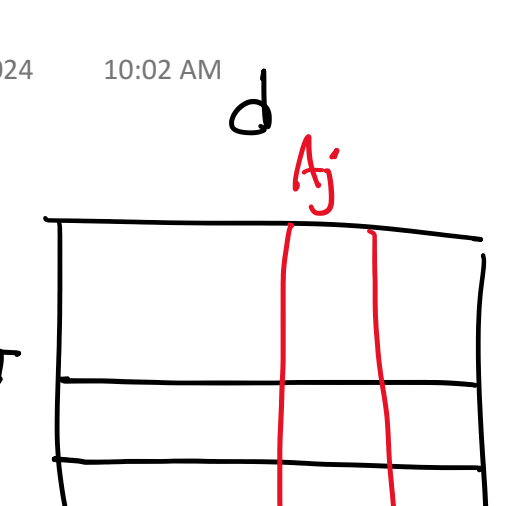
Lecture 2

Thursday, September 5, 2024 10:02 AM



$\vec{x}_i \in \mathbb{R}^d$ ← points/instance

$A_j \in \mathbb{R}^n$ ← attribute/variable

d: dimensionality (extrinsic)

n: sample size

$D \in \mathbb{R}^{n \times d}$

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

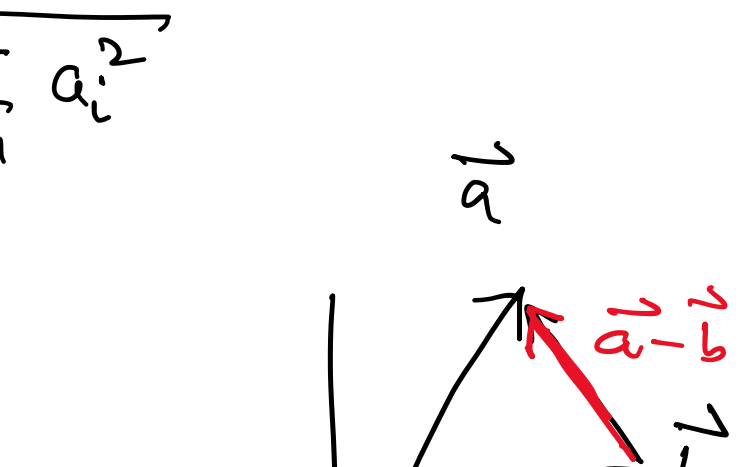
dot/inner product $\vec{a}^T \vec{b} = \underbrace{\sum_{i=1}^{m} a_i b_i}_{\text{Scalar } \mathbb{R}} \equiv \langle \vec{a}, \vec{b} \rangle$

$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}$ $\vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$

norm: $\|\vec{a}\| = \sqrt{\underline{\underline{\vec{a}^T \vec{a}}}} = \sqrt{\sum_{i=1}^{m} a_i^2}$

distance

$\|\vec{a} - \vec{b}\|$



angle $\cos\theta = \left(\frac{\vec{a}}{\|\vec{a}\|}\right)^T \left(\frac{\vec{b}}{\|\vec{b}\|}\right)$

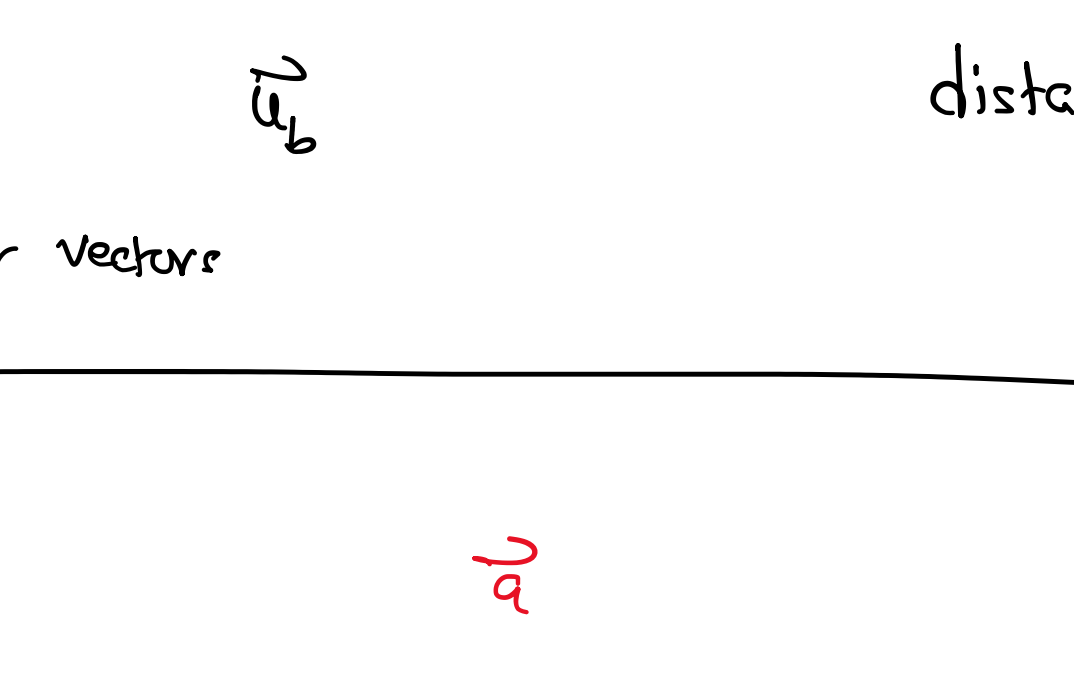
$\vec{u}_a$ $\vec{u}_b$

Unit vectors

Similarity

distance

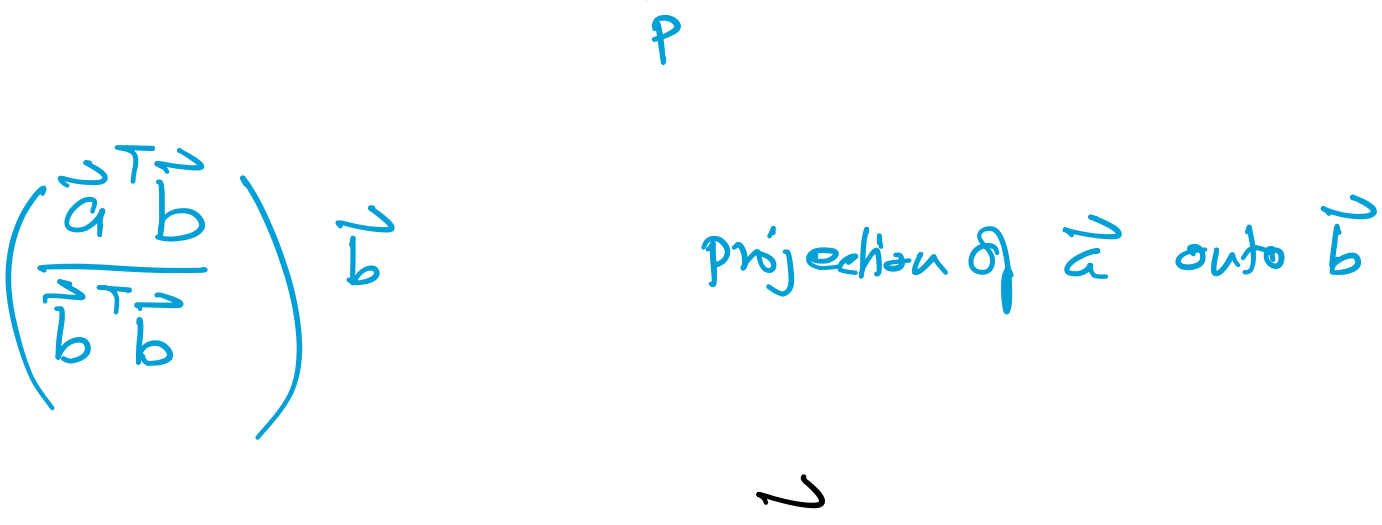
Orthogonal Projection



Project $\vec{a}$ onto $\vec{b}$

Q: how well does $\vec{b}$ approximate $\vec{a}$

$\vec{p} = \left(\frac{\vec{a}^T \vec{b}}{\vec{b}^T \vec{b}}\right) \vec{b}$ projection of $\vec{a}$ onto $\vec{b}$



Proj of $\vec{b}$ onto $\vec{a}$ = $\underbrace{\left(\frac{\vec{b}^T \vec{a}}{\vec{a}^T \vec{a}}\right)}_{\text{Scalar projection}} \vec{a}$ ← direction (+ mag) of $\vec{a}$

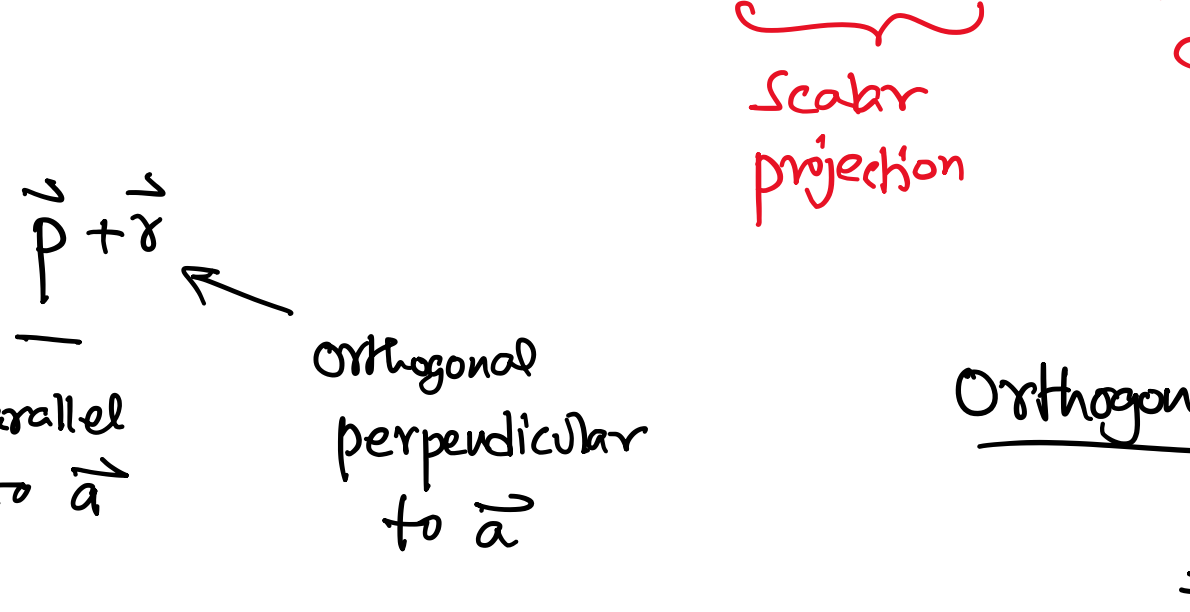
$\vec{b} = \vec{p} + \vec{r}$ ← orthogonal perpendicular to $\vec{a}$

parallel to $\vec{a}$

Orthogonal! $\vec{a}^T \vec{b} = 0$

$\Rightarrow \cos\theta = 0$

$\Rightarrow \theta = 90^\circ$ or $\frac{\pi}{2}$

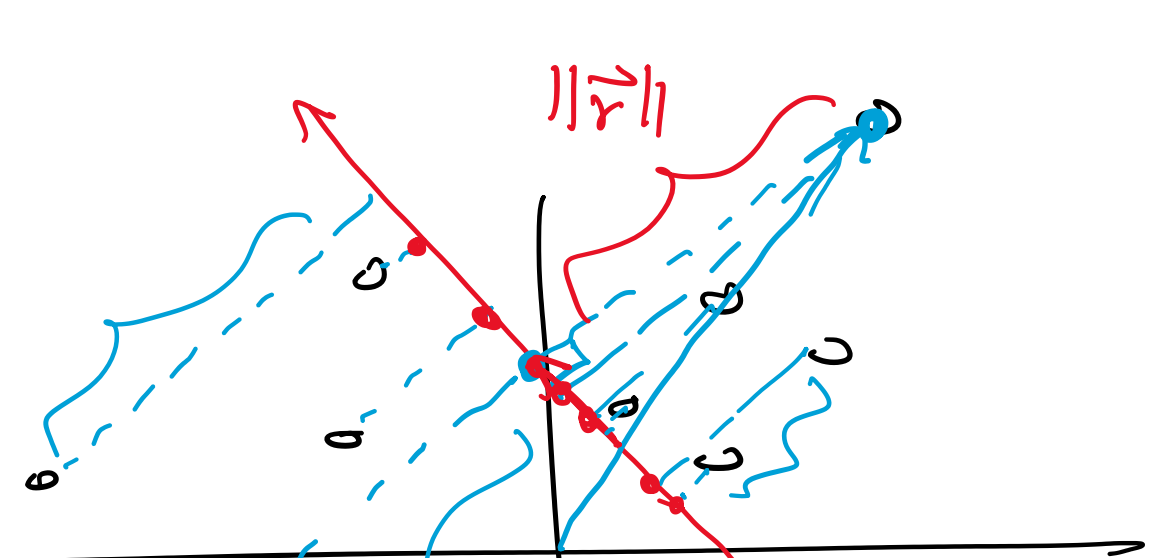


Q: what line best approximate the sample

⇓

Project all points/vector onto the line and sum up the norms of the "error" (perpendicular component)

Objective:



'All' of ML/DM is about finding good direction/line

D

	A_1	
	70	x_1
	70.5	x_2
n = 5	80	⋮
	80	
	90.5	x_n

d = 1

n: sample size

$x_i \in \mathbb{R}$ scalar

A_1: sample vector $\in \mathbb{R}^n$

A_1: random variable denoting the temp

$f: \Omega \Rightarrow \mathbb{R}$

inference/learning: learn the distribution of A_1

true distribution is unknown

true mean/expected value

$E[A_1] \rightarrow$ discrete, $\sum_x x \cdot \underline{p(x)}$ $p(x) \equiv$ Prob mass function for A_1

↗ true expected value

unknown

e.g. fair dice

$E[A_1] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6}$

$= \boxed{3.6}$

↘ Continuous $\cancel{\int x \cdot p(x)\,dx}$

$\boxed{p(x) = 0 \;\; \forall x}$ → Prob of observing a single value

$= \int x \cdot \boxed{f(x)}\,dx$

↘ Prob density function

$f(x)\,dx$

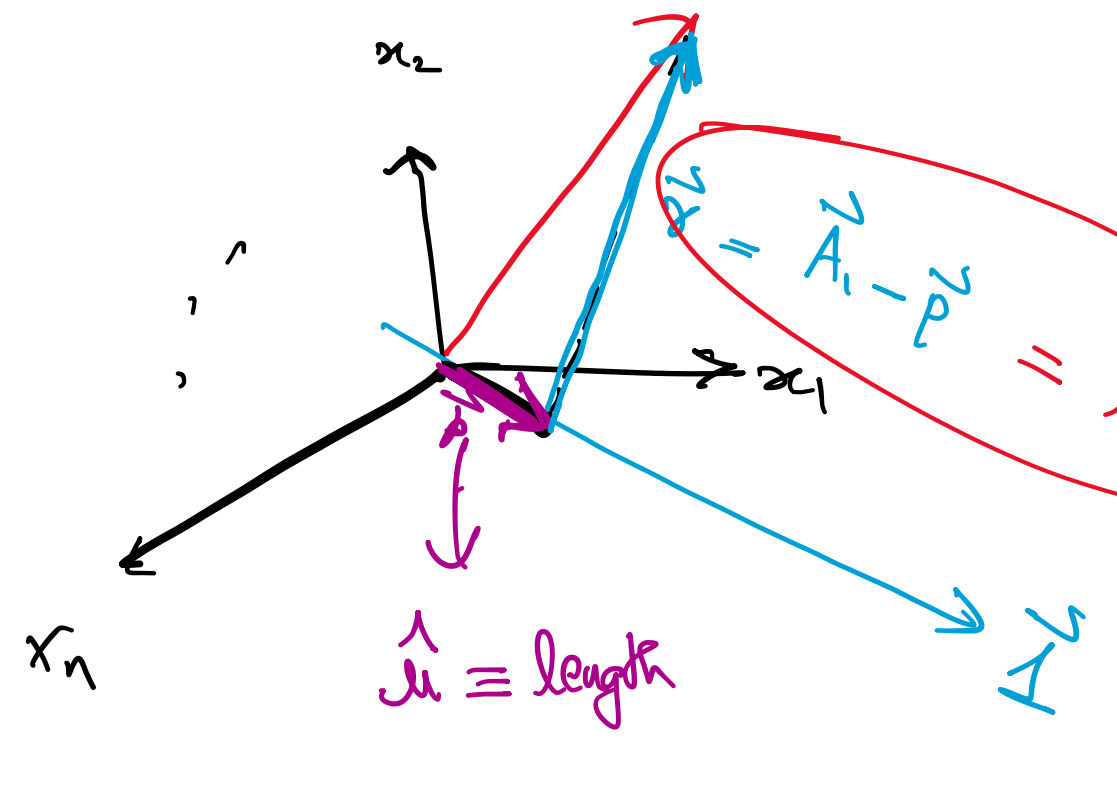
$\underbrace{P[v_1 \leq A_1 \leq v_2]}_{\text{actual probability of a range of values}} = \int_{v_1}^{v_2} f(x)\,dx$

	A_1
x_1	90
x_2	90.5
⋮	70
	100
x_n	10.2

Sample mean $= \hat{\mu} = \underbrace{\frac{1}{n}\sum x_i}_{\text{Point view}}$

true mean $E[A_1] = \mu$

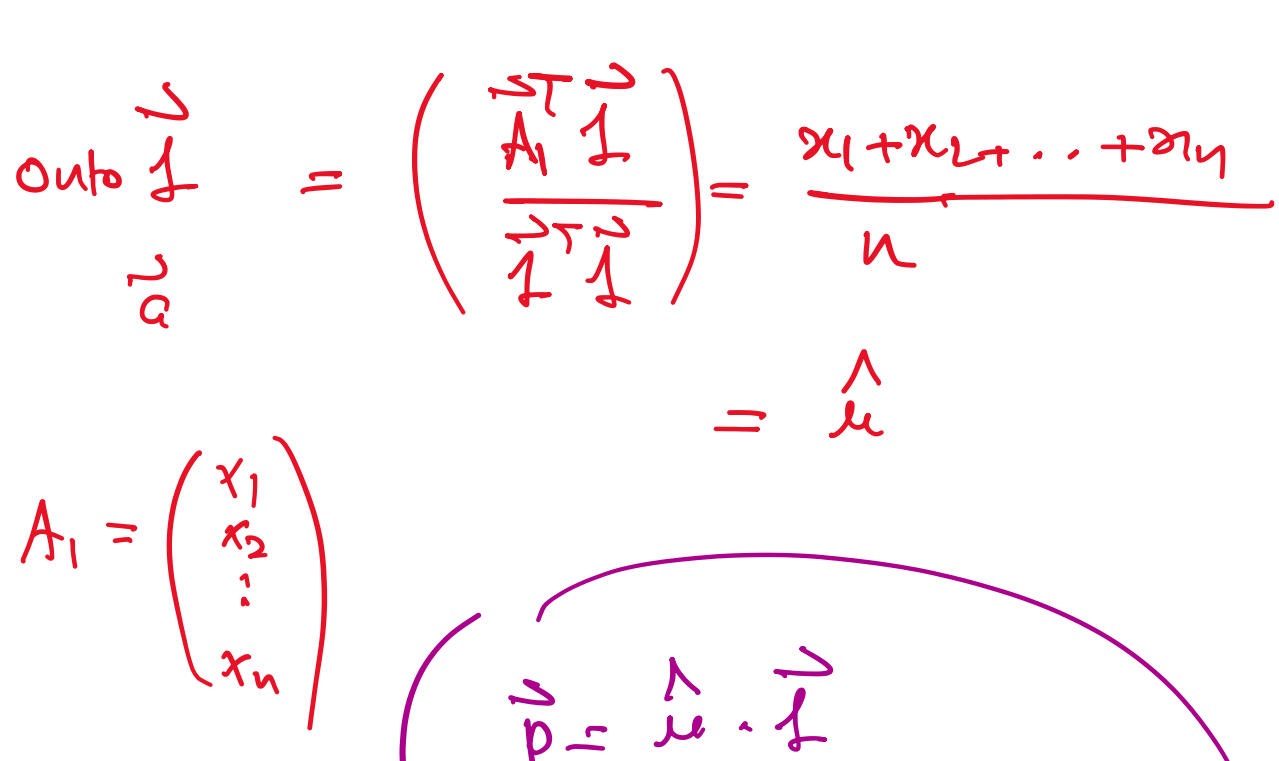
$\hat{\mu} = \frac{90 + 90.5 + 70 + 100 + 10.2}{5}$



Col-view/attribute view

$A_1 \in \mathbb{R}^n$

n-dimensional space


x_2
x_1
x_n
$\vec{A}_1$
$\vec{A}_1 - \vec{p} = \vec{A}_1^c$
$\hat{\mu} \equiv$ length
$\vec{1}$

$\vec{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$

Ones-vector

Scalar Proj $\vec{A}_1$ onto $\vec{1}$ = $\left(\frac{\vec{A}_1^T \vec{1}}{\vec{1}^T \vec{1}}\right) = \frac{x_1 + x_2 + \dots + x_n}{n}$

$\vec{b}$ $\vec{a}$

$= \hat{\mu}$

$\vec{1}^T \vec{1} = 1 \cdot 1 + 1 \cdot 1 + \dots + 1 \cdot 1$

$= 1 + 1 + \dots + 1$

$= n$

$\|\vec{1}\|^2$

$A_1 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

$\vec{p} = \hat{\mu} \cdot \vec{1}$

Scalar direction

Proj of $\vec{b}$ onto $\vec{a}$ = $\underbrace{\left(\frac{\vec{b}^T \vec{a}}{\vec{a}^T \vec{a}}\right)}_{\text{Scalar}} \vec{a}$ ← direction

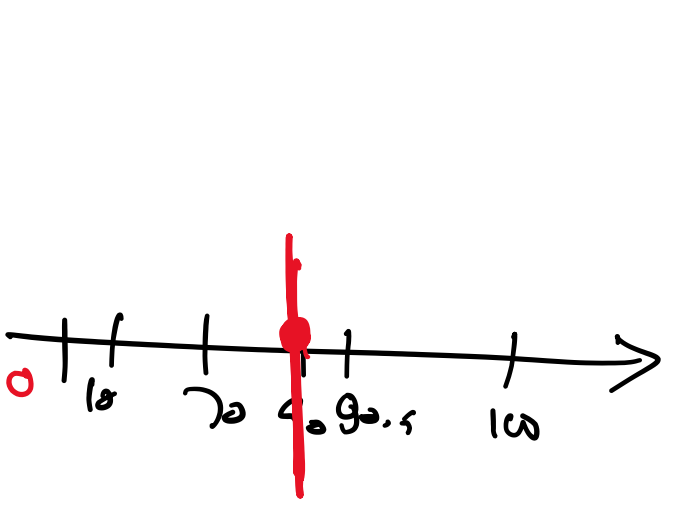
---

$D = \begin{pmatrix} A_1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ $\hat{\mu} \equiv$ Sample mean

original data

$D_c = \begin{pmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_n - \hat{\mu} \end{pmatrix}$ $\vec{A}_1^c$

Centered data



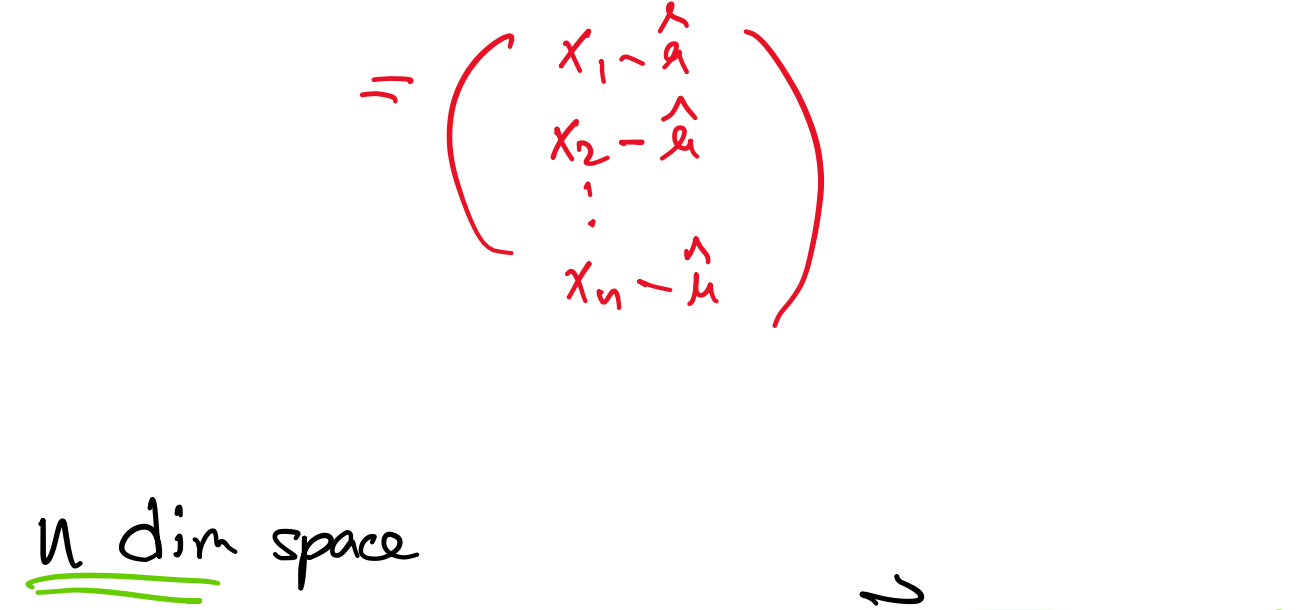
$\hat{\mu}_c = 0$

$\vec{A}_1 - \hat{\mu}\vec{1} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \hat{\mu}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

$= \begin{pmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_n - \hat{\mu} \end{pmatrix}$

$\vec{A}_1 = \underbrace{\text{mean vector}}_{\substack{\text{Parallel} \\ \vec{1}}} + \underbrace{\text{centered attribute}}_{\substack{\text{Orthogonal} \\ \text{remaining data}}}$

n dim space



$\mathbb{R}^n$

⇓

dof = n

degree of freedom

$\vec{A}_1^{cT} \vec{1} = 0$

$\vec{p} = \hat{\mu} \cdot \vec{1}$

$\vec{r} = \vec{A}_1 - \hat{\mu} \cdot \vec{1}$

$\vec{r}^T \vec{1} = 0$

$\vec{r}$ lies in the orthogonal complement of $\vec{1}$

Variance ← deviation from the mean (squared)

$\sigma^2 = E[\underbrace{(A_1 - \mu)^2}_{\text{r.v}}] = E[Z]$

$E[g(x)] = \sum g(x) \cdot p(x)$

function ↑ Prob

$E[A_1] = \sum_x \underline{\underline{x}} \cdot p(x)$

$E[(A_1 - \mu)^2] \rightarrow \sum_x (x - \mu)^2 \cdot p(x)$

↘ $\int (x - \mu)^2 \underline{f(x)}\,dx$

↖ prob density function

Sample variance

$\underbrace{\boxed{\hat{\sigma}^2}}_{\text{biased estimator}} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{\boxed{n}}$

$\hat{\sigma}_u^2 = \frac{\sum_i (x_i - \boxed{\hat{\mu}})^2}{\boxed{n-1}}$

Unbiased variance

$\boxed{E[\hat{\theta}] \neq \theta}$

then $\hat{\theta}$ is biased

$E[\hat{\sigma}_u^2] = \sigma^2$

unbiased

$E[\hat{\sigma}^2] \neq \sigma^2$

biased

Column view

$\vec{A}_1 = \vec{p} + \vec{r}$

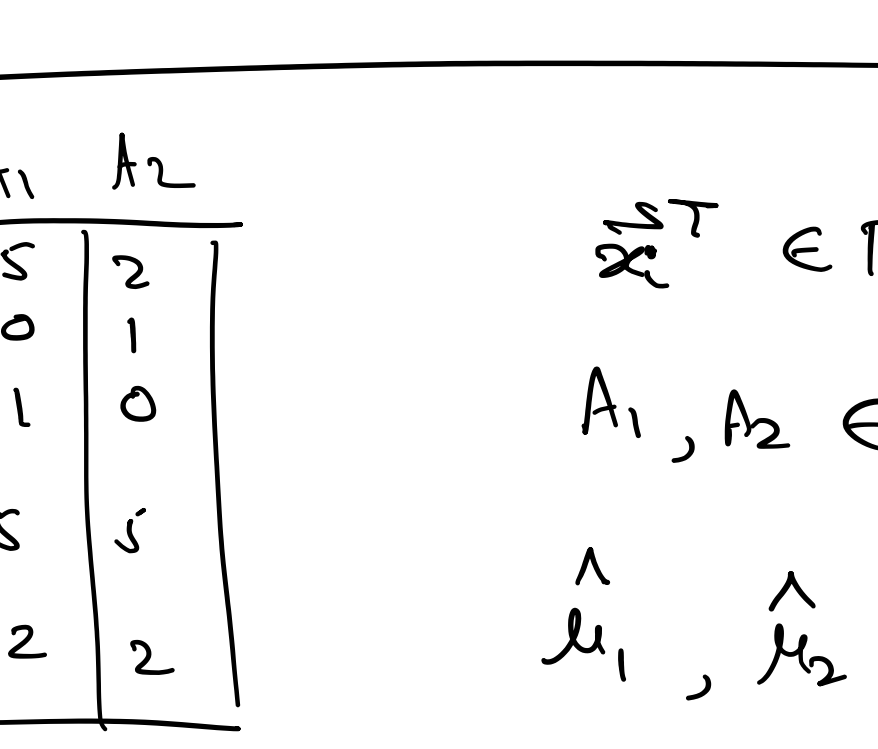
$\vec{r} = \vec{A}_1 - \vec{p} = \vec{A}_1 - \hat{\mu} \cdot \vec{1} = \vec{A}_1^c$

↖ Centered attribute vector

$\hat{\sigma}^2 = \frac{\|\vec{A}_1^c\|^2}{n}$

Variance ≡ squared length of centered attribute normalized by sample size

(avg. sq. length / dim)



$\vec{A}_1^c \in \mathbb{R}^{n-1}$ living in $\mathbb{R}^n$

dof ≡ dim of the subspace that where the statistical vector lies

---

	A_1	A_2
$\vec{x}_1^T$	5	2
$\vec{x}_2^T$	10	1
⋮	-1	0
	5	5
$\vec{x}_n^T$	2	2

$\vec{x}_i^T \in \mathbb{R}^2$ d = 2

$A_1, A_2 \in \mathbb{R}^n$

$\hat{\mu}_1, \hat{\mu}_2$

$\hat{\mu}^T = (\hat{\mu}_1, \hat{\mu}_2)$ $\hat{\vec{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} \in \mathbb{R}^2$

$\hat{\sigma}_1^2, \hat{\sigma}_2^2$: variance along A_1 & A_2

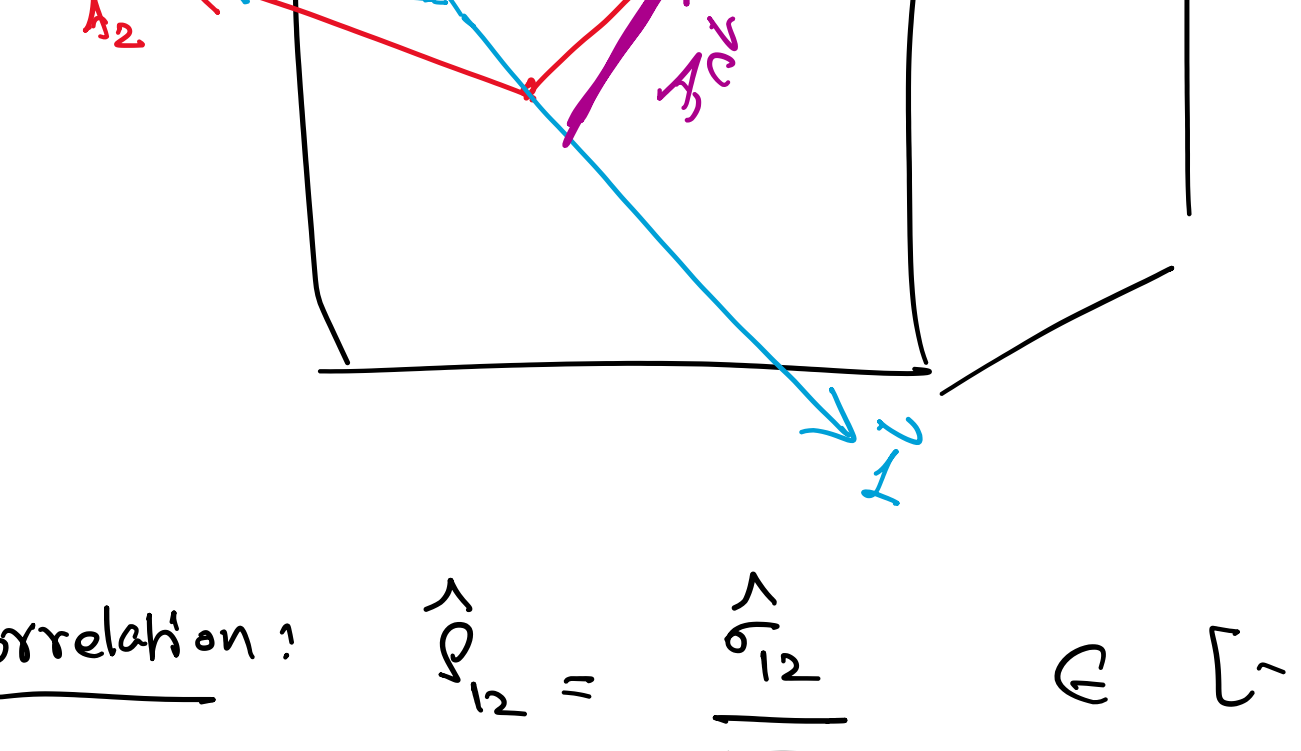
Covariance: $\sigma_{12} = \mathrm{Cov}(A_1, A_2) = E[(A_1 - \mu_1)(A_2 - \mu_2)]$

true means

Sample cov: $\frac{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{n}$ $\vec{x}_i^T = (x_{i1}, x_{i2})$

Col-view: $\hat{\sigma}_{12} = \widehat{\mathrm{Cov}}(A_1, A_2) = \frac{\vec{A}_1^{cT} \vec{A}_2^c}{n}$

$\mathbb{R}^n$



$\frac{\vec{A}_1^{cT} \vec{A}_2^c}{n} = \widehat{\mathrm{Cov}} = \hat{\sigma}_{12}$

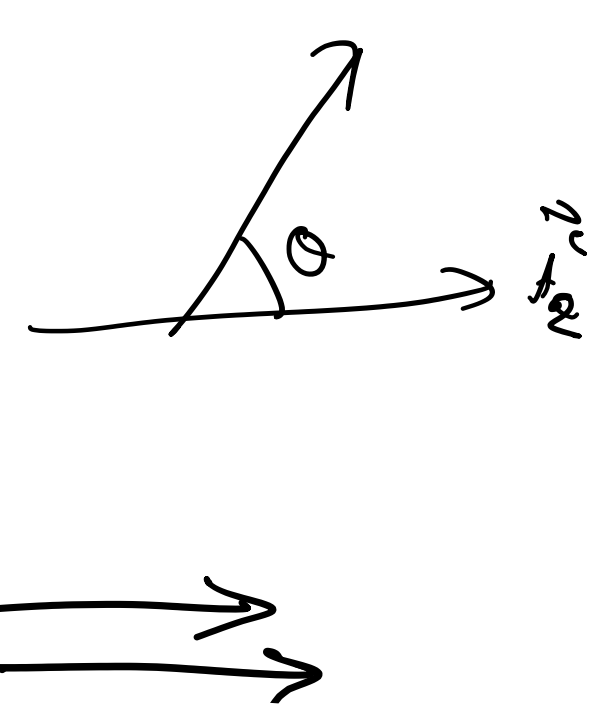
Correlation: $\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\sqrt{\hat{\sigma}_1^2 \hat{\sigma}_2^2}} \in [-1, 1]$

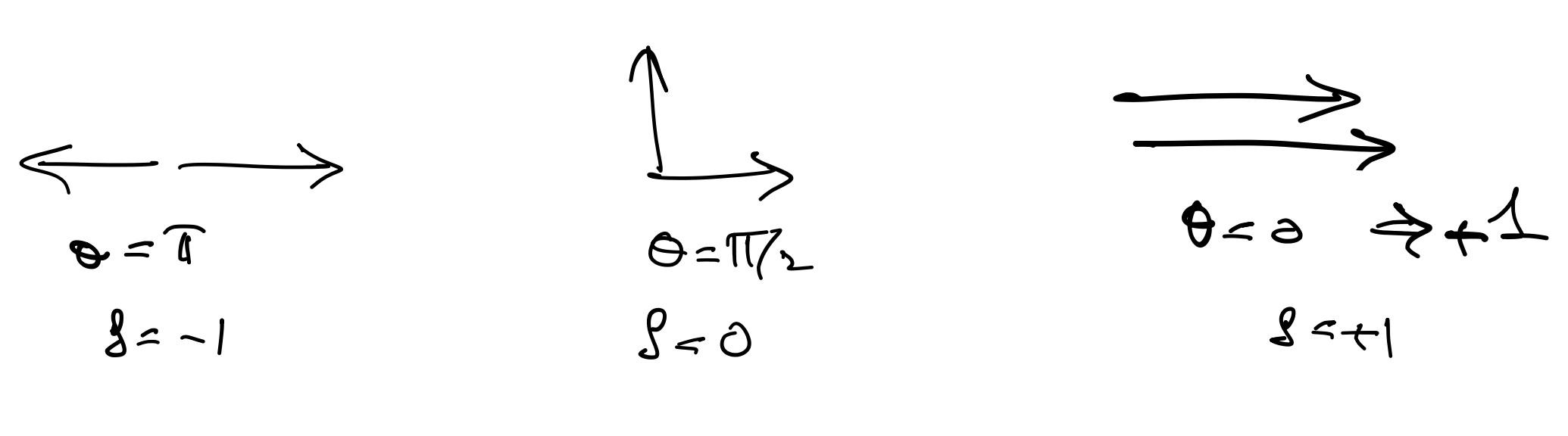
$\underbrace{\sqrt{\hat{\sigma}_1^2}}_{\text{Var}} = \underbrace{\hat{\sigma}_1}_{\text{Std standard dev}}$

$= \frac{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)^2 \cdot \sum_{i=1}^{n}(x_{i2} - \hat{\mu}_2)^2}}$

$\hat{\rho}_{12} = \cos\theta = \left(\frac{\vec{A}_1^c}{\|\vec{A}_1^c\|}\right)^T \left(\frac{\vec{A}_2^c}{\|\vec{A}_2^c\|}\right)$

↑ Cosine of the angle between the two centered vectors





θ = π, ρ = -1

θ = π/2, ρ = 0

θ = 0 → +1, ρ = +1

$\hat{\Sigma} = \begin{pmatrix} \boxed{\hat{\sigma}_1^2} & \boxed{\hat{\sigma}_{12}} \\ \boxed{\hat{\sigma}_{12}} & \boxed{\hat{\sigma}_2^2} \end{pmatrix}$ $\hat{\sigma}_{12} = \hat{\sigma}_{21}$

Cov matrix

2x2 matrix, symmetric $\hat{\Sigma}^T = \hat{\Sigma}$

$\hat{\Sigma}$ is PSD: positive semi-definite

eigen value are real & non-negative

$\boxed{\lambda_1 \geq \lambda_2 \geq 0}$

u_1 u_2