13.

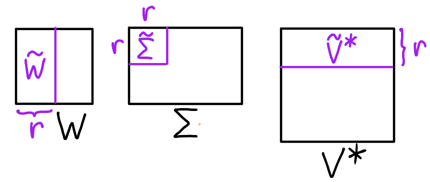
<u>Thm</u>: $A \in \mathbb{F}^{m\times n}$ has SVD $A = W\Sigma V^*$ with

- $V^* \in O_n(\mathbb{F})$
- $\Sigma \in \mathbb{F}^{m\times n}$ all 0 except $\sigma_1, \dots, \sigma_r$ on main diagonal
- $W \in O_m(\mathbb{F})$.

$\Leftrightarrow$ $[\lambda_A]_{\mathcal{V},\mathcal{W}} = \Sigma$ for orthonormal bases $\mathcal{V} = v_1, \dots, v_n$ of $\mathbb{F}^n$ and $\mathcal{W} = w_1, \dots, w_m$ of $\mathbb{F}^m$.

emphasize: simple geometry

<u>Pf</u>: Complete bases in Schmidt decomposition to orthonormal bases of $\ker A$ (for V^*) and $\ker A^*$ (for W). $\square$



<u>Cor</u>: $A \in \mathbb{F}^{n\times n} \Rightarrow A = U|A|$ for some $U \in O_n(\mathbb{F})$.

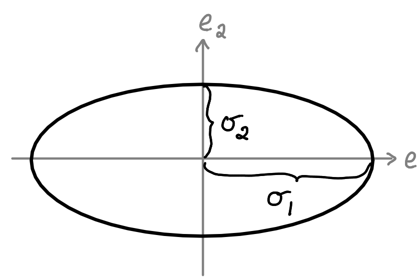
<u>Pf</u>: $A = W\Sigma V^* = \underbrace{WV^*}_{U}\underbrace{V\Sigma V^*}_{|A|}$ $\qquad$ $A^*A = \sum_{i=1}^{r} \sigma_i^2 v_i v_i^* = V\Sigma^2 V^* = \overbrace{V\Sigma \cancel{W^*}}^{A^*}\overbrace{\cancel{W}\Sigma V^*}^{A}$

$= U \ |A|$. $\square$

<u>Note</u>: SVD efficient numerically: fast + accurate

Q. How big can $\|Ax\|$ be, given that $\|x\| = 1$?

$B = \{x \in \mathbb{F}^n \mid \|x\| \le 1\}$ has image $= ?$



A. If $\Sigma = \mathrm{diag}(\sigma_1 \ge \cdots \ge \sigma_n)$ with $\sigma_{r+1} = \cdots = \sigma_n = 0$ then

$$\begin{aligned} y = \Sigma x \text{ for } x \in B \ &\Leftrightarrow\ y_i = \sigma_i x_i \text{ and } x \in B \\ &\Leftrightarrow\ y_i = \sigma_i x_i \text{ and } |x_1|^2 + \cdots + |x_n|^2 \le 1 \\ &\Leftrightarrow\ y_i = 0 \text{ for } i > r \text{ and } \left|\tfrac{y_1}{\sigma_1}\right|^2 + \cdots + \left|\tfrac{y_r}{\sigma_r}\right| \le 1 \quad \underline{\text{ellipsoid}} \end{aligned}$$

- $\mathbb{F} = \mathbb{R}$: <u>principal axes</u> of length $2\sigma_1, \dots, 2\sigma_r$ along $e_1, \dots, e_r$
- $\mathbb{F} = \mathbb{C}$: i^{th} real and imaginary principal axes of length $2\sigma_i$.

A arbitrary $\Rightarrow [\lambda_A]_{\mathcal{V},\mathcal{W}} = \Sigma$
$\qquad A = W\Sigma V^*$ $\qquad$ unitary $\Rightarrow$ do not alter $\|\cdot\|$

<u>Thm</u>: $A(\text{unit ball}) = $ ellipsoid in $\mathrm{im}(A)$ with principal half-axes along $w_1, \dots, w_r$ of lengths $\sigma_1 \ge \cdots \ge \sigma_r$. $\square$ $\qquad$ $\perp$normal basis

<u>Cor</u>: A has <u>operator norm</u> $\|A\| = \max_{x\in B} \|Ax\| = \sigma_1$. $\Leftrightarrow \|Ax\| \le C\|x\|\ \forall x$, and $C = \|A\|$ is smallest such.

<u>Lemma</u>: $A \mapsto \|A\|$ is a norm on $\mathbb{F}^{m\times n}$.

<u>Pf</u>:
- $\|\alpha A\| = |\alpha|\|A\|$
- $\|A + B\| \le \|A\| + \|B\|$
- $\|A\| \ge 0 \quad \forall A$
- $\|A\| = 0 \Leftrightarrow A = 0$.

Compare <u>Frobenius norm</u> $\|A\|_2 = \sqrt{\mathrm{tr}(A^*A)}$ $\qquad$ $\mathrm{tr}(A^*A) = \sum_{i,j} \underbrace{\bar{a}_{ij} a_{ij}}_{|a_{ij}|^2}$

(or <u>Hilbert-Schmidt</u>) $\qquad$ huh? $\qquad$ $= \text{“}\langle A, A\rangle\text{”}$

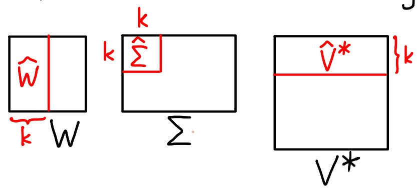
<u>Prop</u>: $\|A\| \le \|A\|_2$.

<u>Pf</u>: $\|A\| = \sigma_1 \le \sqrt{\sigma_1^2 + \cdots + \sigma_n^2} = \|A\|_2$ since $\mathrm{tr}(A^*A) = \sum \text{eigenvalues}(A^*A)$. $\square$

application: in computation, statistics, economics,... often better to have low-rank approximation of A

Eckart-Young Thm: Given $A \in \mathbb{F}^{m\times n}$, the $\hat{A}$ of rank $\leqslant k$ minimizing $\|A-\hat{A}\|_2$ is

$\hat{W}\hat{\Sigma}\hat{V}^*$ where



Pf: omitted for time, though we could totally do it

Principal Component Analysis (PCA)

$$A = {}_{m}\overset{n}{\begin{bmatrix} - A_1 - \\ \vdots \\ - A_m - \end{bmatrix}} \quad \longleftrightarrow \quad m \text{ points in } \mathbb{F}^n$$

sample

sample size

number of features

weight
height
temperature
death rate
response rate to drug
or other stimulus
⋮

unit vector

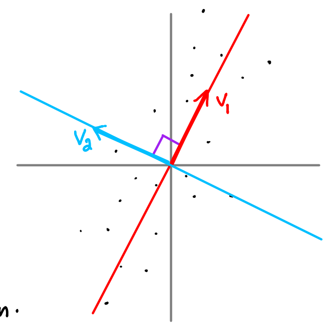
PC 1 = direction $v_1 \in \mathbb{F}^n_{col}$ maximizing sample variance: $|A_1v_1|^2+\cdots+|A_mv_1|^2$

$\hat{A}_1$ = project rows of A orthogonally to $v_1^\perp$

PC 2 = direction $v_2 \in v_1^\perp \subseteq \mathbb{F}^n_{col}$ maximizing $\hat{A}_1$-sample variance

$\hat{A}_2 = \hat{A}_1 / v_2$

⋮



score matrix

Def: The PC decomposition of A is $T = AV$, where V has columns $v_1,\dots,v_n$.

ij entry is score of sample i along PC j.

Interpretation: $\mathrm{cols}(V) \longleftrightarrow$ alternative features

- linear combinations of original features
- explain variance in uncorrelated ($\perp$) way

Thm: $A = W\Sigma V^* \Rightarrow$ $v_1,\dots,v_n$ are the columns of V and

$T = W\Sigma$ is polar decomposed

Pf: (sample variance in direction v) $= \|Av\|^2$ for $v \in \mathbb{F}^n_{col}$.

v maximizes $\|Av\|^2 \Leftrightarrow v \overset{A}{\mapsto}$ longest principal axis! (by Cor: $\|A\| = \sigma_1$)

$\Rightarrow V$ is SVD: $\mathrm{cols}(V)$ $\perp$normal basis of eigenvectors of A^*A (by induction)

$\Rightarrow T = AV = W\Sigma V^*V = W\Sigma$. $\square$

PCA $\rightsquigarrow$ low-rank projection of data: use only PC 1, ..., PC k

$$\mathbb{F}^n \xrightarrow{\perp} \mathbb{F}^k$$

variation in directions PC $k+1$, ..., PC n is small

Note: PC 1, PC 2, ..., PC n $\rightsquigarrow$ flag of best approximating subspaces.