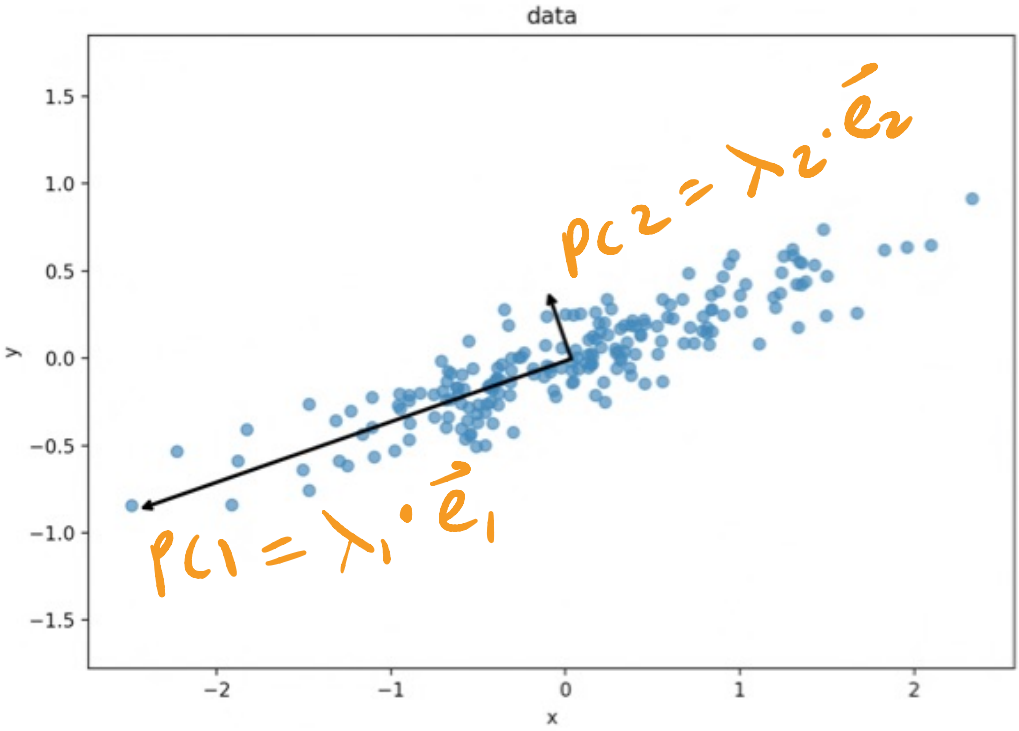


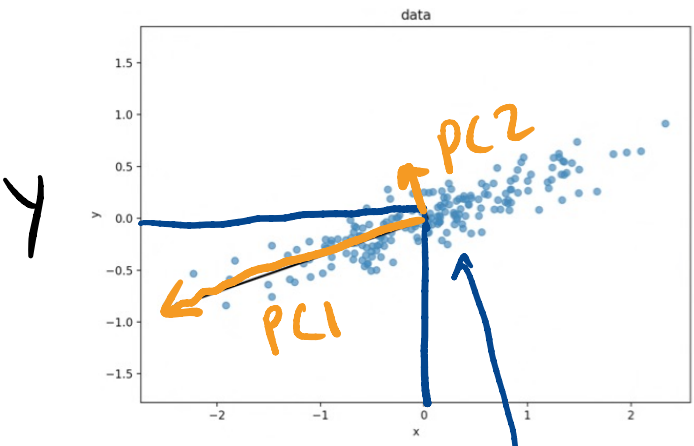
Here:

$\lambda_1 > \lambda_2$

Lecture 19

data space

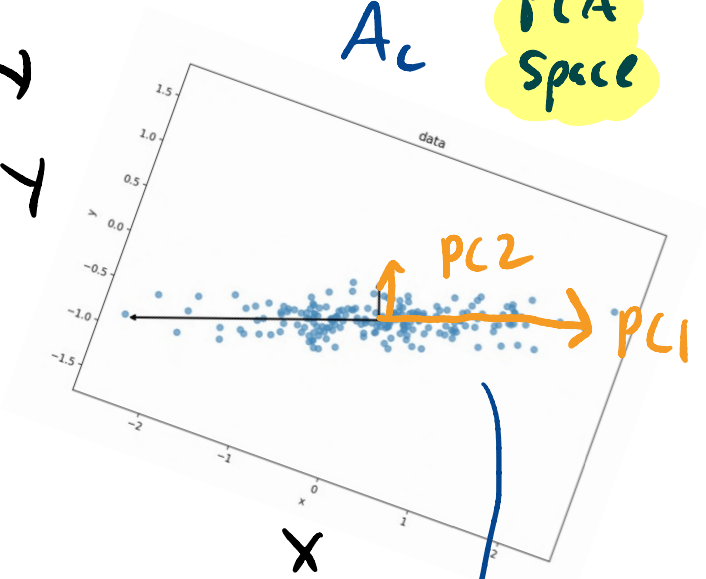
A_c



PCA: rotation

$\hat{A}_c$

PCA Space

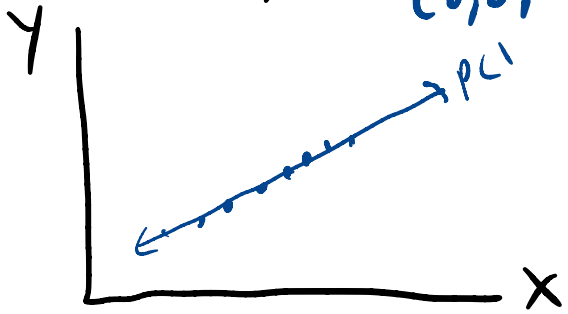


drop PC2:

$\hat{A}_c[:, 0]$

undo R + T

reconstruction



Data space → PCA space:

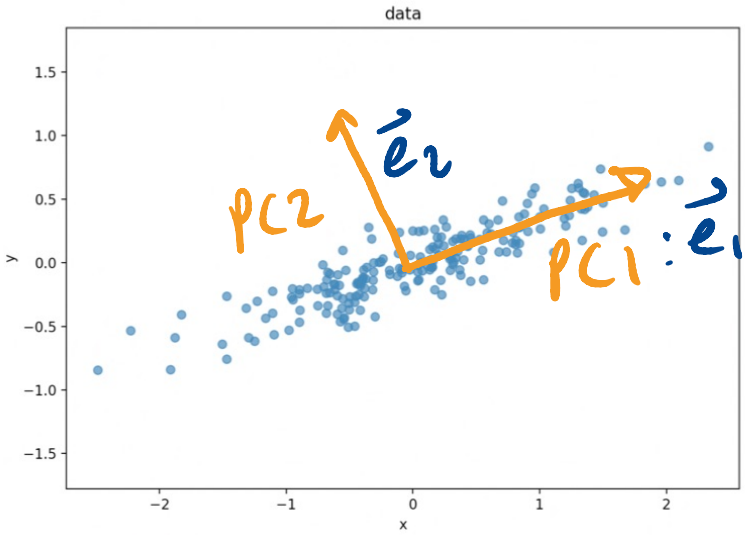
$$\hat{A}_c = (P^T @ A_c)^T$$

P: eigenvectors of Σ (covariance matrix) of the centered data A_c

eigenvectors ↓ columns of P

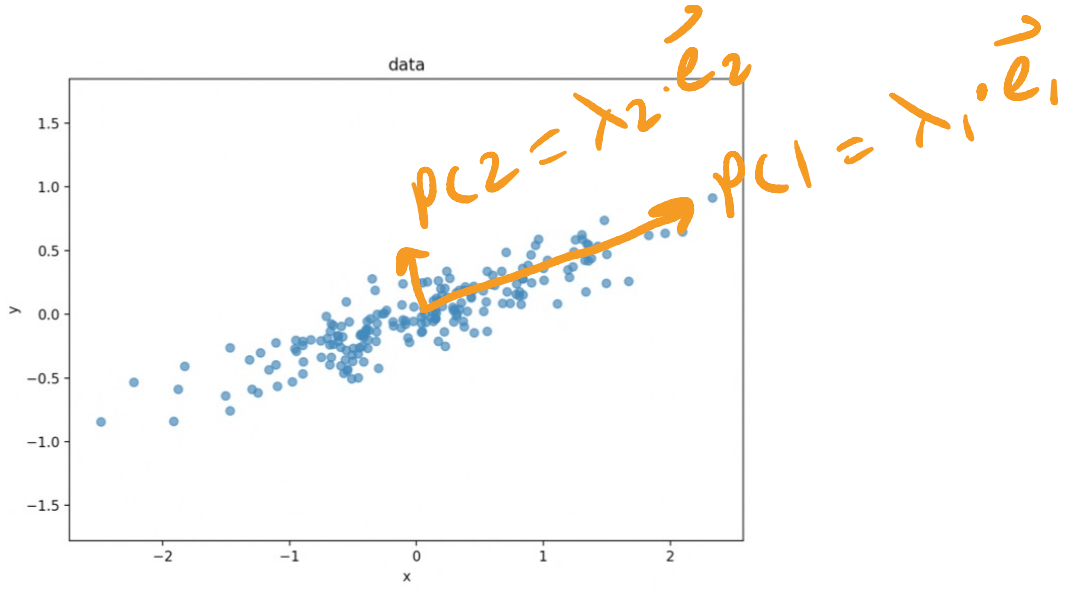
$$P = [\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_m]$$

all unit length (1)



Scalers ↓

Eigenvalues: $[\lambda_1, \lambda_2, \ldots, \lambda_m]$



In numpy:

P — orthogonal matrix

e_vals, e_vecs = np.linalg.eig(Σ)

Strategy for dropping insignificant PCs

M = 4

e_vals = [0.1, 3, 0.7, 5]

Small variance → maybe drop these PCs.

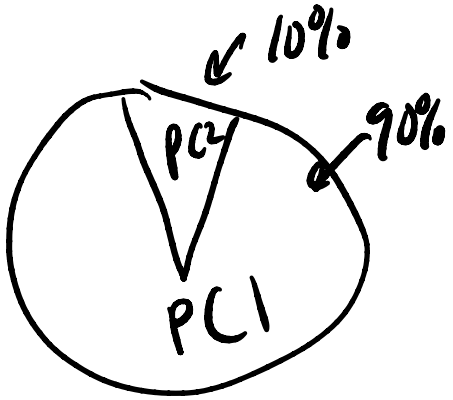
Sort e_vals list high-to-low:

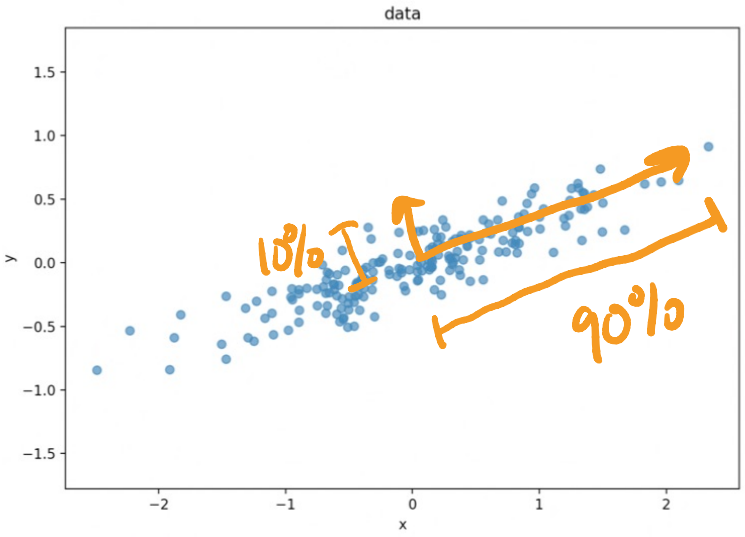
PC1 PC2 PC3 PC4

[5, 3, 0.7, 0.1]

& up to us to decide how many PCs to drop.

proportion of total variance accounted for by each PC:





PC1 PC2 PC3 PC4

$[5, 3, 0.7, 0.1]$

total variance $= 5 + 3 + 0.7 + 0.1 = \underline{8.8}$

prop_var $= [5/8.8, 3/8.8, 0.7/8.8, 0.1/8.8]$

57% 34% 8% 1%

Cumulative Variance accounted for by PCs:

e.g. Cut off / threshold: want to keep ≥ 90% data variance.

Cum_prop_var =

PC1 | PC1+PC2 | PC1+PC2+PC3 | all PCs

$$\left[5/8.8, \quad (5+3)/8.8, \quad (5+3+0.7)/8.8, \quad (5+3+0.7+0.1)/8.8 \right]$$

PC1 | PC1+2 | keep → | ← toss

$$= \left[0.57, \quad 0.9, \quad 0.99, \quad 1.0 \right]$$

PC3 | PC4

Summary of PCA

1) [optional] Normalize data.

- per-variable: useful (maybe) if data have vastly different units (e.g. miles, cm)
- global: same units, variables comparable. (e.g. miles, miles)

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

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

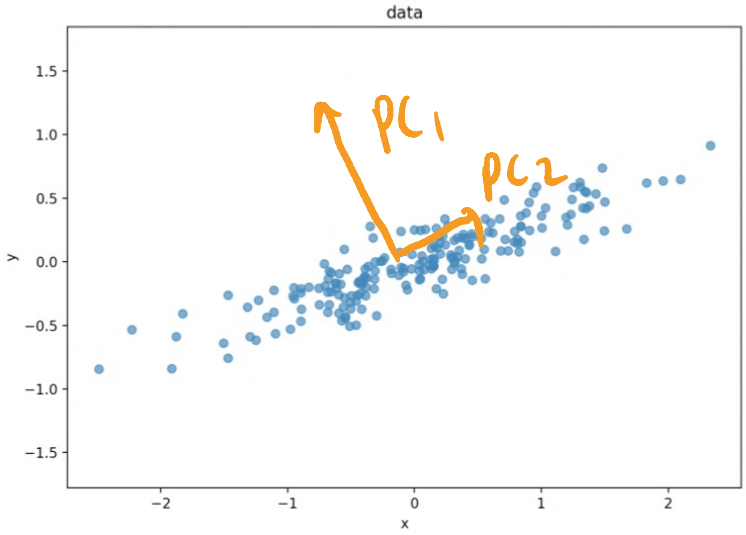
4) Compute e_vecs, e_vals — via numpy
   (e_vecs = P : shape: (M, M))

5) Sort high → low the e_vals

* also sort **columns** of e_vecs (P) in same way/order as the e_vals

problem (no sorting of PCs):

don't want this



6a) Compute prop-var by each PC.

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 cum-prop-var:

$\left[3.2/4.33 \;,\; (3.2+1.1)/4.33 \;,\; (3.2+1.1+0.03)/4.33 \right]$

6c) Threshold: drop some # of PCs to preserve X% of info/var in data

keep k PCs (after dropping)

7) project data to PCA space:

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

$\hat{P}$: $(M, M) \rightarrow (M, K)$ ← cols for top K PCs

A_c: shape: (N, M)

$\hat{A}_c$: (N, K) ← reduced dims $M \rightarrow K$

8) [optional] Reconstruct data. PCA space → data space

if no normalization:

$$A_{reconstruct} = \hat{A}_c @ \hat{P}.T + \vec{\mu}$$

$A_{reconstruct}$: (N, M)

$\hat{A}_c$: (N, K)

$\hat{P}$: (M, K), $\hat{P}.T$: (K, M)

$\hat{A}_c @ \hat{P}.T$: (N, M)

$\vec{\mu}$: uncenter data, $(1, M)$

if we normalized:

$$A_{reconstruct} = \vec{S}\left(\hat{A}_c \,@\, \hat{P}.T\right) + \vec{\mu}$$

range of original data