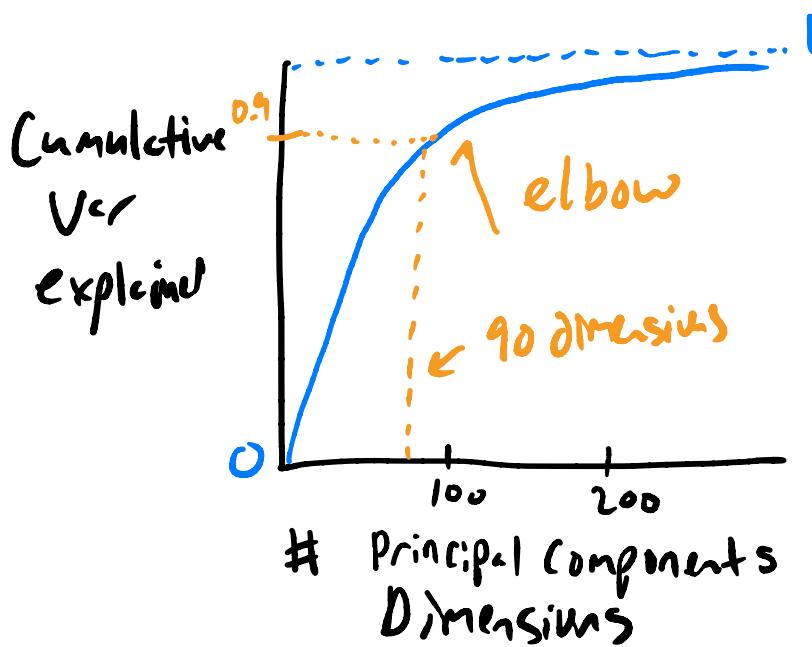


Elbow plot

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

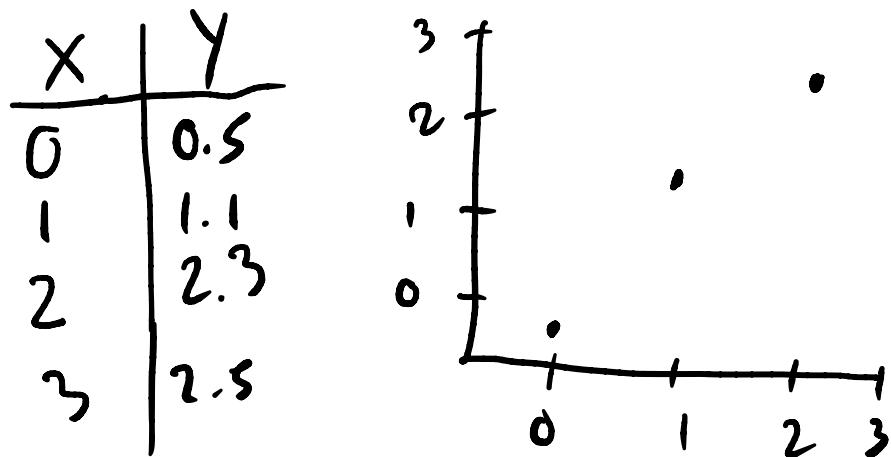
⇒ useful to plot to discover reasonable threshold on % Variance retained.

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



1) Normalize (skip — assume data homogeneous)

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

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

| 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 |

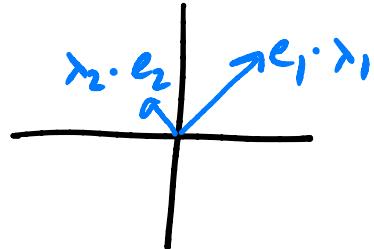
3) Compute covariance matrix : $\frac{A^T A}{N-1}$

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

4) Compute eigenvals/vecs of Σ .

Eigenvalues: $= \left[\frac{\lambda_1}{2.55}, \frac{\lambda_2}{0.04} \right]$

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



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).

eigenvalues

$$\text{prop-var}_i = \frac{\lambda_i}{\sum_{j=1}^m \lambda_j} = \left[\frac{2.55}{2.55+0.04}, \frac{0.04}{2.55+0.04} \right] = [0.98, 0.02]$$

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

5 b) Compute Cumulative Variance accounted for by top K PCs

$$\text{Cum Prop} = \frac{\sum_{j=1}^K \lambda_j}{\sum_{j=1}^M \lambda_j} = \sum_{j=1}^K \text{prop-Var-}K$$

1st PC accounts for 98% of Variance

$$= \begin{bmatrix} \frac{0.98}{0.98+0.02} \\ \frac{0.98+0.02}{0.98+0.02} \end{bmatrix} = \begin{bmatrix} 0.98 \\ 1.0 \end{bmatrix}$$

always should ≈ 1.0

$k=1$

$k=2$

top 2 PCs
Account for 100% of data Variance

6) Threshold out \vec{e}_i corresponding to low cum-prop-var:

$$\text{e.g. } \hat{P} = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix} \quad [\text{could threshold out } \vec{e}_2].$$

For this example, we won't threshold (i.e. $\hat{P} = P$).

7. Project data into PCA Space: $A_c = [\vec{x}_c \vec{y}_c]$

$$\hat{A}_c = A_c @ \hat{P}$$

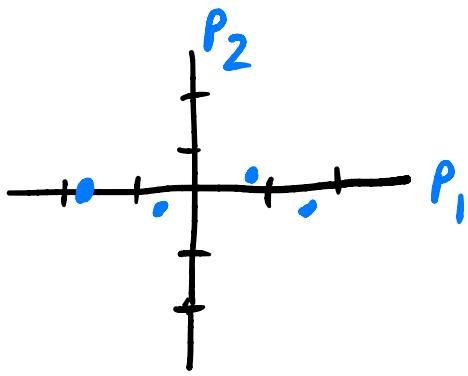
$\underbrace{(N, M)}_{\text{rows}}$ $\underbrace{(M, K)}_{\text{cols}}$ $K = \# \text{ surviving dimensions}$
 $K \leq M$

$$\hat{A}_c = \begin{bmatrix} -1.5 & -1.1 \\ -0.5 & -0.5 \\ 0.5 & 0.7 \\ 1.5 & 0.9 \end{bmatrix} \begin{bmatrix} 0.81 & -0.59 \\ 0.59 & 0.81 \end{bmatrix}$$

$\underbrace{4 \times 2}_{(N, M)}$ $\underbrace{2 \times 2}_{(M, K)}$

$$\hat{A}_c = \begin{bmatrix} -1.86 & 0 \\ -0.7 & -0.1 \\ 0.82 & 0.27 \\ 1.74 & -0.16 \end{bmatrix}$$

$\underbrace{}_{PC1}$ $\underbrace{}_{PC2}$



What would happen if we thresholded out \vec{e}_2 ?

- all dots on x axis (P_1 axis)

8. Reconstruct data from top K PCs.

$$A_R = \hat{A}_C @ \hat{P.T} + \vec{\mu}] \quad \begin{matrix} \text{Want: } A_R: (N, M) \\ = \text{shape}(A) \end{matrix}$$

$\underbrace{(N, K)}_{\text{rows}} \times \underbrace{(K, M)}_{\text{cols}} \quad \underbrace{(M,)}_{\text{cols}}$

- Note if $\hat{P} = P$ (no \vec{e}_i dropped) $\Rightarrow \hat{A}_C = A_C @ P$

$$A_R = A_C @ P @ P.T + \vec{\mu}$$

$\underbrace{\text{Orthogonal}}$
 $\Rightarrow I$

$$A_R = A_C + \vec{\mu}] \quad \begin{matrix} \text{recovers raw Data!} \\ \text{mean } \vec{0} - \text{centered} \end{matrix}$$

* Need to Scale $\hat{A}_C \hat{E} \hat{P} \hat{T}$ by original data range if we normalize (skip 1). Shape will look ok, but range will be incorrect.

• Show slides of reconstructed diamond data.