Análise de componentes principais (Principal Component Analysis)

Tarefa de leitura

Seção 10.2 do livro
Introduction to Statistical Learning
por Gareth James, et al.

Background

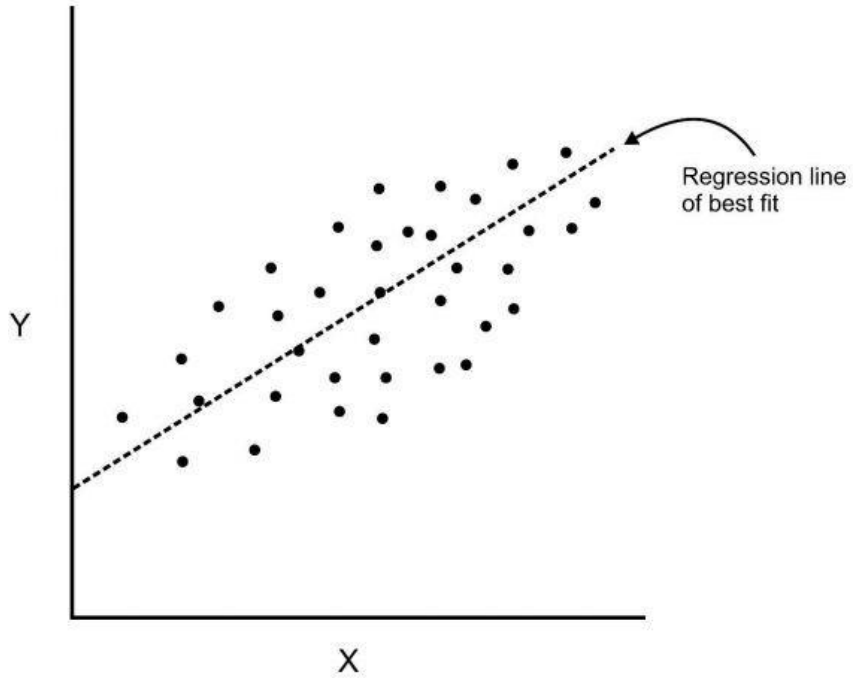
- Vamos discutir a ideia básica por trás da análise de componentes principais (principal component analysis).
- É uma técnica estatística não supervisionada usada para examinar as inter-relações entre um conjunto de variáveis a fim de identificar a estrutura subjacente dessas variáveis.
- Também é conhecida como análise dos fatores gerais (general factor analysis).

Background

- Enquanto a regressão determina uma linha que melhor se ajusta a um conjunto de dados, a análise dos fatores determina várias linhas ortogonais de melhor ajuste para o conjunto de dados. Ortogonal significa “em ângulos retos”.
  - As linhas são perpendiculares entre si em um espaço n-dimensional.
- Espaço n-dimensional é o espaço amostral das variáveis.
  - Existem tantas dimensões quanto existem variáveis, portanto, em um conjunto de dados com 4 variáveis o espaço amostral é 4-dimensional.

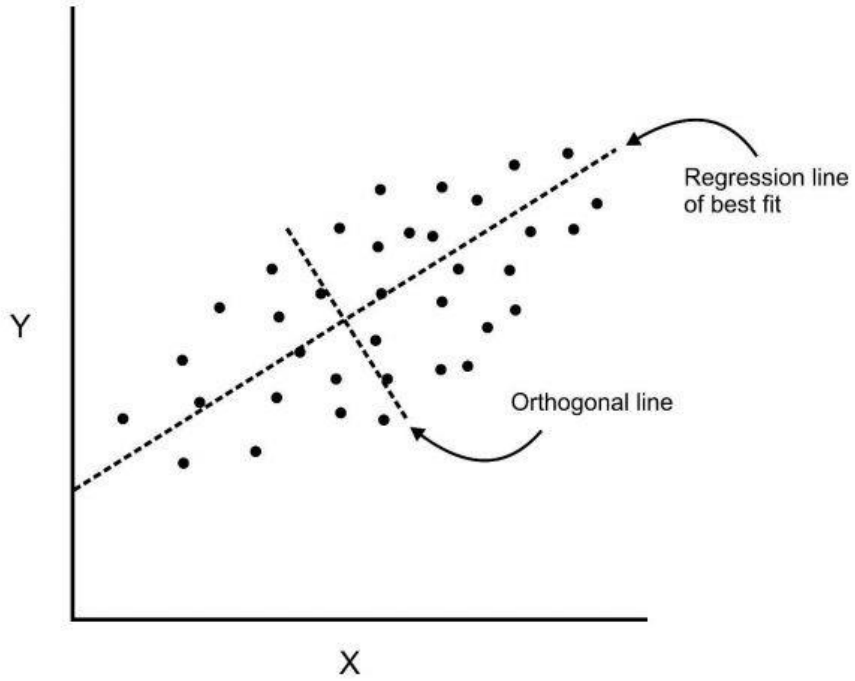
Background

- Aqui temos alguns dados plotados ao longo de duas características (features), x e y.



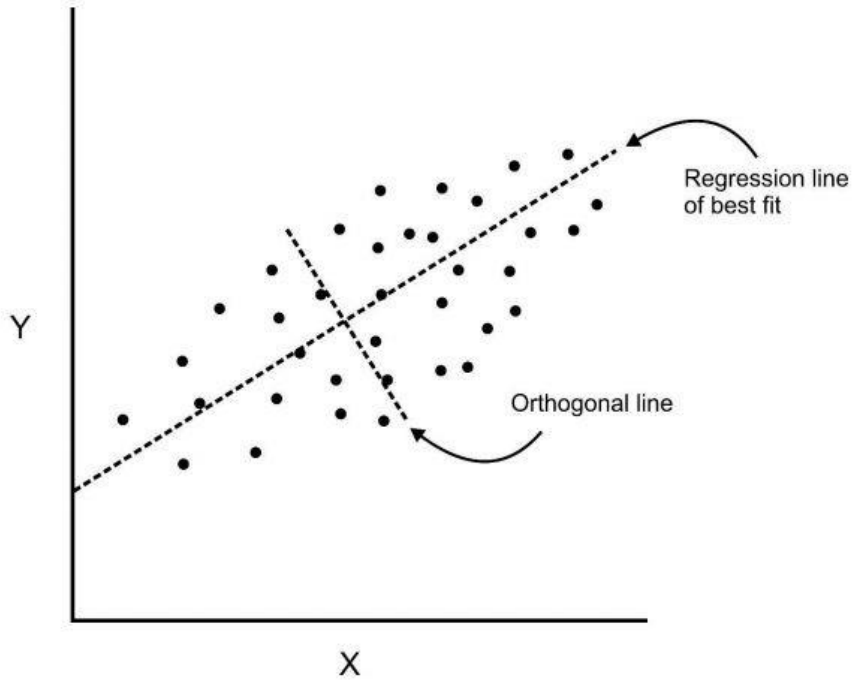
Background

- Podemos adicionar uma linha ortogonal.
- Agora podemos começar a entender os componentes (components).



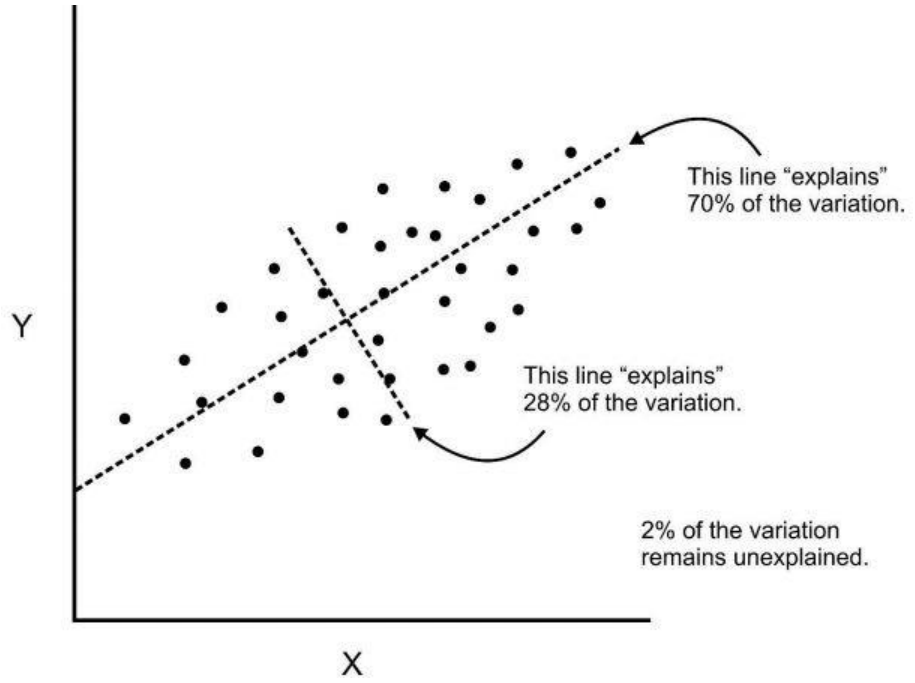
Background

- Os componentes (components) são uma transformação linear que escolhe um sistema de variáveis para o conjunto de dados de modo que a maior variação do conjunto de dados fique no primeiro eixo



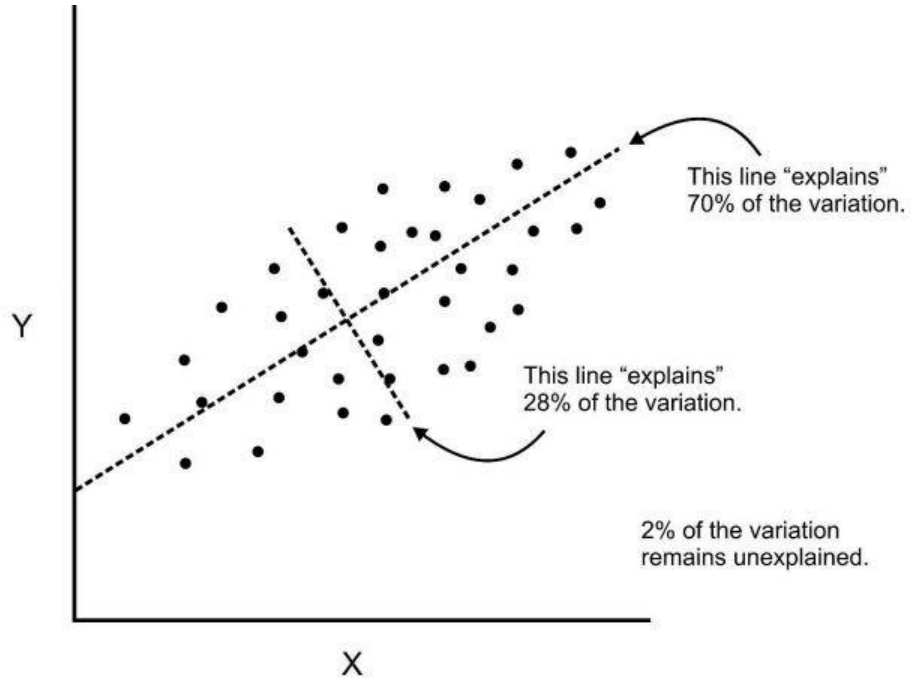
Background

- A segunda maior variação no segundo eixo, e assim por diante...
- Este processo permite-nos reduzir o número de variáveis usadas em uma análise.



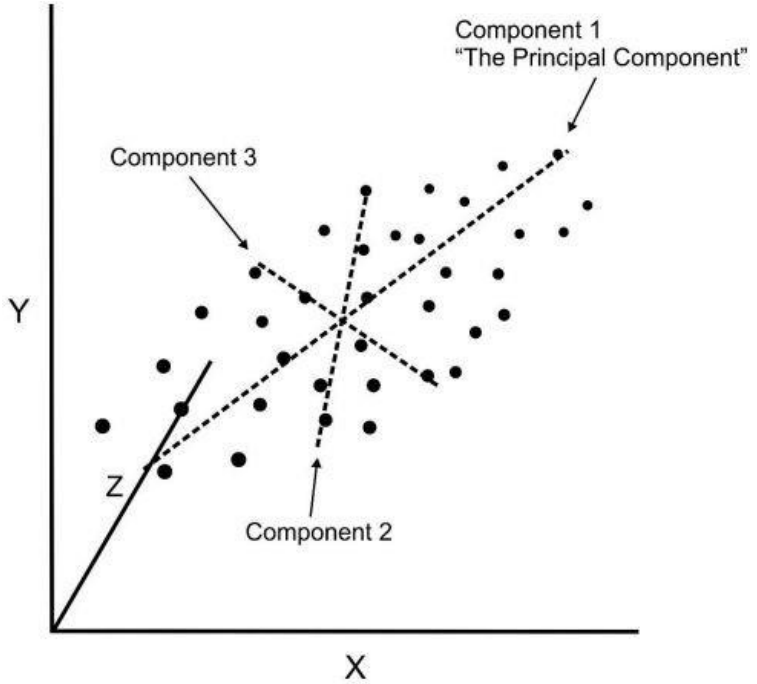
Background

- Observe que as componentes não são correlacionadas, pois no espaço amostral elas são ortogonais entre si.



Background

- Podemos continuar esta análise em mais dimensões.



Background

- Se usarmos essa técnica em um conjunto de dados com um grande número de variáveis, podemos comprimir a quantidade de variáveis necessárias para explicar a variação destes dados para apenas alguns componentes.
- A parte mais desafiadora do PCA é interpretar os componentes.