Summary of PCA

1) [optional] Normalize data if appropriate so that each data value in range $[0, 1]$.

Note: Can be done in vectorized way without matrix multiplication with one line of code.

- Normalize per-variable (separately) may be appropriate if data variables have many different units (e.g. miles & cms)
  - ⇒ You lose relative scales among vars
- normalize globally (together) may be appropriate if variables have similar units (e.g. inches)
  - ⇒ won't distort relative scales among vars

2) Center data by subtracting each variable by its mean $\vec{\mu}$

$$A_c = A - \vec{\mu}$$

3) Compute covariance matrix: $\Sigma = \frac{1}{N-1} A_c.T \circledast A_c$

4) Compute eigenvectors P and eigenvalues via spectral decomposition of Σ — via numpy.

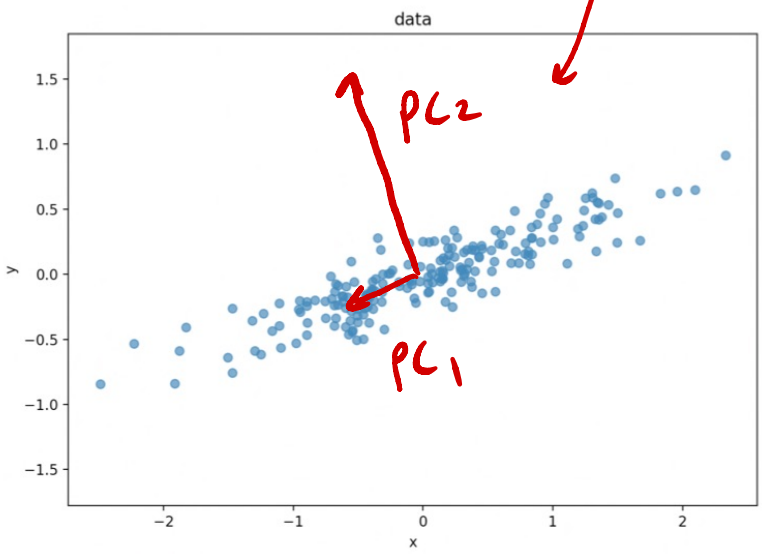
Shape: (M, M)

5) Sort e_vals high → low

e.g. $[3.2, 1.1, 0.03]$

Sort P (columns) in the same order otherwise PC lengths & directions get scrambled!

problem if we don't sort P & evals the same way!



6a) Compute proportion variance accounted for by each PC

e.g total = $3.2 + 1.1 + 0.03 = 4.33$

$\Rightarrow \left[3.2/4.33,\ 1.1/4.33,\ 0.03/4.33\right]$

6b) Compute cumulative variance accounted for by PC_i:

e.g $\left[3.2/4.33,\ (3.2+1.1)/4.33,\ (3.2+1.1+0.03)/4.33\right]$

6c) Threshold PCs based on desired amount of info/variance to keep. We keep K PCs.

e.g. ≥90%. If PC1 + PC2 = 90%, toss out PC3.

7) project data into PCA space (rotating data to align coordinate axes to PCs)

$\hat{P}$ is P with certain PCs/cols dropped

$$\underbrace{\hat{A}_C}_{(N,K)} = \underbrace{A_C}_{(N,M)} @ \underbrace{\hat{P}}_{(M,K)}$$

keep top K PCs

8) [optional] Reconstruct data [rotating data after projecting away insignificant PCs to align coordinate axes with original data variables]

if you didn't normalize:

$$\underbrace{A_{reconstut}}_{(N,M)} = \underbrace{\hat{A}_C}_{(N,K)} @ \underbrace{\hat{P}.T}_{(K,M)} + \underbrace{\vec{\mu}}_{(1,M)}$$

if you normalized:

original data range in each variable

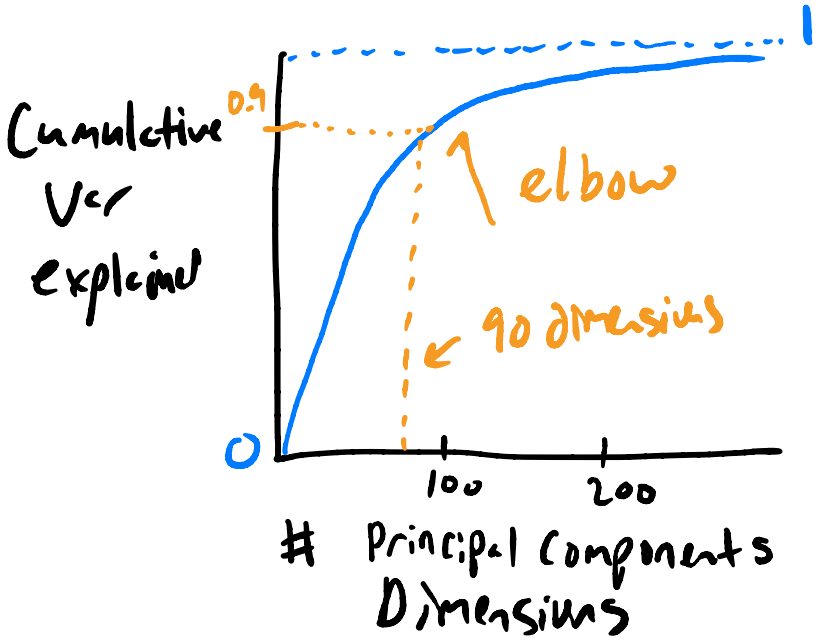
$$A_{reconstut} = \vec{S} \cdot \left(\hat{A}_C @ \hat{P}.T\right) + \vec{\mu}$$

Elbow plot

Step 6b was simple with 2D data, but generally get M values — Cumulative proportion variance explained by top K dimensions — $K \leq M$.

⇒ useful to plot to discover reasonable threshold on % variance retained.

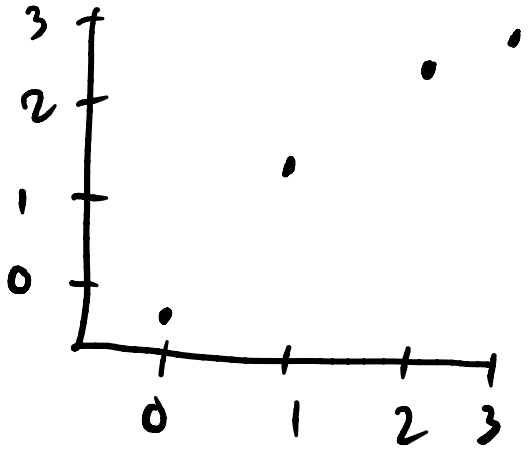
Elbow plot:



plot to determine point of "diminishing gains" visually — Cut off point of # PCs that explain good chunk of variance before plateauing

Example of PCA workflow by hand

X	Y
0	0.5
1	1.1
2	2.3
3	2.5



1) Normalize (skip — assume data homogeneous)

2) Center data. $\mu_x = \frac{0+1+2+3}{4} = \frac{6}{4} = \frac{3}{2} = 1.5$

$\vec{\mu} = (\mu_x, \mu_y)$ $\mu_y = \frac{0.5 + 1.1 + 2.3 + 2.5}{4} = \frac{6.4}{4} = 1.6$

X_c	Y_c
0 - 1.5	0.5 - 1.6
1 - 1.5	1.1 - 1.6
2 - 1.5	2.3 - 1.6
3 - 1.5	2.5 - 1.6

=

X_c	Y_c
-1.5	-1.1
-0.5	-0.5
0.5	0.7
1.5	0.9

3) Compute covariance matrix: $\frac{A_c^T A}{N-1}$

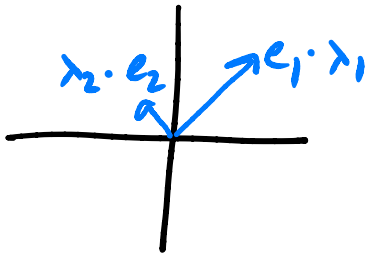
$$\frac{1}{3}\begin{bmatrix} -1.5 & -0.5 & 0.5 & 1.5 \\ -1.1 & -0.5 & 0.7 & 0.9 \end{bmatrix} \begin{bmatrix} -1.5 & -1.1 \\ -0.5 & -0.5 \\ 0.5 & 0.7 \\ 1.5 & 0.9 \end{bmatrix} = \begin{bmatrix} 1.67 & 1.2 \\ 1.2 & 0.92 \end{bmatrix}$$

4) Compute eigenvals/vecs of Σ.

eigenvalues: $= [\underbrace{2.55}_{\lambda_1}, \underbrace{0.04}_{\lambda_2}]$

eigenvectors: $P = \begin{bmatrix} 0.81 & -0.59 \\ 0.59 & 0.81 \end{bmatrix}$

$\vec{e}_1$ $\vec{e}_2$



unit vectors (length = 1)
b/c Σ is symmetric — see Spectral theorem coming soon

5) Sort eigenvalues high-to-low: $[2.55, 0.04]$

6) a) Compute proportion variance accounted for by each principal component (PC).

eigendimension

$$\text{prop\_var}_i = \frac{\lambda_i}{\sum_{j=1}^{M} \lambda_j} = \left[\frac{2.55}{2.55+0.04}, \frac{0.04}{2.55+0.04} \right]$$

$$= [0.98, 0.02]$$

$\vec{e}_1$ accounts for 98% of variance