Chapter III: The singular value decomposition

Singular value decomposition (SVD) of a matrix $A \in \mathbb{R}^{m\times n}$ is factorization into three matrices

$$A = U \Sigma V^T$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal

$$\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_k) \, . \qquad k = \min\{m,n\}$$

$$= \begin{cases} \left(\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_m & \end{array}\right) & : \quad m \leq n \\[2ex] \left(\begin{array}{ccc} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ \hline & 0 & \end{array}\right) & \quad n < m \end{cases}$$

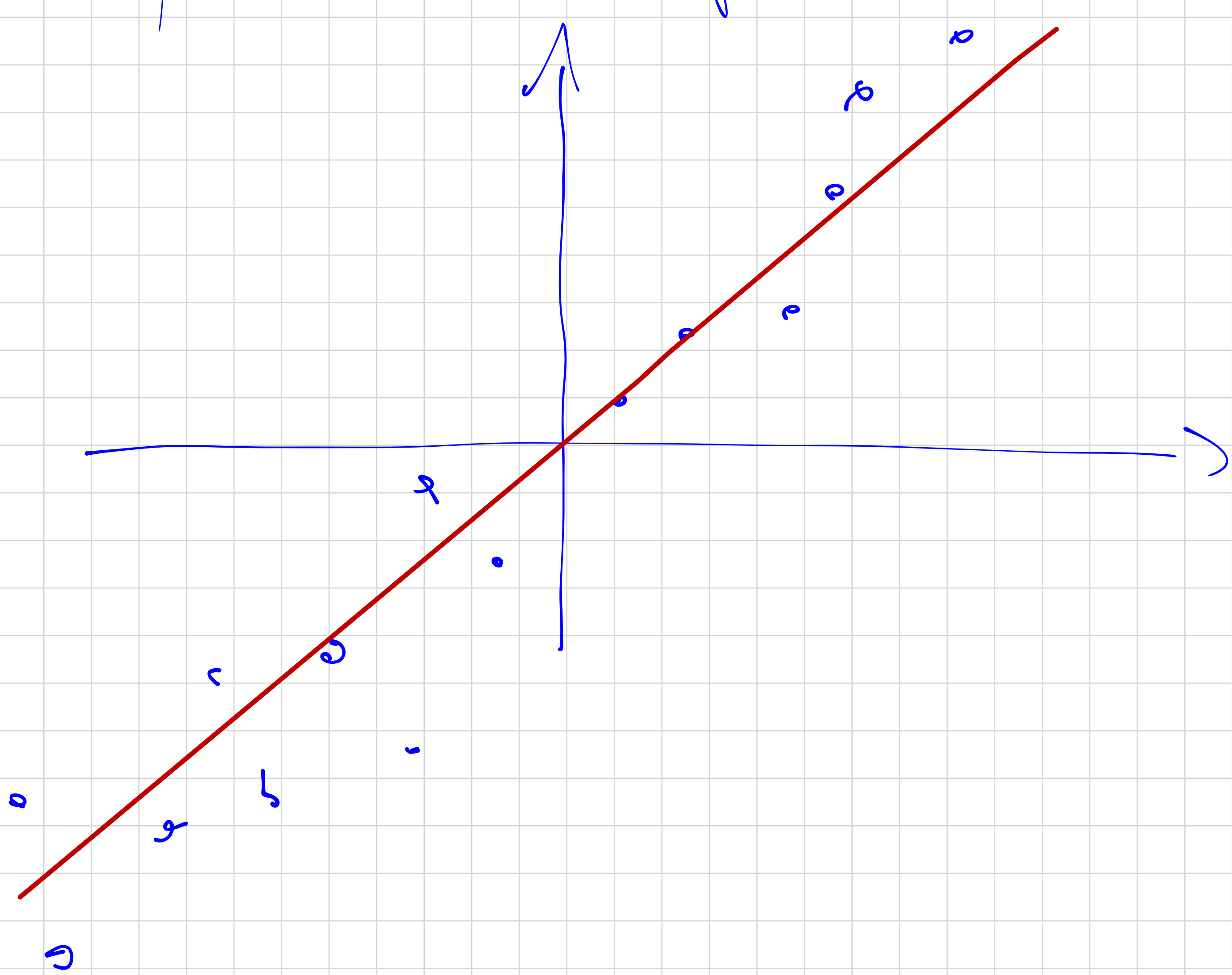
with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k \geq 0$

In contrast to eigenvalue decomposition it also exists for rectangular matrices.

One possible application: In many applications the data matrix is close to a low rank matrix. It is easy to compute the best approximation of rank k to A based on its SVD.

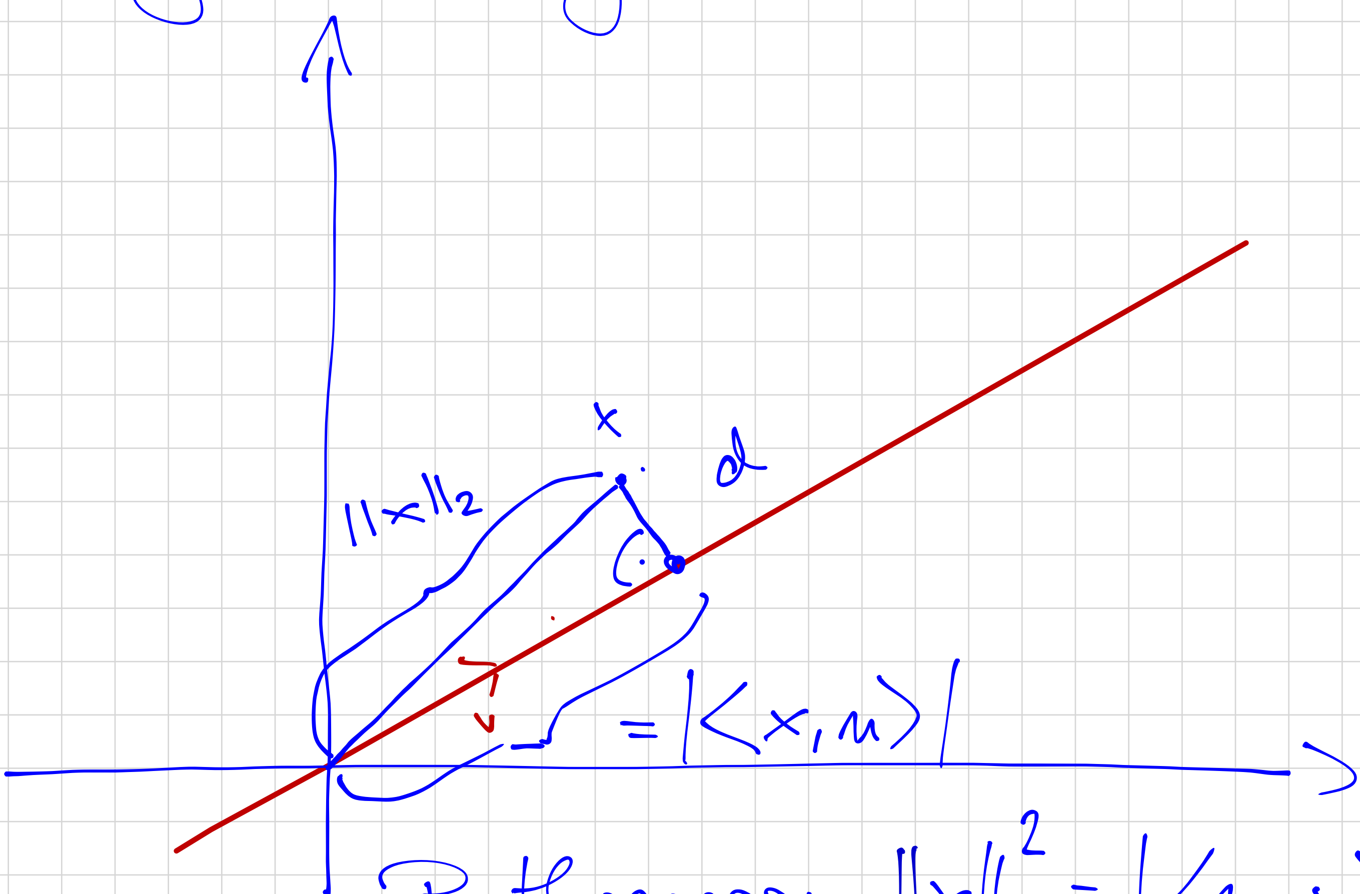
Approach: $A = (A_{(1)}, \dots, A_{(n)})$, $A_{(n)} \in \mathbb{R}^m$

Task: Find k-dimensional subspace that best fits the data points $A_{(1)}, \dots, A_{(n)}$



$k=1$: Find line that best fits the data.

- Distance of a point $x \in \mathbb{R}^m$ to a line through the origin in direction $u \in \mathbb{R}^m$, $\|u\|_2 = 1$.



Pythagoras: $\|x\|_2^2 = |\langle u, x\rangle|^2 + d^2$

$$d^2 = \|x\|_2^2 - |\langle u, x\rangle|^2$$

$\Rightarrow$ Minimizing d is equivalent to maximizing $|\langle u, x\rangle|^2$ because $\|x\|_2^2$ is constant.

$\Rightarrow$ Minimizing sum of square distances of the points $A_{(1)}, \ldots, A_{(n)}$ to the line generated by v is equivalent to maximizing

$$\sum_{j=1}^{n} |\underbrace{\langle A_{(j)}, u\rangle}_{(A^T u)_j}|^2 = \|A^T u\|_2^2 \text{ over } u \in \mathbb{R}^m \text{ with } \|u\|_2 = 1$$

Remark: Minimizing the squared distances may seem arbitrary. In fact, also other approaches may be pursued, e.g. minimize $\sum_{j=1}^{n} d(A_{(j)}, \langle u\rangle)^p$ for some $p \geq 1$ and $d(A_j, \langle u\rangle) = \min_{t\in\mathbb{R}} \|A_j - tu\|$ for some norm $\|\cdot\|$.

$\langle u\rangle$: span$\{u\}$

In general, these are computationally much harder, but may have some advantages in certain cases.

First (left) singular vector u_1 of A is defined as maximizer of

$$\max_{\|u\|_2=1} \|A^T u\|_2$$

the value $\sigma_1 = \max_{\|u\|_2=1} \|A^T u\|_2 = \|A^T\|_{\ell^2\to\ell^2}$ is called the first singular value.

Note that by duality

$$\|x\|_2 = \max_{\|z\|_2=1} |\langle x, z\rangle| .$$

Indeed, by the Cauchy-Schwarz inequality

$$|\langle x, z\rangle| \leq \|x\|_2 \|z\|_2 = \|x\|_2$$ for any $z \in \mathbb{R}^m$ with $\|z\|_2 = 1$

and for $z = \frac{x}{\|x\|_2}$ (so that $\|z\|_2 = 1$)

$$\langle x, z\rangle = \left\langle x, \frac{x}{\|x\|_2}\right\rangle = \frac{\|x\|_2^2}{\|x\|_2} = \|x\|_2$$

Therefore,

$$\sigma_1 = \max_{\|u\|_2=1} \|A^T u\|_2 = \max_{\|u\|_2=1} \max_{\|v\|_2=1} |\underbrace{\langle A^T u, v\rangle}|$$

$$= (A^T u)^T v = u^T A v = \langle u, Av\rangle$$

$$= \max_{\|v\|_2=1} \max_{\|u\|_2=1} |\langle u, Av\rangle|$$

$$= \max_{\|v\|_2=1} \|Av\|_2 = \|A\|_{\ell^2 \to \ell^2} = \|A\|$$

[Note: Hereby, we proved that $\|A^T\|_{\ell_2 \to \ell_2} = \|A\|_{\ell_2 \to \ell_2}$

$$\Longrightarrow \qquad \sigma_1 = \|A\|_{\ell^2 \to \ell^2} = \|A^T\|_{\ell^2 \to \ell^2}$$

Next step, $k=2$: Find best subspace of dimension 2 approximating $A_{(1)}, \ldots, A_{(n)}$.

Greedy approach: Choose 2-dim. subspace containing u_1 and optimize over second basis vector u_2 orthogonal to u_1, leads to subspace $\mathcal{U} = \operatorname{span}\{u_1, u_2\}$.

Then continue in this way to add further basis vectors $u_3, u_4, \ldots$ orthogonal to the previously selected ones.

Second singular vector

$$u_2 = \underset{\substack{\langle u, u_1 \rangle = 0 \\ \|u\|_2 = 1}}{\operatorname{argmax}} \|A^T u\|_2 \qquad \longleftrightarrow \qquad u_2 \perp u_1$$

$$\sigma_2 = \max_{\substack{\langle u, u_1 \rangle = 0 \\ \|u\|_2 = 1}} \|A^T u\|_2 \qquad \text{second singular vector}$$

continue in this way, so that recursively

$$u_r = \underset{\substack{u \perp u_1, u_2, \ldots, u_{r-1} \\ \|u\|_2 = 1}}{\operatorname{argmax}} \|A^T u\|_2$$

$$\sigma_r = \max_{\substack{u \perp u_1, \ldots, u_{r-1} \\ \|u\|_2 = 1}} \|A^T u\|_2$$

stop once $\sigma_{r+1} = 0$ or $r = m$ $(u_1, \ldots, u_r \in \mathbb{R}^m)$,

so that r is maximal with $\boxed{\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0}$.

At this point it is not clear whether for $k \geq 2$ this greedy strategy identifies the best k-dimensional subspace $\mathcal{U}_k = \operatorname{span}\{u_1, \ldots, u_k\}$ approximating the data points in the sense that $d(A, \mathcal{U}) := \left(\sum_{j=1}^{n} d(A_{(j)}, \mathcal{U})^2 \right)^{1/2}$ is minimized over all possible k-dim. subspaces $\mathcal{U}$, where

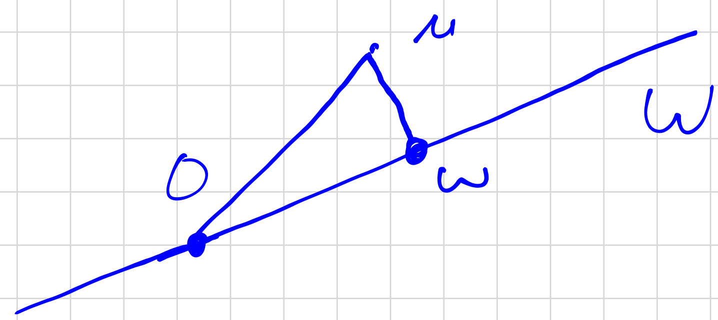
$$d(x, \mathcal{U}) = \min_{u \in \mathcal{U}} \|x - u\|_2 \qquad (*)$$

Interlude: Orthogonal projections

Proposition 3.1: Let $W \subset \mathbb{R}^m$ be a subspace of dimension k with orthonormal basis (ONB) $w_1, \dots, w_k \in \mathbb{R}^m$ and let $u \in \mathbb{R}^m$.

a) The minimizer $\hat{w}$ of

$$\min_{w \in W} \|u - w\|_2$$



exists, is unique and is given by

$$\hat{w} = \sum_{j=1}^{k} \langle u, w_j \rangle w_j .$$

Consequently, the orthogonal projection

$$P_W : \mathbb{R}^m \to W, \quad u \mapsto \operatorname*{argmin}_{w \in W} \|u - w\|_2 = \hat{w}$$

is linear, $P_W u = \sum_{j=1}^{k} \langle u, w_j \rangle w_j$.

b) The difference vector $u - \hat{w}$ and $\hat{w}$ are orthogonal, $u - \hat{w} \perp \hat{w}$ $\left(\langle u - \hat{w}, \hat{w} \rangle = 0 \right)$.

c) It holds $\|\hat{w}\|_2^2 = \|P_W u\|_2^2 = \sum_{j=1}^{k} \langle u, w_j \rangle^2$

and $\|u - \hat{w}\|_2^2 = \|u\|_2^2 - \sum_{j=1}^{k} \langle u, w_j \rangle^2$

Proposition 3.2: Let $A \in \mathbb{R}^{m \times n}$ with left singular vector $u_1, \dots, u_r$. For $k = 1, \dots, r$ define $\mathcal{U}_k = \operatorname{span}\{u_1, \dots, u_k\}$. Then $\mathcal{U}_k$ minimizes $d(A, \mathcal{U})$ over all k-dim. subspaces $\mathcal{U}$, i.e., $\mathcal{U}_k$ is the best k-dim. approximation to the columns of A.

Proof: **$k=1$:** The statement is true by construction of u_1.

$k=2$: Let W be a 2-dimensional best fitting subspace, i.e. W minimizes $(*)$,

$$d(A, W)^2 = \sum_{j=1}^{n} d^2(A_{(j)}, W) = \sum_{j=1}^{n} \| A_{(j)} - P_W A_j \|_2^2$$

Choose an orthonormal basis w_1, w_2 of W such that $w_2 \perp u_1$. This is possible because $u_1^\perp \cap W$ is non-trivial $(\neq \{0\})$ since $\dim u_1^\perp = m-1$ and $\dim W = 2$. In fact if $u_1^\perp \cap W = \{0\}$ then $\dim(u_1^\perp + W) = \dim u_1^\perp + \dim W = m+1$

which is impossible as $u_1^\perp + W \subseteq \mathbb{R}^m$.

Since u_1 maximizes $\|A^T u\|_2$ over the unit sphere, it holds $\|A^T w_1\|_2 \leq \|A^T u_1\|_2$

Moreover, since u_2 is maximizer of

$$\max_{\substack{\|u\|_2 = 1 \\ u \perp u_1}} \|A^T u\|_2$$ and $w_2 \perp u_1$ it holds

$$\|A^T w_2\|_2 \leq \|A^T u_2\|_2 .$$

With $\mathcal{U} = \operatorname{span}\{u_1, u_2\}$, Proposition 3.1 implies

$$d^2(A, U) = \sum_{j=1}^{n} \|A_{(j)} - P_U A_{(j)}\|_2^2 =$$

$$= \sum_{j=1}^{n} \left(\|A_{(j)}\|_2^2 - \langle A_{(j)}, u_1 \rangle^2 - \langle A_{(j)}, u_2 \rangle^2 \right)$$

$$= \|A\|_F^2 - \|A^T u_1\|_2^2 - \|A^T u_2\|_2^2$$

$$\leq \|A\|_F^2 - \|A^T w_1\|_2^2 - \|A^T w_2\|_2^2$$

$$= d^2(A, W)$$

$\Rightarrow$ $\mathcal{U}$ minimizes $d(A, W)$ over all 2-dim.

subspaces $W \subset \mathbb{R}^m$.

For general k, we use induction. Assume that the claim is true for $k-1$, i.e., $U_{k-1} = \mathrm{span}\{u_1, \ldots, u_{k-1}\}$ minimizes $d(A, U)$ over all $k-1$-dim subspaces U. Let W be a best k-dimensional subspace for A, i.e., minimizing $d(A, U)$ over k-dim subspaces U. Choose ONB $w_1, \ldots, w_k$ of W such that $w_k \perp u_1, \ldots, u_{k-1}$. This is possible since $\dim W = k$ and $\dim \mathrm{span}\{u_1, \ldots, u_{k-1}\}^\perp = m-k+1$. Since U_{k-1} is optimal we have

$$\|A^T u_1\|_2^2 + \ldots + \|A^T u_{k-1}\|_2^2 = \|A\|_F^2 - d^2(A, U_{k-1})$$
$$\geq \|A\|_F^2 - d^2(A, \mathrm{span}\{w_1, \ldots, w_{k-1}\})$$
$$= \|A^T w_1\|_2^2 + \ldots + \|A^T w_{k-1}\|_2^2$$

Moreover, since $w_k \perp u_1, \ldots, u_{k-1}$ the definition of u_k implies

$$\|A^T w_k\|_2^2 \leq \|A^T u_k\|_2^2$$

Hence,

$$d^2(A,\mathcal{U}) = \|A\|_F^2 - \sum_{j=1}^{k} \|A^T u_j\|_2^2$$
$$= \|A\|_F^2 - \sum_{j=1}^{k-1} \|A^T u_j\|_2^2 - \|A^T u_k\|_2^2$$
$$\leq \|A\|_F^2 - \sum_{j=1}^{k-1} \|A^T w_j\|_2^2 - \|A^T w_k\|_2^2$$
$$= d^2(A, W)$$

$\Rightarrow$ $\mathcal{U}$ minimizes $d^2(A,W)$ over all k-dim. subspaces W of $\mathbb{R}^m$. $\square$

Remark: • By definition $\sigma_1 \geq \sigma_2 \geq \dots \sigma_r \geq 0$.

Lemma 3.3: For any matrix $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1, \dots, \sigma_r$ it holds
$$\|A\|_F = \sqrt{\sum_{j=1}^{r} \sigma_r^2}.$$

Proof: For each column $A_{(j)}$ of A it holds $\langle A_{(j)}, u\rangle = 0$ for all vectors $u \in \mathbb{R}^m$ that are orthogonal to the left singular vectors $u_1, \dots, u_r$. Completing $u_1, \dots, u_r$ to an ONB

$u_1, \dots, u_r, u_{r+1}, \dots, u_m$ we therefore get

$$\|A_{(j)}\|_2^2 = \sum_{i=1}^{m} \langle A_{(j)}, u_i \rangle^2 = \sum_{i=1}^{r} \langle A_{(j)}, u_i \rangle^2$$

$\langle A_{(j)}, u_i \rangle = 0 \quad i = r+1, \dots, m$

$\Longrightarrow$

$$\|A\|_F^2 = \sum_{i,j} |A_{ij}|^2 = \sum_{j=1}^{n} \|A_{(j)}\|_2^2$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{r} \langle A_{(j)}, u_i \rangle^2 = \sum_{i=1}^{r} \sum_{j=1}^{n} \langle A_{(j)}, u_i \rangle^2$$

$$= \sum_{i=1}^{r} \|A^T u_i\|_2^2 = \sum_{i=1}^{r} \sigma_i^2$$

definition of σ_i, u_i ! $\square$

Right singular vectors

Let $A \in \mathbb{R}^{m \times n}$ with singular values $\sigma_1, \dots, \sigma_r > 0$ and left singular vectors $u_1, \dots, u_r$. If $r < m$ we can complete $u_1, \dots, u_r$ to an ONB $\{u_1, \dots, u_r, u_{r+1}, \dots, u_m\}$ of $\mathbb{R}^m$. Since $u_i \perp u_1, \dots, u_r$ for $i > r$, the definition of the singular values & vectors yields

$A^T u_i = 0 \,, \quad i = r+1, \dots, m .$

Therefore, the image of A^T is given by

$$\operatorname{ran}(A^T) = \operatorname{span}\{A^T u_1, \dots, A^T u_r, A^T u_{r+1}, \dots, A^T u_m\}$$
$$= \operatorname{span}\{A^T u_1, \dots, A^T u_r\}$$

Normalization by $\|A^T u_j\| = \sigma_j \quad j = 1, \dots, r$

yields the <u>right singular vectors</u>

$$v_j = \frac{1}{\sigma_j} A^T u_j \qquad j = 1, \dots, r$$

$$\operatorname{ran}(A^T) = \operatorname{span}\{v_1, \dots, v_r\}$$

Orthogonality of the vectors $v_1, \dots, v_r$ is established next:

<u>Lemma 3.4:</u> Let $A \in \mathbb{R}^{m \times n}$ with singular values $\sigma_1, \dots, \sigma_r$, left singular vector $u_1, u_2, \dots, u_r \in \mathbb{R}^m$ and right singular vectors $v_1, \dots, v_r \in \mathbb{R}^n$. Then the vectors $v_1, \dots, v_r$ are orthonormal.

<u>Proof:</u> Induction on r:

<u>$r = 1$:</u> Only one v_1, so claim is trivial.

$r \mapsto r+1$: Consider the matrix

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

$$\rightsquigarrow B^T = A^T - \sigma_1 v_1 u_1^T$$

Claim: B has singular values $\sigma_2, \dots, \sigma_r$, left singular vectors $u_2, \dots, u_r$ and right singular vectors $v_2, \dots, v_r$.

Indeed, observe first that $B^T u_1 = \underbrace{A^T u_1}_{= \sigma_1 v_1} - \sigma_1 v_1 \underbrace{u_1^T u_1}_{=1}$

$= \underline{0}$

Let $z \in \mathbb{R}^m$ be the first left singular vector of B and consider $z = z_1 + z_2$ where $z_2 \perp u_1$.

$z_1 = \alpha_1 u_1$

Then $\left\| B^T \frac{z - z_1}{\|z - z_1\|_2} \right\|_2 = \left\| B^T \frac{z}{\|z - z_1\|_2} \right\|$

$$= \frac{\|B^T z\|}{\|z - z_1\|_2} = \frac{\|B^T z\|_2}{(\|z\|_2^2 - \alpha_1^2)^{1/2}} \begin{cases} = \|B^T z\|_2 & \text{if } \alpha_1 = 0 \\ > \|B^T z\|_2 & \text{if } \alpha_1 \neq 0 \end{cases}$$

Since $\|B^T z\|_2$ is maximal, this implies that $\alpha_1 = 0$ and $z = z_2 \perp u_1$, i.e., z is orthogonal to u_1.

$\Longrightarrow$ z is a second left singular vector of A. In the same way it follows that a second left singular vector of B is a third left singular vector of A etc. This shows the claim.

By the induction hypothesis, the right singular vectors $v_2, \ldots, v_r$ are orthogonal. We need to show that v_1 is orthogonal to $v_2, \ldots, v_r$.

Assume that this is not true, i.e. $\langle v_1, v_i \rangle \neq 0$ for some $i \in \{2, \ldots, r\}$. For $t \in \mathbb{R}$, set

$$\hat{u}_t = \frac{u_1 + t u_i}{\|u_1 + t u_i\|_2} \qquad \|\hat{u}_t\|_2 = 1$$

$$A^T \hat{u}_t = A^T \left(\frac{u_1 + t u_i}{\|u_1 + t u_1\|_2} \right) = \frac{\sigma_1 v_1 + \sigma_i t v_i}{\sqrt{1+t^2}}$$

By Cauchy-Schwarz

$$\|A^T \hat{u}_t\|_2 = \|A^T \hat{u}_t\|_2 \|v_1\|_2 \geq |\langle A^T \hat{u}_t, v_1 \rangle|$$

$$= \frac{1}{\sqrt{1+t^2}} \left| \sigma_1 \langle v_1, v_1 \rangle + \sigma_i t \langle v_i, v_1 \rangle \right|$$

Taylorexpansion of $\frac{1}{\sqrt{1+t^2}}$ in 0.

$$= \left| \sigma_1 + t \sigma_i \langle v_i, v_1 \rangle \right| \left(1 - \frac{t^2}{2} + \mathcal{O}(t^4) \right)$$

$$= \sigma_1 + t \sigma_i \langle v_i, v_1 \rangle + \mathcal{O}(t^2) \quad \text{for } t \text{ small enough}$$

In particular, there exists $t \neq 0$ such that $\| A^T \hat{u}_t \|_2 > \sigma_1$. $(\| \hat{u}_t \|_2 = 1)$

This is a contradiction to σ_1 being the largest singular value.

$$\Rightarrow \langle v_1, v_i \rangle = 0 \quad i = 2, \dots, r. \qquad \square$$

<u>Corollary 3.5</u> Let $A \in \mathbb{R}^{m \times n}$ be a matrix with singular values $\sigma_1, \dots, \sigma_r > 0$. Then $\operatorname{rank}(A) = r$.

<u>Proof:</u> We have already seen that $\operatorname{ran}(A) = \operatorname{span}\{v_1, \dots, v_r\}$.

Since $v_1, \dots, v_r$ are orthonormal, they form an ONB for $\operatorname{ran}(A)$ and hence

$\operatorname{rank}(A) = \dim \operatorname{ran}(A) = r$ $\square$

Corollary 3.6: Let $A \in \mathbb{R}^{m \times n}$ have singular values $\sigma_1, \dots, \sigma_r$, left singular vectors $u_1, \dots, u_r$ and right singular vectors $v_1, \dots, v_r$. Then we can write

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T$$

or in matrix form

$$A = U \Sigma V^T$$

where $U = (u_1 | \dots | u_r) \in \mathbb{R}^{m \times r}$, $V = (v_1 | \dots | v_r) \in \mathbb{R}^{n \times r}$ and $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r \times r}$

$$= \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix}.$$

Proof: We show that $A^T = \left(\sum_{j=1}^{r} \sigma_j u_j v_j^T \right)^T$.

If $r < m$ we complete $u_1, \dots, u_r$ to an ONB $u_1, \dots, u_r, u_{r+1}, \dots, u_m$ of $\mathbb{R}^m$.

For $i = 1, \dots, r$:

$$A^T u_i = \sigma_i v_i$$

$$\left(\sum_{j=1}^{r} \sigma_j u_j v_j^T\right)^T u_i = \sum_{j=1}^{r} \sigma_j v_j \underbrace{u_j^T u_i}_{=\delta_{ji}} = \sigma_i v_i$$

If $r < m$ then

for $i = r+1, \dots, m$

$$A^T u_i = 0$$

$$\left(\sum_{j=1}^{r} \sigma_j u_j v_j^T\right)^T u_i = \sum_{j=1}^{r} \sigma_j v_j \underbrace{u_j^T u_i}_{=0} = 0$$

The action of A^T and $\left(\sum_{j=1}^{r} \sigma_j u_j v_j^T\right)^T$ on the basis $u_1, \dots, u_m$ is the same and therefore $A = \sum_{j=1}^{r} \sigma_j u_j v_j^T$. $\square$

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T = U \Sigma V^T \qquad \text{write } \Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix}$$

is called the <u>reduced singular value</u> decomposition.

If $r < m$ complete $u_1, \dots, u_r$ to an ONB $u_1, \dots, u_r, u_{r+1}, \dots, u_m$ of $\mathbb{R}^m$.

Set $\tilde{U} = (u_1 | \dots | u_m) \in \mathbb{R}^{m \times m}$ (orthogonal)

If $r < n$ complete $v_1, \dots, v_r$ to an ONB

$v_1, \dots, v_r, v_{r+1}, \dots, v_n$ of $\mathbb{R}^n$ and set

$$\tilde{V} = (v_1 | \dots | v_n) \in \mathbb{R}^{n \times n} \quad \text{(orthogonal)}$$

Then the singular value decomposition of A is given as

$$A = \tilde{U} \tilde{\Sigma} \tilde{V}^T$$

$$\text{with } \tilde{\Sigma} = \underbrace{\left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right)}_{n} \Big\} m$$

Observation:

$$\underline{A^T A} = (\tilde{U} \tilde{\Sigma} \tilde{V}^T)^T (\tilde{U} \tilde{\Sigma} \tilde{V}^T) = \tilde{V} \tilde{\Sigma}^T \underbrace{\tilde{U}^T \tilde{U}}_{= I} \tilde{\Sigma} \tilde{V}^T$$

$$= \tilde{V} \tilde{\Sigma}^T \tilde{\Sigma} \tilde{V}^T = \underline{\tilde{V} D \tilde{V}^T}$$

$$\text{with } D = \tilde{\Sigma}^T \tilde{\Sigma} = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_r^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

$$= \operatorname{diag}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0)$$

$$AA^T = \tilde{u}\tilde{\Sigma}\tilde{V}^T(\tilde{u}\tilde{\Sigma}\tilde{V}^T)^T = \tilde{U}\tilde{\Sigma}\tilde{\Sigma}^T\underbrace{\tilde{V}^T\tilde{V}}_{=I}\tilde{\Sigma}^T\tilde{u}^T$$

$$= \tilde{U}\tilde{\Sigma}\tilde{\Sigma}^T\tilde{u}^T = \tilde{U}\hat{D}\tilde{U}^T$$

with $\hat{D} = \tilde{\Sigma}\tilde{\Sigma}^T = \begin{pmatrix} \sigma_1^2 & & & & \\ & \ddots & & & \\ & & \sigma_r^2 & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} \in \mathbb{R}^{m \times m}$

Lemma 3.7: Let $A \in \mathbb{R}^{m \times n}$. The singular values $\sigma_1, \dots, \sigma_r$ of A are the square-roots of the non-zero eigenvalues of A^TA and AA^T ordered by decreasing magnitude and counted with multiplicities, i.e.,

$$\sigma_j = \sqrt{\lambda_j(A^TA)} = \sqrt{\lambda_j(AA^T)}.$$

The corresponding left singular vectors $u_1, \dots, u_r$ are the eigenvectors corresponding to $\lambda_1(AA^T), \dots, \lambda_r(AA^T)$ of AA^T. The corresponding right singular vectors $v_1, \dots, v_r$ are the eigenvectors of

A^TA corresponding to $\lambda_1(A^TA), \dots, \lambda_r(A^TA)$.

Proof: exercise (based on above observation)

Remark: The singular value decomposition can be computed via eigenvalue decompositions of A^TA and AA^T as indicated in Lemma 3.7.

Corollary 3.8: If $A = A^T$ is symmetric with non-negative eigenvalues $\lambda_j(A) \geq 0$ then its eigenvalue decomposition coincides with its singular value decomposition.

Corollary 3.9: Let $A \in \mathbb{R}^{m \times n}$

- The sequence of singular values $\sigma_1, \dots, \sigma_r$ is unique.
- If $\sigma_j \neq \sigma_{j+1}$ (if $j<r$) & $\sigma_j \neq \sigma_{j-1}$ (if $j>1$) then the pair of singular vectors (u_j, v_j) is unique up to a sign change $(-u_j, -v_j)$.

- If $\sigma_{j-1} > \sigma_j = \dots = \sigma_{j+t} > \sigma_{j+t+1}$ then the spaces $\mathcal{U} = \operatorname{span}\{u_j, \dots, u_{j+t}\}$ and $V = \operatorname{span}\{v_j, \dots, v_{j+t}\}$ are unique and any ONB of these spaces may be chosen as set of singular vectors.

<u>Proof:</u> ~~Follows~~ Follows from Lemma 3.7 by uniqueness of eigenvalue decomposition (up to the described invariances). $\square$

<u>Corollary 3.10:</u> Let $A \in \mathbb{R}^{m \times n}$ have SVD

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T.$$

Then the SVD of A^T is given by

$$A^T = \sum_{j=1}^{r} \sigma_j v_j u_j^T.$$

In particular, A and A^T have the same singular values, the left singular vectors of A are the right singular vectors of A^T and the right singular vectors of A are the left singular vectors of A^T.

<u>Remark:</u> As a consequence the singular values and right singular vectors of A may equivalently be defined as

$$\sigma_1 = \max_{\|v\|_2=1} \|Av\|_2 \quad , \quad v_1 = \operatorname{argmax}_{\|v\|_2=1} \|Av\|_2$$

$$\sigma_j = \max_{\substack{\|v\|_2=1 \\ v \perp v_1, v_2, \ldots, v_{j-1}}} \|Av\|_2 \quad , \quad v_j = \operatorname{argmax}_{\substack{\|v\|_2=1 \\ v \perp v_1, \ldots, v_{j-1}}} \|Av\|_2 .$$

The left singular vectors of A satisfy

$$u_j = \frac{1}{\sigma_j} A v_j \qquad j = 1, \ldots, r$$

Best low-rank approximations

For $A \in \mathbb{R}^{m \times n}$ let

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T \qquad \text{be its SVD.}$$

For $k \in \{1, \ldots, r\}$ let

$$A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T \quad ; \quad \underline{\operatorname{rank}(A_k) = k}$$

Let $U_k = (u_1 | \dots | u_k) \in \mathbb{R}^{m\times k}$

$V_k = (v_1 | \dots | v_k) \in \mathbb{R}^{k \times n}$

$\Sigma_k = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} = \operatorname{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{k\times k}$

$$A_k = U_k \Sigma_k V_k^T = \boxed{m\times k}\;\boxed{k\times k}\;\boxed{k\times n}$$

storage cost of A : $\boxed{m\cdot n}$

″ ″ ″ A_k, i.e. of U_k, Σ_k, V_k : $\boxed{m\cdot k + k\cdot n + k}$

$= k(m+n+1)$

$\ll \underline{m\cdot n}$ if $k < m,n$

Application of A to $x \in \mathbb{R}^n$: $\sim \underline{m\cdot n}$ operations

Computation of $A_k x = U_k \Sigma_k V_k^T x$: $\sim \underline{m\cdot k + k\cdot n + k}$ op.

Approximation quality $\|A - A_k\|_F$, $\|A - A_k\|_{\ell^2 \to \ell^2}$?

<u>Lemma 3.11:</u> The columns $(A_k)_{(j)}$ of A_k are the orth. projections of the columns $A_{(j)}$ of A onto the subspace $\mathcal{U}_k = \operatorname{span}\{u_1, \dots, u_k\}$, i.e.,

$(A_k)_j = P_{\mathcal{U}_k} A_{(j)}$, $j = 1, \dots, n$.

Proof: exercise

Proof: Since $u_1, \dots, u_k$ is an ONB of U_k

$$P_{U_k} A_{(j)} = \sum_{i=1}^{k} \langle A_{(j)}, u_i \rangle u_i = \sum_{i=1}^{k} u_i u_i^T A_{(j)}$$

$$\Rightarrow \left(P_{U_k} A_{(1)} \,\middle|\, \dots \,\middle|\, P_{U_k} A_{(n)} \right) = \sum_{i=1}^{k} u_i u_i^T A$$

$$= \sum_{i=1}^{k} u_i u_i^T \sum_{j=1}^{r} \sigma_j u_j v_j^T$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{r} \sigma_j u_i \underbrace{u_i^T u_j}_{=\delta_{ij}} v_j^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T = A_k$$

$\square$

Theorem 3.12: Let $A \in \mathbb{R}^{m \times n}$ and A_k its rank-k truncation defined via the SVD, for $k \leq r$. Then

$$(*) \quad \|A - A_k\|_F \leq \|A - B\|_F \quad \text{for all } B \in \mathbb{R}^{m\times n} \text{ with } \operatorname{rank}(B) \leq k.$$

Proof: Let B be a minimizer of the right hand side of $(*)$ with rank at most k.

(We assume that such B exists, if not we would need to make an additional approximation argument: Let $\varepsilon > 0$ and choose B such that $\|A - B\|_F \leq \inf_{\operatorname{rank}(B) \leq k} \|A - B\|_F + \varepsilon$ .....)

Write $B = (B_{(1)}, \ldots, B_{(n)})$ and let

$W_k = \operatorname{span}\{B_{(1)}, \ldots, B_{(n)}\}$

It holds $\dim W_k = \operatorname{rank}(B) \leq k$.

Since B minimizes $\|A - B\|_F^2 = \sum_{j=1}^{n} \|A_{(j)} - B_{(j)}\|_2^2$,

$B_{(j)}$ minimizes $\|A_{(j)} - u\|$ over all $u \in W_k$.

In other words, $B_{(j)} = P_{W_k} A_{(j)}$.

Since $U_k = \operatorname{span}\{u_1, \ldots, u_k\}$ minimizes

$$d(A,\mathcal{U}) = \sum_{j=1}^{n} \|A_{(j)} - P_U A_{(j)}\|^2$$

over all possible subspaces $\mathcal{U}$ of dim. k

by Proposition 3.2 it follows that

$$d(A,U) \le d(A,W_k) \text{ and}$$

$$\|A - A_k\|_F^2 \le \|A - B\|_F^2 \qquad \square$$

Next we show that A_k is also the best rank-k approximation to A in the <u>spectral norm</u> $\|\cdot\|_{\ell^2\to\ell^2}$.

<u>Lemma 3.13:</u> Let $A \in \mathbb{R}^{m\times n}$ with rank-k truncation A_k, $k\le r$. Then $\|A - A_k\|_{\ell^2\to\ell^2} = \sigma_{k+1}$.

<u>Proof:</u> Let $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ be the SVD of A.

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

Let u be the top left singular vector of $A - A_k$ and write $u = \sum_{i=1}^{m} \alpha_i u_i$ (where $u_1, \dots, u_r$ is completed to an ONB $u_1, \dots, u_m$)

Then $\|(A - A_k)^T u\|_2 = \|\sum_{i=k+1}^{r} \sigma_i v_i u_i^T \sum_{j=1}^{m} \alpha_j u_j\|_2$

$$= \Big\| \sum_{i=k+1}^{r} \sum_{j=1}^{m} \sigma_i \alpha_j v_i \underbrace{u_i^T u_j}_{=\delta_{ij}} \Big\|_2$$

$$= \Big\| \sum_{i=k+1}^{r} \sigma_i \alpha_i v_i \Big\|_2 = \sqrt{\sum_{i=k+1}^{r} \alpha_i^2 \sigma_i^2}$$

$\{v_1, \dots, v_r\}$ orthonormal

The vector u maximizing the last expression under the constraint

$1 = \|v\|_2^2 = \sum_{i=1}^{m} \alpha_i^2 = 1$ is the one occurring

for $\alpha_{k+1} = 1$, $\alpha_i = 0$, $i = 1, \dots, m$, $i \neq k+1$,

i.e. $u = u_{k+1}$ and $\|(A - A_k) u_{k+1}\|_2 = \sigma_{k+1}$.

$\blacksquare$

<u>Theorem 3.14</u>. Let $A \in \mathbb{R}^{m \times n}$ and A_k its rank-k truncation defined via the SVD for $k \leq r$. Then

(*) $\|A - A_k\|_{\ell^2 \to \ell^2} \leq \|A - B\|_{\ell^2 \to \ell^2}$ for all $B \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(B) \leq k$.

<u>Proof</u>: By Lemma 3.13, $\|A - A_k\|_{\ell^2 \to \ell^2} = \sigma_{k+1}$.

Assume that there exists some matrix $B \in \mathbb{R}^{m \times n}$

with rank at most k such that

$$\|A-B\|_{\ell^2\to\ell^2} < \sigma_{k+1}.$$

Note that $\dim \ker \overline{B}^T \geq m-k$. Therefore, there exist $z \neq 0$, w.l.o.g. $\|z\|_2 = 1$, such that

$$z \in \ker(\overline{B}^T) \cap \operatorname{span}\{u_1, \dots, u_{k+1}\}$$

since otherwise $\dim\left(\ker(\overline{B}^T) + \operatorname{span}\{u_1,\dots,u_{k+1}\}\right)$

$$= \dim\left(\ker \overline{B}^T\right) + k+1 \geq m-k+k+1 = m+1.$$

By the definition of the operator norm,

since $\overline{B}^T z = 0$ and $z \in \operatorname{span}\{u_1, \dots, u_{k+1}\}$

$$\|A-B\|^2_{\ell^2\to\ell^2} = \|\overline{A}^T - \overline{B}^T\|^2_{\ell^2\to\ell^2} \geq \|(\overline{A}^T - \overline{B}^T) z\|_2^2$$

$$= \|\overline{A}^T z\|_2^2 = \Big\|\sum_{i=1}^{r} \sigma_i v_i \underbrace{\overline{u_i}^T z}_{=0 \text{ if } i > k+1}\Big\|^2$$

$$= \sum_{i=1}^{k+1} \sigma_i^2 (\overline{u}_i^T z)^2$$

$$\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} \langle u_i, z\rangle^2 = \sigma_{k+1}^2 \|z\|_2^2 = \sigma_{k+1}^2$$

This is a contradiction, which proves the theorem. $\square$

The Moore-Penrose Pseudo-Inverse

Generalization of the usual inverse, existing for any matrix.

Definition 3.15: Let $A \in \mathbb{R}^{m\times n}$ of rank r with reduced SVD
$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T = U \Sigma V^T$$
Then its Moore-Penrose pseudo-inverse $A^+ \in \mathbb{R}^{n\times m}$ is given by
$$A^+ = \sum_{j=1}^{r} \sigma_j^{-1} v_j u_j^T = V \Sigma^{-1} U^T$$

If $A \in \mathbb{R}^{m\times m}$ is invertible (i.e. rank$(A) = m$) then $A^+ = A^{-1}$, since then $A = U\Sigma V^T$ with orthogonal $U, V \in \mathbb{R}^{m\times m}$ and $\Sigma \in \mathbb{R}^{m\times m}$ inv., so that
$$A^+ A = V\Sigma^{-1} U^T U \Sigma V^T = \mathrm{Id}$$
and
$$A A^+ = U \Sigma V^T V \Sigma^{-1} U^T = \mathrm{Id}$$

If rank$(A) = r$ then rank$(A^+) = r$ &
$$\sigma_1(A^+) = \|A^+\|_{\ell^2\to\ell^2} = (\sigma_r(A))^{-1}$$

If $A^T A \in \mathbb{R}^{n\times n}$ is invertible ($\Rightarrow m \geq n$) then

$$A^+ = (A^T A)^{-1} A^T$$

Indeed,

$$(A\ A)^{-1} A = (V \Sigma^2 V^T)^{-1} V \Sigma U^T$$

$$= V \Sigma^2 V^T V \Sigma U^T = V \Sigma^{-1} U^T = A^+$$

If $A A^T \in \mathbb{R}^{m\times m}$ is invertible ($\Rightarrow n \geq m$) then (in the same way)

$$A^+ = A^T (A A^T)^{-1}$$

Least Squares Problems

Connection of Moore-Penrose pseudo-inverse to least square problems

Proposition 3.16: Let $A \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$.

Define $\mathcal{M} := \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \|Ax - y\|_2$, i.e.

the set of minimizers of $\|Ax-y\|_2$. The optimization problem

$$\min_{x\in\mathcal{M}} \|x\|_2$$

possesses the unique minimizer $\boxed{x^\# = A^+ y}$.

Proof: The (full) SVD of A can be written in the form $A = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ with $\tilde{U} \in \mathbb{R}^{m\times m}$, $\tilde{V} \in \mathbb{R}^{n\times n}$ orthogonal and

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma = \underbrace{\operatorname{diag}(\sigma_1, \dots, \sigma_r)}_{\in \mathbb{R}^{r\times r}}$$

$r = \operatorname{rank}(A)$.

Let $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \tilde{V}^T x$, $z_1 \in \mathbb{R}^r$
$z_2 \in \mathbb{R}^{n-r}$

$b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \tilde{U}^T y$ $\quad b_1 \in \mathbb{R}^r, b_2 \in \mathbb{R}^{m-r}$

Since the ℓ_2-norm is invariant under orthogonal transformations, it holds

$$\|Ax - y\|_2 = \|\tilde{U}^T (Ax - y)\|_2$$
$$= \|\tilde{U}^T \tilde{U} \tilde{\Sigma} \tilde{V}^T x - \tilde{U}^T y\|_2$$
$$= \|\tilde{\Sigma} \tilde{V}^T x - \tilde{U}^T y\|_2 = \left\| \begin{pmatrix} \Sigma z_1 - b_1 \\ -b_2 \end{pmatrix} \right\|_2$$

Therefore, $\|Ax - y\|_2$ is minimized for

$z_1 = \Sigma^{-1} b_1$ and arbitrary z_2, i.e.,

$$\mathcal{M} = \left\{ x = \tilde{V} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \tilde{V} \begin{pmatrix} \Sigma^{-1} b_1 \\ z_2 \end{pmatrix} : z_2 \in \mathbb{R}^{n-r} \right\}.$$

For $x \in \mathcal{M}$, in particular, for z_1 fixed,

$$\|x\|_2 = \|\tilde{V}^T x\|_2^2 = \|z\|_2^2 = \|z_1\|_2^2 + \|z_2\|_2^2$$

Hence, $\|x\|_2^2$ is minimized for $z_2 = 0$, i.e.,

for $x = \tilde{V}\begin{pmatrix} \Sigma^{-1} b_1 \\ 0 \end{pmatrix} = \tilde{V}\begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} b$

$$= \tilde{V}\begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} \tilde{U}^T y = A^+ y$$

by definition of the Moore-Penrose pseudo-inverse. $\square$

<u>Corollary 3.17</u>: Let $A \in \mathbb{R}^{m \times n}$, $m \geq n$, be of full rank ($\text{rank}(A) = n$), and let $y \in \mathbb{R}^m$. Then the least squares problem

$$\min_{x \in \mathbb{R}^n} \|Ax - y\|_2$$

has the unique solution $x^\# = A^+ y$.

<u>Proof</u>: Note that $Ax^\#$ is the orthogonal projection of y onto $\text{ran}(A)$. Since A is of full rank n it is injective and therefore $x^\#$ is unique by uniqueness of the orth. projection.

The statement follows therefore from Prop. 3.16 □

Remark: • The proof reveals that

AA^+ is the orthogonal projection onto $\operatorname{ran}(A)$ (even if A is not of full rank).

• If $A \in \mathbb{R}^{m\times n}$, $m \geq n$, has full rank, then A^*A is invertible, so that $A^+ = (A^*A)^{-1}A^*$

Therefore $x^\# = A^+ y$ is equivalent to the normal equation

$$\boxed{A^T A x^\# = A^T y}$$

Corollary 3.18: Let $A \in \mathbb{R}^{m\times n}$, $n \geq m$, have full rank m, and let $y \in \mathbb{R}^m$. Then the least squares problem

$$\min_{x\in\mathbb{R}^n} \|x\|_2 \quad \text{subject to} \quad Ax = y$$

has the unique solution $x^\# = A^+ y$

Proof: Since A has full rank m, a solution to $Ax = y$ exists and therefore a minimizer

of $\|Ax - y\|_2$ satisfies $Ax = y$.
The statement follows then from Prop. 3.16.B

Remark: If $A \in \mathbb{R}^{m \times n}$, $m \le n$, is of full rank m, it holds $A^+ = A^T(AA^T)^{-1}$. Therefore $x^\#$ satisfies the normal equation of the second kind

$$x^\# = A^T b \quad \text{where} \quad AA^T b = y \quad .$$

The singular value decomposition exists also in the complex case $A \in \mathbb{C}^{m \times n}$ and takes the form

$$A = U \Sigma V^*$$

where the columns of $U \in \mathbb{C}^{m \times r}$, $V \in \mathbb{C}^{n \times r}$ are orthonormal and $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r \times r}$ with $\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_r > 0$.

All results above extend literally when replacing "orthogonal" by "unitary" and "V^T" by "V^*".

Application: Principal Component Analysis (PCA)

$A \in \mathbb{R}^{m \times n}$ data matrix, columns $A_{(j)}$ are data points

Idea: Fit an m-dimensional ellipsoid to data such that each axis of the ellipsoid represents a principal component.

Consider $A_{(j)}$, $j = 1, \dots, n$, as realizations of random vector X in $\mathbb{R}^m$. Assume X has mean zero, otherwise, subtract mean $\mathbb{E}X$, or in practice, empirical mean $\hat{E} = \frac{1}{n} \sum_{k=1}^{n} A_{(k)}$ from each column $A_{(j)}$

PCA transforms data via an orthogonal basis transformation $U = (u_1 | \dots | u_m)$ so that the principal component scores

$$T_{(j)} = \begin{pmatrix} \langle A_{(j)}, u_1 \rangle \\ \vdots \\ \langle A_{(j)}, u_m \rangle \end{pmatrix} = \left(A^T U\right)_{(j)}$$

are such that the first components $(T_{(j)})_1$, $j = 1, \dots, n$

have the largest (empirical) variance, the second components $(T_{(j)})_{(2)}$, $j = 1, \dots, n$, have the second largest variance, etc., i.e.

$$u_1 = \underset{\|u\|_2 = 1}{\operatorname{argmax}} \sum_{j=1}^{n} (T_{(j)})_1^2 = \underset{\|u\|_2 = 1}{\operatorname{argmax}} \sum_{j=1}^{n} \langle A_{(j)}, u \rangle^2$$

$$= \underset{\|u\|_2 = 1}{\operatorname{argmax}} \|A^T u\|_2$$

$$u_2 = \underset{\substack{\|u\|_2 = 1 \\ u \perp u_1}}{\operatorname{argmax}} \|A^T u\|_2$$

$$\vdots$$

In other words, the principle components of A are the left singular vectors of A.

Moreover, the principle components (scores)

$$T = (T_{(1)}, \dots, T_{(n)}) = (U^T A_{(1)}, \dots, U^T A_{(n)})$$

$$T^T = A^T U = (U \Sigma V^T)^T U = V \Sigma^T$$

where $A = U \Sigma V^T$ is the SVD of A, so that $T = \Sigma V^T$ and $A = UT$.

Replacing U by $U_k = (u_1 | \ldots | u_k) \in \mathbb{R}^{m \times k}$

T by $T_k = V_k \Sigma_k^T = V_k \operatorname{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{n \times k}$

yields a low rank approximation $A_k = U_k T_k$

in explicit factorized form.

The scaled basis vectors $w_j = \sigma_j u_j$ are called <u>loadings</u> and correspond to the ellipsoid axes of the "ellipsoid approximation" of the data points.

<u>Covariances</u>

covariance matrix of mean-zero random vector X defined as $C = \mathbb{E}\, X X^T$.

Empirical covariance matrix

$$\hat{C} = \frac{1}{n} \sum_{j=1}^{n} A_{(j)} \overline{A_{(j)}^T} = \frac{1}{n} A A^T = \frac{1}{n} \left(\langle A_{(h)}, A_{(e)} \rangle \right)_{h,e=1}^{m}$$

Since $u_1, \ldots, u_m$ are eigenvectors of AA^T the empirical covariance between two principle components (entries h,e of $T_{(:)}$)

$$\frac{1}{n}\sum_{j=1}^{n} \left(T_{(j)}\right)_k \left(T_{(j)}\right)_\ell = \frac{1}{n}\left(T T^T\right)_{k,\ell}$$

$$= \frac{1}{n}\left\langle (U^T A)_k, (U^T A)_\ell \right\rangle = \frac{1}{n} \Sigma V^T V \Sigma^T$$

$$= \frac{1}{n} \Sigma \Sigma^T = \frac{1}{n} \operatorname{diag}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0)$$

This means that the basis transformation with U diagonalizes the covariance matrix. If $\frac{1}{n} A A^T$ would be true covariance matrix of X then $U^T X$ would have diagonal covariance matrix, in particular, $E\,(U^T X)_k (U^T X)_j = 0$ if $j \neq k$. If X is Gaussian this would imply independence of the entries of $U^T x$.