Session 7: Spectral Theorem, PCA and SVD

Optimization and Computational Linear Algebra for Data Science

Léo Miolane

Contents

Midterm

- The Midterm exam is in 2 weeks.
- **Scope:** everything that we have seen so far (this week's video included).
- **Knowing is not enough!** You need to practice: review problems available on the course's webpage.
- Past years midterms also available, with solutions.
- **Important:** when working on a problem, take **at least** 10min on it before looking at the solution (in case you are stuck).
- The midterm is open books/notes, but **if you think that you need them for the exam, this probably means that you are not prepared enough**.

The Spectral Theorem

The spectral theorem

$Id\,x = 1 \cdot x$

Theorem

Let $A \in \mathbb{R}^{n \times n}$ be a **symmetric** matrix. Then there is a orthonormal basis of $\mathbb{R}^n$ composed of eigenvectors of A.

$$P = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix} \qquad D = \begin{pmatrix} \lambda_1 & & (0) \\ & \ddots & \\ (0) & & \lambda_n \end{pmatrix}$$

Theorem (Matrix formulation)

Let $A \in \mathbb{R}^{n \times n}$ be a **symmetric** matrix. Then there exists an orthogonal matrix P and a diagonal matrix D of sizes $n \times n$ such that

$$\boxed{A = PDP^{\mathsf{T}}.}$$

Geometric interpretation

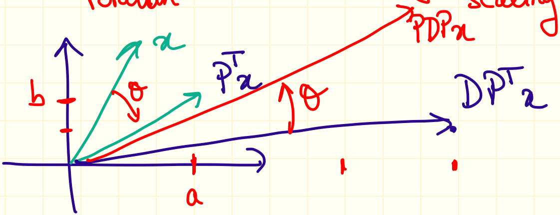
Ex:

$$A = \underbrace{P}_{\text{"inverse rotation"}} \underbrace{\begin{pmatrix} 3 & 0 \\ 0 & 0{,}5 \end{pmatrix}}_{D} \underbrace{P^T}_{\text{rotation}}$$

"scaling matrix"

$$\begin{pmatrix} a \\ b \end{pmatrix}$$

$$\begin{pmatrix} 3 & 0 \\ 0 & 0{,}5 \end{pmatrix} \begin{pmatrix} 3a \\ 0{,}5\,b \end{pmatrix}$$



The Theorem behind PCA

Theorem

by Spec. Thm

Let A be a $n \times n$ symmetric matrix and let $\lambda_1 \geq \cdots \geq \lambda_n$ be its n eigenvalues and $v_1, \ldots, v_n$ be an associated orthonormal family of eigenvectors. Then

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

$$\lambda_1 = \max_{\|v\|=1} v^\top A v \quad \text{and} \quad v_1 = \arg\max_{\|v\|=1} v^\top A v \,.$$

Moreover, for $k = 2, \ldots, n$:

$$\lambda_k = \max_{\|v\|=1,\, v \perp v_1, \ldots, v_{k-1}} v^\top A v \,, \quad \text{and} \quad v_k = \arg\max_{\|v\|=1,\, v \perp v_1, \ldots, v_{k-1}} v^\top A v.$$

$k=2$

$$\lambda_2 = \max_{\substack{\|v\|=1 \\ v \perp v_1}} v^\top A v \qquad v_2 = \arg\max_{\substack{\|v\|=1 \\ v_1 \perp v}} v^\top A v$$

Proof

- Let $v \in \mathbb{R}^n$ such that $\|v\| = 1$. Let $\boxed{(\alpha_1 \dots \alpha_n)}$ be the coordinates of v in $\mathcal{B} = (v_1, \dots v_n)$.
- $Av = A(\alpha_1 v_1 + \dots + \alpha_n v_n)$

$$= \alpha_1 A v_1 + \dots + \alpha_n A v_n$$

$$= \alpha_1 \lambda_1 v_1 + \dots + \alpha_n \lambda_n v_n \rightarrow \begin{pmatrix} \alpha_1 \lambda_1 \\ \vdots \\ \alpha_n \lambda_n \end{pmatrix}$$

are the coords. of Av in $\mathcal{B}$.

- $\boxed{v^T A v} = v \cdot (Av) = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \lambda_1 \\ \vdots \\ \alpha_n \lambda_n \end{pmatrix}$

$$= \alpha_1^2 \lambda_1 + \dots + \alpha_n^2 \lambda_n.$$

Proof

Maximize $\boxed{\alpha_1^2 \lambda_1 + \dots + \alpha_n^2 \lambda_n}$ subject to $\|v\|^2 = 1$

$v^T A v$

$1 = \alpha_1^2 + \dots + \alpha_n^2$

- Since $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, the maximum is achieved for $\alpha_1 = 1, \ \alpha_2 = \dots \alpha_n = 0$.
- This corresponds to $v = v_1$
- The corresponding value of $v^T A v$ is then $1^2 \cdot \lambda_1 = \lambda_1$

Proof

if now we maximize $v^T A v$ subject to $\begin{cases} \|v\| = 1 \\ v \perp v_1 \end{cases}$

we maximize $\underbrace{\langle v, v_1 \rangle}_{\alpha_1} = 0$

$$\boxed{\cancel{\alpha_1^2 \lambda_1} + \alpha_2^2 \lambda_2 + \dots + \alpha_n^2 \lambda_n \quad \text{subj. to} \begin{cases} \alpha_1^2 + \dots + \alpha_n^2 = 1 \\ \underline{\underline{\alpha_1 = 0}}\,. \end{cases}}$$

The maximum is now achieved for $\boxed{\begin{array}{l} \alpha_1 = 0 \\ \alpha_2 = 1 \\ \alpha_3 = \dots = \alpha_n = 0 \end{array}}$

v : $\boxed{\begin{array}{l} \alpha_1 = \langle v_1, v \rangle \\ \vdots \qquad \vdots \\ \alpha_n = \langle v_n, v \rangle \end{array}}$

Principal Component Analysis

Empirical mean and covariance

We are given a dataset of n points $a_1, \ldots, a_n \in \mathbb{R}^d$

$\underline{d = 1}$

- **Mean**

$$\mu = \frac{1}{n} \sum_{i=1}^{n} a_i \quad \in \mathbb{R}$$

- **Variance**

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (a_i - \mu)^2 \quad \in \mathbb{R}$$

Empirical mean and covariance

We are given a dataset of n points $a_1, \ldots, a_n \in \mathbb{R}^d$

<u>$d = 1$</u>

- **Mean**

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i \qquad \in \mathbb{R}$$

- **Variance**

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (a_i - \mu)^2 \qquad \in \mathbb{R}$$

<u>$d \geq 2$</u>

- **Mean**

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i \qquad \in \mathbb{R}^d$$

- **Covariance matrix**

$$S = \frac{1}{n}\sum_{i=1}^{n} \boxed{(a_i - \mu)(a_i - \mu)^\mathsf{T}} \qquad \in \mathbb{R}^{d\times d}$$

$$= \frac{1}{n}\sum_{i=1}^{n} a_i a_i^\mathsf{T} \qquad \text{if } \mu = 0.$$

PCA

- We are given a dataset of n points $a_1, \dots, a_n \in \mathbb{R}^d$, where d is «large».
- **Goal:** represent this dataset in lower dimension, i.e. find $\tilde{a}_1, \dots, \tilde{a}_n \in \mathbb{R}^k$ where $k \ll d$.
- Assume that the dataset is centered: $\sum_{i=1}^n a_i = 0$.
- Then, S can be simply written as:

" cov. matrix without the 1/n "

$$\boxed{S = \sum_{i=1}^n a_i a_i^\mathsf{T} = A^\mathsf{T} A.} = BB^\mathsf{T}$$

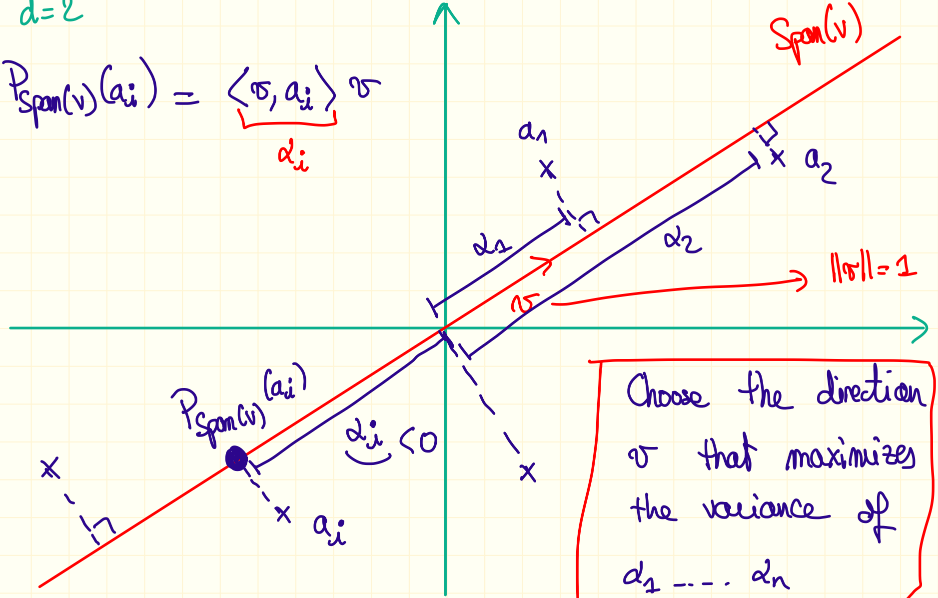
where A is the $n \times d$ "data matrix":

$$B = \begin{pmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{pmatrix} = A^\mathsf{T} \qquad \boxed{A = \begin{pmatrix} - & a_1^\mathsf{T} & - \\ & \vdots & \\ - & a_n^\mathsf{T} & - \end{pmatrix}.}$$

Direction of maximal variance

$d=2$

$$P_{\mathrm{Span}(v)}(a_i) = \underbrace{\langle v, a_i \rangle}_{\alpha_i} v$$



$\boxed{\text{Choose the direction } v \text{ that maximizes the variance of } \alpha_1 \dots \alpha_n}$

Direction of maximal variance

Mean:
$$\frac{\alpha_1 + \dots + \alpha_n}{n} = \frac{\langle v, a_1\rangle + \dots + \langle v, a_n\rangle}{n}$$
$$= \left\langle v, \frac{\overbrace{a_1 + \dots + a_n}^{=0}}{n} \right\rangle = 0$$

Variance:
$$\frac{1}{n}\sum_{i=1}^{n} \langle v, a_i\rangle^2 = \frac{1}{n}\sum_{i=1}^{n} v^T a_i a_i^T v$$
$$= \frac{1}{n} v^T \left(\sum_{i=1}^{n} a_i a_i^T \right) v$$
$$= \frac{1}{n} \boxed{v^T S v}$$

$$\langle v, a_i\rangle = v^T a_i$$
$$= a_i^T v$$

Direction of maximal variance

Good news: $S = A^{\mathsf{T}} A$ is symmetric. $\leftarrow$ $S = A^T A$ $\leftarrow$ • sym. • PSD

Spectral Theorem: let $\boxed{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d}$ be the eigenvalues of S ≥ 0
and $(v_1, \ldots, v_d)$ an associated orthonormal basis of eigenvectors.

- By the theorem we saw before a unit vector v that maximises $v^T S v$ is $v = v_1$
- and the corresponding $\boxed{\text{“variance” } v^T S v \text{ is}}$ equal to $\boxed{\lambda_1}$.
- The dimensionally-reduced dataset is then

$$\tilde{a}_1 = \langle v_1, a_1 \rangle \quad - - - - \quad \tilde{a}_n = \langle v_1, a_n \rangle$$

2nd direction of maximal variance

- We would like to find another vector v, such that the variance of $\langle v, a_1 \rangle \dots \langle v, a_n \rangle$ is large.

- Goal: find v that maximizes

$$\boxed{v^T S v \quad \text{subject to} \quad \begin{cases} \|v\| = 1 \\ v \perp v_1 \end{cases}}$$

- An optimal vector is $\boxed{v = v_2}$ $\quad \lambda_1 > \lambda_2 > \dots \quad \lambda_2 = \lambda_3$

1st princ. comp. of a_1

$$\tilde{a}_1 = \begin{pmatrix} \langle v_1, a_1 \rangle \\ \langle v_2, a_1 \rangle \end{pmatrix} \quad \dots \quad \tilde{a}_n = \begin{pmatrix} \langle v_1, a_n \rangle \\ \langle v_2, a_n \rangle \end{pmatrix}$$

j^{th} direction of maximal variance

- The « j^{th} direction of maximal variance » is v_j since v_j is solution of

$$\boxed{\text{maximize } \underline{v^\mathsf{T} S v}, \qquad \text{subject to } \|v\| = 1,\ v \perp v_1, v \perp v_2, \ldots, v \perp v_{j-1}.}$$

- The dimensionally reduced dataset is then

$$\underbrace{\begin{pmatrix} \langle v_1, a_1 \rangle \\ \langle v_2, a_1 \rangle \\ \vdots \\ \langle v_k, a_1 \rangle \end{pmatrix}}, \underbrace{\begin{pmatrix} \langle v_1, a_2 \rangle \\ \langle v_2, a_2 \rangle \\ \vdots \\ \langle v_k, a_2 \rangle \end{pmatrix}}, \underbrace{\begin{pmatrix} \langle v_1, a_3 \rangle \\ \langle v_2, a_3 \rangle \\ \vdots \\ \langle v_k, a_3 \rangle \end{pmatrix}} \cdots \underbrace{\begin{pmatrix} \langle v_1, a_n \rangle \\ \langle v_2, a_n \rangle \\ \vdots \\ \langle v_k, a_n \rangle \end{pmatrix}}. \quad \in \mathbb{R}^k$$

Recap

① Center your dataset → get a dataset such that $\sum_{i=1}^{n} a_i = 0$.

② Compute the covariance matrix $S = \sum_{i=1}^{n} a_i a_i^T$ $\left(= A^T A\right)$

③ Compute the eigenvalues $\lambda_1 \text{ —— } \lambda_d$ of S and associated eigenvectors $v_1 \text{ —— } v_d$ of S.

④ Sort eigenvalues / eigenvectors.

⑤ Select some k.

⑥ Compute $\tilde{a}_1 \text{ - - - - } \tilde{a}_n$.

Which value of k should we take?

1st way

When using the k first principal components we capture a fraction

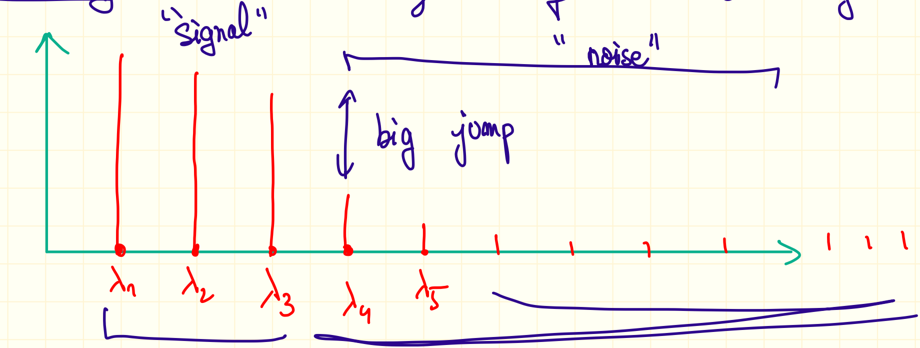
$$f_k = \boxed{\frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_k + \dots + \lambda_d}}$$

of the total variance.

Choose k such that $f_k \geq 80\%$

Which value of k should we take?

2nd way : Plot the eigenvalues of S in decreasing ordr

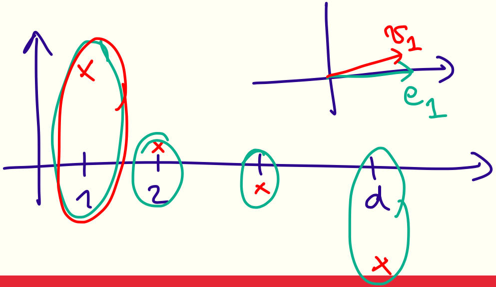


$\longrightarrow$ select $k = 3$.

10 000

$$v_1 = \begin{pmatrix} 0,8 \\ 0,1 \\ 0,2 \end{pmatrix} \qquad a_1 = \begin{pmatrix} a_{11} \\ \vdots \\ a_{1d} \end{pmatrix}$$

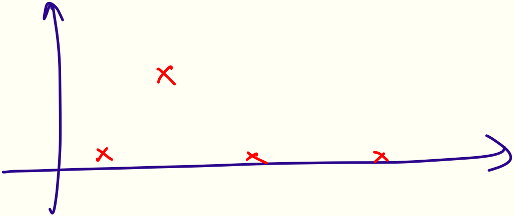
entries of v_1



$$a_1 = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_d \end{pmatrix}$$

Singular Value Decomposition

v_2



$$f_1 = \langle a_1, e_1 \rangle$$

Singular values/vectors

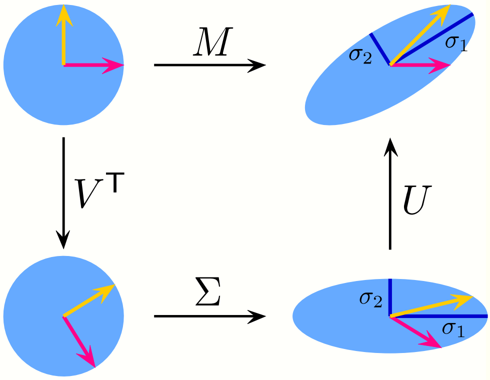
Singular Value decomposition

Theorem

Let $A \in \mathbb{R}^{n\times m}$. Then there exists two orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ and a matrix $\Sigma \in \mathbb{R}^{n\times m}$ such that $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \cdots \geq 0$ and $\Sigma_{i,j} = 0$ for $i \neq j$, that verify

$$A = U\Sigma V^{\mathsf{T}}.$$

Geometric interpretation of $U\Sigma V^\top$



$$M = U\Sigma V^\top$$

Questions?

$$A = \begin{pmatrix} - & a_1^T & - \\ - & a_n^T & - \end{pmatrix}$$

$$\frac{1}{n} S = \frac{1}{n} A^T A = \frac{1}{n} \begin{pmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{pmatrix} \begin{pmatrix} - & a_1^T & - \\ & \vdots & \\ - & a_n^T & - \end{pmatrix} = \boxed{\sum_{i=1}^{n} a_i a_i^T}$$

square of the

$$\frac{1}{n} \sum_{i=1}^{n} (a_i)_1 (a_i)_1 = \text{average of the } 1^{st} \text{ feature over the dataset}$$

Questions?

$$\underbrace{A^T A}_{d \times d} \stackrel{?}{=} \left(\sum_{i=1}^{n} a_i a_i^T \right)^{d \times d}$$

$$A = \begin{pmatrix} — a_1^T — \\ — a_n^T — \end{pmatrix}$$

$$\underline{\underline{(A^T A)_{k,l}}} = \sum_{i=1}^{n} \overbrace{(A^T)_{k,i}}^{A_{i,k}} A_{i,l}$$

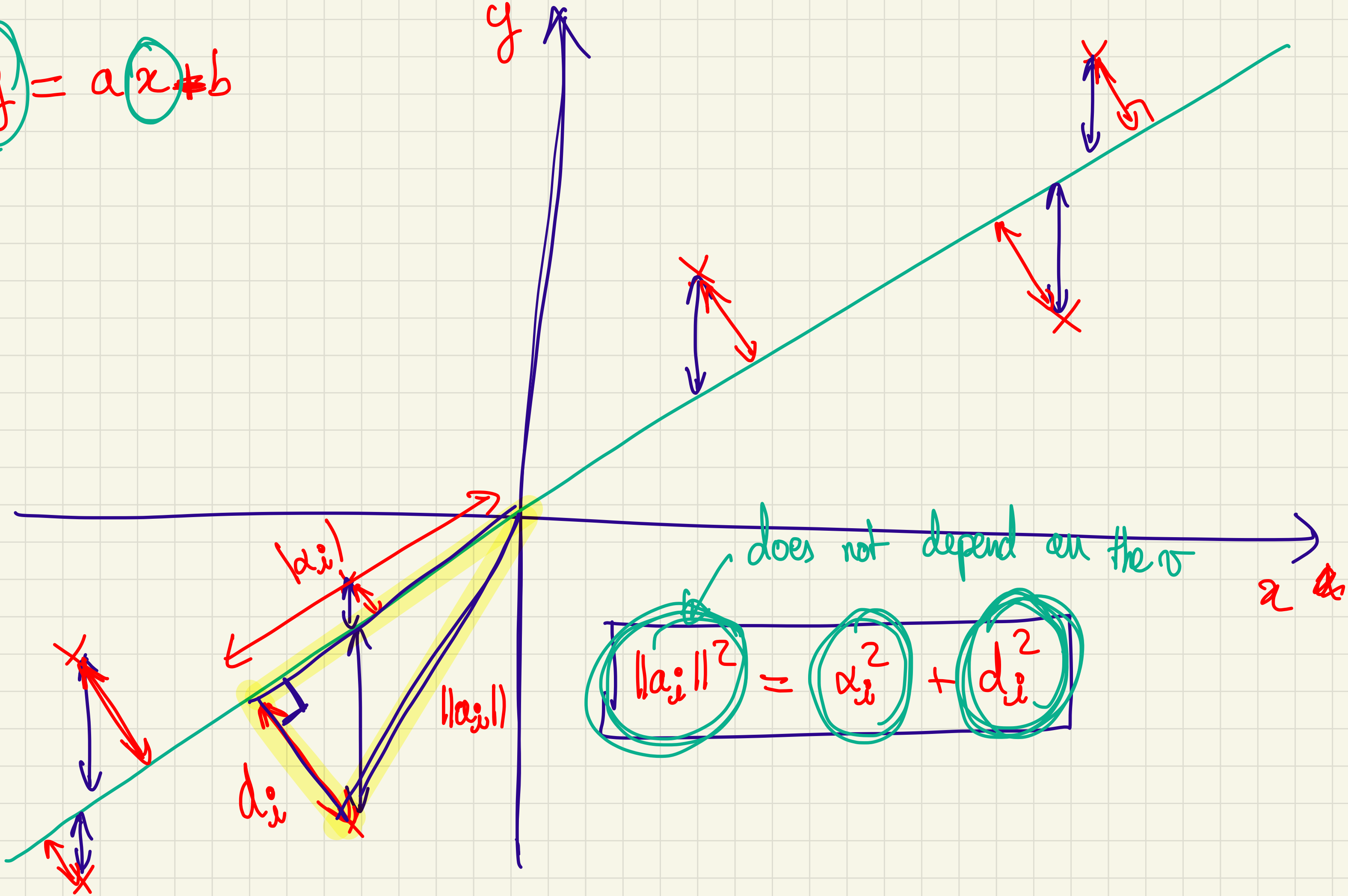
$$(— a_i^T —)$$

$$\left(a_i \right) \left(\underline{d \times d} \right)$$

$$= \sum_{i=1}^{n} \underbrace{(a_i)_k (a_i)_l}_{} = (a_i a_i^T)_{k,l}.$$

$$= \left(\sum_{i=1}^{n} \overbrace{a_i a_i^T}^{d \times d} \right)_{k,l}.$$

$y = ax + b$



does not depend on the v

$$\boxed{\|a_i\|^2 = \alpha_i^2 + d_i^2}$$