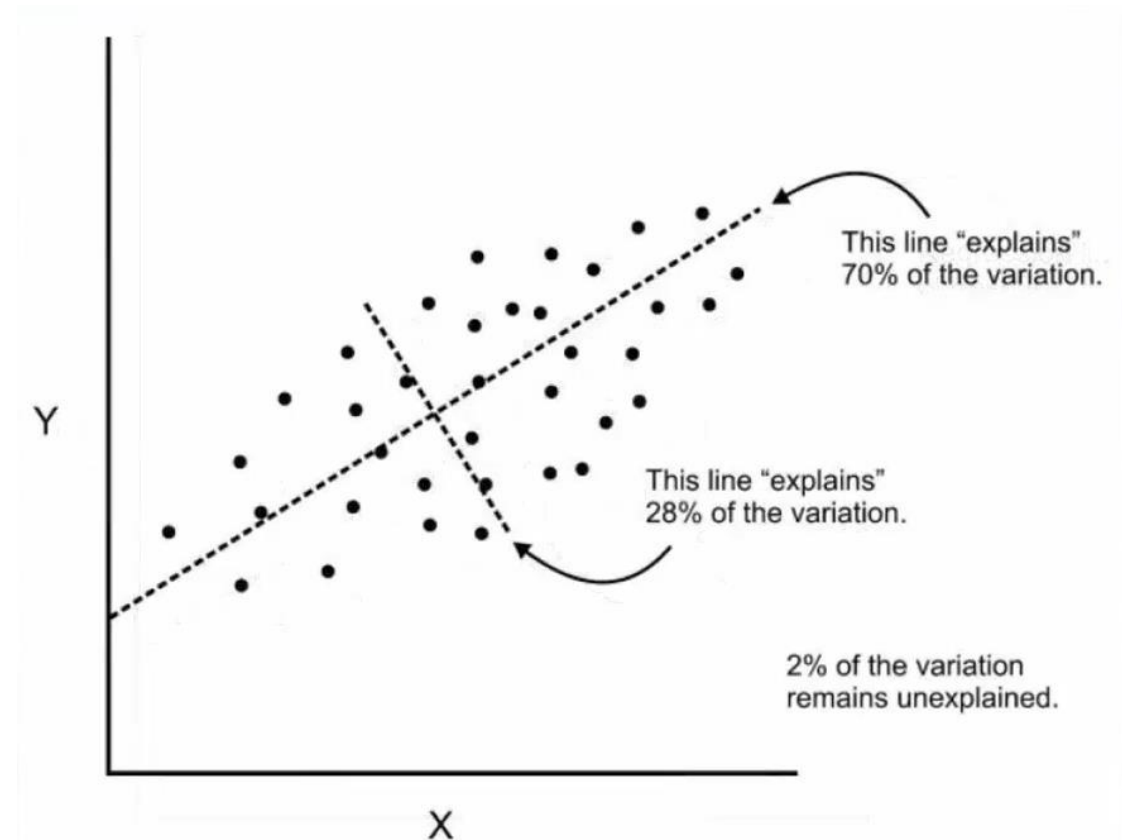
- Os componentes são transformações lineares dos parâmetros iniciais, de forma que a maior variância no conjunto de dados apareça no primeiro eixo, a segunda, maior variância no segundo eixo, etc ...





# Análise do Componente Principal (PCA)

- Isso permite reduzir o número de variáveis usadas em nossa análise
- Note que os componentes são descorrelacionados, uma vez que são ortogonais uns aos outros.



# Análise do Componente Principal (PCA)

- Se usarmos essa técnica em um conjunto de dados com um número muito grande de variáveis, podemos comprimir a variância explicada para apenas alguns componentes.
- A parte mais difícil no PCA é interpretar os componentes.

# Análise do Componente Principal (PCA)

- Para nosso trabalho em python, mostraremos como utilizar no scikit-learn
- Nós normalmente padronizamos os dados para o PCA performar melhor
- O algoritmo é usado normalmente para análise e não para a criação de um modelo por si só.