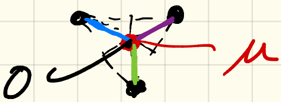
Data: Probabilistic View

9/24

mean of data matrix is the average of our points

$$\text{mean}(D) = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

mean as **center of mass**



total variare avg sq dist of each pt from mean

$$\text{var}(D) = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \mu\|^2$$

rewrite as (see book for details)

$$\frac{1}{n}\left(\sum_{i=1}^{n} \|x_i\|^2\right) - \|\mu\|^2$$

Center data matrix

move origin to mean
e.g. subtract mean from each point

$\mathbb{1} \in \mathbb{R}^n$ vect of all 1's

$$\bar{D} = D - \mathbb{1} \cdot \mu^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} - \begin{pmatrix} \mu^T \\ \mu^T \\ \vdots \\ \mu^T \end{pmatrix} = \begin{pmatrix} x_1^T - \mu^T \\ x_2^T - \mu^T \\ \vdots \\ x_n^T - \mu^T \end{pmatrix} = \begin{pmatrix} \bar{x}_1^T \\ \bar{x}_2^T \\ \vdots \\ \bar{x}_n^T \end{pmatrix}$$

Ortogonal Projection

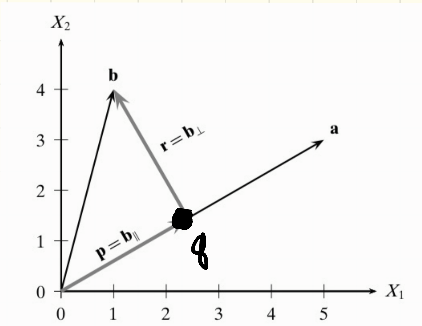
Let $a, b \in \mathbb{R}^m$

ortho decomp of b in the direction of vector a

$b = b_{\parallel} + b_{\perp}$ $p = b_{\parallel}$ $r = b_{\perp}$

$= p + r$

ortogonal projection



$q \in \mathbb{R}^m$

closest point on line $\overleftrightarrow{Oa}$ to b

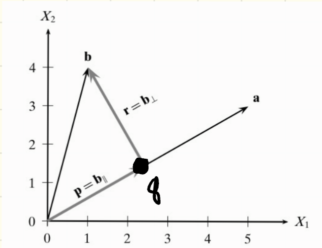
$\|r\|$ is prep dist between b and a

(aka error or residual)

How to get p?

Note: $p = \gamma a \quad \gamma \in \mathbb{R}$

$\Rightarrow \; r = b - p = b - \gamma a$



$p \perp r \;\Rightarrow\; 0 = p \cdot r$

$$= (\gamma a) \cdot (b - \gamma a)$$

$$= \gamma a \cdot b - \gamma^2 a \cdot a$$

solve for γ

$$\gamma = \frac{a \cdot b}{a \cdot a} = \mathrm{proj}_a(b)$$

$$p = \gamma a = \mathrm{proj}_a(b)\, a$$

E.g Iris dataset

5.9	3.0	4.2	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
6.6	2.9	4.6	1.3	Iris-versicolor
4.6	3.2	1.4	0.2	Iris-setosa
6.0	2.2	4.0	1.0	Iris-versicolor
4.7	3.2	1.3	0.2	Iris-setosa
6.5	3.0	5.8	2.2	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica

$\Downarrow$

proj to e_1 & e_2

5.9	3.0
6.9	3.1
6.6	2.9
4.6	3.2
6.0	2.2
4.7	3.2
6.5	3.0
5.8	2.7

$\Leftarrow D$

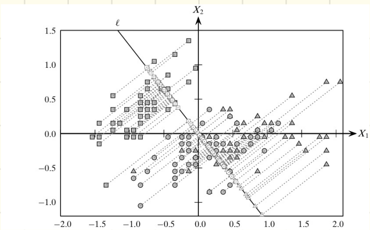
⋮

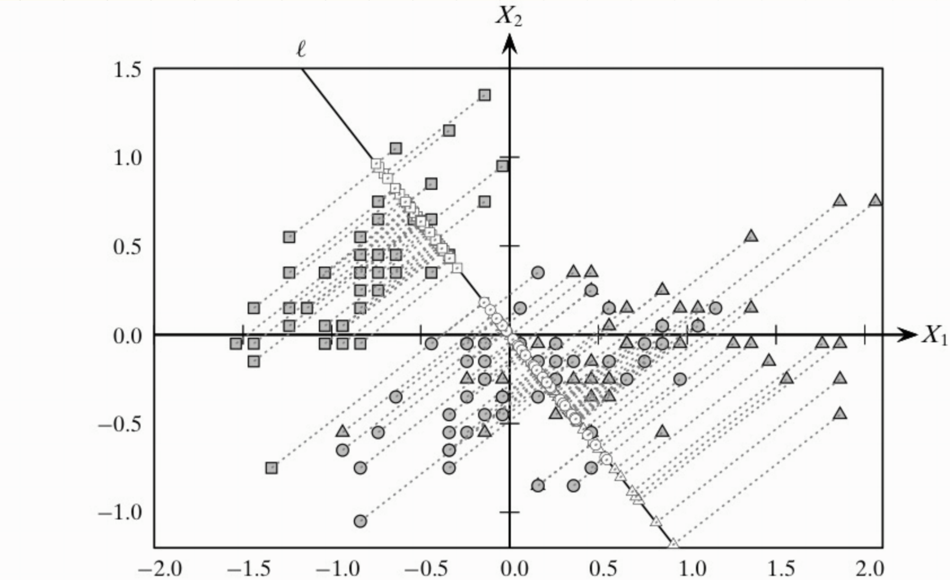
$\Rightarrow$ compute mean & var

$$\text{mean}(D) = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

$$\text{var}(D) = .868$$

$\Downarrow$ project D onto $\text{mean}(D)$





X_2
ℓ
1.5
1.0
0.5
0.0
−0.5
−1.0
X_1
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0