- How to compute covariance matrix for any data set?

entry (i,j)

$$\begin{bmatrix} \square & \\ & \end{bmatrix} = \Sigma$$

$\boxed{\text{Scalar equation}}$

$$\Sigma(i,j) = \sum_{k=1}^{N} \frac{(x_{ki} - \mu_i)(x_{kj} - \mu_j)}{N-1}$$

$x: (N,M)$

diagonals:

$$\Sigma(i,i) = \sum_{k=1}^{N} \frac{(x_{ki} - \mu_i)(x_{ki} - \mu_i)}{N-1}$$

$$= \frac{(x_{ki} - \mu_i)^2}{N-1}$$

$$= \sigma_i^2 \Rightarrow \text{diagonal entries are ordinary variances}$$

$\Sigma(x,y)$

2D example: $$\Sigma = \begin{bmatrix} \underbrace{\text{cov}(x,x)}_{\sigma_x^2} & \text{cov}(x,y) \\ \text{cov}(y,x) & \underbrace{\text{cov}(y,y)} \end{bmatrix}$$

$\text{cov}(y,y) = \sigma_y^2$

vectorized equation:

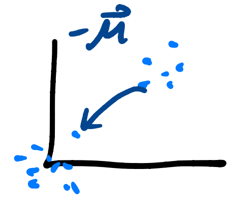
$$\Sigma = \frac{1}{N-1}\sum_{k=1}^{N} \underbrace{(\vec{x}_k - \vec{\mu})^T}_{\text{shape: }(1,M)} \underbrace{(\vec{x}_k - \vec{\mu})}_{\text{shape: }(M,1)}$$

shape = (M, 1)

e.g. M=3

$\vec{\mu} = (1, 2, 3)$

Centers data so that $\vec{x}_k$ has $\vec{0}$ mean:



Call centered data matrix A_c : A_c shape = (N, M)

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

$$A \; \begin{array}{ccc} 1 & 2 & 3 \\ \end{array} \left[\begin{array}{ccc} \vdots & \vdots & \vdots \end{array}\right] - [\mu_1, \mu_2, \mu_3] = A_c \left[\begin{array}{ccc} \vdots & \vdots & \vdots \end{array}\right]$$

$$A \; \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \end{array} - [2, 5, 8] = A_c \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

if data centered:

[matrix equation]

$$\boxed{\Sigma = \frac{1}{N-1} A_c.T @ A_c}$$

$\vec{x}_i$

Example:

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 10 & 2 & 1 \\ 3 & 4 & 5 \\ 3 & 2 & 1 \end{bmatrix}$$

$$- \left(\mu_1, \mu_2, \mu_3\right)$$

$$A_c = \begin{bmatrix} -3.25 & -1/2 & -2.5 \\ 5.75 & -1/2 & -0.5 \\ -1.25 & 1.5 & 3.5 \\ -1.25 & -1/2 & -0.5 \end{bmatrix}$$

$$\vec{\mu} = \left(\frac{1+10+3+3}{4}, \frac{2+2+4+2}{4}, \frac{-1+1+5+1}{4}\right) = (4.25, 2.5, 1.5)$$

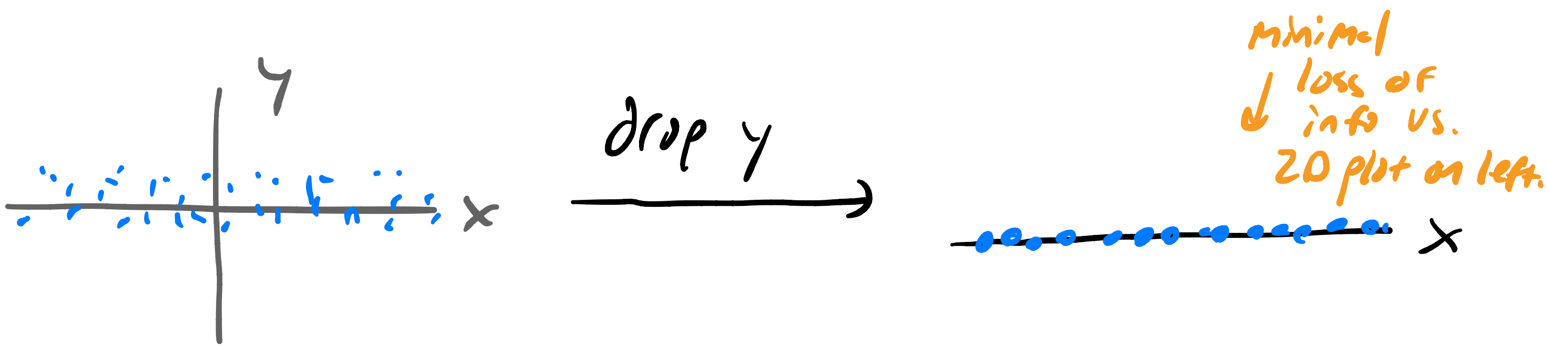
$$\frac{1}{N-1} \overset{(3,4)}{A_c^T} \overset{(4,3)}{A} \longrightarrow (3,3)$$

$$\frac{1}{3} \begin{bmatrix} -3.25 & 5.75 & -1.25 & -1.25 \\ -1/2 & -1/2 & 1.5 & -1/2 \\ -2.5 & -0.5 & 3.5 & -0.5 \end{bmatrix} \begin{bmatrix} -3.25 & -1/2 & -2.5 \\ 5.75 & -1/2 & -0.5 \\ -1.25 & 1.5 & 3.5 \\ -1.25 & -1/2 & -0.5 \end{bmatrix}$$

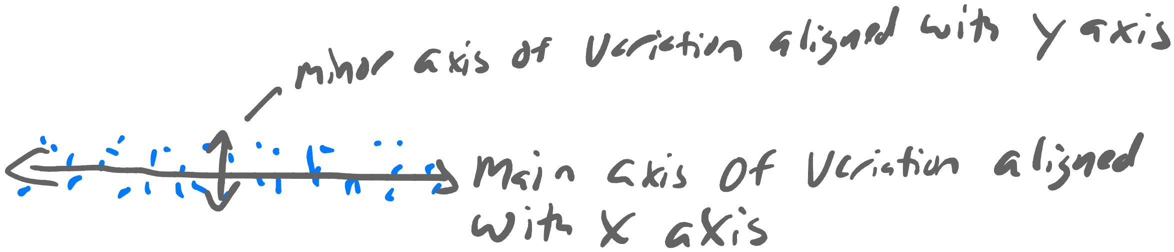
$$\Sigma = \begin{bmatrix} 15.8 & -0.83 & 0.5 \\ -0.83 & 1 & 2.3 \\ 0.5 & 2.3 & 6.3 \end{bmatrix}$$

Principal Component Analysis (PCA)

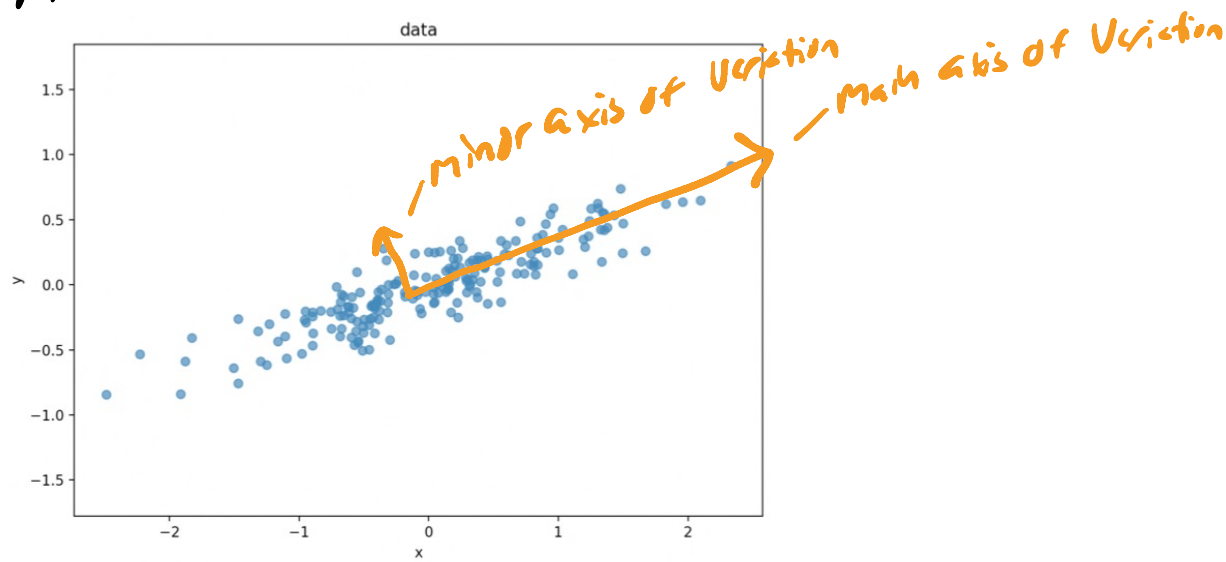
Recall the example where we could toss out a variable (y) without lossing much information:



In this case, the main axes of variation of data are aligned with the cordinate axes



⇒ clearly this is not true in general! Axes of variation do not need to correspond to variables in dataset e.g. x, y, etc:

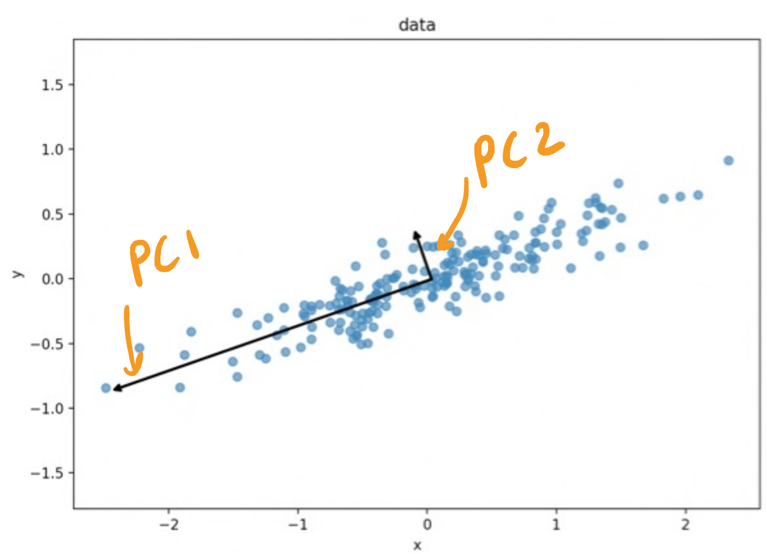


Above main variation axis != x axis!
minor variation axis != y axis!

We call these intrinsic axes of variation in data
principal components (PCs)

⇒ we number them <u>in order</u> based on the amount of variation along each intrinsic axis

⇒ larger variation → smaller number.



Goal of PCA Reduce dimensionality of dataset (e.g. $M = 500 \to 30$) by dropping intrinsic axes with little variation (least amount of info lost)

e.g. PC2 above, y variable in