Foundation of Data Analysis

Note: The dates on the lecture notes correspond with the dates on the videos, which were recorded one year prior to SS21.

Lecture 1

21.04.20

Chapter 1 : Preliminaries on Linear Algebra

1.1 Matrix Notations

Assumptions / Notations / Definitions

(i) I, J, L are finite index sets.

(ii) We mainly consider fields $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$,
for $\alpha \in \mathbb{C}$, $\alpha = a + i \cdot b$, $a, b \in \mathbb{R}$, we consider the complex conjugate

$$\overline{\alpha} := a - i \cdot b \in \mathbb{C}.$$

(iii) Define $\mathbb{K}^{I \times J}$ the vector space of matrices $A \in \mathbb{K}^{I \times J}$. (I: rows, J: columns)
The entries of A are denoted by A_{ij}, $i \in I$ row, $j \in J$ column.
For $a_{ij} \in \mathbb{K}$ we can form $A := (a_{ij})_{i \in I, j \in J} \in \mathbb{K}^{I \times J}$.

(iv) Transposed matrix $A^T \in \mathbb{K}^{J \times I}$ for $A \in \mathbb{K}^{I \times J}$.

(v) $A \in \mathbb{K}^{I \times I}$ is symmetric if $A^T = A$.

(vi) $A \in \mathbb{K}^{I \times J}$ is hermitian if $A^H = A$ for $A^H = \overline{A^T}$.
For $\mathbb{K} = \mathbb{R}$: symmetric $\Leftrightarrow$ hermitian.

(vii) For $A \in \mathbb{K}^{I \times J}$ denote the ith row of A by $A^{(i)} = (a_{ij})_{j \in J}$
and the jth column of A by $A_{(j)} = (a_{ij})_{i \in I}$.

(viii) Matrix-vector-multiplication: $A \in \mathbb{K}^{I \times J}$, $x \in \mathbb{K}^J$, then

$$(Ax)_i = \sum_{j \in J} a_{ij} x_j.$$

Bsp:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11} \cdot x_1 + a_{12} \cdot x_2 \\ a_{21} \cdot x_1 + a_{22} \cdot x_2 \end{pmatrix}$$

(ix) Matrix-Matrix-multiplication: $A \in \mathbb{K}^{I \times J}$, $B \in \mathbb{K}^{J \times L}$, then

$$(AB)_{il} = \sum_{j \in J} A_{ij} \cdot B_{jl}$$

Bsp:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} b_{11} + a_{12} b_{21} & * \\ * & * \end{pmatrix}$$

(x) Define the Kronecker Symbol: $\delta: I \times I \to \{0,1\}$

$$\delta_{ij} := \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

(xi) Denote the unit vector $e^{(i)} \in \mathbb{K}^I$ by $e^{(i)} := (\delta_{ij})_{j \in I}$, the identity matrix by $I := (\delta_{ij})_{i \in I, j \in I}$.

(xii) Denote the range of $A \in \mathbb{K}^{I \times J}$ by

$$\text{range}(A) := \{Ax : x \in \mathbb{K}^J\} = \text{span}\{A_{(j)} : j \in J\}.$$

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \cdot a_{11} + x_2 \cdot a_{12} \\ x_1 \cdot a_{21} + x_2 \cdot a_{22} \end{pmatrix}$$

$$= x_1 \cdot \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix}$$

$$= x_1 \cdot A_{(1)} + x_2 \cdot A_{(2)}$$

(xiii) For $x, y \in \mathbb{K}^I$, we define the Euclidean scalar product by

$$\langle x, y \rangle = y^H x = \sum_{i \in I} x_i \overline{y_i}.$$

We can rewrite for $A \in \mathbb{K}^{I \times J}$, $x \in \mathbb{K}^J$ ("undo the conjugate of definition")

$$(Ax)_i = \langle A^{(i)}, \overline{x} \rangle \quad \text{and} \quad (AB)_{il} = \langle A^{(i)}, \overline{B_{(l)}} \rangle.$$

(xiv) $x, y \in \mathbb{K}^I$ are orthogonal, if $\langle x, y \rangle = 0$, short $x \perp y$.

The sets $X, Y \subset \mathbb{K}^I$ are mutually orthogonal, short $X \perp Y$, if

$$\forall x \in X \text{ and } y \in Y: \langle x, y \rangle = 0, \text{ i.e. } x \perp y.$$

For two subspaces $X, Y \subset \mathbb{K}^I$, it is sufficient to check if their bases are mutually orthogonal.

(scalar product is linear, i.e. $\langle ax, y \rangle = a \langle x, y \rangle$
$\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
and each element $x \in X$, $y \in Y$ is linear combination of the basis.)

(xv) A family of vectors $X = \{x_\nu\}_{\nu \in F} \subset \mathbb{K}^I$ is called orthogonal if the vectors x_ν are pair-wise orthogonal, i.e.

$$\langle x_\nu, x_{\nu'} \rangle = 0 \quad \forall \nu, \nu' \in F, \nu \neq \nu'.$$

The family is called orthonormal, if it is orthogonal and

$$\langle x_\nu, x_\nu \rangle = 1 \quad \forall \nu \in F.$$

A matrix $A \in \mathbb{K}^{I \times J}$ is called <u>orthogonal</u>, if the columns of A are orthonormal, i.e.

$$A^H A = I \quad \in \mathbb{K}^{J \times J}.$$

An orthogonal, square matrix is called <u>unitary</u>, then

$$A^H A = A A^H = I,$$

i.e. $A^{-1} = A^H$.

(xvi) For either $J \subset I$ or $I \subset J$, a matrix $A \in \mathbb{K}^{I \times J}$ is called <u>diagonal</u>, if $A_{ij} = 0 \quad \forall\, i \in I,\ j \in J,\ i \neq j$. ←

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

1.2. Matrix Rank

$A =$  rows I

columns J

Proposition 1.1 Let $A \in \mathbb{K}^{I \times J}$. The following statements are equivalent and define <u>rank</u> $r = \operatorname{rank}(A) \in \mathbb{N}$.

(i) $r = \dim(\operatorname{range}(A))$ — $\operatorname{range}(A) = \{Ax : x \in \mathbb{K}^J\}$

(ii) $r = \dim(\operatorname{range}(A^H))$ — dim is cardinality of lin. basis

(iii) r is the max. number of linear independent rows of A.

(iv)

(v) r is minimal with the property

$$A = \sum_{i=1}^{r} \underbrace{a_i b_i^H}_{\text{matrix}}, \quad \text{where } a_i \in \mathbb{K}^I,\ b_i \in \mathbb{K}^J.$$

$$ab^H = \begin{pmatrix} a_1 \\ \vdots \\ a_I \end{pmatrix} \cdot \underbrace{\begin{pmatrix} b_1 & \cdots & b_J \end{pmatrix}}_{J} = \begin{pmatrix} & & \\ & I \times J & \\ & & \end{pmatrix} = A \in \mathbb{K}^{I \times J}$$

$$\Rightarrow \quad A = \underbrace{\begin{pmatrix} & & \\ & I \times J & \\ & & \end{pmatrix} + \cdots + \begin{pmatrix} & & \\ & I \times J & \\ & & \end{pmatrix}}_{r \text{ times (minimum)}}$$

is sum of tensor products $a \otimes b := ab^H$.

(vi) r is maximal with the property that there exists an invertible $r \times r$ submatrix of A.

(vii) r is the number of positive singular values (proven later).

Remark:

- $r_{max} = \min\{\#I, \#J\}$.
- rank of real valued matrix does not change when considered in $\mathbb{C}$.
- Denote the set of matrices of bound rank $r \leq k$ by
$$\mathcal{R}_k = \{A \in \mathbb{K}^{I \times J} : \operatorname{rank}(A) \leq k\}.$$
$\mathcal{R}_k$ is <u>not</u> a vector space. e.g. $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

1.3. Matrix Norms

1.3.1. Norms

Let $\mathbb{K}$ be a field, V a vector space over $\mathbb{K}$.

We call a function $\|\cdot\| : V \to [0, \infty)$ a norm, if

(i) $\|v\| = 0 \iff v = 0$, $v \in V$.

(ii) $\|\lambda v\| = |\lambda| \|v\|$, $v \in V$, $\lambda \in \mathbb{K}$.

(iii) $\|v + w\| \leq \|v\| + \|w\|$, $v, w \in V$, called triangle inequality.

Rem: Norms are always continuous as a consequence of the (inverse) triangle inequality.

$$|\|v\| - \|w\|| \leq \|v - w\|, \text{ for all } v, w \in V.$$

We call the pair $(V, \|\cdot\|)$ normed vector space.

1.3.2. Scalar Products, Hilbert spaces, projections

A normed vector space $(V, \|\cdot\|)$ is a pre-Hilbert space if its norm is defined by

$$\|v\| = \sqrt{\langle v, v \rangle}, \quad v \in V,$$

(vector with itself)

where $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{K}$ is a scalar product on V, i.e. it fulfills the following properties:

"How does the vector v look like, seen from y?"

https://math.stackexchange.com/questions/1829952/when-is-a-norm-induced-by-an-inner-product

Note:
An inner product always gives a norm, by taking the square root
But a norm is only induced by an inner product if the parallelogram law holds:

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2)$$

(i) Positive definiteness

$\langle v, v \rangle > 0$ for $v \neq 0$

(ii) Conjugate symmetry

$\langle v, w \rangle = \overline{\langle w, v \rangle}$ for $v, w \in V$.

(iii) Linearity in first argument

$\langle u + \lambda v, w \rangle = \langle u, w \rangle + \lambda \langle v, w \rangle$ for $u, v, w \in V$, $\lambda \in \mathbb{K}$.

(ii)+(iii) gives: $\langle w, u + \lambda v \rangle = \langle w, u \rangle + \overline{\lambda} \langle w, v \rangle$ for $u, v, w \in V, \lambda \in \mathbb{K}$.

We refer to this pre-Hilbert space by $(V, \langle \cdot, \cdot \rangle)$.

Remark: In a pre-Hilbert space $(V, \|\cdot\|)$ over $\mathbb{K}$, the triangle inequality follows from the Schwarz inequality

$$|\langle v, w \rangle| \leq \|v\| \|w\|, \quad v, w \in V.$$

Example: Euclidean norm on $\mathbb{K}^I$ (generated by $\langle v,w\rangle = v w^H$):

$$\|v\|_2 = \sqrt{\sum_{i\in I} (v_i)^2}$$

We call a pre-Hilbert space $(V, \langle\cdot,\cdot\rangle)$ which is complete, i.e. each Cauchy-Sequence is convergent wrt. its induced norm $\|\cdot\|$, a Hilbert space.

Remark: Each finite dimensional pre-Hilbert space is a Hilbert space.

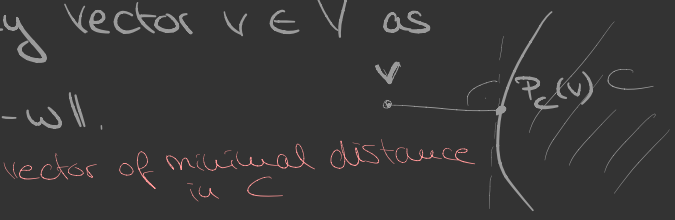
A set $C \subseteq V$, V vector space, is called convex, if

$$\forall v,w \in C,\ t \in [0,1]: \quad tv + (1-t)w \in C.$$

Given a closed, convex set C in a Hilbert space $(V, \langle\cdot,\cdot\rangle)$, we define a projection on C of any vector $v \in V$ as

$$P_C(v) = \operatorname*{argmin}_{w\in C} \|v-w\|.$$

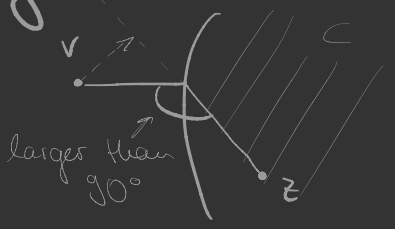
(vector of minimal distance in C)



Such projections fulfill the following inequality:

$$\mathrm{Re}\left(\langle z - P_C(v),\ v - P_C(v)\rangle\right) \leq 0.$$

for all $z \in C$.



Note: Convexity of C ensures uniqueness of projection.
This projection is also called orthogonal projection.

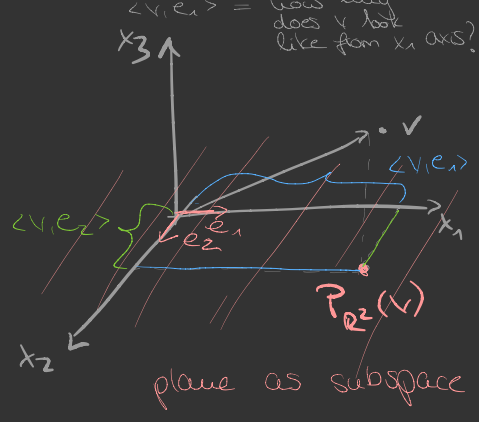
Calculation of $P_C(v)$: In case $C = W < V$ is a closed linear subspace of V and let $\{w_\nu\}_{\nu\in F}$ be an orthonormal basis of W, then

$$P_W(v) = \sum_{\nu\in F} \langle v, w_\nu\rangle w_\nu \quad \text{for all } v \in V.$$

Theorem: Pythagoras-Fourier-Theorem

$$\|P_W(v)\|^2 = \sum_{\nu\in F} |\langle v, w_\nu\rangle|^2$$

for all $v \in V$.



Remark: Applied to the special case $V = W$, we have that $P_W = id_W$ and hence the orthonormal expansion

$$v = \sum_{\nu \in J} \langle v, w_\nu \rangle w_\nu$$

and the norm equivalence

$$\|v\|^2 = \sum_{\nu \in J} |\langle v, w_\nu \rangle|^2.$$

1.3.3. Trace and matrix norms

For $A \in \mathbb{K}^{I \times I}$, we define the trace of A by

$$tr(A) = \sum_{i \in I} A_{ii}.$$

Proposition:

a) Let $A \in \mathbb{K}^{I \times J}$ and $B \in \mathbb{K}^{J \times I}$. Then

$$tr(AB) = tr(BA).$$

b) Circularity property

$\Downarrow$ For A, B, C: $tr(ABC) = tr(BCA)$.

c) Invariance under unitary transformations, i.e.

$$tr(A) = tr(UAU^*)$$

for $A \in \mathbb{K}^{I \times I}$ and unitary matrix $U \in \mathbb{K}^{I \times I}$.

d) $tr(A) = \sum_{i \in I} \lambda_i$ where $\{\lambda_i \mid i \in I\}$ set of eigenvalues of A.

Definition: Frobenius-Norm (aka Schur norm, Hilbert-Schmidt norm)

For $A \in \mathbb{K}^{I \times J}$ we define the Frobenius norm $\|\cdot\|_F$ as

$$\|A\|_F := \sqrt{\sum_{i \in I, j \in J} A_{ij}^2}.$$

It is generated by scalar product $\langle \cdot, \cdot \rangle_F$ defined by

$$\langle A, B \rangle_F := \sum_{i \in I} \sum_{j \in J} A_{ij} \overline{B_{ij}}$$

$$= tr(AB^*) = tr(B^*A).$$

$$AB^* = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \overline{b_{11}} & \overline{b_{21}} \\ \overline{b_{12}} & \overline{b_{22}} \end{pmatrix} = \begin{pmatrix} a_{11} \cdot \overline{b_{11}} + a_{12} \cdot \overline{b_{12}} & * \\ * & a_{21} \cdot \overline{b_{21}} + a_{22} \cdot \overline{b_{22}} \end{pmatrix}.$$

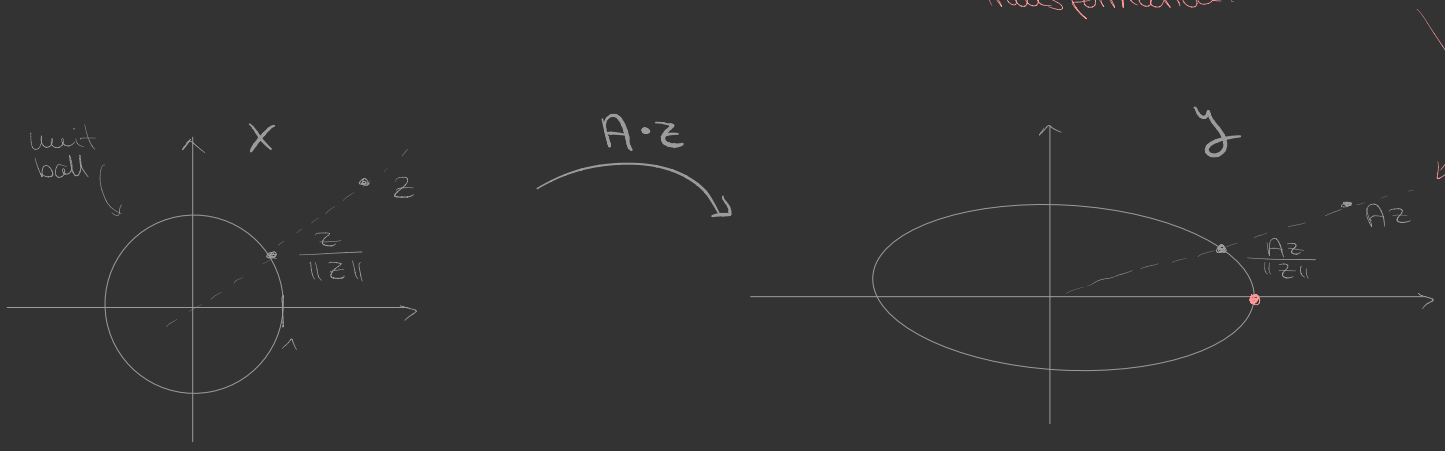
In particular $\|A\|_F^2 = \operatorname{tr}(AA^H) = \operatorname{tr}(A^HA) = \langle A, A\rangle_F$.

Definition:

Let $X = \mathbb{K}^I$, $Y = \mathbb{K}^J$ and $(X, \|\cdot\|_X), (Y, \|\cdot\|_Y)$ normed vector spaces. Then the associated matrix norm is

$$\|A\| := \|A\|_{X\to Y} := \sup_{z\neq 0} \frac{\|Az\|_Y}{\|z\|_X}, \quad A \in \mathbb{K}^{I\times J}$$

search for the 'worst possible' transformation.



Example: For $\|\cdot\|_X = \|\cdot\|_Y = \|\cdot\|_2$, the associated matrix norm $\|\cdot\|_\infty := \|\cdot\|_{X\to Y}$ is called spectral norm, often denoted by $\|A\|$.

Definition: A matrix norm $\|\cdot\|$ is said to be

(i) unitary invariant, if

$$\|A\| = \|UAV^*\|$$

for unitary matrices U, V.

(ii) submultiplicative, if

$$\|AB\| \leq \|A\|\cdot\|B\|.$$

Proof uses the property of the trace.

! unitary matrices represent rotations
⇒ no change of the magnitude.

Remark: Both $\|\cdot\|_F$ and $\|\cdot\|_\infty$ are unitary invariant and submultiplicative.

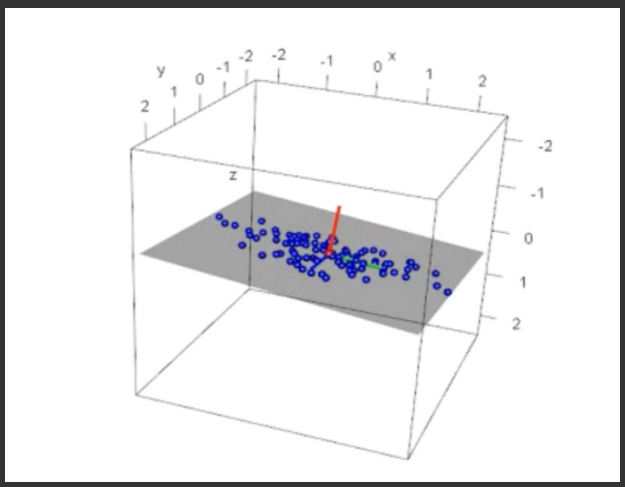
Also, $\|AB\|_F \leq \|A\|_\infty \cdot \|B\|_F \leq \|A\|_F \cdot \|B\|_F$.

Chapter 2: Singular Value Decomposition

Example:
- We can understand data as a stream of numbers, e.g. representation of color in an image in a matrix / table or spacial movement represented by geological coordinates.
- If we want to extract some information by simultaneously looking at all the data together
  ⇝ want to find correlation between the independent streams (images, paths)...
- For this, combine all the data in a single matrix and study its algebraic properties.
  These properties encode fundamental properties of the data.
- What we don't do: multidimensional tensors.
  ⇝ often transformed into "flat" matrix due to theoretical methods

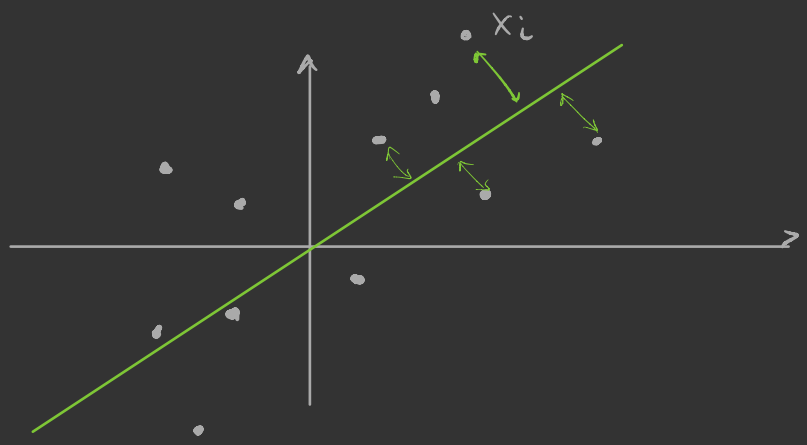
Idea: $A \in \mathbb{R}^{I \times J}$ with $n = |I|$ and $d = |J|$, where each row represents a data point in $\mathbb{R}^d$.
Want to now find the best k-dim. subspace to represent the data points

Aim: dimension reduction.



⇝ cut nicely through the "data cloud"

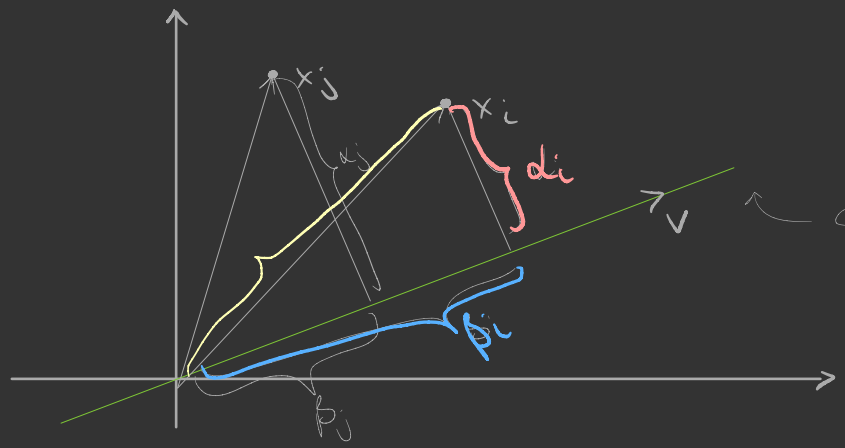
Do this by minimizing sum of squares between subspace and data points.
⇒ find the direction of maximal spread of the points in the data cloud.



① First step: Try to find a 1-dim subspace, i.e. line through the origin.

↝ later: finding best fitting k-dim subspace is found by applying ① k-times.

Therefore: $x_i = A^{(i)}$, i.e. the i-th row of matrix A.



Pythagoras Theorem gives:

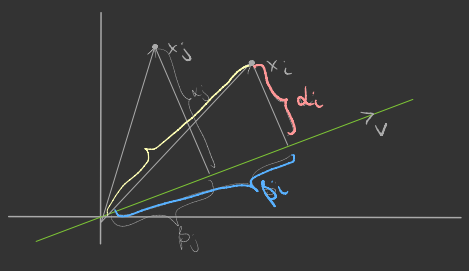
$a^2 + b^2 = c^2$

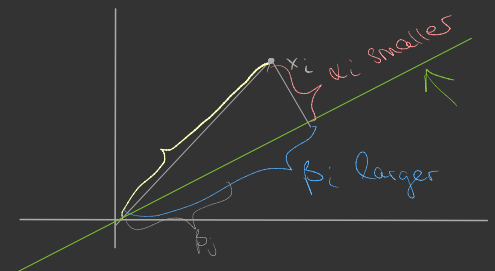
$\|x\|^2 = (\text{lenght of projection})^2 + (\text{distance of point to line})^2$

Want to find v to

minimize the distance of points to the line

$\Leftrightarrow$ maximize the length of the projection for each point





Let now v be the unit vector along the line.

Then the length of the projection is given by

$$|\langle A^{(i)}, v \rangle|$$

To do this for all data points simultaneously, we want to maximize $\|Av\|_2^2 = \sum_{i \in I} |\langle A^{(i)}, v \rangle|$,

verzerrt → meaning find the vector $v \in \mathbb{R}^d$ which gets maximally distorted by matrix A (since we start with $\|v\|_2^2 = 1$)
or want to find the direction in which the d-dimensional sphere gets maximally deformed.

We call this vector the first singular vector $v_1 \in \mathbb{R}^d$ or first principle component, thus

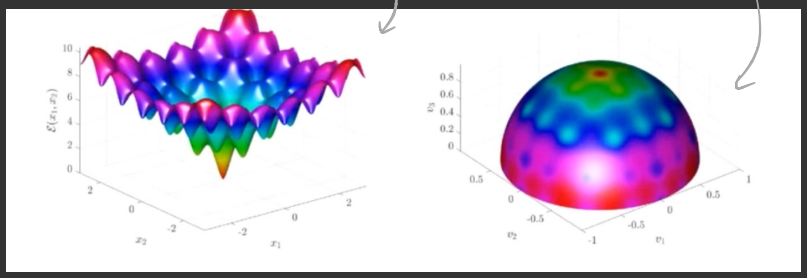
$$v_1 = \underset{\substack{v \in \mathbb{R}^d \\ \|v\|_2 = 1}}{\operatorname{argmax}} \|Av\|_2$$

compare spectral norm

and value $\sigma_1(A) = \|Av_1\|_2$ is called the first singular value of A.

Examples

1) function we want to find the minimizer for → set of potential minimizers



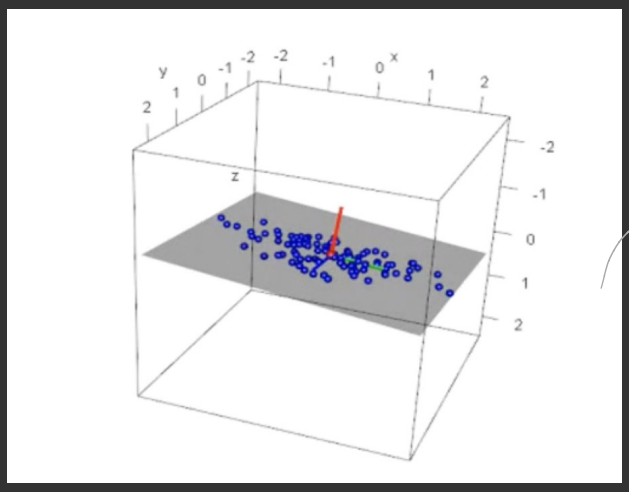
2)

Figure 12: Samples from the *10K US Adult Faces Database* [7] and one instance of outlier.

Want to compute the "most typical face"

↳ "Eigenface" computed with SVD.

② Go from 1-dim. subspace to 2-dim. subspace.



↪ plane with plane

greedy approach: we keep v_1 and then try to find a new vector v which is perpendicular to v_1 and again maximizes the distortion:

$$v_2 = \underset{\substack{\|v\|=1 \\ \langle v_1, v\rangle = 0}}{\operatorname{argmax}} \|Av\|_2.$$

maximally "different" to v_1

Then $\sigma_2(A) = \|Av_2\|_2$ is second singular value of A

↪ continue like this for kth vector v_k, s.t.

$$v_k = \underset{\substack{\|v\|_2 = 1 \\ \langle v_1, v\rangle = 0, \ldots, \langle v_{k-1}, v\rangle = 0}}{\operatorname{argmax}} \|Av\|_2.$$

Note: Process stops when we have found $v_1, \ldots, v_r$ singular vectors and

$$0 = \max_{\substack{\|v\|_2 = 1 \\ \langle v_1, v\rangle = \ldots = \langle v_r, v\rangle = 0}} \|Av\|_2.$$

Reason: $r = \operatorname{rank}(A)$, i.e. dimension of $\operatorname{range}(A)$.

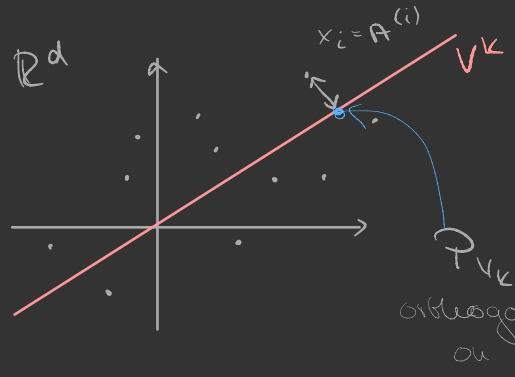
Problem: Hard optimization algorithm, but can be made much easier: "Power Method" (⇝ later)

And: Does one-by-one optimization actually give the optimal subspace?
Yes, due to Pythagoras theorem and properties of the Euclidean distance. ⇒ Proposition 2.1.

Vorlesung 4 | 29.04.20

Proposition 2.1: Let $A \in \mathbb{K}^{I \times J}$ and $v_1, v_2, \dots, v_r$ be singular values defined above. For $1 \le k \le r$, let V_k be the subspace spanned by $v_1, v_2, \dots, v_k$. Then for each k, V_k is the best fit subspace for A.

Proof:



u-point data cloud,
$\dim V_k = k \le r = \operatorname{rank}(A)$

$$\sum_{i=1}^{u} \| A^{(i)} - P_V(A^{(i)}) \|_2^2 = \delta(V) \quad \text{(distance of } x_i = A^{(i)} \text{ to } V\text{)}$$

$$V_k = \operatorname*{argmin}_{\substack{V \\ \dim(V) \le k}} \delta(V)$$

Pythagoras

$$= \operatorname*{argmax}_{\dim(V) \le k} \sum_{i=1}^{u} \| P_V(A^{(i)}) \|_2^2$$

$$= \hat{V} \text{ optimal subspace}; \quad V_k = \operatorname{span}\{v_1, \dots, v_k\}.$$

to be proved

by induction:

$k = 1$: $V_1 = \operatorname{span}\{v_1\} = \hat{V}$, true by construction of v_1.

$k=2$: $\hat{V}$ is the optimal subspace. Fix ONB $\{w_1, w_2\}$ of V_1, i.e.

$$\hat{V} = \operatorname{span}\{w_1, w_2\} = \left\{ \sum_{i=1}^{2} \lambda_i w_i \right\}$$

and $\|w_i\|_2 = 1$, $\langle w_1, w_2 \rangle = 0$

For V_r with ONB $\{v_1, \dots, v_r\}$ over $\mathbb{R}$:

$$\sum_{i=1}^{m} \| P_{V_r}(A^{(i)}) \|_2^2 \quad \leftarrow \text{Sum over all data points contained in } A$$

Pythagoras-Fourier-Theorem $\rightarrow$

$$\sum_{i=1}^{m} \sum_{j=1}^{r} |\langle A^{(i)}, v_j \rangle|^2 \quad \leftarrow \text{Sum over all basis vectors}$$

Switching the sums $\rightarrow$

$$\sum_{j=1}^{r} \sum_{i=1}^{m} |\langle A^{(i)}, v_j \rangle|^2 \quad \leftarrow c_i \mapsto \text{Rotation of a matrix multiple}$$

Pythagoras-Fourier-Theorem $\rightarrow$

$$\sum_{j=1}^{r} \| A v_j \|^2$$

This ONB maximizes

$$\sum_{i=1}^{2} \| P_{\hat{V}}(A^{(i)}) \|_2^2 \overset{\text{Ex.}}{=} \| A w_1 \|_2^2 + \| A w_2 \|_2^2 .$$

We choose w_1, w_2 in such a way that w_2 is orthogonal to V_1.

Case 1: If $\hat{V} \perp \{V_1\}$, then any vector $w_2 \in \hat{V}$ will be orthogonal to V_1.

Case 2: If $\hat{V} \not\perp \{V_1\}$, then we simply pick $w_2 \perp P_{\hat{V}}(V_1)$.

$\overset{\text{Ex.}}{\Rightarrow} w_2 \perp V_1$.

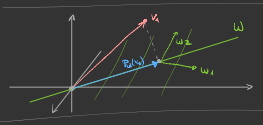
$w_2 \perp P_{\hat{V}}(V_1) \Rightarrow w_2 \perp V_1$

We use the unique decomp. of V_1 as

$V_1 = \underbrace{P_{\hat{V}}(V_1)}_{\in \hat{V}} + \underbrace{(V_1 - P_{\hat{V}}(V_1))}_{\in \hat{V}^\perp}$

and because (by definition of $\hat{V}^\perp$)

$\langle \underbrace{w_2}_{\in \hat{V}}, \underbrace{V_1 - P_{\hat{V}}(V_1)}_{\in \hat{V}^\perp} \rangle = 0$



V_1 maximizes $\| AV \|_2^2$ by construction of V_1.

$$\Rightarrow \| A w_1 \|_2^2 \leq \| A V_1 \|_2^2 \qquad (1)$$

V_2 maximizes $\| AV \|_2^2$ among all vectors s.t. $V \perp V_1$.

In particular $w_2 \perp V_1$

$$\Rightarrow \| A w_2 \|_2^2 \leq \| A V_2 \|_2^2 \qquad (2)$$

(1), (2)

$$\Rightarrow \| A w_1 \|_2^2 + \| A w_2 \|_2^2 \leq \| A V_1 \|_2^2 + \| A V_2 \|_2^2 .$$

$\Rightarrow V_2$ is an optimal subspace.

This reasoning can be extended to the case of $k > 2$.

We learned: $V_k = \operatorname{span}\{v_1, \dots, v_k\}$ with v_k iteratively computed s.t.

$$v_k = \underset{\substack{\|v\|_2^2 = 1 \\ v \perp v_i,\ i = 1, \dots, k-1}}{\operatorname{argmax}} \| A v \|_2^2$$

gives best fitting subspace.

Remark: With the definitions above, we get that

$$\sigma_1(A) \geq \sigma_2(A) \geq \dots \geq \sigma_r(A),$$

(simply by construction, otherwise V_i would be V_k instead)

Note: We can understand AV_i as a list of lengths (with sign) of the projection of the rows of A onto V_i.

$$AV_i = \begin{pmatrix} — A^{(1)} — \\ — A^{(2)} — \\ \vdots \\ — A^{(n)} — \end{pmatrix} \begin{pmatrix} | \\ V_i \\ | \end{pmatrix} = \begin{pmatrix} \langle A^{(1)}, V_i \rangle \\ \langle A^{(2)}, V_i \rangle \\ \vdots \\ \langle A^{(n)}, V_i \rangle \end{pmatrix}$$

length of projection of a_n onto V_i.

Remember: $v = \sum_{w \in F} \langle v, w_v \rangle w_v$ the decomposition of vector v into components w.r.t. $w_v \in F$.

and $\|v\|^2 = \sum_{v \in F} \|\langle v, w_v \rangle\|$

Hence: Think of $\sigma_i(A) = \|AV_i\|_2$.

Summing up all the components of A along V_i will give us the whole content of A.

Proof: Note that $\text{span}(\{A^{(i)}, i = 1, \dots, n\}) = \text{span}\{V_1, \dots, V_r\}$ (due to $\text{rank}(A) = r$). Therefore

$$\langle A^{(i)}, V \rangle = 0 \iff V_1, \dots, V_k \perp V.$$

($A^{(i)}$ linear combination of $V_1, \dots, V_k$)

Pythagoras $\Rightarrow$ $\|A^{(i)}\|_2^2 \overset{ONB}{=} \sum_{k=1}^{r} |\langle A^{(i)}, V_k \rangle|^2$

$$\Rightarrow \sum_{i \in I} \|A^{(i)}\|_2^2 = \sum_{k=1}^{r} \sum_{i \in I} |\langle A^{(i)}, V_k \rangle|^2 \overset{\text{Def. Matrix-Multipl.}}{=} \sum_{k=1}^{r} \|AV_k\|_2^2 \overset{\text{Def. Singular Values}}{=} \sum_{k=1}^{r} \sigma_k(A)^2.$$

Relations to norms:

(1) $\sum_{i \in I} \|A^{(i)}\|_2^2 = \sum_{i \in I} \sum_{j \in J} |A_{ij}|^2$ is the **Frobenius norm**.

and hence

$$\|A\|_F = \sqrt{\sum_{k=1}^{r} \sigma_k(A)^2}.$$

(2) $\sigma_1(A) = \max_{\|v\|_2=1} \|Av\|_2 = \max_{v \neq 0} \frac{\|Av\|_2}{\|v\|} = \|A\|$

is the spectral norm of A.

(3) Define Schatten-p-norms for matrices for $1 \leq p < \infty$ by

$$\|A\|_p = \left(\sum_{k=1}^{r} \sigma_k(A)^p \right)^{1/p}$$

and

$$\|A\|_\infty = \max_{k=1,\dots,r} \sigma_k(A) \overset{\text{by construction}}{=} \sigma_1(A).$$

[$p=2$: Frobenius-Norm, $p=\infty$: spectral norm, $p=1$: nuclear norm]

SVD in its full glory

Def: (i) $v_1, \dots, v_r$ are called (right) singular vectors.

(ii) we define the left singular vectors $u_1, \dots, u_r$ by

$$u_i = \frac{1}{\sigma_i(A)} A v_i$$

$\sigma_i(A) = \|Av_i\|_2$ i.e. normalization

Thm 2.2: $u_1, \dots, u_r$ are orthogonal.

[This is surprising since for ONB with $e_i = (0,\dots,0,\underset{\text{i-th entry}}{1},0,\dots,0)$ and $A = (A^{(1)} \dots A^{(d)})$ where columns are not orthogonal, e.g.

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Then $A \cdot e_j = A^{(j)}$ which are not orthogonal.]

Proof: We proof the orthonormality of $\{u_1, \dots, u_r\}$ by induction over r.

$r=1$ $\rightsquigarrow$ $\{u_1\}$ is ON. ✓

$r>1$, we try to apply the induction hypothesis on a deflated matrix

$$B = A - \sigma_1 u_1 v_1^H$$

(we hope that B has lower rank)

$$Bv_1 = Av_1 - \sigma_1 u_1 \underbrace{\langle v_1, v_1 \rangle}_{=1} \overset{\text{def. of } u_1, v_1}{\equiv} 0.$$

Hence, $v_1 \in \ker(B)$.

We claim that the first singular vector z of B is orthogonal to v_1:

If this were not the case, then $0 \neq z_1 = \langle z, v_1 \rangle v_1$.

$$\left\| B \frac{z - z_1}{\|z - z_1\|_2} \right\|_2 \overset{\text{lin.}}{=} \left\| \frac{Bz - Bz_1}{\|z - z_1\|_2} \right\|_2$$

$Bz_1 = \langle z, v_1 \rangle \underbrace{Bv_1}_{=0} = 0$

$$= \left\| \frac{Bz}{\|z - z_1\|_2} \right\|_2 > \|Bz\|_2$$

$\|z - z_1\|_2 < \|z\|_2 < 1$

z maximizes of $\|Bz\|_2$.

$\Rightarrow z \perp v_1$.

Let v be orthogonal to v_1, then

$$Bv = Av - \sigma_1 u_1 \underbrace{\langle v_1, v \rangle}_{=0} = Av$$

Hence, the singular vectors of B are those of A orthogonal to v_1, i.e. $\{v_2, \dots, v_r\}$

$\Rightarrow \operatorname{rank}(B) = r - 1$.

$\Rightarrow \{u_2, \dots, u_r\}$ ONB.

Remains to show: $u_j \perp u_1$, $j = 2, \dots, r$.

Assume $\exists j$ s.t. $\langle u_j, u_1 \rangle \neq 0$. For simplicity $\mathbb{K} = \mathbb{R}$. Wlog we assume $\langle u_j, u_1 \rangle > 0$.

Then for $\varepsilon > 0$

$$A \left(\frac{v_1 + \varepsilon v_j}{\|v_1 + \varepsilon v_j\|_2} \right) = \frac{Av_1 + \varepsilon \cdot Av_j}{\sqrt{\underbrace{\|v_1\|_2^2}_{=1} + \varepsilon^2 \underbrace{\|v_j\|_2^2}_{=1} + 2\varepsilon \underbrace{\langle v_1, v_j \rangle}_{=0}}}$$

construction of unit vector

$$= \frac{\sigma_1 u_1 + \varepsilon \sigma_j u_j}{\sqrt{1 + \varepsilon^2}} =: w$$

Pythagoras Thm.

$$\|w\|_2 \geq |\langle w, u_1 \rangle| = \left\langle u_1, \frac{\sigma u_1 + \varepsilon \sigma_j u_j}{\sqrt{1 + \varepsilon^2}} \right\rangle$$

Taylor expansion

$$= \left(\sigma_1 + \varepsilon \sigma_j \underbrace{\langle u_1, u_j \rangle}_{>0} \right) \left(1 - \frac{\varepsilon^2}{2} + \mathcal{O}(\varepsilon^4) \right)$$

$$= \sigma_1 + \varepsilon \underbrace{\sigma_j \langle u_1, u_j \rangle}_{>0} - \mathcal{O}(\varepsilon^2) > \sigma_1.$$

} following the induction iteration

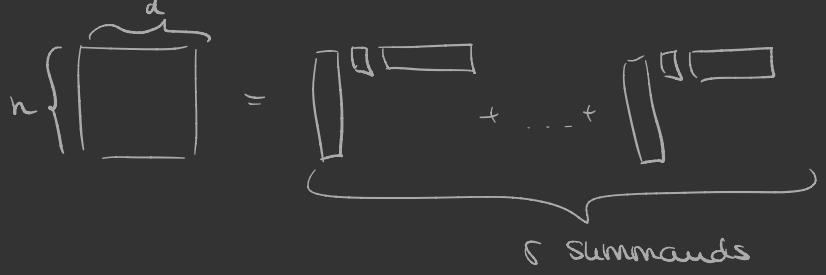
with larger norm than $\arg\max$ value

Theorem 2.3: Let $A \in \mathbb{K}^{I \times J}$ with right singular vectors $v_1, \dots, v_r$, left singular vectors $u_1, \dots, u_r$ and $\sigma_1, \dots, \sigma_r$ the corresponding singular values. Then

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^H.$$

Interpretation:

$$A - \sigma_1 u_1 v_1^H - \sigma_2 u_2 v_2^H - \dots - \sigma_r u_r v_r^H = 0.$$



To spell out the matrix for $A \in \mathbb{R}^{n \times d}$, one needs $\boxed{n \cdot d}$ scalar entries.
With this decomposition we need $\boxed{r \cdot (n + d + 1)}$ scalar entries.

$\Rightarrow$ $(n + d + 1) r \ll n \cdot d$ whenever $r \ll \min\{n, d\}$,
i.e. we save lots of space for sparse, low rank matrizes,
we are compressing the information of A.

$\Rightarrow$ enables us to extract relevant information from our data.

Example: image as a matrix containing pixel-wise color values.

neglect summands with smaller singular values

even though very pixelated, but still contain sufficient information

Vorlesung 5

5.5.20

What can we use it for?

Given $A = \sum_{i=1}^{r} \sigma_i u_i v_i^H$ for $A \in \mathbb{K}^{[x]}$.

▷ **Best rank-k approximation** remember: singular values are ordered by magnitude

Idea: discard information with "negligible" singular values

↳ compress matrix

$$A_k = \sum_{l=1}^{k} \sigma_l u_l v_l^H \qquad \text{"truncated SVD"}$$

Note: ▷ A_k has rank k by construction

Lemma 2.4: Rows of A_k are orthogonal projections of rows of A onto subspace $V_k = \text{span}\{\underbrace{v_1, \dots, v_k}_{\text{first } k \text{ singular vectors of } A}\}$

Proof: Assume $\mathbb{K} = \mathbb{R}$.

Consider $A^{(i)}$ arbitrary row of A and $V_k = \text{span}\{v_1, \dots, v_k\}$ with V_k optimal k-dim best fitting space, i.e.

$$V_k = \underset{\dim V \le k}{\arg\min} \sum_{i=1}^{n} \| A^{(i)} - P_V(A^{(i)}) \|_2^2 .$$

We know that

$$\underbrace{P_{V_k}(A^{(i)})^T}_{\text{PF-Theorem for orthogonal projection}} = \sum_{l=1}^{k} \langle A^{(i)}, v_l \rangle v_l^T$$

and

$$\langle A^{(i)}, v_l \rangle = (Av_l)_i$$

With this we get that

$$\sum_{l=1}^{k} Av_l \cdot v_l^T = \sum_{l=1}^{k} \underbrace{\sigma_l u_l}_{= Av_l} v_l^T = A_k ,$$

i.e. the rows of A_k are the orthogonal projections of the rows of A onto V_k.

Theorem 2.5: For any matrix B of rank at most k

$$\|A - A_k\|_F \leq \|A - B\|_F$$

i.e. A_k performs the best possible approximation w.r.t. $\|\cdot\|_F$ of rank $\leq k$.

Proof: $R_k := \{C \mid \operatorname{rank} C \leq k\}$, $B \in R_k$.

Assume B is minimizing $\|A - B\|_F^2$.

Let V be the space spanned by the rows of B, clearly $\dim V = \operatorname{rank} B \leq k$.

$$\underbrace{\|A - B\|_F^2}_{\text{it's minimal}} \overset{(*)}{=} \sum_{i \in I} \|A^{(i)} - B^{(i)}\|_2^2$$

$\|\cdot\|_2$ ← Euclidean norm

$(*)$ $v \in \mathbb{R}^d$,

$$\begin{cases} \|v\|_p = \left(\sum_{j=1}^{d} |v_j|^p\right)^{1/p} \\ \|v\|_\infty = \max_{j=1,\dots,d} |v_j| \end{cases}$$

defines the p-norm for $1 \leq p < \infty$.

$B^{(i)} = P_V(A^{(i)})$ is the best choice to minimize $\|A^{(i)} - B^{(i)}\|$ by definition.

B minimizes

$\Rightarrow$ $\|A - B\|_F^2 = \sum_{i \in I} \|A^{(i)} - P_V(A^{(i)})\|_2^2$

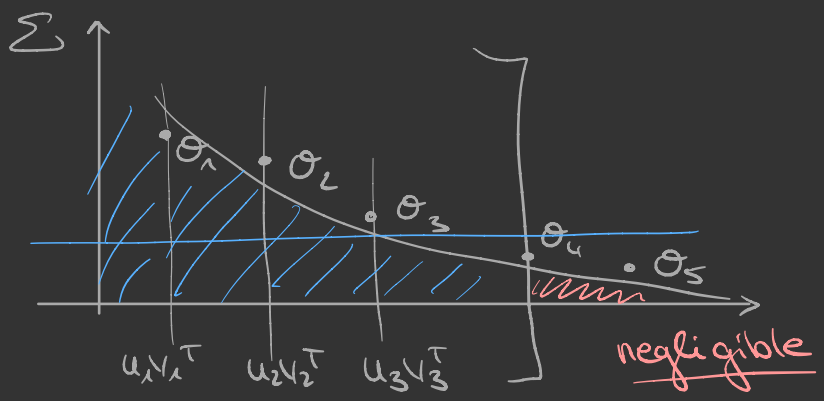
which precisely is the definition of V_k.

$\rightarrow$ best k-dim. fitting subspace $V = V_k$.

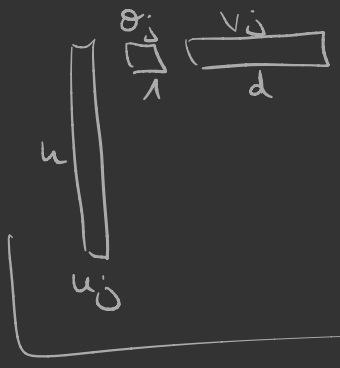
Lemma 2.4

$\Rightarrow$ $B = A_k$.

Interpretation:



$(d+n+1)(r-k) =$



$\Rightarrow$ compress a matrix which has in principle $n \times d$ components to a matrix defined by only

$$(n+d+1)k$$

parameters.

$\Rightarrow$ For $k \ll r \ll \min\{n,d\}$, A_k is truely low complexity i.e.

$$(n+d+1)k \ll (n+d+1)r \ll n \times d.$$

Approximating A in terms of $\|\cdot\|_F$ is done optimally by A_k.

Remark: A_k is the k-rank best approximation to A in any Schatten-p-norm

$$\|A\|_p = \underbrace{\sum_{i=1}^{r} \sigma_i^p}_{\text{p-norm of singular values}}{}^{1/p} = \|\Sigma\|_p$$

with $\Sigma = (\sigma_1, \dots, \sigma_r)^T$.

This means for all p:

$$A_k = \underset{B \in R_k}{\operatorname{argmin}} \|A - B\|_p$$

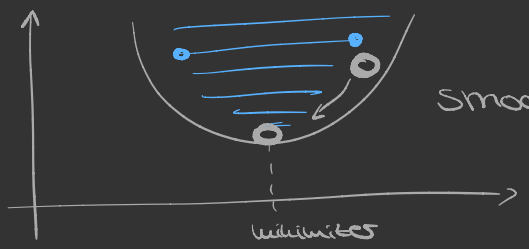
Optimization problems

▷ "Easy" optimizations:

↳ typically unconstrained

↳ objective function to be minimized is smooth
(differential with differential which is at least Lipschitz-continous)

↳ objective function is convex.



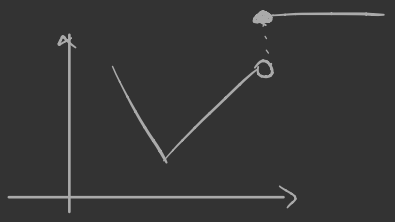
smooth + epigraph is convex

use gradient descent for minimization

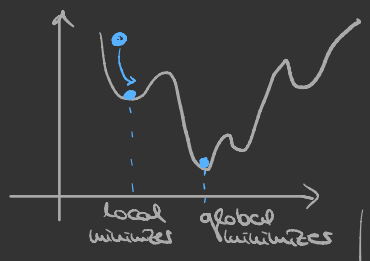
▷ "Hard" optimizations:

↳ constraint domain which is very complicated + not convex.

↳ non-smooth objective function
gradient information is not available



↳ non-convex functions
with multiple local minimizers



<u>in our case</u>: $A_k = \underset{B \in R_k}{\operatorname{argmin}} \|A - B\|_p$

- norms are convex
- depending on p, the norm might not be diff'able ($p=1$)
- set of R_k is <u>no</u> subspace i.e. $A + B = C$ for $\operatorname{rank}(C) > k$, with $A \in R_k$, $B \in R_k$

<u>but</u> solveable in closed form with SVD.

Best k-rank approximation w.r.t. $\|\cdot\|_\infty$ spectral norm

Lemma 2.6: $\|A - A_k\|_\infty^2 = \sigma_{k+1}^2$

Proof: $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

$$\Rightarrow A - A_k = \sum_{i=k+1}^{r} \sigma_i u_i v_i^T.$$

Assume v is the first right singular vector of $A - A_k$. Then v can be expressed as linear combination of $v_1, \dots, v_r$:

$$v = \sum_{\ell=1}^{r} \alpha_\ell v_\ell.$$

Then

$$\|(A - A_k)v\|_2 = \Big\| \sum_{i=k+1}^{r} \sigma_i u_i v_i^T \cdot \sum_{\ell=1}^{r} \alpha_\ell v_\ell \Big\|_2$$

orthogonality of v_ℓ's

$$= \Big\| \sum_{\ell=k+1}^{r} \alpha_\ell \sigma_\ell u_\ell \underbrace{v_\ell^T \cdot v_\ell}_{=1} \Big\|_2$$

$$= \Big\| \sum_{\ell=k+1}^{r} \alpha_\ell \cdot \sigma_\ell u_\ell \Big\|_2$$

Fourier-Pythagoras

$$= \Big(\sum_{\ell=k+1}^{r} \alpha_\ell^2 \cdot \sigma_\ell^2 \Big)^{1/2}$$

Since $\sigma_{k+1} \geq \sigma_{k+2} \geq \dots \geq \sigma_r$, we get that

$$(0, \dots, 0, 1, 0, \dots, 0) = \underset{\|\alpha\|_2 = 1}{\operatorname{argmax}} \sum_{\ell=k+1}^{r} \alpha_\ell^2 \cdot \sigma_\ell^2$$

($(k+1)$-th entry)

and hence

$$v = v_{k+1}$$

such that

$$\|A - A_k\| = \|(A - A_k) v_{k+1}\| = \sigma_{k+1}. \qquad \square$$

Theorem 2.7: $A_k = \underset{B \in R_k}{\operatorname{argmin}} \|A - B\|_\infty$.

Proof: Assume $\operatorname{rank} A > k$.

By Lemma 2.6. $\|A - A_k\|_2^2 = \theta_{k+1}^2$.

Assume $\exists B \in R_k$ s.t. $\|A - B\|_\infty < \theta_{k+1} = \|A - A_k\|_\infty^2$.

Since $\operatorname{rank} B \leq k$, then

$$\dim(\ker B) \geq d - k \qquad \text{for } d = |J|.$$

Consider $V_{k+1} = \operatorname{Span}\{v_1, \dots, v_{k+1}\}$, then $\dim V_{k+1} = k+1$ and

$$\dim\Big(\underbrace{\ker B}_{\dim \geq d-k} \cap \underbrace{V_{k+1}}_{\dim = k+1}\Big)$$

$\dim(V+W) = \dim(V) + \dim(W) - \dim(V \cap W)$

$$= \underbrace{\dim(\ker B)}_{\geq d-k} + \underbrace{\dim(V_{k+1})}_{=k+1} - \underbrace{\dim(\ker B + V_{k+1})}_{\leq d} \geq 1$$

Hence, $\exists z \in (\ker B \cap V_{k+1})$ with $\|z\|_2 = 1$. We want to show that

$$\|(A-B)z\|_2 \geq \theta_{k+1}.$$

We have that

$$\|(A-B)\|^2 \geq \|(A-B)z\|_2^2$$

$$\overset{z \in \ker B}{=} \|Az\|_2^2$$

$$= \Big\| \sum_{l=1}^{u} \theta_l u_l v_l^T z \Big\|_2^2$$

$$= \sum_{l=1}^{u} \theta_l^2 \, |v_l^T z|^2$$

$$\overset{z \in V_{k+1}}{=} \sum_{l=1}^{k+1} \theta_l^2 \, |v_l^T z|^2$$

$$\geq \theta_{k+1}^2 \underbrace{\sum_{l=1}^{k+1} |v_l^T z|^2}_{\overset{z \in V_{k+1}}{=} \|z\|_2^2 = 1} = \theta_{k+1}^2$$

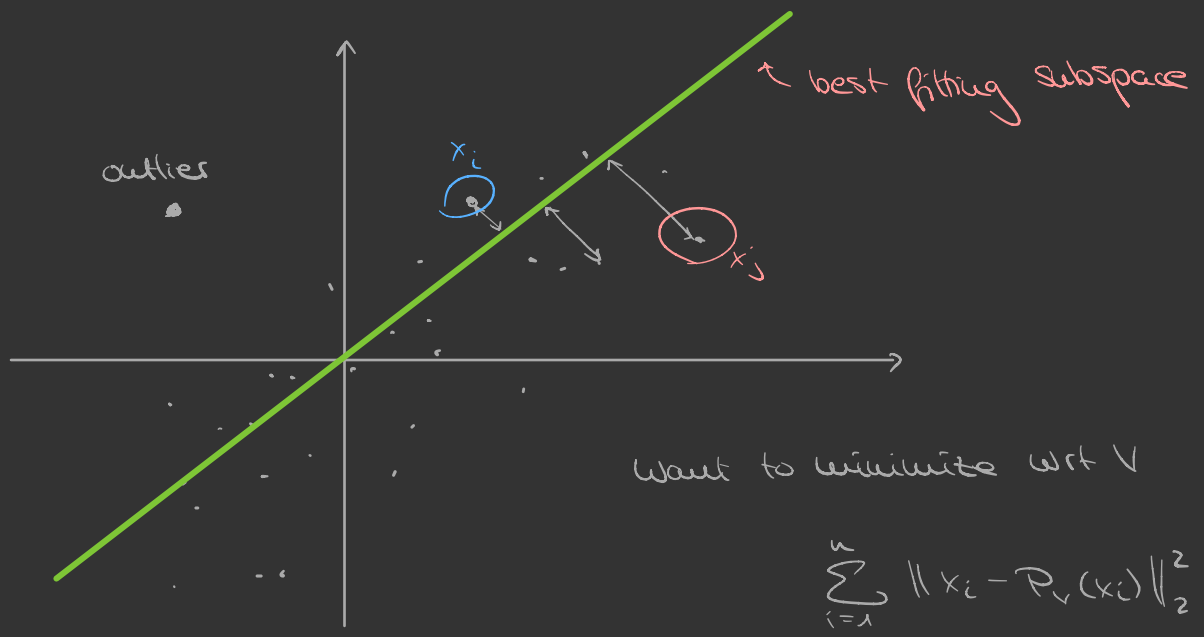
We have proven that

$$\theta_{k+1}^2 > \|A-B\|^2 \geq \|(A-B)z\|_2^2 = \|Az\|_2^2 \geq \theta_{k+1}^2$$

(assumption)

$$\Rightarrow \|A-B\| \geq \|A-A_k\|. \qquad \blacksquare$$

Repetition on the intuition

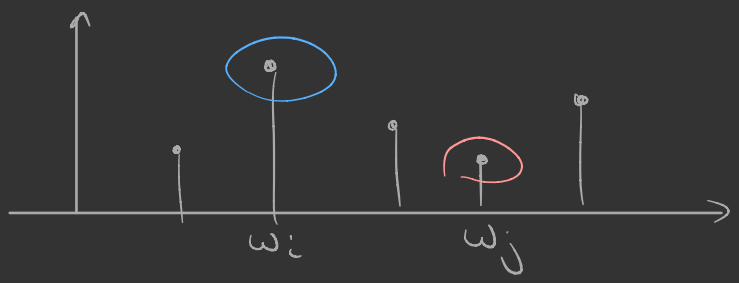


Want to minimize wrt V

$$\sum_{i=1}^{n} \|x_i - P_V(x_i)\|_2^2$$

But: there might be some points that are more important than others

↝ introduce a sequence of weights $w_j \geq 0$, i.e.



then x_i is better approximated than x_j.

↝ good way to handle outliers and counterweights that e.g. $\|\cdot\|_2$ strongly considers outliers.

↝ alternative: use norms with $0 < p < 1$.

but this gives us a non-smooth, non-convex and constraint optimization problem.

Also: we can use other distances than the perpendicular distance, but it heavily depends on the application (would give more weight to a specific feature).

Vorlesung 6

06.05.20

Everything so far boils down to the calculation of the SVD.

However: calculation of right singular vectors is itself an optimization problem
⇒ maximize over convex function

$$v_k = \underset{\|v\|_2 = 1}{\operatorname{argmax}} \|Av\|_2$$

$\|Av\|_2$: convex objective function ✓

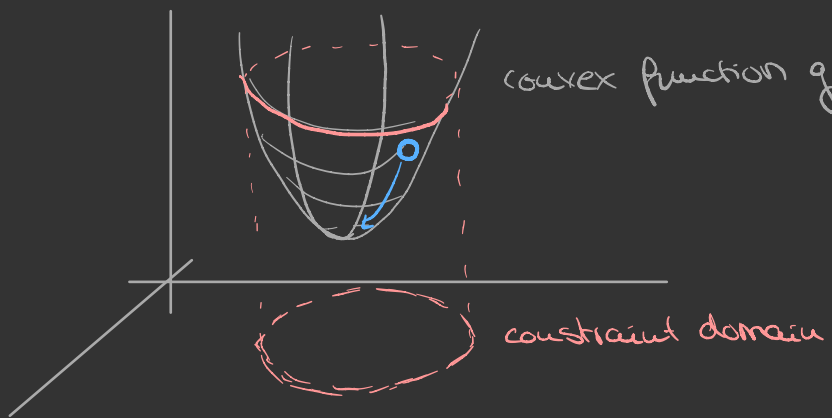
Ball is convex, but we need to stay on sphere → surface of ball ⇒

$\|v\|_2 = 1$

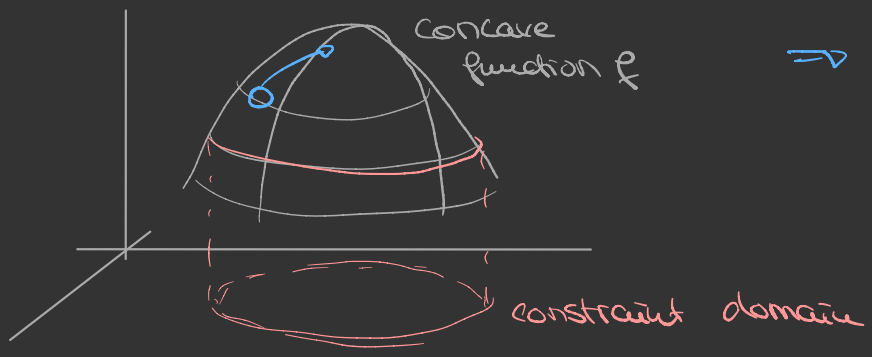
$\langle v_1, v \rangle = \langle v_2, v \rangle = \dots = \langle v_{k-1}, v \rangle = 0$: linear constraints ✓

⇒ we could optimize over convex ball if minimizing, since it's scaleable, but here we need to maximize.

More details on this assertion:



⇒ good for minimization



⇒ good for maximization

i.e. minimize convex function $-f$

$\Rightarrow$ Since scaling any solution that already is close to a maximum to the boarder of the ball will only further increase the function value (function is convex), we can simply consider the relaxation

$$v_k = \underset{\substack{\|v\| \leq 1 \\ \langle v_1, v \rangle = \ldots = 0}}{\operatorname{argmax}} \|Av\|_2$$

Additional challenges: "curse of dimensionality"

- We usually work with Big Data, i.e. several billions of data points in millions of dimensions with pictures for example
- the more dimensions, the more directions v we can chose from, so the complexity of our algorithm will scale depending on our dimension d.

Goal: Find an algorithm with complexity linear in d.

Method: "Power Method"

1) Easy case first:
Assume $A \in \mathbb{R}^{n \times n}$ and symmetric. Then the left singular vectors equal the right singular vectors. Then

$$A = \sum_{k=1}^{r} \sigma_k v_k v_k^T$$

For SVD we have

$$A^2 = \left(\sum_{k=1}^{r} \sigma_k v_k v_k^T \right) \left(\sum_{l=1}^{r} \sigma_l v_l v_l^T \right)$$

$$= \sum_{k,l} \sigma_k \sigma_l v_k \underbrace{v_k^T v_l}_{= \delta_{kl} \text{ due to orthogonality}} v_l^T$$

$$= \sum_{k=1}^{r} \sigma_k^2 v_k v_k^T.$$

Induction shows: $A^m = \sum_{k=1}^{r} \sigma_k^m v_k v_k^T.$

In case $\sigma_1 \gg \sigma_2$, we would get that

$$\lim_{m \to \infty} \frac{A^m}{\sigma_1^m} = v_1 v_1^T.$$

Indeed:

$$\frac{A^m}{\sigma_1^m} = \sum_{k=1}^{r} \frac{\sigma_k^m}{\sigma_1^m} v_k v_k^T = v_1 v_1^T + \sum_{k=2}^{r} \underbrace{\left(\frac{\sigma_k}{\sigma_1}\right)^m}_{<<1 \xrightarrow{m \to \infty} 0} v_k v_k^T.$$

We can easily extract v_1 from $v_1 v_1^T$ by taking the square root of the diagonal of the matrix:

$$\begin{pmatrix} v_1^{(1)} \\ v_1^{(2)} \\ \vdots \\ v_1^{(d)} \end{pmatrix} \cdot \begin{pmatrix} v_1^{(1)} & v_1^{(2)} & \dots & v_1^{(d)} \end{pmatrix} = \begin{pmatrix} (v_1^{(1)})^2 & v_1^{(1)} v_1^{(2)} & \dots & v_1^{(1)} v_1^{(d)} \\ v_1^{(2)} v_1^{(1)} & (v_1^{(2)})^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ v_1^{(d)} v_1^{(1)} & \dots & & (v_1^{(d)})^2 \end{pmatrix}.$$

While calculation of σ_1 is an difficult optimization problem corresponding with the spectral norm A_1, it is easy to compute the Frobenius norm $\|A\|_F$.

Spectral norm:
$$\sigma_1 = \max_{\|v\|_2 = 1} \|Av\|_2 = \max_{v \neq 0} \frac{\|Av\|_2}{\|v\|_2}$$

Frobenius norm:
$$\|A\|_F = \left(\sum_i \sum_j |A_{ij}|^2 \right)^{1/2}$$

With this

$$\|A^m\|_F = \left(\sum_{i,j} \left(\underbrace{A \times \dots \times A}_{m\text{-times}} \right)_{ij} \right)^{1/2}$$

we get

$$\lim_{m \to \infty} \frac{A^m}{\|A^m\|_F} = v_1 v_1^T \quad \text{as soon as } \sigma_1 \gg \sigma_2.$$

Proof: Exercise.

Problem with this calculation

- need to calculate A^m for $m \to \infty$.
  $A_{m+1} = A \times A^m$ ⇝ polynomial complexity wrt. dim. of A i.e. $C \geq n.d.$
- need to calculate the Frobenius norm
  $$v_{m+1} = \frac{A_{m+1}}{\|A_{m+1}\|_F}$$

2) If $A \in \mathbb{R}^{n \times d}$ is not symmetric, $n \neq d$, we use

$$A^T A = \ldots = \sum_{k=1}^{r} \sigma_k^2 v_k v_k^T$$

which itself is again squared and symmetric.

Then $$\lim_{m \to \infty} \frac{(A^T A)^m}{\|A^T A\|_F} \overset{\sigma_1 \gg \sigma_2}{=} v_1 v_1^T .$$

Now: To reduce complexity we fix any vector $x \in \mathbb{R}^d$, $\|x\|_2 = 1$ and calculate

$$(A^T A)^m \cdot x = A^T A \left(A^T A \cdot \ldots \cdot (\underbrace{A^T A x}_{\in \mathbb{R}^d}) \ldots \right)$$

which only requires a single matrix-matrix-multiplication, followed by significantly less complex matrix-vector-multipl.

Then $$\lim_{m \to \infty} \frac{(A^T A)^m x}{\|A^T A^m\|_F} = v_1 v_1^T \cdot x = \underbrace{\langle v_1, x \rangle}_{\in \mathbb{R}} \cdot v_1 .$$

Two issues at this point

- we need $\sigma_1 \gg \sigma_2$ for it to work
- What if we pick x, s.t. $\langle v_1, x \rangle = 0$? (numerically)

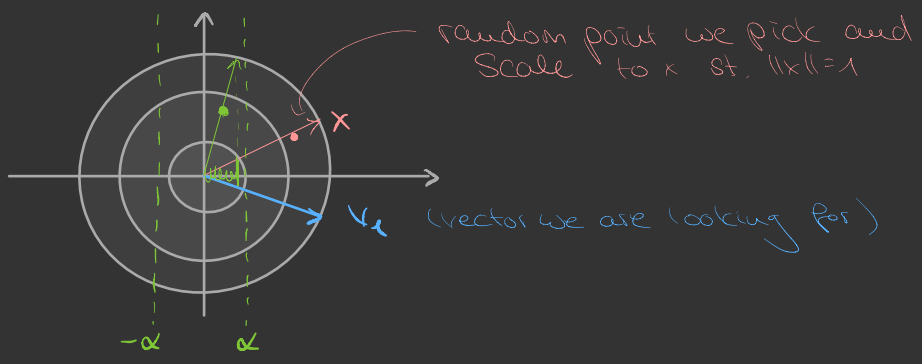
Then our limit will approx. 0, and we have no useful result.

⇒ To solve the "curse of dimension", we will start using probability (randomness) (adding to LA + optimization)

From this point on, we enter the space of randomized algorithms.

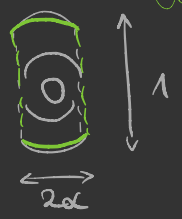
Example: Playing darts:

Terrible dart player will have same probability for any point on the target.



$\Rightarrow$ Probability to pick a point x which would be scaled to be v_1 exactly is equal to zero

$\Rightarrow$ Probability to pick a point such that $|\langle v_1, x\rangle| \leq \alpha$ for $\alpha \in \mathbb{R}$ small is volume of the area



as part of $S^{d-1} \subseteq \mathbb{R}^d$.

Remember: $V(d)$ = volume of B^d

$$\simeq \frac{1}{\sqrt{d\pi}} \left(\frac{2\pi e}{d}\right)^{d/2}$$

Considering the probability to choose $x \in S$, $S \subseteq B^d$, then

$$P(x \in S) = \frac{Vol(S)}{V(d)} \in [0,1].$$

With this, we get that probability to pick $x \in \mathbb{R}^d$ at random, such that $|\langle v_1, x\rangle| \leq \alpha$, α small, is given by

Lemma 28: Let $x \in \mathbb{R}^d$ be a unit d-dimensional vector of components $x_1, \dots, x_d \in \mathbb{R}$. We pick this vector at random. (*)
Then, for $\alpha > 0$

$$P(|x_1| \geq \alpha) \leq 1 - 2\alpha\sqrt{d-1}. \qquad (1)$$

Note: If α is small enough, the probability in (1) is very close to 1.

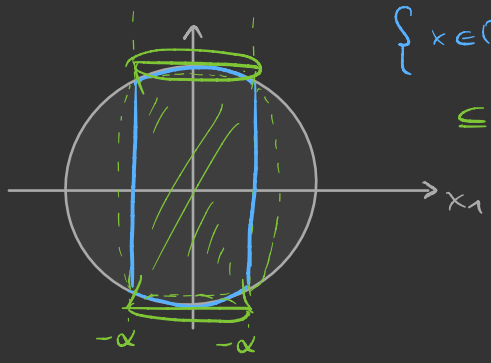
Remark: This corresponds with the assumption that $v_1 = e_1$.
We can use this assumption wlog, because we can always rotate the coordinate system accordingly.

Then $\langle x, v_1 \rangle = \langle x, e_1 \rangle = x_1$.

Proof: Firstly, note that $\mathbb{P}(|x_1| \geq \alpha) = 1 - \mathbb{P}(|x_1| < \alpha)$.
Hence, we will prove that

$$\mathbb{P}(|x_1| < \alpha) \leq 2\alpha\sqrt{d-1}.$$

For this, lets picture the situation first

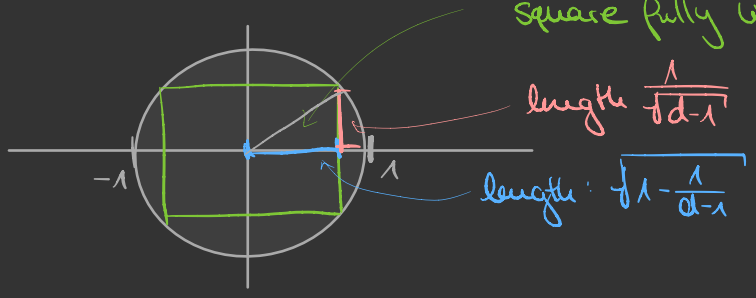


$$\{x \in \mathbb{R}^d \mid \|x\|_2 \leq 1, |x_1| \leq \alpha\}$$
$$\subseteq \{z \in \mathbb{R}^d \mid \|(z_2, \dots, z_d)\| \leq 1, |z_1| \leq \alpha\}$$
cylinder

$$\Rightarrow \mathrm{Vol}(\{\}) = V(d-1)2\alpha$$
(Volume cylinder unit $(d-1)$ dim. Basis)

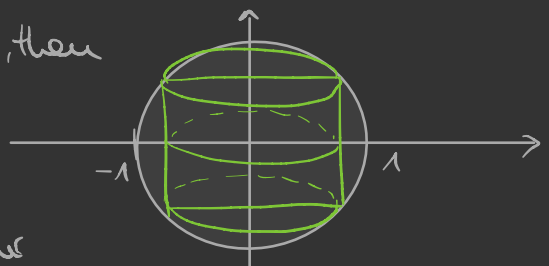
$$\Rightarrow \mathbb{P}(x \in \{\}) = \frac{\mathrm{Vol}(\{\})}{V(d)} \leq \frac{\mathrm{Vol}(\{\})}{V(d)} = \frac{2\alpha \cdot V(d-1)}{V(d)}.$$

One more geometrical idea:



It is fully contained, since $\left(\frac{1}{\sqrt{d-1}}\right)^2 + \left(\sqrt{1-\frac{1}{d-1}}\right)^2 = 1$.
Create a cylinder around the square, then

$$V(d) \geq 2 \underbrace{\frac{1}{\sqrt{d-1}}}_{\text{radius of cylinder}} \cdot \underbrace{\left(1-\frac{1}{d-1}\right)^{\frac{d-1}{2}} \cdot V(d-1)}_{\text{base of cylinder}}$$



using $V(R \cdot B^{d-1}) = V(B^{d-1}) R^d$ with R radius.

With $(1-\varepsilon)^a \geq 1 - a \cdot \varepsilon$, for ε small and a large, we get that (Bernoullische Ungleichung?)

$$\left(1 - \frac{1}{d-1}\right)^{\frac{d-1}{2}} \geq 1 - \left(\frac{d-1}{2}\right)\frac{1}{d-1} = \frac{1}{2}.$$

Simplifying further, we get that

$$V(d) \geq \frac{1}{\sqrt{d-1}} \cdot V(d-1)$$

denominator $\geq$ lb

$$\frac{\text{numerator}}{\text{denominator}} \leq \frac{\text{numerator}}{\text{lb}}$$

With this we get

$$P(x \in \{\}) = \frac{2\alpha \cdot V(d-1)}{V(d)} \leq \frac{2\alpha \cdot V(d-1)}{\frac{1}{\sqrt{d-1}} \cdot V(d-1)} = 2\sqrt{d-1}\,\alpha.$$

□

Remark 2.9: Lemma 2.8. shows that independently of d, the $x_1 = \langle x, u_1 \rangle$ component of a random unit vector x w.r.t. any orthogonal basis $\{u_1, \ldots, u_d\}$ is bounded away from zero with overwhelming probability.

Remark 2.10: Identifying $\mathbb{C}$ with $\mathbb{R}^2$, we can extend the result of Lemma 2.8 to random unit vectors in $\mathbb{C}^d$ as follows:
The probability that, for a randomly chosen unit vector $z \in \mathbb{C}^d$, $|z_1| \geq \alpha > 0$ holds is at least $1 - 2\alpha\sqrt{2d-1}$.

Vorlesung 7

12.05.

Theorem 2.11 (speed of convergence of the power method)

Let $A \in \mathbb{K}^{I \times J}$ and $x \in \mathbb{K}^I$ be a random, unit length vector.
Let V be the space spanned by the left singular vectors of A corresponding to singular values $\sigma_k \geq (1-\varepsilon)\sigma_1$, $k \in [\![1, r]\!]$.

Let $m \in \mathbb{N}$, $m \in \Omega\left(\frac{\ln(d^\beta/\varepsilon)}{\varepsilon}\right)$ for $\beta \geq 1/2$. Let w^* be the unit vector after m iterations of the power method, i.e.

$$w^* := \frac{(AA^H)^m \cdot x}{\|(AA^H)^m x\|_2}$$

Then, the probability that w^* has a component of at most $\underbrace{O\left(\frac{\varepsilon}{\alpha d^\beta}\right)}_{\text{error of the approximation}}$ orthogonal to V is $\geq 1 - 2\alpha\sqrt{2d-1}$.

Note: For α small, we get a better probability, but a worse approximation.

Proof: Let $A = \sum_{k=1}^{r} \sigma_k u_k v_k^H$ be the SVD of A. If $r \leq n = |I|$, complete $\{u_1, \dots, u_r\}$ to a orthonormal basis $\{u_1, \dots, u_r, \dots, u_n\}$ of $\mathbb{K}^n$.
Then for $x \in \mathbb{K}^n$:

$$x = \sum_{i=1}^{n} \langle x, u_k \rangle u_k.$$

Since

$$AA^H = \sum_{i=1}^{r} \sigma_k u_k v_k^H \cdot \sum_{i=1}^{r} \sigma_i v_k u_k^H$$

Cauchy-product $\rightarrow$

$$= \sum_{k=1}^{r} \sum_{j=0}^{r-1} \sigma_k u_k \underbrace{v_k^H \cdot \sigma_{k-j} \cdot v_{k-j}}_{\delta_{k,k-j} \text{ due to orthogonality.}} u_{k-j}^H$$

$$= \sum_{k=1}^{r} \sigma_k^2 u_k u_k^H$$

we get that

$$(AA^H)^2 = \left(\sum_{k=1}^{r} \sigma_k^2 u_k u_k^H\right)^2$$

$$= \left(\sum_{k=1}^{r} \sum_{j=0}^{r-1} \sigma_k^2 u_k u_k^* \cdot \sigma_{k-j}^2 u_{k-j} u_{k-j}^* \right)$$

($\delta_{k,k-j}$)

$$= \left(\sum_{k=1}^{r} \sigma_k^{2 \cdot 2} u_k u_k^* \right).$$

By induction, this yields

$$(AA^*)^m = \sum_{k=1}^{r} \sigma_k^{2m} u_k u_k^*.$$

Setting $\sigma_k = 0$ for $r+1 \le k \le n$, we write the extended sum

$$(AA^*)^m = \sum_{k=1}^{n} \sigma_k^{2m} u_k u_k^*.$$

and hence for $x \in \mathbb{K}^I$ picked at random

$$(AA^*)^m x = \sum_{k=1}^{n} \sigma_k^{2m} u_k \langle x, u_k \rangle.$$

We note that since u_k's are fixed, but x was random, it is as we would pick $\langle x, u_k \rangle$ at random.

By Lemma 2.8 and Remark 2.9 and 2.10, the probability that $|\langle x, u_1 \rangle| \ge \alpha > 0$ is at least $1 - 2\alpha\sqrt{2d-1}$.

Suppose now that $\sigma_1, \sigma_2, \dots, \sigma_{r_\varepsilon}$ are the singular values of A that are $\ge (1-\varepsilon)\sigma_1$, i.e.

$$\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_{r_\varepsilon} \ge (1-\varepsilon)\sigma_1 \ge \sigma_{r_\varepsilon + 1} \ge \dots \ge \sigma_n.$$

[For $\sigma_1 > \sigma_2$ $\exists \varepsilon > 0$ small enough s.t. $r_\varepsilon = 0$.]

By Pythagoras-Fourier theorem, we have

$$\| (AA^*)^m x \|_2^2 = \| \sum_{k=1}^{u} \sigma_k^{2m} \langle x, u_k \rangle u_k \|_2^2 =$$

$$= \sum_{k=1}^{u} \sigma_k^{4m} |\langle x, u_k \rangle|^2$$

$$\ge \sigma_1^{4m} |\langle x, u_1 \rangle|^2$$

$$\overset{\text{w.h.p.}}{\ge} \alpha^2 \sigma_1^{4m}$$

w.h.p. = with high prob.

with probability at least $1 - 2\alpha\sqrt{2d-1}$,

Which gives us a lower bound for $\|(AA^*)^m x\|_2^2$.

Now consider the component of $(AA^*)^m x$, which is orthogonal w.r.t. $V = \operatorname{span}\{u_1, \dots, u_{r_\varepsilon}\}$. We write

$$(AA^*)^m x = \sum_{k=1}^{r_\varepsilon} \theta_k^{2m} \langle x, u_k \rangle u_k \qquad \in V$$

$$+ \sum_{k=r_\varepsilon+1}^{n} \theta_k^{2m} \langle x, u_k \rangle u_k \qquad \in V^\perp.$$

Then

$$\left\| \sum_{k=r_\varepsilon+1}^{n} \theta_k^{2m} |\langle x, u_k \rangle| \right\|_2^2 \overset{\text{P.F.T}}{=} \sum_{k=r_\varepsilon+1}^{n} \theta_k^{4m} |\langle x, u_k \rangle|^2$$

$$\overset{\text{by construction}}{\leq} (1-\varepsilon)^{4m} \theta_1^{4m} \underbrace{\sum_{k=r_\varepsilon+1}^{n} |\langle x, u_k \rangle|^2}_{\leq 1}$$

$$\leq (1-\varepsilon)^{4m} \theta_1^{4m}$$

Since $\sum_{k=r_\varepsilon+1}^{n} |\langle x, u_k \rangle|^2 \leq \sum_{k=1}^{n} |\langle x, u_k \rangle|^2 = \|x\|_2^2 = 1$ by the Pythagoras-Fourier theorem.

In summary, we get

$$\frac{\|P_{V^\perp}((AA^*)^m x)\|_2^2}{\|(AA^*)^m x\|_2^2} \xrightarrow{\text{upper bound}} \leq \frac{\theta_1^{4m} (1-\varepsilon)^{4m}}{\alpha^2 \cdot \theta_1^{4m}} = \frac{(1-\varepsilon)^{4m}}{\alpha^2}.$$

$\xrightarrow{\text{lower bound}}$

Using the Taylor expansion on $\varepsilon \simeq 0$, we get $(1-\varepsilon) \simeq e^{-\varepsilon}$ and hence we can rewrite

$$\frac{(1-\varepsilon)^{4m}}{\alpha^2} = \Theta(\alpha^{-2} e^{-4\varepsilon m})$$

$$m = c \cdot \left(\frac{\ln(d^\beta/\varepsilon)}{\varepsilon}\right) = \Theta\left(\alpha^{-2} \exp\left(\underbrace{-4 \cdot \varepsilon \cdot c \ln(d^\beta/\varepsilon) \cdot \varepsilon^{-1}}_{= \ln\left(\left(\frac{d^\beta}{\varepsilon}\right)^{-4c}\right)}\right)\right)$$

$$= \Theta\left(\frac{\varepsilon}{\alpha^2 \cdot d^{4\beta c}}\right).$$

$\square$

Relation between left & right singular vectors:

v's are <u>right</u> singular vectors

$$\frac{(A^TA)^m x}{\|(A^TA)^m x\|_2} \longrightarrow v_1 v_1^T x$$

and u's are <u>left</u> singular vectors

$$\frac{(AA^T)^m x}{\|(AA^T)^m x\|_2} \longrightarrow u_1 u_1^T x$$

For $A \in \mathbb{R}^{u \times d}$:

$A^TA = \begin{matrix} d \\ \updownarrow \\ \square \end{matrix} \begin{matrix} \overset{d}{\longleftrightarrow} \\ \square \end{matrix} \in \mathbb{R}^{d \times d}$

$\Rightarrow v_1 \in \mathbb{R}^d$

$AA^T = \begin{matrix} u \\ \updownarrow \\ \square \end{matrix} \begin{matrix} \overset{u}{\longleftrightarrow} \\ \square \end{matrix} \in \mathbb{R}^{u \times u}$

$\Rightarrow u_1 \in \mathbb{R}^u$

i.e.

$$A^TA = \sum_{k=1}^{d} \sigma_k^2 v_k v_k^T.$$

and

$$AA^T = \sum_{k=1}^{d} \sigma_k^2 u_k u_k^T.$$

Having v_1, then $u_1 = \frac{Av_1}{\|Av_1\|}$, <u>or</u>

having u_1, then $v_1 = \frac{A^Tu_1}{\|A^Tu_1\|}$.

Calculation of $\sigma_2, \dots, \sigma_k$, once we have σ_1:

Compute $A_1 = A - \sigma_1 u_1 v_1^T$ "<u>deflation of A</u>"
and again execute the SVD-power method again.

↝ Problem: not as stable as it can be.

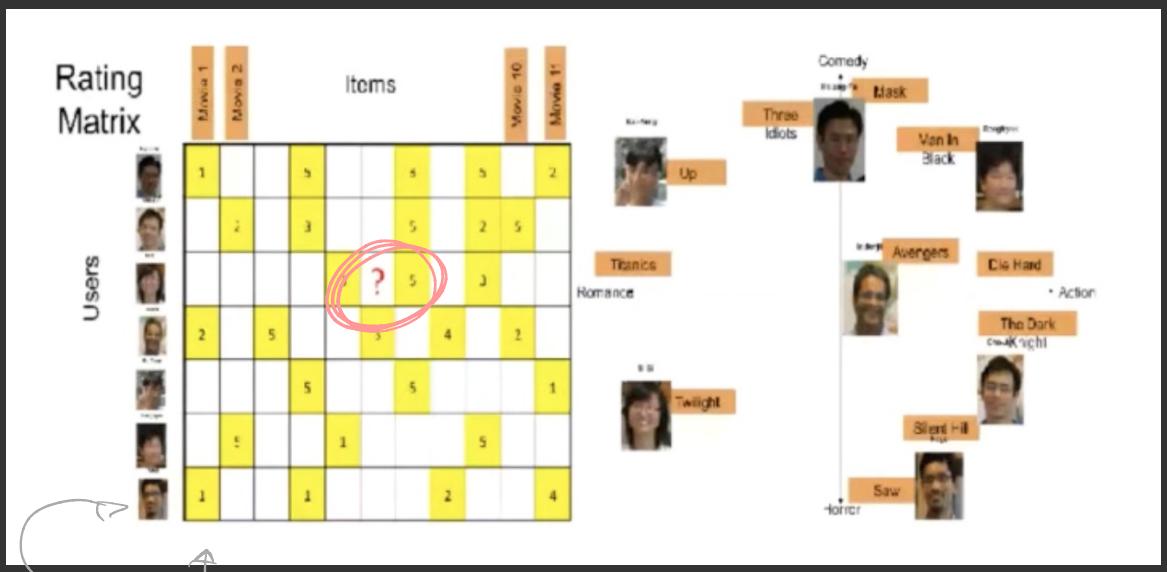
Vorlesung 8

13.05.2020

2.5. Applications of SVD

2.5.1. Principal Component Analysis

Example: Customer Product Matrix



↑ customer n bought product 1 a single time.

Goal: complete this matrix

Problem: Let A with elements a_{ij} represent the probability of customer i to buy product j.

Assume: ▷ k underlying basic factors like age, income, family... to determine customer's behavior (weighted combination)

⇒ look for these factors in our matrix.

▷ characterize the ith customers behavior by

$u_i = (u_{\ell,i})_{\ell=1,\dots,k} \in \mathbb{R}^k$ } how relevant are the factors for a customer?

weights for each basic factor ℓ.

▷ vector of probabilities

$v_\ell = (v_{\ell j})_{j=1,\dots,d}$ } how does each factor effect the prob. to buy a product?

representing the probability of purchasing product j when only looking on basic factor ℓ.

$\Rightarrow$ $\exists\, U \in \mathbb{C}^{n \times k}$ and $V \in \mathbb{C}^{k \times d}$ matrices s.t.

$$A = UV^H$$

↳ often no equality here

while all factors (columns) of U and V should represent distinct factors (being orthogonal).

Then

$$(A^{(i)})^H = \sum_{\ell=1}^{k} u_{\ell,i}\, v_\ell^H$$

} prob. of i-th customer to buy each product

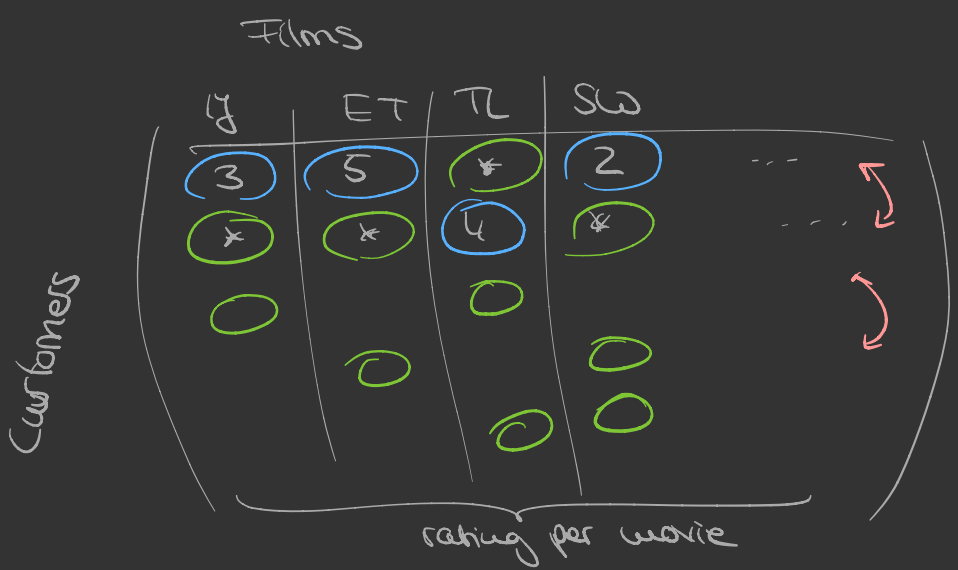
$\Rightarrow$ Taking the best rank-k approximation A_k from SVD gives such U, V.

Problems: ▷ Interpretation of the vectors in terms of e.g. identify the corresponding customer properties.

▷ Noise is often contained in the information of A

▷ Don't know how large the k should be chosen

Real-Life Example: Netflix-Recommendation



* for unseen movies

↓

want to predict the rating in a collaborative fashion for unseen movies

$\Rightarrow$ "Matrix completion problem" (low-rank matrix completion)

small

$\Omega \subsetneq I \times J$, we know A_{ij} for $(i,j) \in \Omega$.

How can we find A_{ik}, $(i,k) \in \Omega^c$?

(*) similarities in combinations of customer profiles, i.e. the rows are linearly dependent.

$\Rightarrow$ $A^{(i)} \simeq \sum_{\delta \neq i} \theta_\delta A^{(\delta)}$ linear dependent, then

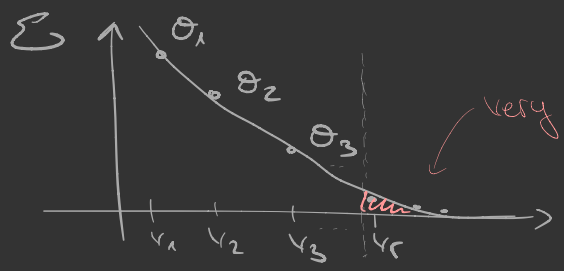
$\{$ #customers

$\dim\left(\text{span}\left\{A^{(i)}, i = 1, \dots, n\right\}\right) = \text{rank}(A) = r \ll n.$

$= \text{first integer, for which } (\theta_{r+1} = 0)$

Assumption: Data matrix A has "essentially" low rank r, i.e. $r \ll \min\{n, d\}$. (relaxation of the concept of rank)

$\Rightarrow$ All the singular values $\theta_k, k \geq r$ are very small



Idea to reconstruct complete matrix A^* out of Y (with missing values)

$Y = \{A^*_{ij}, (i,j) \in \Omega\}$ $\rightsquigarrow$ given a small subset of entries of A^* that we see

$$X = \underset{\{z_{ij}, (i,j) \in \Omega\} = Y}{\arg\min} \ \text{rank}(z)$$

$\rightsquigarrow$ reconstruct original matrix "recovery matrix"

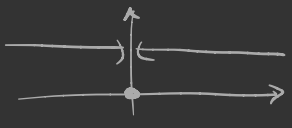
But: optimization of rank is a hard problem. We know, rank

$r =$ Smallest integer for which $\theta_{r+1} = 0$.

In other words, having low rank means having most of the singular values equal to 0.

$$X = \underset{\{z_{ij}, (i,j) \in \Omega\} = Z}{\arg\min} \ \#\{\theta_j(Z) \neq 0\}.$$

Rewrite again: Define $|t|_0 = \begin{cases} 0 & , t = 0 \\ 1 & , t \neq 0 \end{cases}$



- non-convex
- non-smooth