PCA

Principal Component Analysis

PCA - stegen

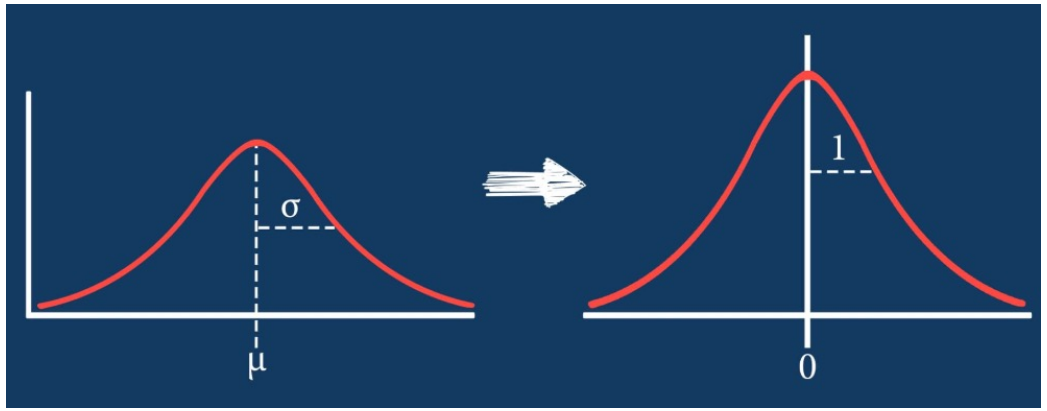
1. Standardize kontinerliga features (varaibler)
2. Beräkna covariance matris för att identifiera korrelationer
3. Beräkna eigenvectors och eigenvalues av covaraiance matrisen för att identifiera principal components
4. Skapa feature vector för att bestämma vilka principal components man ska behålla
5. Recast data kring principal components

PCA

- Dimension – antal variabler/features i datan
- PCA – dimensionsreducering medan man behåller information
- Tradeoff mellan accuracy och simplicity

1. Standardize kontinerliga features (varaibler)

- PCA sensitiv och vi vill att alla variabler ska bidra lika mycket
- Standardisera alla varaibler
- $z = \frac{x-\mu}{\sigma} = \frac{value\ -mean}{standard\ deviation}$



2. Beräkna covariance matris för att identifiera korrelationer

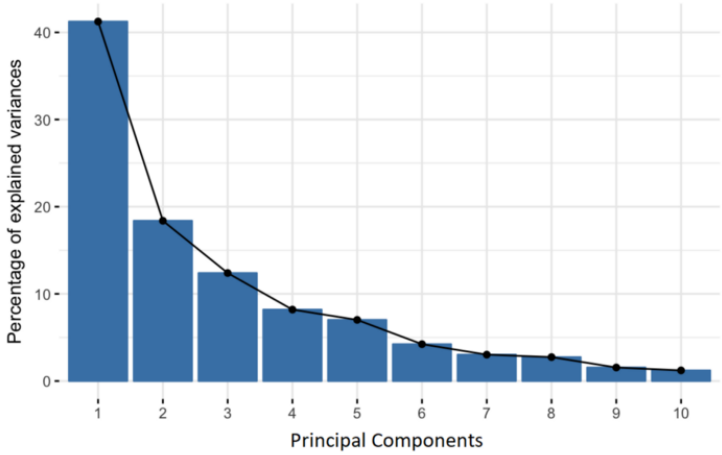
- Hur varierar variablerna från mean med avseende på varandra?
- Hög korrelation mellan variabler (multicollinearity) ger överflödig information
- Skapa en covariance matris som får samma dimension som antal features

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Covariance Matrix for 3-Dimensional Data

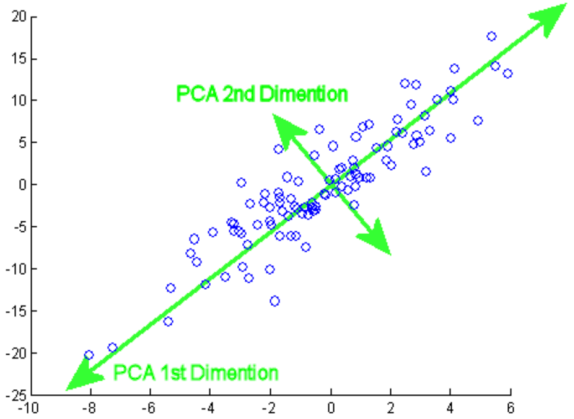
3. Beräkna eigenvectors och eigenvalues av covaraiance matrisen för att identifiera principal components

- **Principal component:** nya variabler skapat av linjärkombination av input variablerna. Principal component kommer vara uncorrelated och informationen från datan finns kvar
- Ex: 6 dimensionell data ger 6 principal components, men PCA kommer maxa informationen i den första principal componenten



3. Beräkna eigenvectors och eigenvalues av covaraiance matrisen för att identifiera principal components

- Behålla information men reducera dimensionen
- Nackdel – svårtolkad
- Vill hitta riktningen till datan som förklarar **max varians**
- Längre linje = större varainsd = mer info
- 1a principal component = störst varains



3. Beräkna eigenvectors och eigenvalues av covaraiance matrisen för att identifiera principal components

- **Eigenvectors** – av covariance matrisen. Riktningen till axlerna där det är störst varians (mest info). Aka principal component
- **Eigenvalues** - koefficient till eigenvektoren. Mängd varians i varje principal component
- Rangera eigenvector efter värdet på eigenvalues och du får den principal component med mest information!
- $Av = \lambda v$

4. Skapa feature vector för att bestämma vilka principal components man ska behålla

- **Feature vector** – ny matris med kolumner av eigenvectors med de principal component vi väljer behålla
- Vi väljer hur många principal component vi ska behålla!
- Vi kan behålla alla vilket betyder vi beskriver datan med nya variabler som är uncorrelated

5. Recast data kring principal components

- Orientera om datan
- $Slutliga\ dataset = feature\ vector^T \cdot standardized\ original\ dataset^T$