PRINCIPAL COMPONENT ANALYSIS

$\downarrow$

Dimensionality reduction technique

$p = (x_1, x_2)$

$P \in \mathbb{R}^{N \times D}$ $\quad K < D$

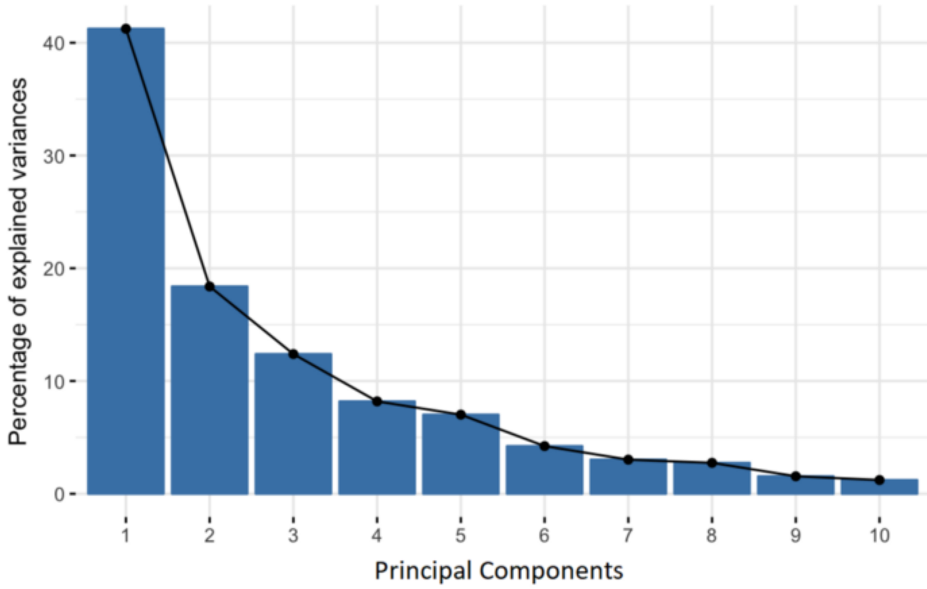
$\bar{P} \in \mathbb{R}^{N \times K}$

find the principle components i.e. directions along which the variance of the data is MAXIMAL

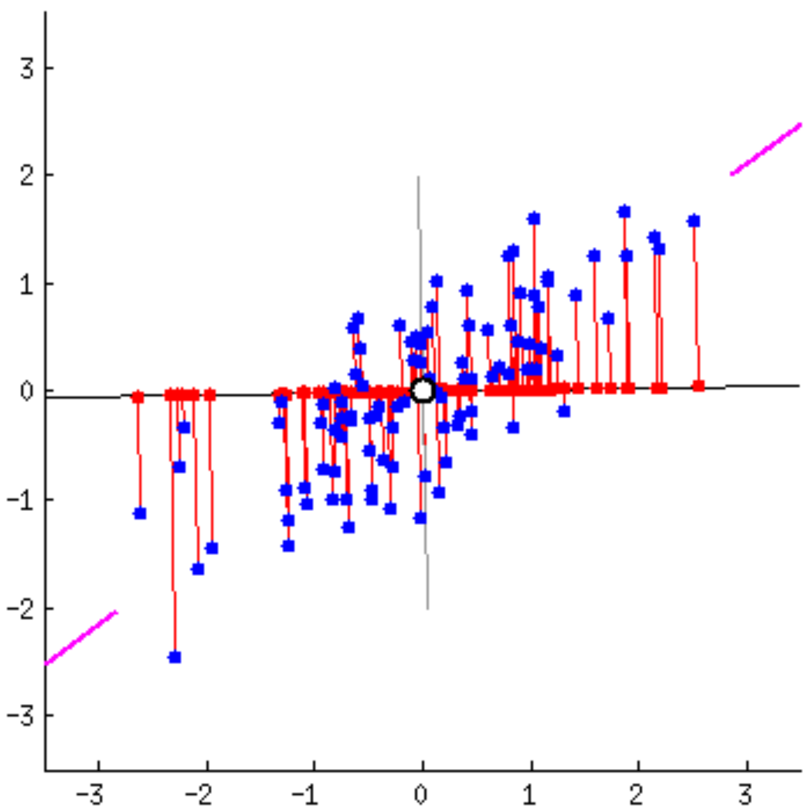
larger variance $\longrightarrow$ larger dispersion of data

$\downarrow$

larger information



Percentage of explained variances
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10
Principal Components



3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

ALGORITHM

Consider a dataset with N data points

Each datapoint $\vec{p_i} \in \mathbb{R}^D$ $\quad 1 \leq i \leq N$ $\quad (P_{N \times D})$

i.e $\quad \vec{p_i} = (x_1, x_2, x_3, \ldots x_D) \quad x_i \in \mathbb{R}$

① Standardize the data

$$x_i' = \frac{x_i - E[x_i]}{\sqrt{var(x_i)}}$$

② Compute the Covariance matrix $C_{D \times D}$

$$C = \begin{bmatrix} cov(x_1, x_1) \cdots & cov(x_1, x_2) & \cdots & cov(x_1, x_D) \\ \vdots & \vdots & & \vdots \\ cov(x_D, x_1) \cdots & cov(x_D, x_2) & \cdots & cov(x_D, x_D) \end{bmatrix}_{D \times D}$$

→ Symmetric

③ Compute the eigenvector and eigen values of the Cov matrix

$$C = V_{D \times D} \Sigma_{D \times D} V^T_{D \times D}$$

OR

$$Cv = \lambda v$$

$$(C - \lambda I) v = 0$$

$$\boxed{\det(C - \lambda I) = 0}$$

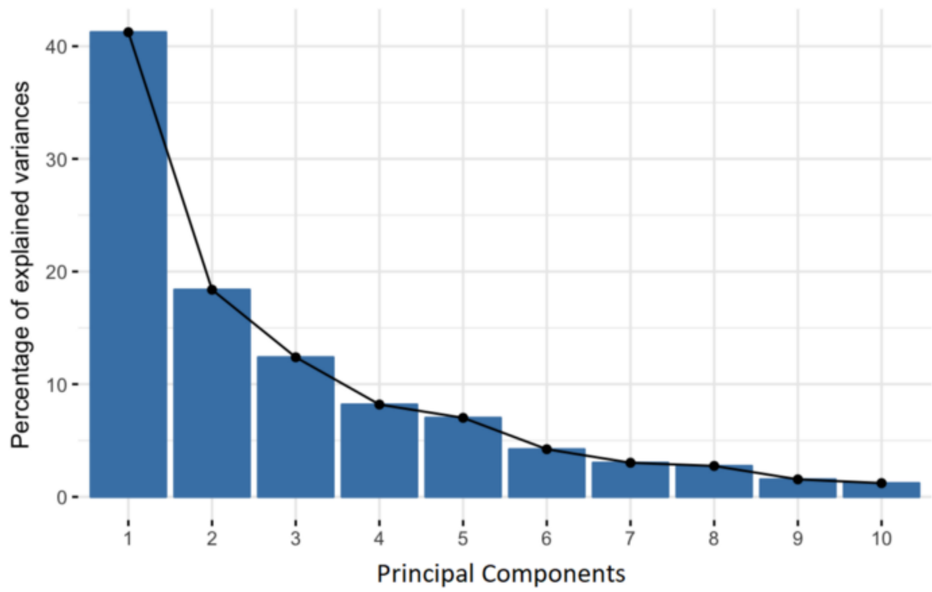
np.linalg.eig

④ Select Top K Eigenvectors are the principal components

$$X = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_k \\ | & | & & | \end{bmatrix}_{D \times K} \qquad v_i \in \mathbb{R}^D$$

$(\lambda_1) \quad (\lambda_2) \quad \quad (\lambda_k)$

such that $\lambda_1 > \lambda_2 > \lambda_3 \ldots > \lambda_k$



⑤ Recast the data along principal components

$$P_{N \times D} \, X_{D \times K} = \bar{P}_{N \times K}$$

$$\boxed{K \leq D}$$