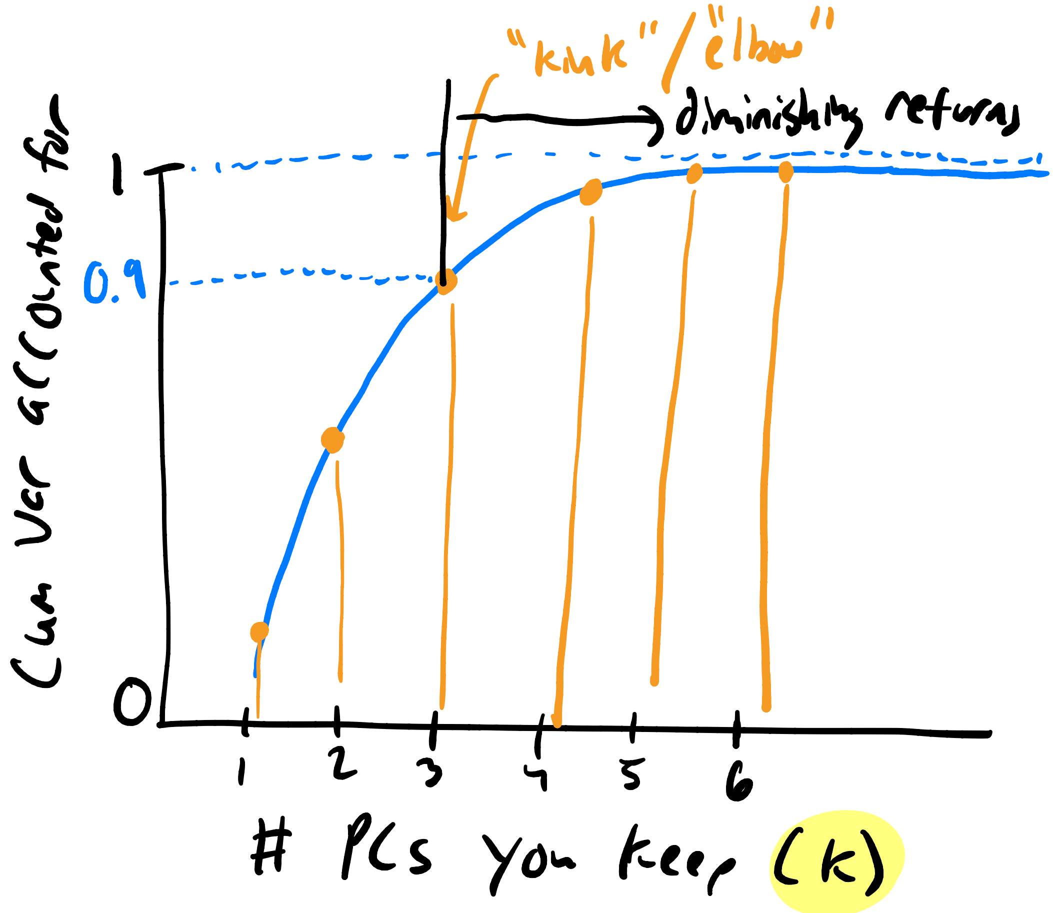
Lecture 20

elbow plot:

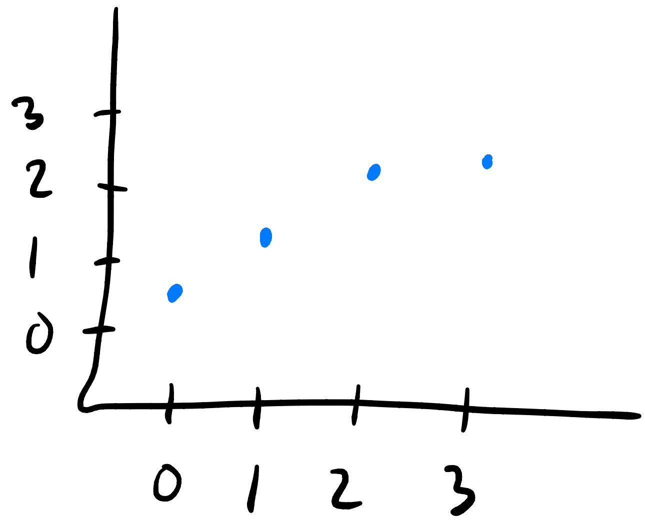


⇒ maybe 90% cutoff threshold in cum_prop_var is good for this dataset.

[illegible] Numerical example

$\boxed{A}$

X	Y
0	0.5
1	1.1
2	2.3
3	2.5



1) Normalize?

2) Center the data : $A_c = A - \vec{\mu}$

$\vec{\mu} = (\mu_x, \mu_y)$: $\mu_x = \dfrac{0+1+2+3}{4} = 1.5$

$\mu_y = \dfrac{0.5+1.1+2.3+2.5}{4} = 1.6$

X	Y
0	0.5
1	1.1
2	2.3
3	2.5

$- [1.5, 1.6]$

$= \boxed{A_c}$

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

3) Compute $\Sigma = \frac{1}{(N-1)} A^T @ A$

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

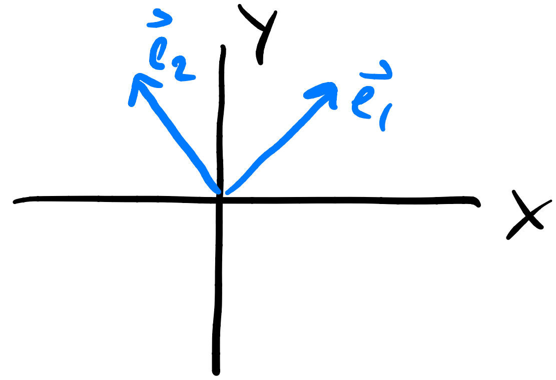
$$= \begin{bmatrix} 1.67 & 1.2 \\ 1.2 & 0.92 \end{bmatrix}$$

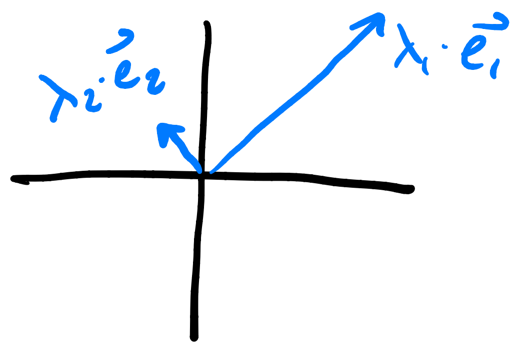
4) Compute eigenvalues + eigenvectors of Σ

$$\text{e\_vals} = \begin{array}{cc} \lambda_1 & \lambda_2 \\ \multicolumn{2}{c}{[\,2.55, \quad 0.04\,]} \end{array}$$

$$P = \begin{array}{cc} \vec{e}_1 & \vec{e}_2 \\ \multicolumn{2}{c}{\begin{bmatrix} 0.81 & -0.59 \\ 0.59 & 0.81 \end{bmatrix}} \end{array}$$

unit length





5) Sort e_vals and e_vecs high → low according to e_vals ⇒ already done

6) a) Compute prop_var by each PC

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

$$= [\,\underbrace{0.98}_{\text{PC1: 98\% total var}}, \underbrace{0.02}_{\text{PC2: 2\% total variance}}\,]$$

b) Compute cum_prop_var accounted for by top K PCs:

$$\text{Cum prop var} = \frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{M} \lambda_j} = \sum_{j=1}^{K} \text{prop\_var}_j$$

Variance retained

$$= \Big[\underbrace{0.98}_{K=1},\ \underbrace{\overbrace{0.98+0.02}^{1.0}}_{K=2}\Big]$$

threshold

c) Threshold. Example: Drop PC 2:

$$\hat{P} = P[:,0] = \underbrace{\begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix}}_{\vec{e}_1}$$

For this example, keep PC2:

$$\hat{P} = \underbrace{P}_{(M,M)}$$

7) Project data into PCA space

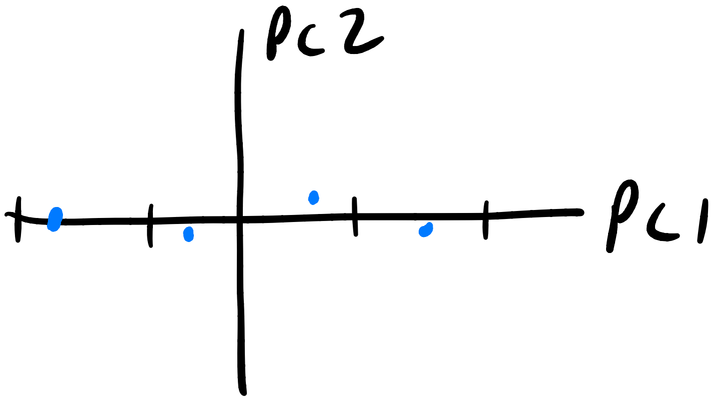
$$\underbrace{\hat{A}_c}_{(N,K)} = \underbrace{A_c}_{(N,M)} @ \underbrace{\hat{P}}_{(M,K)}$$

we ~~keep~~ keep K PCs

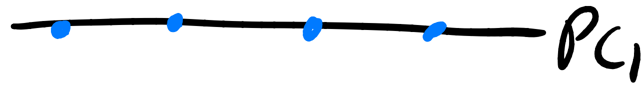
$K \leq M$

$$\underbrace{\hat{A}_c}_{(4,2)} = \underbrace{\begin{bmatrix} -1.5 & -1.1 \\ -0.5 & -0.5 \\ 0.5 & 0.7 \\ 1.5 & 0.9 \end{bmatrix}}_{(4,2)} \underbrace{\begin{bmatrix} 0.81 & -0.59 \\ 0.59 & 0.81 \end{bmatrix}}_{(2,2)}$$

$$= \begin{array}{cc} PC1 & PC2 \\ \end{array}$$
$$\begin{bmatrix} -1.86 & 0 \\ -0.7 & -0.1 \\ 0.82 & 0.27 \\ 1.74 & -0.16 \end{bmatrix}$$



if we dropped $\vec{e}_2$:



8) Reconstruct data: PCA space → data space

$$\underbrace{A_R}_{(N,M)} = \underbrace{\hat{A}_c}_{(N,K)} @ \underbrace{\overbrace{\hat{P}}^{(M,K)}.T}_{(K,M)} + \underbrace{\vec{\mu}}_{(1,M)}$$

$$\text{shape}(A_R) = \text{shape}(A)$$

if $\hat{P} = P$ [$K = M$, we did not drop any e_vecs/PCs]

P is orthogonal matrix

$$A_R = \hat{A}_c @ P.T + \vec{\mu}$$
$$= (A_c @ P) @ P.T + \vec{\mu}$$

$A_R = A_C + \vec{\mu}$] recover raw data!