Principal Component Analysis

Victor Moreli dos Santos

Universidade de São Paulo

victormoreli@usp.br

May 31, 2024

Overview

Principal Component Analysis

Considerando um vetor $\vec{X}$ de dados como um vetor coluna em que cada componente equivale a uma certa variável de interesse:

$$\vec{X} = \begin{bmatrix} x_a \\ x_b \\ x_c \\ x_d \end{bmatrix}$$

A matriz X de dados é formado pelos vetores colunas de dados ($\vec{X}$). O PCA busca uma melhor forma de representar esses dados, fazendo uma mudança de base:

$$PX = Y$$

- X: Dataset original
- Y: Melhor representação do dataset
- P: transformação linear que leva X a Y, fazendo uma mudança de base

Principal Component Analysis

Sendo $p_1, p_2, ..., p_m$ as linhas de P e $x_1, ..., x_n$ as colunas de X, então Y é da forma:

$$Y = \begin{bmatrix} p_1 \cdot x_1 & \dots & p_1 \cdot x_n \\ \vdots & \ddots & \vdots \\ p_m \cdot x_1 & \dots & p_m \cdot x_n \end{bmatrix}$$

Ou seja, para uma i-ésima coluna de Y:

$$y_i = \begin{bmatrix} p_1 \cdot x_i \\ \vdots \\ p_m \cdot x_i \end{bmatrix}$$

Nota-se que o j-ésimo componente de y_i é a projeção de x_i na base P, ou seja, y_i **é a projeção de** x_i **na base P.**
Os vetores-linha de P $\{p_1, ..., p_m\}$ **são os componentes principais de X.**

Questões:

- O que seria uma boa escolha para P?
- O que significa representar os dados de X de uma maneira 'melhor'?

Redundância e Covariância

Para fins de simplificação da representação, é evidente que representar variáveis com alto grau de correlação se mostra algo redundante. Essa redundância se dá pelo fato de que uma variável pode ser facilmente inferida pela outra, sendo assim necessária a representação de somente uma delas.

Figure: Casos de covariância

Covariância

Sendo A e B dois sets de diferentes medidas (indicadas pelos subíndices) de duas variáveis (A e B) com média nula:

$$A = \{a_1, \ldots, a_n\}$$
$$\mathrm{B} = \{\mathrm{b}_1, \ldots, b_n\}$$

A variância dos conjuntos é calculada da seguinte forma:

$$\sigma_A^2 = \frac{1}{n} \sum_i a_i^2$$
$$\sigma_B^2 = \frac{1}{n} \sum_i b_i^2$$

Já a **covariância** é calculada entre duas variáveis, **buscando medir o grau de relacionamento entre elas**.

Pré-requisitos e noções introdutórias

Covariância x Correlação

1. Ambas descrevem o relacionamento entre duas variáveis. Se duas variáveis "progridem" na mesma direção, ou seja, ambas aumentam ou ambas diminuem, é dito que há uma covariância e correlação positiva.
2. Se o comportamento é inverso, é dito que há uma covariância e correlação negativa.
3. O terceiro caso, de total independência das variáveis, é de covariância e correlação nulas.

Diferença: O coeficiente de correlação é padronizado: vai de -1 a 1. Ou seja, além da noção de relacionamento entre as variáveis, traz também um significado de força dessa relação.

Covariância

A covariância de A e B é calculada da seguinte forma:

$$\sigma^2_{AB} = \frac{1}{n} \sum_i a_i b_i$$

Passando para a notação de matrizes:

$$a = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$$

$$b = \begin{bmatrix} b_1 & b_2 & \dots & b_n \end{bmatrix}$$

$$\sigma^2_{ab} = \frac{1}{n} a b^T$$

Essa ideia pode ser generalizada para mais dimensões, culminando assim na definição de **matriz de covariância**.

Matriz de Covariância

Definição

Sendo X uma matriz de dados tal que **cada linha é igual a uma série de medidas de uma variável** (total de m variáveis), e **cada coluna é uma amostra contendo medidas de todas as variáveis** (total de n amostras). A matriz de Covariância de X é definida tal que:

$$C_x = \frac{1}{n} X X^T$$

- C_x é quadrada (m x m) e **simétrica**
- Na **diagonal** há a **variância de cada variável**
- **Fora da diagonal há a covariância entre diferentes variáveis.** O elemento na linha i e coluna j de C_x é igual a covariância entre a variável i (linha i de X) e a variável j (coluna j de X^T = linha j de X)

Principal Component Analysis

Voltando para a linha principal de argumentação, com $Y = PX$, o que seria uma boa matriz de covariância para Y?
Para evitar redundância, deseja-se que os termos fora da diagonal sejam zero, ou seja, $C_y = \frac{1}{n} YY^T$ é diagonal. Mas, tem-se que:

$$C_y = \frac{1}{n} YY^T = \frac{1}{n}(PX)(PX)^T = \frac{1}{n} PXX^T P^T = P(\frac{1}{n} XX^T)P^T$$

$$\Rightarrow C_y = PC_X P^T$$

Sendo X uma matriz simétrica, ela também é diagonalizável, tal que:

$$C_x = EDE^T \Rightarrow C_y = (PE)D(E^T P^T)$$

Ou seja, escolhendo $P = E^T$:
$C_y = (PE)D(E^T P^T) = (PP^T)D(PP^T) = D.$

Principal Component Analysis

Definição

- **Os componentes principais de uma matriz X são os autovetores de sua matriz de covariância.**
- Sendo Y a nova representação dos dados após a mudança de base ($Y = PX$), o i-ésimo elemento da diagonal de C_Y é a variância de X ao longo do componente principal p_i.

Definição

Pelo SVD: $\frac{1}{\sqrt{n}}X = U\Sigma V^T$. Como U é a matriz dos autovetores de $\frac{1}{n}XX^T$:

- **U tem os componentes principais de $\frac{1}{\sqrt{n}}X$ nas colunas**
- $\frac{1}{\sqrt{n}}U^T X = \Sigma V^T \Rightarrow P = U^T$ e $Y = \Sigma V^T$

Ou seja, os componentes principais podem ser obtidos diretamente calculando o SVD.

Principal Component Analysis

O **primeiro componente principal** é a direção na qual os dados tem **maior variância**. Os componentes principais subsequentes são direções **ortogonais** com variância menor.
É dessa forma que o PCA realiza uma **redução de dimensionalidade** enquanto mantém as **informações mais importantes**, tornando-as mais explícitas.

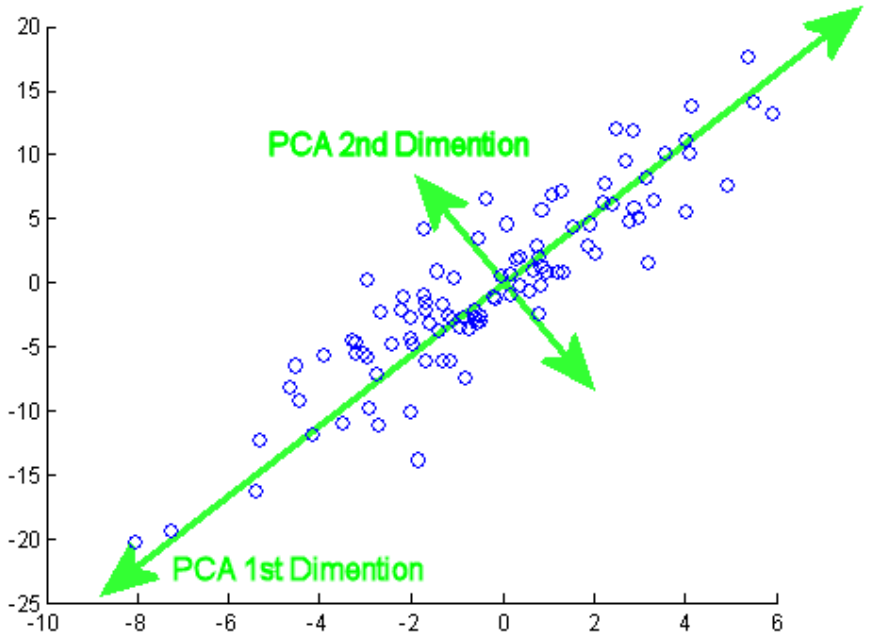

Figure: Enter Caption

Principal Component Analysis

**O PCA é uma técnica sensível à média dos valores das variáveis, logo deve-se subtrair essas médias do dataset original (X) antes de passar para os outros passos.