Dimensionality reduction

2D, 3D : Scatter plot

4D, 5D, 6D : pair plot   784-dim data

10-D, 100-D, 1000-D - :

nD ⟶ 3D or 2D

(PCA)

Principle component Analysis & t-SNE

Row Vector & Column Vector

flower: $\underbrace{[SL, PL, SW, PW]}_{\text{real values}}$

ith Point:

$x_i \in \mathbb{R}^{d}$ → d-dim column vector

$\mathbb{R}$ → real numbers

x_i → column vector

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{id} \end{bmatrix}_{d \times 1} : \text{column vector}$$

$x_i = [2.1, 3.2, 4.6, 1.2]_{1 \times 4}$: row vector

Dataset

$D : \{x_i, y_i\}_{i=1}^{n}$ → data points

$x_i \in \mathbb{R}^d$; $x_i \in \mathbb{R}^4$

$y_i \in \{$setosa, versicolor, virginica$\}$

Dataset as a data-matrix:

$$X = \begin{array}{c|cccc|} & f_1 & f_2 & f_3 & \cdots & f_d \\ 1 & & & & & \\ 2 & & & & & \\ 3 & & & & & \\ \vdots & & & \longleftarrow x_i^T \longrightarrow & & \\ n & & & & & \end{array}_{n \times d}$$

$\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$

$x_i \in \mathbb{R}^d$

$y_i \in \{s, v, ve\}$

d-features

each data point = row

each column = feature

$$X = \begin{array}{c|cccccc|} & 1 & 2 & 3 & 4 & \cdots\, i \,\cdots & n \\ f_1 & & & & & \uparrow & \\ f_2 & & & & & & \\ f_3 & & & & & x_i & \\ f_j & & & & & & \\ f_d & & & & & \downarrow & \end{array}_{d \times n}$$

each column = data point

row = feature

f_1 = PL

f_2 = PW

f_3 = SL

f_4 = SW

Data pre-processing: column-normalization

$$X = \begin{array}{c|cccccc|} & f_1 & f_2 & f_3 & \cdots f_j \cdots & f_d \\ x_1 & & & & & \\ x_2 & & & & & \\ \vdots & & & & & \\ x_i & & \longleftarrow & x_i^T & \longrightarrow & \\ \vdots & & & & & \\ x_n & & & & & \end{array}_{n \times d} \rightarrow \quad Y = \begin{bmatrix} \\ y_i \\ \\ \end{bmatrix}_{n \times 1}$$

obtain data → pre-processing → data modelling (dim reduction)

→ column norm

$$
X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ 150 = n \end{array}
\begin{bmatrix}
f_1 & f_2 & f_3 & \cdots & f_j & \cdots & f_d \\
 & & b_1 & & a_1 & & \\
 & & b_2 & & a_2 & & \\
 & & \vdots & & a_3 & & \\
 & & & & \vdots & & \\
\longleftarrow & & x_i^T & & a_i & & \longrightarrow \\
 & & \vdots & & \vdots & & \\
 & & b_n & & a_n & &
\end{bmatrix}_{n \times d}
\quad
Y = \begin{bmatrix} \\ \\ \\ y_i \\ \\ \\ \end{bmatrix}_{n \times 1}
$$

Column normalization

column: $a_1, a_2, \ldots\ldots a_n \rightarrow$ n values of feature f_j

$Max(a_i) = \cancel{max} (a_{max}) \geq a_i \quad (i: 1 - n)$

$min(a_i) = a_{min} \leq a_i \quad (i: 1 \rightarrow n)$

$$\boxed{a_i' = \frac{a_i - a_{min}}{a_{max} - a_{min}}} \qquad a_i' \in [0,1]$$

$a_1', a_2', a_3' \ldots\ldots a_i' \ldots\ldots a_n'$

$a_1, a_2, \ldots\ldots a_i \ldots\ldots a_d$

$\downarrow$ coulumn - noumalization

$a_1', a_2', \ldots\ldots a_i' \ldots\ldots a_d'$ st $a_i' \in [0,1]$

why?

	$f_1 = h$	$W = f_2$	
	162	56	student 1
	172	72	2
	182	84	⋮
	⋮	⋮	n
	↑ cm	↑ kg	

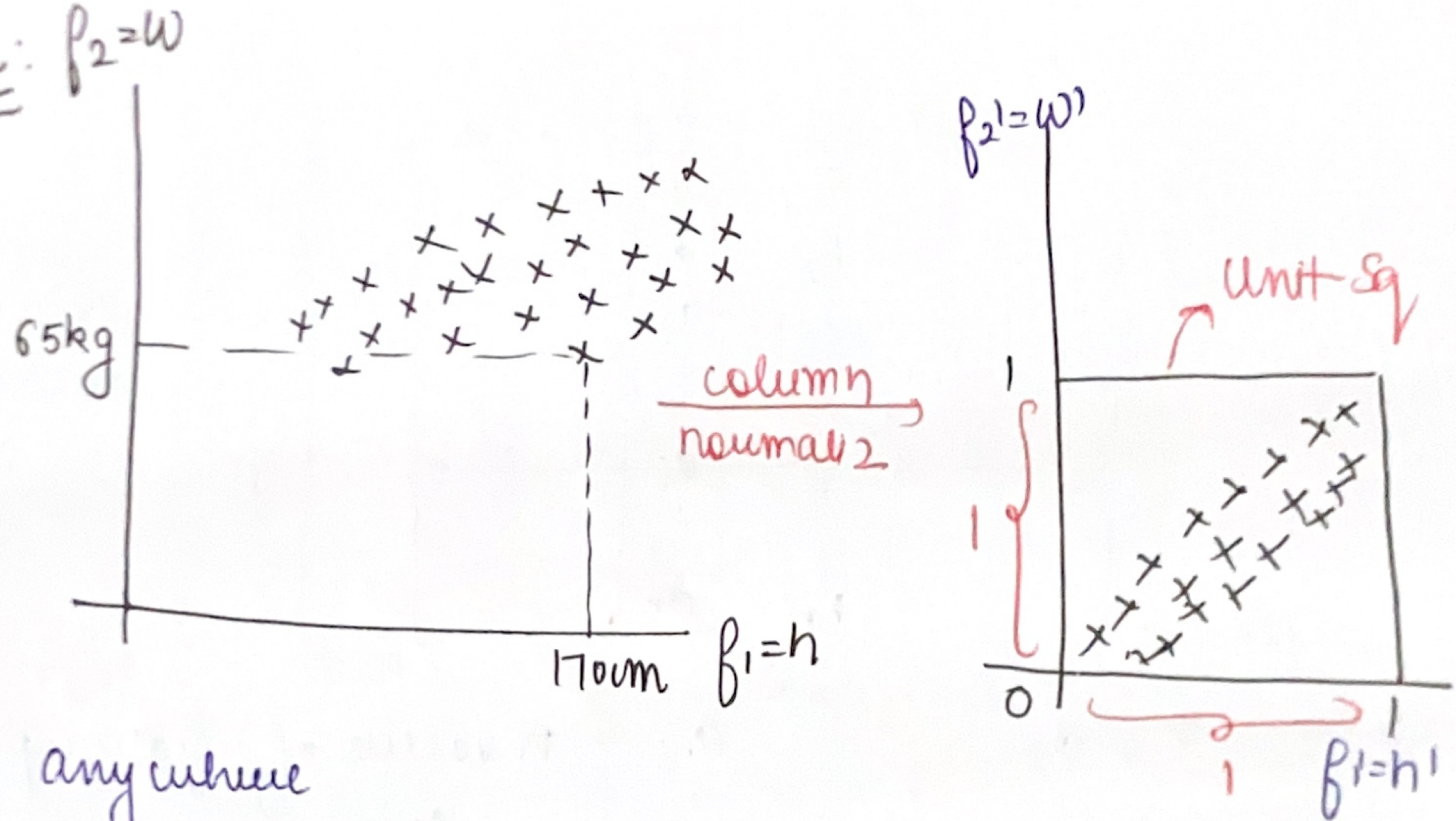
colun - nou

getting rid of scale (sme scale)

	h_i'	w'
1		
2		
3		
	↑ [0,1]	↑ [0,1]

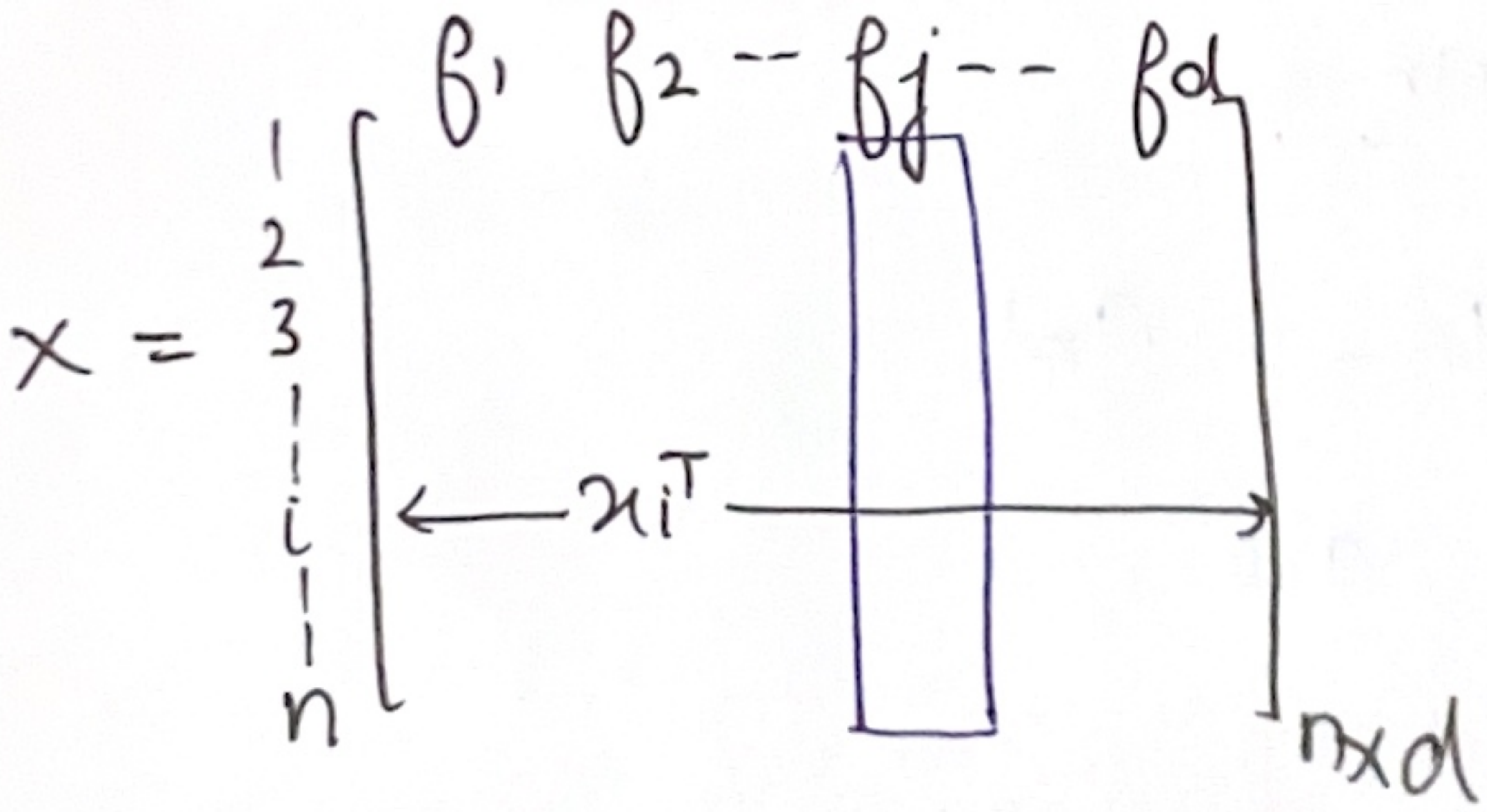
getting all points in one statand foumat $\in [0,1]$

Geom:



anywhere n-dim space $\xrightarrow[\text{norma}]{\text{col}}$ unit hyp cube in n-dim space

Mean Vector :



Let $x_i \in \mathbb{R}^d$

$x_1 = [2.2, 4.2] \in \mathbb{R}^3$

$x_2 = [1.2, 3.2] \in \mathbb{R}^2$

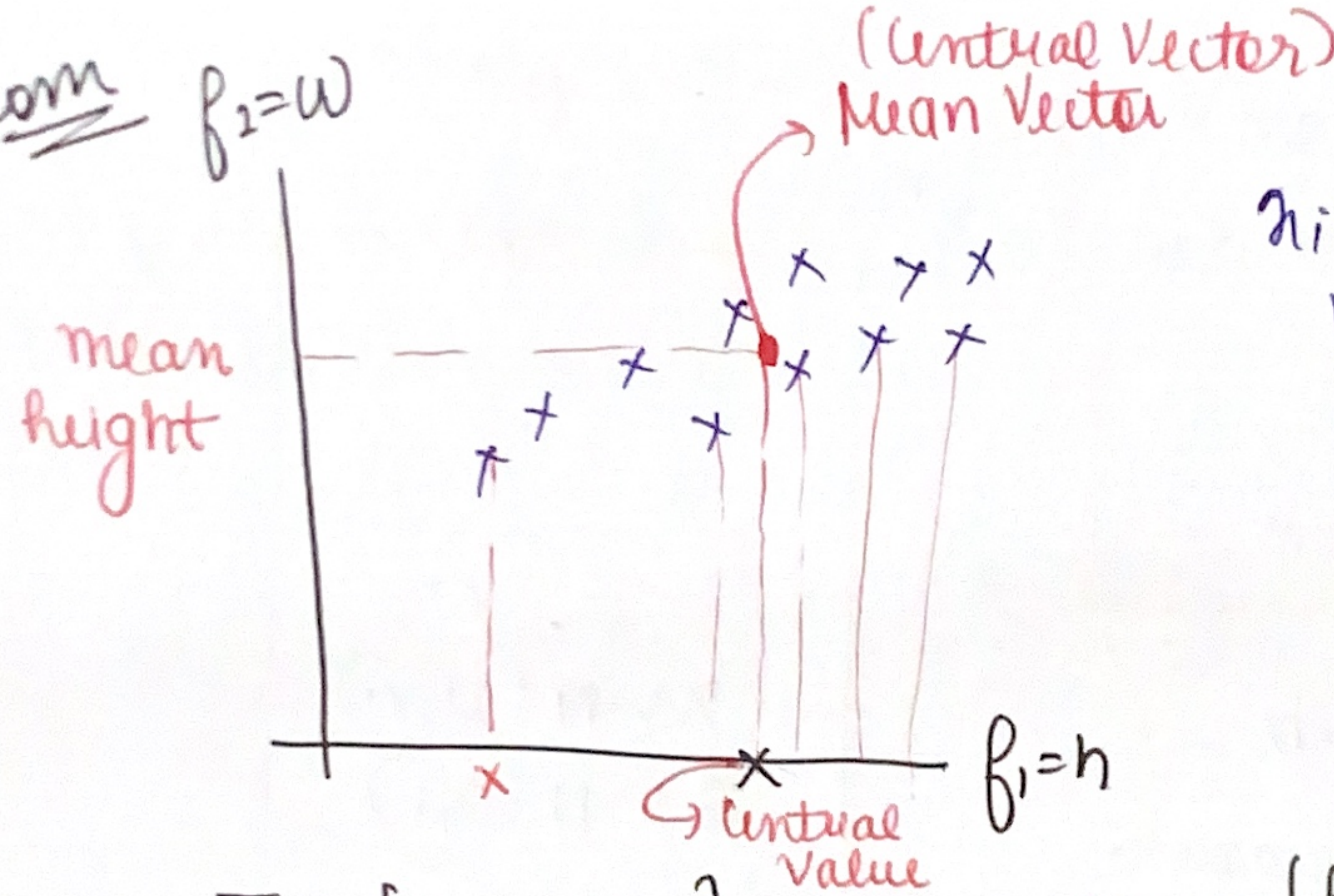
$x_1 + x_2 = [3.4, 7.4]$

$x_i \in R_d$

$\bar{x} \in \mathbb{R}^d$ $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 - - - - + x_n)$

Mean Vector

Geom



$x_i \in \mathbb{R}^2$

$[h, w]$
$\ \ f_1 \quad f_2$

$$\bar{x} = [h_{\bar{x}}, w_{\bar{x}}] \qquad h_{\bar{x}} = \text{mean}(h_i)_{i=1}^{n}$$

$$w_{\bar{x}} = \text{mean}(w_i)_{i=1}^{n}$$

Data-preprocessing: Column standardization

column normalization: $[0,1]$ ← get rid of scales of each feature

column standarization: more often used

$$X = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} f_1 \quad f_2 \cdots f_j \cdots f_d \\ \left[\begin{array}{ccc} & \boxed{\begin{array}{c} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_n \end{array}} & \\ \leftarrow x_i^T & & \rightarrow \end{array}\right]_{n \times d} \end{array}$$

$a_1, a_2, a_3 \cdots a_n$ ← n values of f_j

$\big\downarrow$ col. std

$a_1', a_2', a_3' \cdots a_n'$ ← $\text{mean}\{a_i'\}_{i=1}^{n} = 0$

$\text{std dev}\{a_i'\}_{i=1}^{n} = 1$

$\bar{a} = \text{Mean}\{a_i\}_{i=1}^{n}$ ← sample mean

$s = \text{std dev}\{a_i\}_{i=1}^{n}$ ← sample std-dev

$$\boxed{a_i' = \frac{a_i - \bar{a}}{s}}$$

any distb

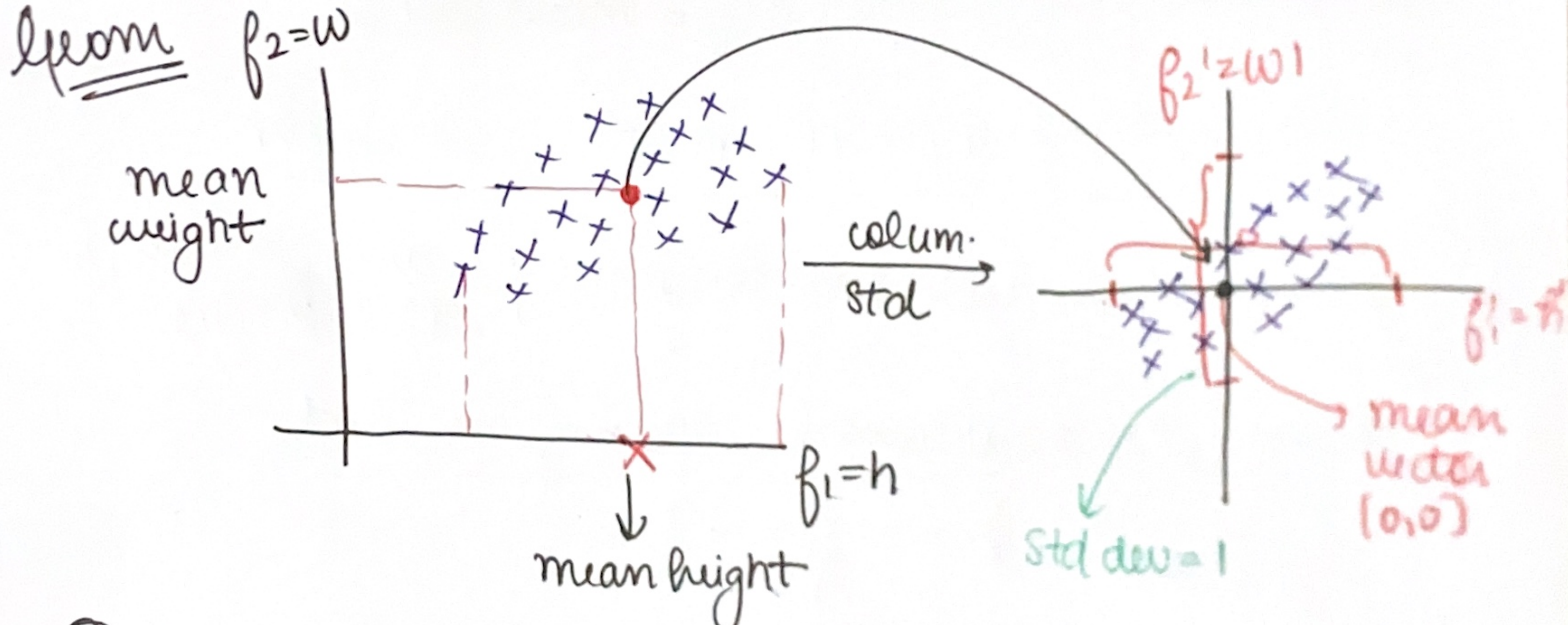
a_i's → mean = 0
→ std dev = 1

Std-normal Variate (z)

$$z = \frac{x - \mu}{\sigma}$$

$x \sim N(\mu, \sigma)$

$z \sim N(0,1)$

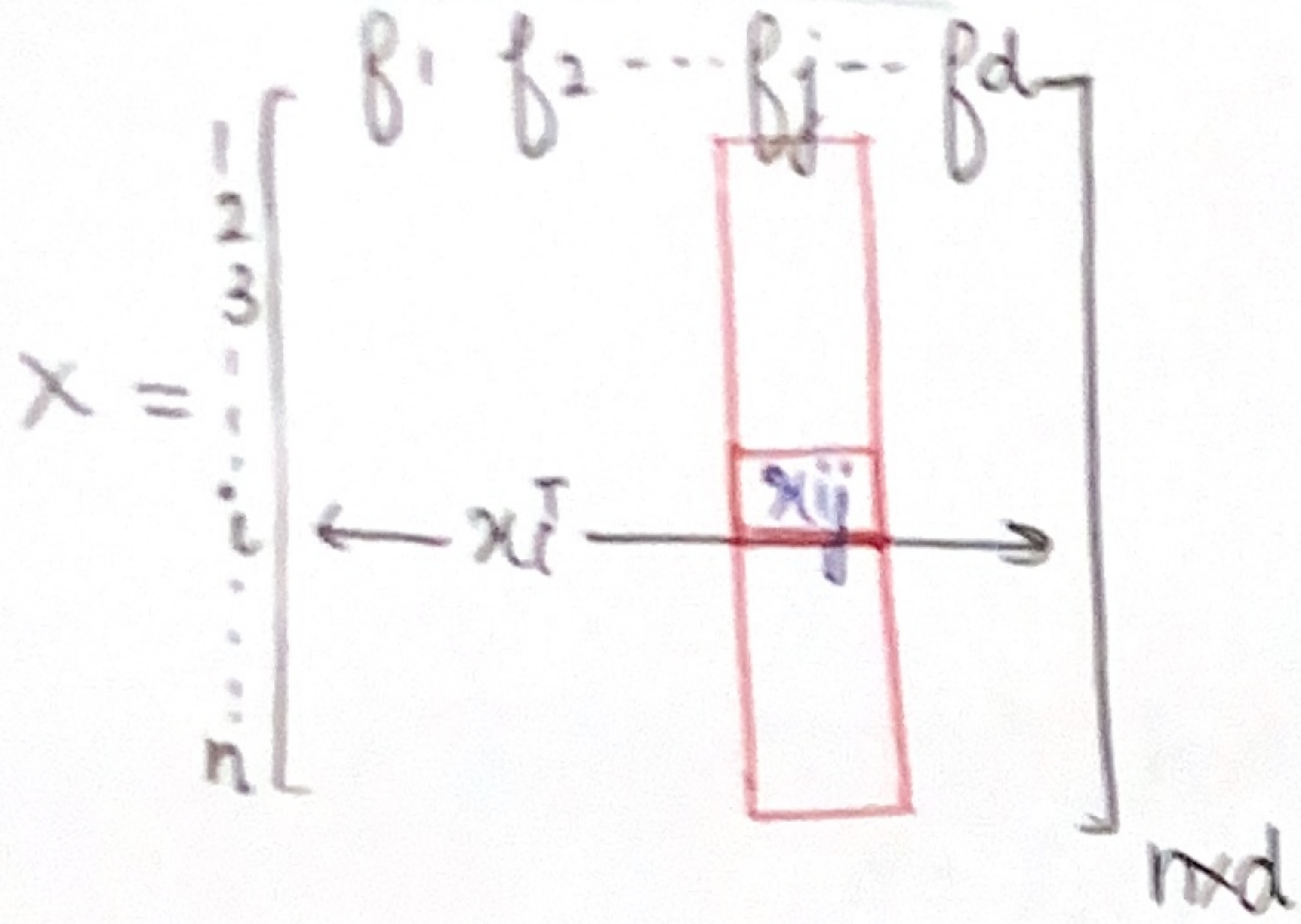


① Moving mean vector to origin

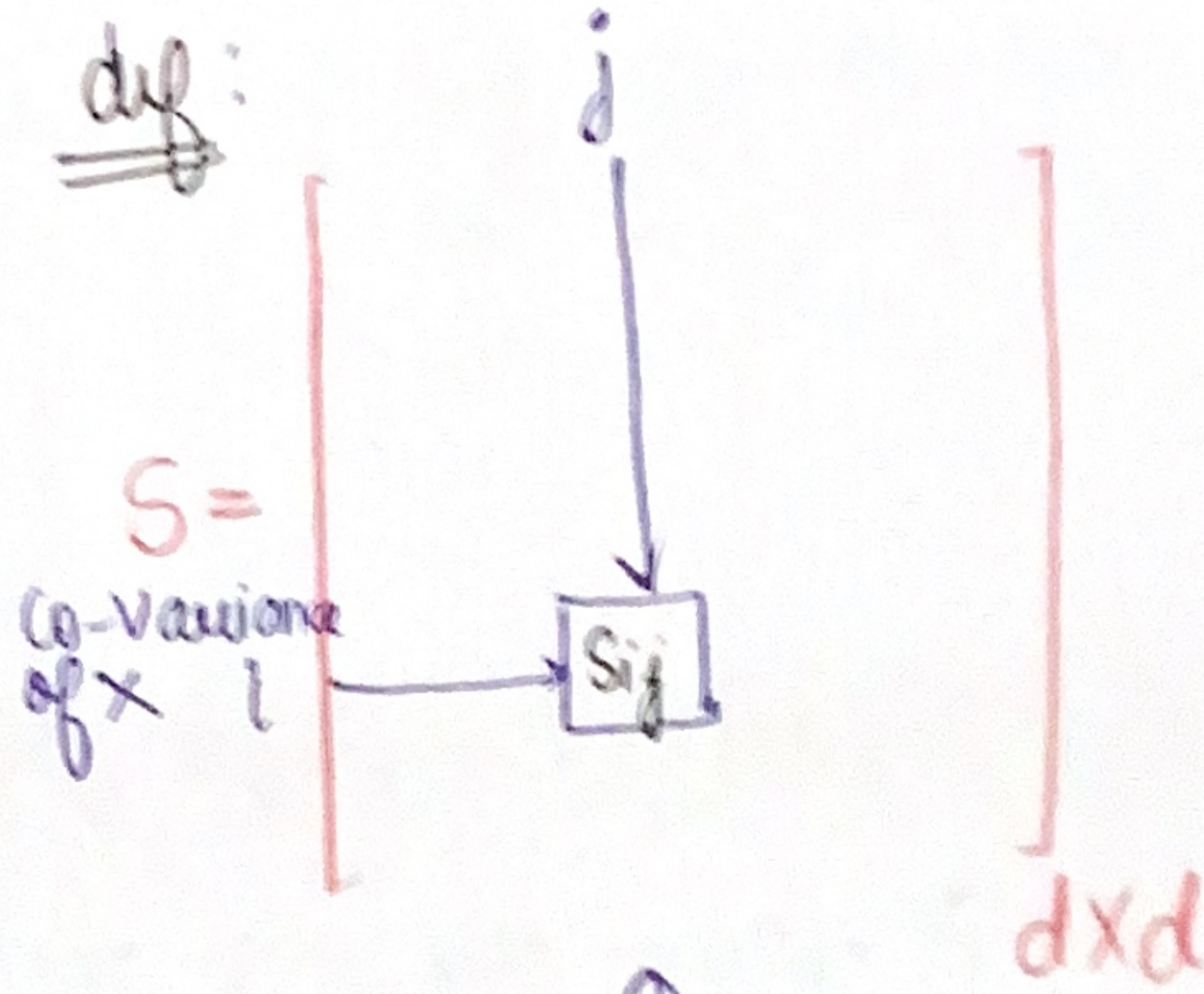
② squishing / expanding s.t std-dev for any feature is 1

Col. Std = mean centering → origin
scaling → std dev = 1 for all features

Co-Variance Matrix



def:



↑ square-matrix

S_{ij} = i^{th} row & j^{th} col element in S

$$S_{ij} = cov(f_i, f_j) \qquad i: 1 \to d, \quad j: 1 \to d$$

f_j = column vector j^{th} feature

x_{ij} = j^{th} feature for i^{th} data point

$$\boxed{\underline{\underline{cov}}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}$$

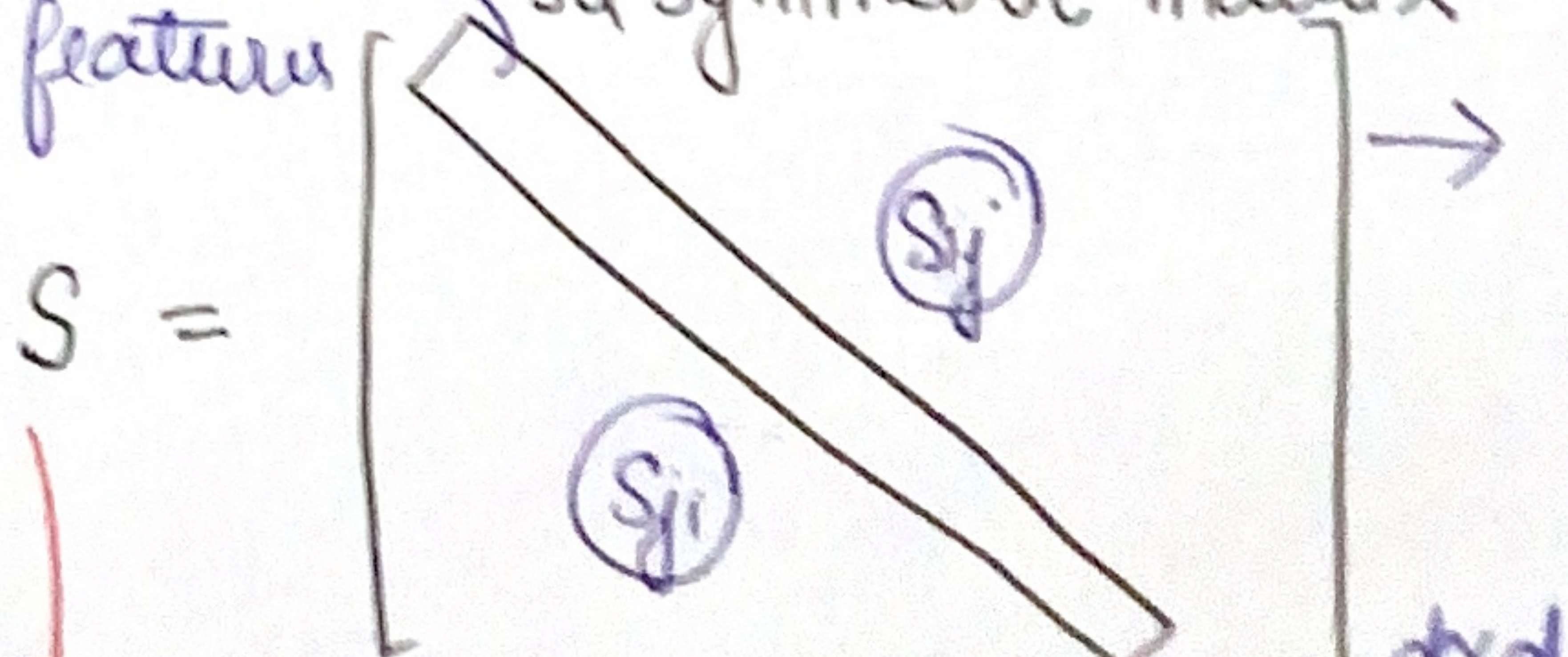
$$cov(f_i, f_j) = var(f_i)$$

$$cov(x, x) = var(x) \quad \text{—— ①}$$

$$cov(f_i, f_j) = cov(f_j, f_i) \quad \text{—— ②}$$

Variance of features

SQ symmetric matrix



→ Symmetric matrix

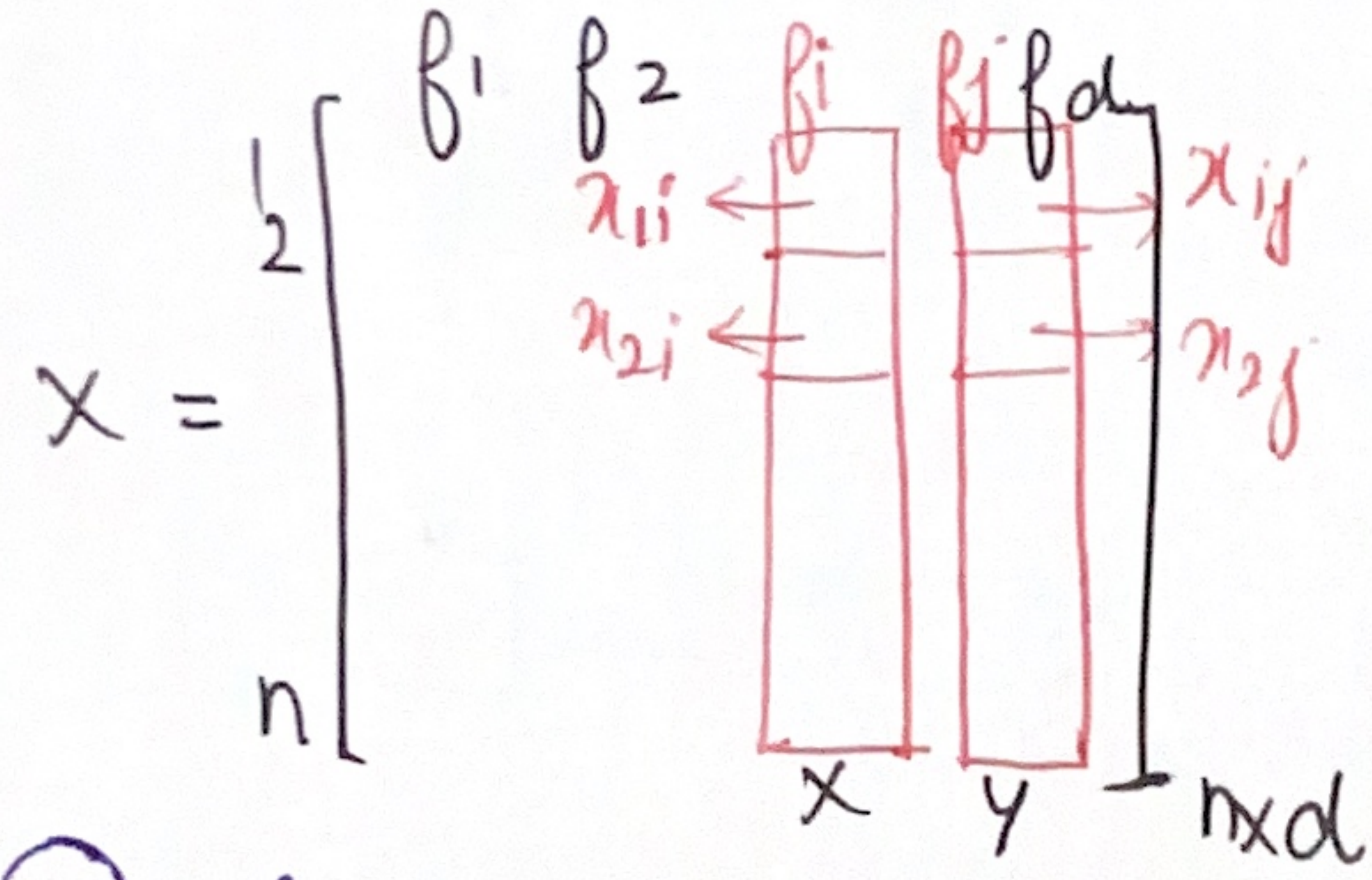
$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 2 & 1 & 2 \\ 2 & 1 & 1 & 5 \\ 3 & 2 & 5 & 3 \end{array}_{3\times 3}$$

$$A_{21} = A_{12}$$

$$A_{ij} = A_{ji} \ \forall \ i,j$$

Symmetric Matrix & Square Matrix (d, d)

f_1 f_2



let X colum std $\rightarrow$ mean $\{f_i\} = 0$

std dev $\{f_i\} = 1$

$$\boxed{\underset{(1)}{cov}(f_i, \underset{(2)}{f_j}) = \frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\mu_1)(x_{i2}-\mu_2)}$$

mean(f_2)

mean(f_1)

$$\boxed{cov(f_1, f_2) = \frac{1}{n}\sum_{i=1}^{n} x_{i1} * x_{i2}}$$

if f_1 & f_2 are std,

then $cov(f_1, f_2) = \frac{f_1^T f_2}{n}$

$$S_{d\times d} = \underset{d\times n}{(X^T)}\underset{n\times d}{(X)} = d\times d$$

* assuming $X \rightarrow$ <u>cold std</u>

$$S_{ij} = cov(f_i, f_j) = \frac{f_i^T f_j}{n}$$

MNIST Dataset

obj: classify the written character into one of 10 numeric characters

$$\mathcal{D} = \{x_i, y_i\}_{i=1}^{60K}$$

$x_i =$ [0] 28 × 28

$y_i \in \{0,1,2,3,4,5,6,7,8,9\}$

$x_i =$ [ ] 28 × 28 $\longrightarrow$ $x_i = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$

$x_i \in \mathbb{R}^d$

$x_i =$ image $= \begin{bmatrix} & \\ & \end{bmatrix}_{28\times 28}$ → NOT data matrix, matrix represen of image

numerical / real matrix

	1	2	3	4	5
1	1	2	4	6	8
2	3	2	1	8	2
3	2	1	6	8	4
4	3	2	1	8	2
5	4	2	6	8	1

5×5

row flattening →

$\begin{bmatrix} 1 \\ 2 \\ 4 \\ 6 \\ 8 \\ 3 \\ 2 \\ 1 \\ 8 \\ \vdots \end{bmatrix}_{25\times 1}$

[ ] 28 × 28 (←28→, ↑28↓)

data point

row flattening →

$28\times 28 = 784$

$\begin{bmatrix} \\ \\ \\ \end{bmatrix}_{784\times 1}$

MNIST

$$X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ 60k \end{array} \begin{array}{c} f_1 \quad f_2 \quad f_3 \cdots f_{784} \\ \left[\begin{array}{c} \\ x_i^T \in \mathbb{R}^{784} \\ \longleftarrow 784 \text{ dim} \longrightarrow \\ \\ \end{array} \right]_{n \times d} \end{array} \quad \left[\begin{array}{c} \\ y_i \in \{0, 1, 2 \cdots 9\} \\ \\ \end{array} \right]_{n \times 1}$$

$n = 60k$

$d = 784$