Singular Value Decomposition

Note Title

- SVD is a matrix factorization that is useful for many applications, e.g., search engines, LS problems, tomographic image reconstruction, ...

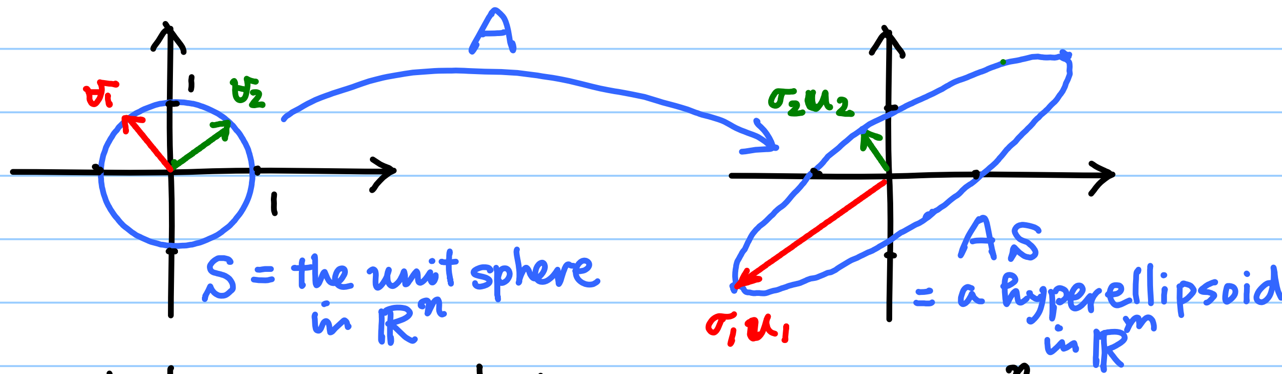
- SVD can be a concepual tool in linear algebra
  - ⇒ via SVD, we can check:
    - a given matrix is near singular
    - rank of the matrix
    - etc.

- ∃ a numerically stable algorithm to compute the SVD of a given matrix (it's expensive though...)
  In fact, one of the hottest topics in numerical linear algebra is how to compute a good approximation to the SVD of a huge matrix fast!

☆ A Geometric Observation

Let $A \in \mathbb{R}^{m \times n}$, and consider how A maps an input vector in $\mathbb{R}^n$ to an output vector in $\mathbb{R}^m$.

"The image of the unit sphere under any $m \times n$ matrix is a hyperellipsoid"



ONB = orthonormal basis

Let $\{v_1, \cdots, v_n\}$ be an ONB of $\mathbb{R}^n$
Let $\{u_1, \cdots, u_m\}$ be an ONB of $\mathbb{R}^m$
Let $\{\sigma_1, \cdots, \sigma_m\}$ be a set of m scalars with $\sigma_i \geq 0$, $i = 1, \cdots m$.
Then, $\sigma_i u_i$ is the ith principal semiaxis with length σ_i in $\mathbb{R}^m$.

Now, if $\text{rank}(A) = r$, then exactly r of $\{\sigma_1, \cdots, \sigma_m\}$ are nonzero, and exactly $m-r$ of σ_i's are zero.
So, if $\underline{m \geq n}$, then $\underline{\text{rank}(A) \leq n}$. ← full rank if $= n$
i.e., at most n of σ_i's are nonzero.

For simplicity, let's assume $m \geq n$ and $\text{rank}(A) = n$ for the time being.

<u>Def.</u> The singular values of A
$\overset{\text{def}}{\Longleftrightarrow}$ The lengths of the n principal semiaxes of the hyperellipsoid AS

Our convention: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \underset{\neq}{\geq} 0$

Def. The n **left singular vectors** of A
$\overset{\text{def}}{\Longleftrightarrow}$ $\{u_1, \cdots, u_n\}$: the unit vectors in $\mathbb{R}^m$ along the principal semiaxes of AS.
So, $\sigma_i u_i$ is the ith largest principal semiaxis of AS.

Def. The n **right singular vectors** of A
$\overset{\text{def}}{\Longleftrightarrow}$ $\{v_1, \cdots, v_n\} \in S$: the preimages of the principal semiaxes of AS, i.e.,

$$A v_i = \sigma_i u_i \qquad i = 1, \cdots, n.$$

☆ Reduced SVD

$$m\underset{n}{\begin{bmatrix} A \end{bmatrix}} \underbrace{\begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}}_{\overset{n}{=V}}{}^{n} = m\underbrace{\begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix}}_{\overset{n}{=\hat{U}}} \underbrace{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix}}_{\overset{n}{=\hat{\Sigma}}}{}^{n}$$

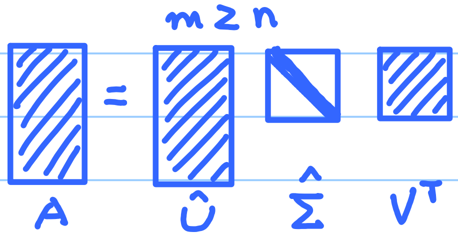
$$\Rightarrow \quad \underset{m\times n}{A}\ \underset{n\times n}{V} = \underset{m\times n}{\hat{U}}\ \underset{n\times n}{\hat{\Sigma}}$$

Since V is an orthogonal matrix,

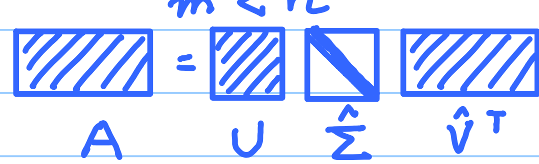
$$\boxed{A = \hat{U}\hat{\Sigma}V^T}$$

The **reduced SVD** of A.

$m \geq n$



$m < n$



☆ Full SVD

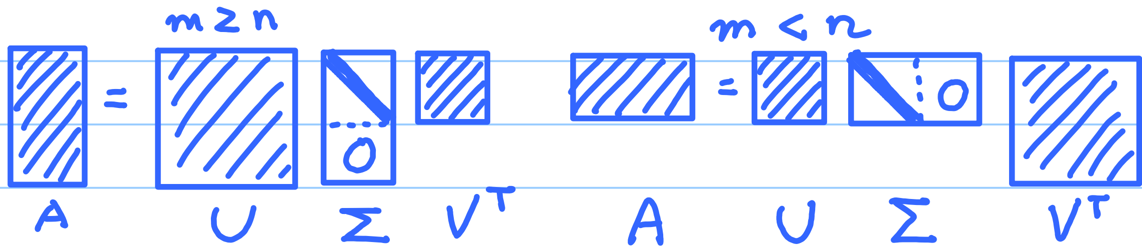
Note $\hat{U} \in \mathbb{R}^{m \times n}$ in the reduced SVD with $m \geq n$.

$\Rightarrow$ The column vectors of $\hat{U}$ do not form an ONB of $\mathbb{R}^m$ unless $m = n$.

$\Rightarrow$ Remedy : adjoin $m-n$ ON vectors to $\hat{U}$ to form an orthogonal matrix U.

Then $\hat{\Sigma}$ must be changed to $\Sigma \in \mathbb{R}^{m \times n}$

$$\boxed{A = U \Sigma V^T}$$ The full SVD of A



For non-full rank matrices, i.e., $\operatorname{rank}(A) = r < \min(m, n)$, $\exists$ only r positive singular values.

So,

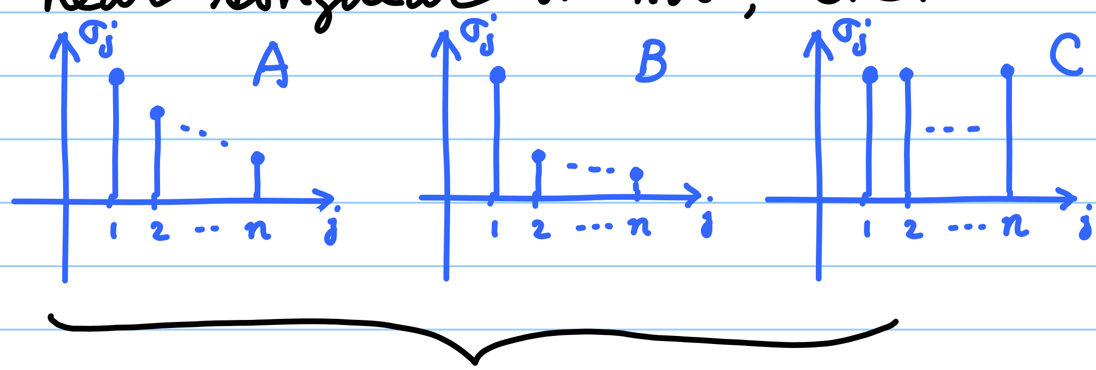
$$\Sigma = \left[\begin{array}{ccccc} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \; 0 \\ \hline & & O & & \end{array}\right] \quad \text{or} \quad \left[\begin{array}{ccccc|c} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & O \\ & & & 0 & & \\ & & & & \ddots \; 0 & \end{array}\right]$$

$m \geq n$ $\qquad\qquad$ $m \leq n$

Let's consider $m = n$ and full rank case.

Theoretically, it's invertible, nonsingular.

However, we can gain more info by checking the distribution of the singular values of A $\Rightarrow$ We can see whether A is near singular or not, etc.



Out of these three scenarios, which matrix do you think behaves best numerically?

$\Rightarrow$ C.

☆ Pseudoinverse via SVD

$$\boxed{A^+ = V \Sigma^+ U^T}$$

where

$$\Sigma^+ := \underbrace{\begin{bmatrix} \frac{1}{\sigma_1} & & & & \\ & \ddots & & & \\ & & \frac{1}{\sigma_r} & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \\ \hline & & & 0 & & \end{bmatrix}}_{m \geq n} \quad \text{or} \quad \underbrace{\left[\begin{array}{cccccc|c} \frac{1}{\sigma_1} & & & & & & \\ & \ddots & & & & & \\ & & \frac{1}{\sigma_r} & & & & 0 \\ & & & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & \end{array}\right]}_{m \leq n}$$

<u>Check</u>: $AA^+ = U\Sigma V^T V \Sigma^+ U^T$

$$= U \Sigma \Sigma^+ U^T$$

$$= U \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & \ddots \ 0 \\ \hline & & O & & \end{bmatrix} U^T$$

$$= [\,u_1 \cdots u_r \ 0 \cdots 0\,] \begin{bmatrix} u_1^T \\ \vdots \\ u_m^T \end{bmatrix}$$

$$= \hat{U}\hat{U}^T$$

→ reduced version.

Similarly, $A^+A = \hat{V}\hat{V}^T$

The Moore-Penrose Conditions

For a given matrix $A \in \mathbb{R}^{m\times n}$, if $X \in \mathbb{R}^{n\times m}$ satisfies the following:

$$\begin{cases} (1) & AXA = A \\ (2) & XAX = X \\ (3) & (AX)^T = AX \\ (4) & (XA)^T = XA \end{cases}$$

then X is called the pseudoinverse (or the Moore-Penrose inverse) of A and written as A^+

$\exists$ many applications using A^+!

<u>Note</u>: If $\| AX - I_m \|_F \to \min$ then $X = A^+$.

SVD

Note Title

★ Formal Definition

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

Then SVD of A is a factorization

full SVD → $$A = U \Sigma V^T$$

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

$$\operatorname{diag}(\Sigma) = [\sigma_1, \sigma_2, \dots, \sigma_p]^T$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0.$$

$$p = \min(m, n)$$

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

A & Σ are the same shape.

Geometrically,

$$\begin{array}{ccccccc} S & \xrightarrow{V^T} & S & \xrightarrow{\Sigma} & \text{hyper-ellipsoid} & \xrightarrow{U} & \text{hyper-ellipsoid} \\ \mathbb{R}^n & & \mathbb{R}^n & & \mathbb{R}^m & & \mathbb{R}^m \end{array}$$

V^T: rotation or reflection in $\mathbb{R}^n$

Σ: stretching or squeezing

U: rotation or reflection in $\mathbb{R}^m$

So if we prove every $A \in \mathbb{R}^{m\times n}$ has an SVD, then we shall have proved that A maps the unit sphere in $\mathbb{R}^n$ to a hyperellipsoid in $\mathbb{R}^m$.

☆ Existence & Uniqueness of SVD

→ We can get peace of mind if we know that $\exists!$ SVD for any given matrix.

<u>Thm</u> Every matrix $A \in \mathbb{R}^{m\times n}$ has an SVD. Furthermore, the singular values $\{\sigma_j\}$ are uniquely determined. If A is square and σ_j's are distinct, then singular vectors $\{u_j\}$, $\{v_j\}$ are uniquely determined up to signs (i.e., ± 1 factor).

(Proof : Existence)

Let's check the longest action of A first, then do induction.

definition

$$\text{Set } \sigma_1 = \|A\|_2 = \sup_{v\in S} \|Av\|_2$$

This is often called "compactness" argument.

Because we are dealing with vectors in $\mathbb{R}^n$ (i.e., finite dimensional space), and $\|A\cdot\|_2$ is a continuous fcn, $\exists\, v_1 \in S \subset \mathbb{R}^n$ s.t. $\|Av_1\|_2 = \sigma_1$ is attained.

Now set $\tilde{u}_1 = Av_1 \in \mathbb{R}^m$, and consider orthogonal matrices

$V_1 = [v_1 \; v_2 \cdots v_n] \in \mathbb{R}^{n\times n}$,

$$U_1 = [\, u_1 \;\; u_2 \;\cdots\; u_m \,] \in \mathbb{R}^{m\times m}$$

where $u_1 = \frac{1}{\sigma_1} \tilde{u}_1$

Note $\| u_1 \| = \frac{1}{\sigma_1} \| \tilde{u}_1 \| = \frac{1}{\sigma_1} \| A v_1 \|$

$$= \frac{1}{\sigma_1} \cdot \sigma_1 = 1 \quad \checkmark$$

Then, $$U_1^T A V_1 = \begin{bmatrix} u_1^T \\ \vdots \\ u_m^T \end{bmatrix} A \, [\, v_1 \cdots v_n \,]$$

$$= \begin{bmatrix} u_1^T \\ \vdots \\ u_m^T \end{bmatrix} [\, \underbrace{A v_1}_{= \tilde{u}_1 = \sigma_1 u_1} \cdots A v_n \,]$$

$$= \underbrace{\left[\begin{array}{c|c} \sigma_1 & w^T \\ \hline 0 & \\ \vdots & B \\ 0 & \end{array}\right]}_{\text{let's call } = \Sigma_1}$$

$u_j^T u_1 = 0$ for $j \geq 2$.

where $w^T = [\, u_1^T A v_2, \cdots, u_1^T A v_n \,] \in \mathbb{R}^{1\times n-1}$

$$B = \begin{bmatrix} u_2^T A v_2 & \cdots & u_2^T A v_n \\ \vdots & & \vdots \\ u_m^T A v_2 & \cdots & u_m^T A v_n \end{bmatrix} \in \mathbb{R}^{m-1\times n-1}$$

$$\Big\| \underbrace{\begin{bmatrix} \sigma_1 & w^T \\ 0 & B \end{bmatrix}}_{= \Sigma_1} \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \Big\|_2 \geq \sigma_1^2 + w^T w$$

$$= \sqrt{\sigma_1^2 + \|w\|^2} \; \Big\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \Big\|_2$$

$$\Rightarrow \quad \|\Sigma_1\|_2 \geq \sqrt{\sigma_1^2 + \|w\|^2} \quad \text{——} \; ①$$

Since U_1, V_1 are orthogonal,

$$\|\Sigma_1\|_2 = \|A\|_2 = \sigma_1 \quad \text{——} \; ②$$

From ① & ②, we can conclude that $w = 0$, i.e.,

$$U_1^T A V_1 = \left[\begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & B \end{array}\right]$$

Hence if $m=1$ or $n=1$, we are done!

In general case, we can use the induction hypothesis:

Suppose an SVD exists for any $m-1 \times n-1$ matrix. Then the above matrix B has its SVD: $B = U_2 \Sigma_2 V_2^T$

Then
$$A = \underbrace{U_1 \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & U_2 \end{array}\right]}_{U} \underbrace{\left[\begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & \Sigma_2 \end{array}\right]}_{\Sigma} \underbrace{\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & V_2 \end{array}\right]^T V_1^T}_{V^T}$$

This is an SVD of A! ///

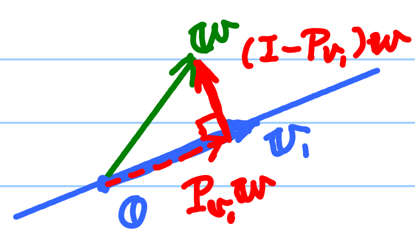
(Proof: Uniqueness)

Let $v_1 \in S \subset \mathbb{R}^n$ s.t.
$$\|A\|_2 = \|\tilde{u}_1\|_2 = \|A v_1\|_2 = \sigma_1$$

Suppose $\exists w \in S$, s.t., $w \neq v_1$, w is linearly independent from v_1, and $\|Aw\|_2 = \sigma_1$.

Let's define a unit vector $v_2 \in S$ by

$$v_2 := \frac{(I - P_{v_1})w}{\|(I - P_{v_1})w\|_2}$$



$v_2 \perp v_1$

Since $\|A\|_2 = \sigma_1$, by definition

$\|Av_2\|_2 \leq \sigma_1$ ---- (a)

We now claim $\|Av_2\|_2 = \sigma_1$.

Exercise: why? Because $w = P_{v_1} w + (I - P_{v_1}) w$

$= c v_1 + s v_2$

why $c^2+s^2=1$? where c, s: constants satisfying $c^2 + s^2 = 1$ --- (b)

$\sigma_1^2 = \|Aw\|_2^2 = \|cAv_1 + sAv_2\|_2^2$

$= c^2\|Av_1\|_2^2 + 2cs\underbrace{(Av_1)^T Av_2}_{=0} + s^2\|Av_2\|_2^2$

$= c^2\sigma_1^2 + s^2\|Av_2\|_2^2 \underset{(a)}{\leq} c^2\sigma_1^2 + s^2\sigma_1^2 \underset{(b)}{=} \sigma_1^2$

This means that the inequality above must be an equality, and hence $\|Av_2\|_2 = \sigma_1$ //

Hence, what we have proved is:
if v_1 is not unique, then the corresp. singular value σ_1 is not simple (i.e., has some multiplicity).
After determining σ_1, u_1, v_1, we can use the induction argument.
In particular, for A: square, $\{\sigma_j\}$ are distinct (no multiple singular values), then it's clear that $\{u_j\}, \{v_j\}$ are uniquely determined up to signs. ///

More about SVD!

☆ "A Change of Bases" viewpoint

$$A = U \Sigma V^T \in \mathbb{R}^{m \times n}$$

Pick any $\mathbf{x} \in \mathbb{R}^n$ and consider

$$\tilde{\mathbf{x}} = V^T \mathbf{x}$$

Then $\tilde{\mathbf{x}}$ is the expansion coefficient of $\mathbf{x}$ w.r.t. the ONB $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$

<u>why?</u> You should know this by now.

But, just in case,

$$\tilde{\mathbf{x}} = V^T \mathbf{x} \iff \mathbf{x} = V \tilde{\mathbf{x}}$$

$$= \underbrace{\tilde{x}_1 \mathbf{v}_1 + \cdots + \tilde{x}_n \mathbf{v}_n}_{\text{linear comb. of } \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}.}$$

//

Now, let $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^m$

Expand $\mathbf{b}$ w.r.t. the ONB $\{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$

$$\tilde{\mathbf{b}} = U^T \mathbf{b} = U^T A \mathbf{x} = U^T A V \tilde{\mathbf{x}}$$

$$= \underbrace{U^T U}_{= I_m} \Sigma \underbrace{V^T V}_{" I_n} \tilde{\mathbf{x}} = \Sigma \tilde{\mathbf{x}}$$

Now, we know that Σ is diagonal!

This again shows that

"Σ represents the essence of A in a much clearer manner!"

☆ SVD vs Eigenvalue Decomposition

Let $A \in \mathbb{R}^{m\times m}$ be diagonalizable, i.e., $\exists$ the eigenvalue decomposition:

$$A = X \Lambda X^{-1}$$

Note: Even if $A \in \mathbb{R}^{m\times m}$, its eigval's & eigvec's may be complex-valued! Ex. $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

where $X = [\mathbb{x}_1 \cdots \mathbb{x}_m] \in \mathbb{C}^{m\times m}$

$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_m) \in \mathbb{C}^{m\times m}$

satisfying $A\mathbb{x}_j = \lambda_j \mathbb{x}_j$, $j = 1, \cdots, m$

$$\Updownarrow$$

$$A X = X \Lambda$$

Note that the eigenvectors $\{\mathbb{x}_1, \cdots, \mathbb{x}_m\}$ form a basis of $\mathbb{C}^m$, but not necessarily orthonormal in general unless $A^* = A$ (unitary)

Here $A^* := (\bar{a}_{ji}) = \bar{A}^T$

conjugate transposition of $A \in \mathbb{C}^{m\times m}$

"unitarity" is a generalization of "symmetry".

With the eigenvalue decomposition,

$\mathbb{b} = A\mathbb{x}$ can be simplified as

$\tilde{\mathbb{b}} = \underset{\text{diagonal}}{\Lambda}\, \tilde{\mathbb{x}}$ via $\begin{cases} \tilde{\mathbb{b}} = X^{-1}\mathbb{b} \\ \tilde{\mathbb{x}} = X^{-1}\mathbb{x} \end{cases}$

change of bases again!

So, we can summarize as follows:

- SVD: Use two different ONB's U, V and work for any matrix.
- EIG: Use one basis (not ONB in general) and work only for square matrices.

☆ Matrix Properties via SVD

Let $A \in \mathbb{R}^{m \times n}$,

$p := \min(m, n)$

$r := \#$ nonzero singular values $\leq p$.

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

(Proof) Let $A = U \Sigma V^T$.

Since U, V are orthogonal matrices, they are of full rank.

Hence, $\operatorname{rank}(A) = \operatorname{rank}(\Sigma)$

$= \#$ nonzero diagonal entries

$= r$ ///

Recall $\langle u_1, \cdots, u_r \rangle := \operatorname{span}\{u_1, \cdots, u_r\}$

Thm $\operatorname{range}(A) = \langle u_1, \cdots, u_r \rangle$

$\operatorname{null}(A) = \langle v_{r+1}, \cdots, v_n \rangle$

(Proof) Since $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal with only r nonzero entries,

$\text{range}(\Sigma) = \langle e_1, \cdots, e_r \rangle \subset \mathbb{R}^m$

$\Leftrightarrow \text{range}(A) = \langle u_1, \cdots, u_r \rangle \subset \mathbb{R}^m$. ✓

On the other hand, it is clear that for any vector $\mathbb{x} \in \mathbb{R}^n$ s.t.

$$\mathbb{x} = [\underbrace{0, 0, \cdots, 0}_{r}, x_{r+1}, \cdots, x_n]^T,$$

$$\Sigma \mathbb{x} = \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ x_{r+1} \\ \vdots \\ x_n \end{bmatrix} = \mathbb{0}.$$

So, $\text{null}(\Sigma) = \langle e_{r+1}, \cdots, e_n \rangle \subset \mathbb{R}^n$

Then, for such $\mathbb{x}$, we have

$$\begin{aligned} A V \mathbb{x} &= U \Sigma V^T V \mathbb{x} \\ &= U \Sigma \mathbb{x} = \mathbb{0} \end{aligned}$$

i.e., any member of null(A) should be of the form $V\mathbb{x}$, $\mathbb{x} \in \text{null}(\Sigma)$

i.e., $\text{null}(A) = \langle v_{r+1}, \cdots, v_n \rangle \subset \mathbb{R}^n$ ///

Thm $\|A\|_2 = \sigma_1$, $\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$

(Proof) Since U, V are orthogonal,

$$\|A\|_2 = \|\Sigma\|_2 = \max_{1\le j\le r} \{|\sigma_j|\} = \sigma_1 \checkmark$$

The Frobenius norm is also invariant w.r.t. rotations (ortho. matrix multiplications)

Hence, $\|A\|_F = \|\Sigma\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$ ///

<u>Thm</u> The nonzero singular values of A are the square roots of the nonzero eigenvalues of A^TA or AA^T.

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

$\Leftrightarrow (A^TA)V = V(\Sigma^T\Sigma)$

" $\mathrm{diag}(\sigma_1^2, \cdots, \sigma_r^2, 0, \cdots, 0)$
$\in \mathbb{R}^{n\times n}$

So, the col's of V are the eigenvectors of A^TA and their nonzero eigval's are $\sigma_1^2, \cdots, \sigma_r^2$
You can show similarly that the col's of U are the eigenvec's of AA^T, and their nonzero eigval's are $\sigma_1^2, \cdots, \sigma_r^2$. ///

<u>Thm</u> $A^T = A \Rightarrow \sigma_i(A) = |\lambda_i(A)|$

(Proof) HW#3 Prob 3 says:
Any symmetric matrix has only real-valued eigenvalues and the eigenvec's form an ONB.
So, $A = Q\Lambda Q^T$, Q: ortho, Λ: diag
$= Q|\Lambda|\,\mathrm{sgn}(\Lambda)\,Q^T$

where $|\Lambda| := \begin{bmatrix} |\lambda_1| & & 0 \\ & \ddots & \\ 0 & & |\lambda_m| \end{bmatrix}$

$\operatorname{sgn}(\Lambda) := \begin{bmatrix} \operatorname{sgn}(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & \operatorname{sgn}(\lambda_m) \end{bmatrix}$

Now, it's clear that $Q \operatorname{sgn}(\Lambda)$ is orthogonal if Q is orthogonal.

<u>why?</u>

$$(Q \operatorname{sgn}(\Lambda))(Q \operatorname{sgn}(\Lambda))^T$$
$$= Q \operatorname{sgn}(\Lambda) \operatorname{sgn}(\Lambda) Q^T$$
$$= Q Q^T = I_m$$

So, $$A = \underbrace{Q}_{U} \underbrace{|\Lambda|}_{\Sigma} \underbrace{(Q \operatorname{sgn}(\Lambda))^T}_{V^T}$$ ///

<u>Thm</u> For $A \in \mathbb{R}^{m \times m}$,

$$|\det(A)| = \prod_{i=1}^{m} \sigma_i = \sigma_1 \cdot \sigma_2 \cdot \cdots \cdot \sigma_m$$

(Proof) We'll use the following facts.

- $\det(AB) = \det(A) \cdot \det(B)$.
- $\det(A^T) = \det(A)$
- $\det(\operatorname{diag}(a_1, \cdots, a_m)) = \prod_{i=1}^{m} a_i$
- For any Q: orthogonal, $|\det(Q)| = 1$.

<u>why?</u> $\det(Q^T Q) = \det(Q^T) \cdot \det(Q) = (\det(Q))^2$
$= \det(I) = 1$. so, $|\det(Q)| = 1$ ✓

Then, $|\det(A)| = |\det(U \Sigma V^T)| = |\det(\Sigma)|$
$= \prod \sigma_i$

///

Low Rank Approximations

Note Title

Recall Outer product in Lecture 3.

Let $u \in \mathbb{R}^m = \mathbb{R}^{m\times 1}$,
$v \in \mathbb{R}^n = \mathbb{R}^{n\times 1}$.

Then, the outer product between u and v is:

$$uv^T = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} [v_1 \cdots v_n] = \begin{bmatrix} u_1v_1 & \cdots & u_1v_n \\ \vdots & & \vdots \\ u_mv_1 & \cdots & u_mv_n \end{bmatrix} \in \mathbb{R}^{m\times n}$$

This matrix has rank 1 because
$uv^T = [v_1 u, \cdots, v_n u]$
i.e., each column is just a scalar multiple of the same vector u.

Now SVD can be viewed as a sum of rank 1 matrices:

Thm $$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T, \quad r = \text{rank}(A)$$

(Proof) just obvious!

$$[u_1 \cdots u_m] \begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ 0 & & & & \ddots \; 0 \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$$

///

Among all possible $m\times n$ matrices of rank k $(k \le r)$,

$\sum_{j=1}^{k} \sigma_j u_j v_j^T$ is the best approximation of A in the following sense:

Thm For any k with $0 \le k \le r$, let
$$A_k := \sum_{j=1}^{k} \sigma_j u_j v_j^T$$
If $k = p = \min(m,n)$, then define $\sigma_{k+1} = 0$. Then,
$$\|A - A_k\|_2 = \inf_{\substack{B \in \mathbb{R}^{m\times n} \\ \mathrm{rank}(B) \le k}} \|A - B\|_2 = \sigma_{k+1}$$

(Proof)
$$\|A - A_k\|_2 = \left\| \sum_{j=k+1}^{p} \sigma_j u_j v_j^T \right\|_2$$
$$= \left\| U \begin{bmatrix} 0 & & & & 0 \\ & \ddots & & & \\ & & 0 & & \\ & & & \sigma_{k+1} & \\ 0 & & & & \ddots \ \sigma_p \\ \hline & & 0 & & \end{bmatrix} V^T \right\|_2$$
$$= \left\| \begin{bmatrix} 0 & & & & 0 \\ & \ddots & & & \\ & & 0 & & \\ & & & \sigma_{k+1} & \\ 0 & & & & \ddots \ \sigma_p \\ \hline & & 0 & & \end{bmatrix} \right\|_2$$
since U, V : orthogonal!

$= \sigma_{k+1}$ by definition of the matrix norm

Prove this as an exercise! ⇒ **Note:** If $D = \mathrm{diag}(d_1, \cdots, d_m) = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_m \end{bmatrix}$

then $\|D\|_p = \max_{1 \le j \le m} |d_j| \quad \forall p \ge 1$

Now, let $B \in \mathbb{R}^{m\times n}$ be any rank k matrix. Then $\dim(\mathrm{null}(B)) = n - k$

why? Because of the following Thm:

For any $A \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(A) + \dim(\mathrm{null}(A)) = n$

Let $W := \text{null}(B) \cap \langle v_1, \cdots, v_{k+1} \rangle$

We know $W \neq \{0\}$ because

$\dim(\text{null}(B)) = n-k$

$\dim(\langle v_1, \cdots, v_{k+1} \rangle) = k+1$

so, if these two do not intersect, $\mathbb{R}^n$'s dimension would become $n-k+k+1 = n+1$

This cannot happen! #

So let $h \in W$, $h \neq 0$.

We can always normalize h, so can assume $\|h\|_2 = 1$.

Then,

$$
\begin{aligned}
\|A-B\|_2 &\geq \|(A-B)h\|_2 \quad \text{by def.} \\
&= \|Ah\|_2 \quad \text{since } h \in \text{null}(B) \\
&= \|U \Sigma V^T h\|_2 \\
&= \|\Sigma V^T h\|_2 \quad \text{since } U \text{: ortho.} \\
&\geq \sigma_{k+1} \|V^T h\|_2 \\
&= \sigma_{k+1} \|h\|_2 = \sigma_{k+1}
\end{aligned}
$$

since $h \in \langle v_1, \cdots v_{k+1} \rangle$

$$
V^T h = \begin{bmatrix} * \\ \vdots \\ * \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \left.\vphantom{\begin{matrix} * \\ \vdots \\ * \end{matrix}}\right\} k+1 \\ \left.\vphantom{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}}\right\} n-k-1 \end{array}
$$

///

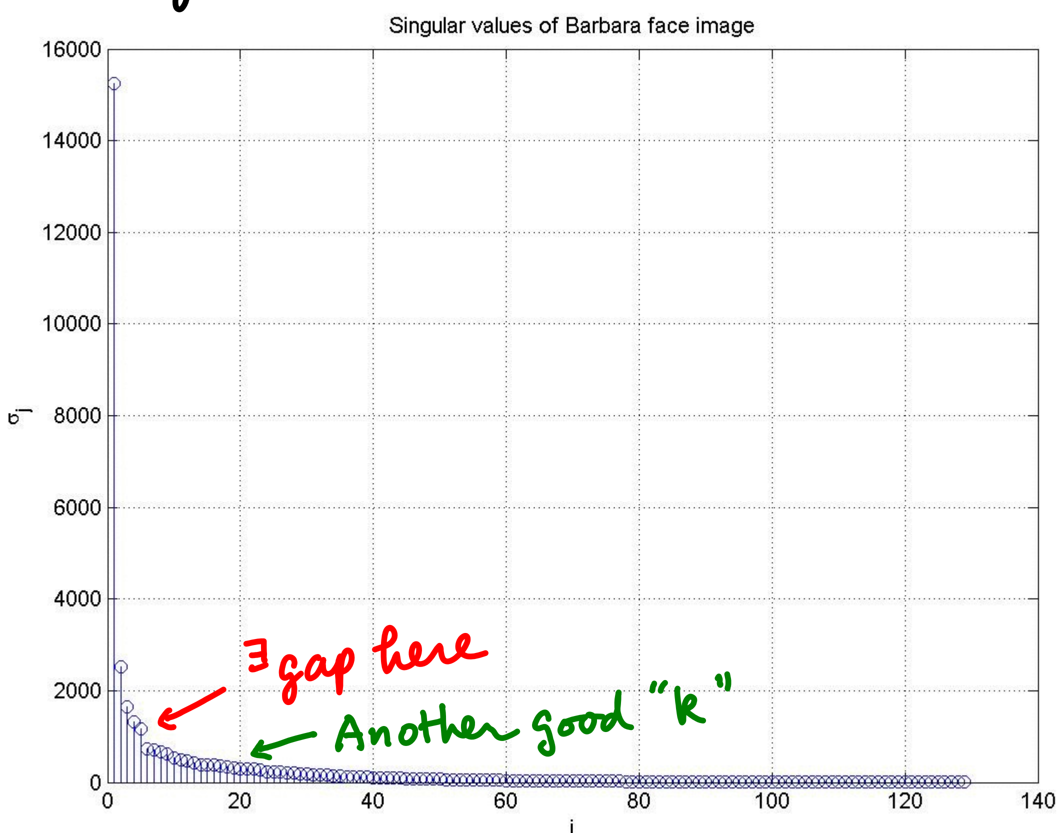
<u>Thm</u> For any k with $0 \leq k \leq r$,

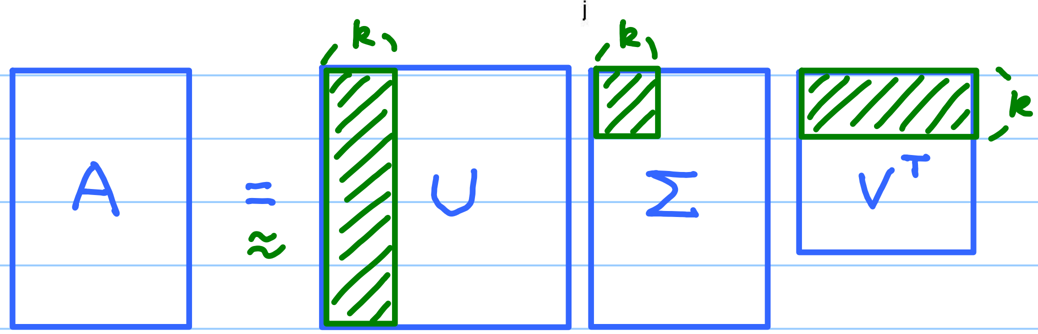
$$
\|A - A_k\|_F = \inf_{\substack{B \in \mathbb{R}^{m \times n} \\ \text{rank}(B) \leq k}} \|A - B\|_F = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2}
$$

(Proof) Exercise!

So, for a given matrix, say, A how to determine a good "k" so that we can efficiently (i.e., compress) A without losing too much info of A?

$\Rightarrow$ Check the distribution of the singular values!





rank k approximation of A only uses ///// portions!

☆ Condition Number and SVD

Recall the condition number for a square nonsingular matrix A:

$$\kappa(A) = \mathrm{cond}(A) := \|A\|_2 \cdot \|A^{-1}\|_2$$

$\kappa(A)$: small $\Rightarrow$ A: well-conditioned.

$\kappa(A)$: large $\Rightarrow$ A: ill-conditioned,

lose $\approx \log_{10}\kappa(A)$ digits to solve $A\mathbf{x} = \mathbf{b}$.

If A: singular, $\kappa(A) = +\infty$.

Using SVD of A, we can nicely compute $\kappa(A)$ as follows.

$\|A\|_2 = \sigma_1$ $\rightarrow$ by definition

$\|A^{-1}\|_2 = 1/\sigma_m$ $\xrightarrow{\text{why?}}$ $A^{-1} = (U\Sigma V^T)^{-1} = V\Sigma^{-1}U^T$

$= V\,\mathrm{diag}(1/\sigma_1, \cdots, \underbrace{1/\sigma_m}_{\text{largest}})U^T$

So, $\kappa(A) = \sigma_1/\sigma_m$

We can generalize the definition of the condition number for a rectangular matrix $A \in \mathbb{R}^{m\times n}$ using the pseudo-inverse $A^\dagger$ and SVDs as

$$\kappa(A) := \|A\|_2 \cdot \|A^\dagger\|_2$$

$$= \sigma_1/\sigma_r \qquad r = \mathrm{rank}(A) \le \min(m,n)$$

SVD and Least Squares Problems

Note Title

✩ LS via SVD

Recall the LS solution via QR factorization:

(1) Compute reduced QR of A.
(2) Compute $y = \hat{Q}^T b$.
(3) Solve $\hat{R} x = y$ — (*)

If A: full rank, then $\hat{R}_{ii} \neq 0$, $1 \le i \le n$, and the triangular system (*) has a unique LS solution.

Now using the reduced SVD of A, i.e., $A = \hat{U}\hat{\Sigma}V^T$, we can also solve the normal eqn:

$$A^T A x = A^T b$$
$$\Leftrightarrow (\hat{U}\hat{\Sigma}V^T)^T(\hat{U}\hat{\Sigma}V^T) x = (\hat{U}\hat{\Sigma}V^T)^T b$$
$$\Leftrightarrow V\hat{\Sigma}^T\hat{U}^T\hat{U}\hat{\Sigma}V^T x = V\hat{\Sigma}\hat{U}^T b$$
$$\Leftrightarrow V\hat{\Sigma}^T\hat{\Sigma}V^T x = V\hat{\Sigma}^T\hat{U}^T b$$
$$\Leftrightarrow \hat{\Sigma}^T\hat{\Sigma}V^T x = \hat{\Sigma}^T\hat{U}^T b \quad \text{since } V\text{: ortho.}$$
$$\Leftrightarrow \hat{\Sigma}V^T x = \hat{U}^T b \quad \text{if } A\text{: full rank, i.e., } \sigma_j > 0,\ 1 \le j \le n$$

This can be solved easily.

(1) Compute reduced SVD of A.
(2) Compute $y = \hat{U}^T b$.
(3) Solve $\hat{\Sigma} w = y$. —— (**)
(4) Set $x = V w$.

Note: (**) is a diagonal system, easier to solve than (*) !!

★ Pseudoinverse and SVD

Recall that if $A \in \mathbb{R}^{m \times n}$ is full rank,

$m > n$: $A^+ = (A^TA)^{-1}A^T$

$m = n$: $A^+ = A^{-1}$

$m < n$: $A^+ = A^T(AA^T)^{-1}$

However, we can define the pseudo inv. using SVD even if A is not full rank!

$$A = U\Sigma V^T, \qquad \Sigma = \begin{array}{c} \begin{array}{cc} \overbrace{\qquad\qquad}^{r} & \overbrace{\quad}^{n-r} \end{array} \\ \left[\begin{array}{cc|c} \sigma_1 & 0 & \\ & \ddots & 0 \\ 0 & \sigma_r & \\ \hline \multicolumn{2}{c|}{0} & 0 \end{array}\right] \begin{array}{l} \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} r \\ \}\, m-r \end{array} \end{array}$$

Define

$$A^+ := V\Sigma^+U^T, \qquad \Sigma^+ := \begin{array}{c} \begin{array}{cc} \overbrace{\qquad\qquad}^{r} & \overbrace{\quad}^{m-r} \end{array} \\ \left[\begin{array}{cc|c} 1/\sigma_1 & 0 & \\ & \ddots & 0 \\ 0 & 1/\sigma_r & \\ \hline \multicolumn{2}{c|}{0} & 0 \end{array}\right] \begin{array}{l} \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} r \\ \}\, n-r \end{array} \end{array}$$

As we discussed before, A^+ satisfies the following Moore-Penrose conditions:

(i) $AXA = A$; (ii) $XAX = X$

(iii) $(AX)^T = AX$; (iv) $(XA)^T = XA$.

Such X is uniquely determined and $X = A^+$!!

★ Pseudoinverse & Orthogonal Projectors

Thm AA^+ is an ortho. proj. onto range(A)

and $AA^+ = U_r U_r^T$

A^+A is an ortho. proj. onto range(A^T)

and $A^+A = V_r V_r^T$

where $U_r \in \mathbb{R}^{m\times r}$, $V_r \in \mathbb{R}^{n\times r}$ consist of the first r columns of U, V, respectively.

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

(Proof) Let $P_A := AA^+$, $P_{A^T} := A^+A$.

Now, $P_A = U\Sigma V^T V \Sigma^+ U^T$

$$= U\Sigma\Sigma^+U^T = U \left[\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right] U^T$$

$$= U_r U_r^T \quad ✓$$

$$P_A^2 = U_r \underbrace{U_r^T U_r}_{= I_r} U_r^T = U_r U_r^T = P_A \quad ✓$$

so it's a proj.!

$$P_A^T = (U_r U_r^T)^T = (U_r^T)^T U_r^T = U_r U_r^T = P_A \quad ✓$$

So it's an ortho. proj.!

Finally, it's also clear that P_A maps onto range(A) since range $(A) = \langle u_1, \cdots, u_r \rangle$. ✓

You can do similarly for P_{A^T} ///

Note: Consider any $x \in \text{range}(A)$.

Then $\exists y \in \mathbb{R}^n$ s.t. $x = Ay$.

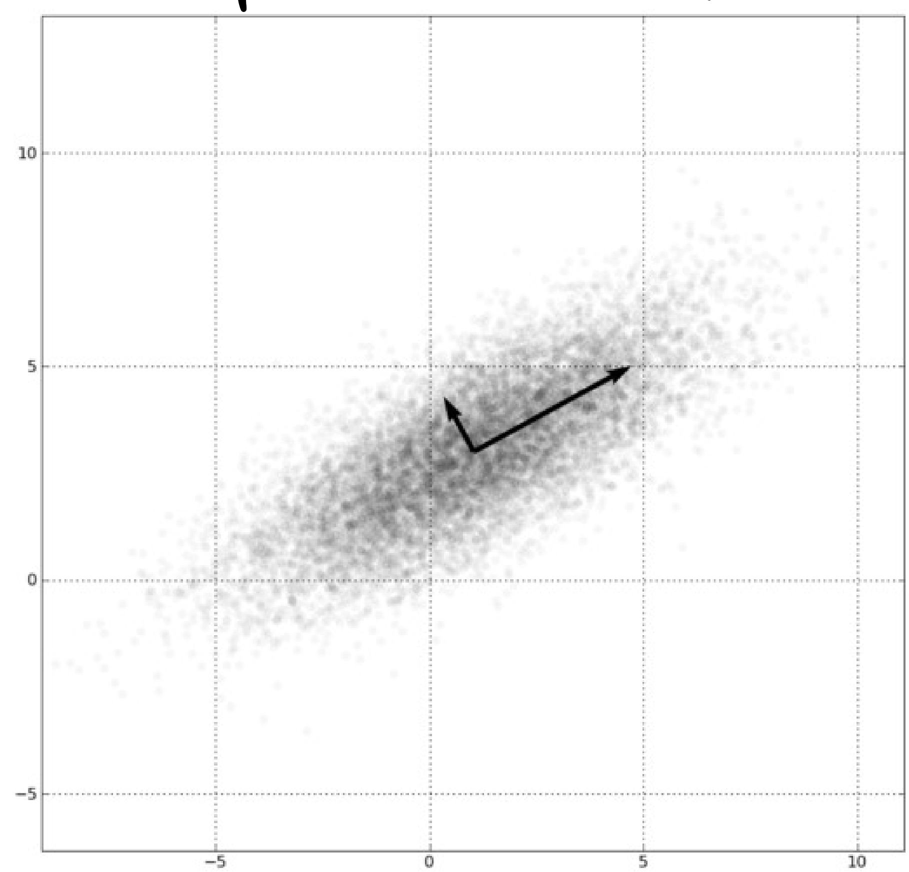
Now $P_A x = AA^+x = \underbrace{AA^+A}\, y$

$$= Ay = x.$$

"A via Moore-Penrose (i)

★ Principal Component Analysis (PCA)
(a.k.a. Karhunen-Loève Transform)
is a data analysis technique that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called "principal components."

2D example (from Wikipedia)



One can understand PCA using SVD! But before doing so, we need a bit of statistics.

often these are viewed as n realizations of some stochastic process.

Suppose we are given a set of vectors (observations)

$$\mathbb{x}_1, \mathbb{x}_2, \cdots, \mathbb{x}_n$$

and each $\mathbb{x}_j \in \mathbb{R}^d$. d : could be huge (ex. a face image database).

Let $X := [\mathbb{x}_1 \ \mathbb{x}_2 \ \cdots \ \mathbb{x}_n] \in \mathbb{R}^{d \times n}$

You know the mean (or average) of this data set

$$\bar{\mathbb{x}} := \frac{1}{n} \sum_{j=1}^{n} \mathbb{x}_j$$

And define the centered data matrix

$$\tilde{X} := [\mathbb{x}_1 - \bar{\mathbb{x}} \ \ \mathbb{x}_2 - \bar{\mathbb{x}} \ \cdots \ \mathbb{x}_n - \bar{\mathbb{x}}]$$

<u>Note</u> : $\tilde{X} = X \left(I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T \right)$

↳ Good exercise!

Now the sample covariance matrix S is defined as

$$S := \frac{1}{n} \tilde{X} \tilde{X}^T \in \mathbb{R}^{d \times d}$$

S_{ij} indicates the covariance or mutual correlation between the ith and jth entries of data vectors.

PCA is nothing but an eigenvalue decomposition of S, i.e.,

$$S = \Phi \Lambda \Phi^T, \quad \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$$

Let's sort λ_i's as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

Because $S^T = S$, and $S = \frac{1}{n} \tilde{X} \tilde{X}^T$,
we can show that $\lambda_i \geq 0$. $1 \leq i \leq d$

$\Phi = [\phi_1 \cdots \phi_d] \in \mathbb{R}^{d \times d}$
is a matrix containing the eigenvectors.
Also thanks to $S^T = S$, Φ is an orthogonal matrix whose columns form an ONB of $\mathbb{R}^d$.
The change of the bases from $[e_1 \cdots e_d]$ to $[\phi_1 \cdots \phi_d]$ is achieved simply by $\Phi^T \tilde{X}$.

$\phi_j^T \tilde{X}$ is called the j th principal components of X.

PCA was known for a long time, e.g., since the time of Pearson (1901) and Hotelling (1933).
Those days, the measurement dimension d was much smaller than the number of samples n, i.e. $d \ll n$.
This is called the "classical" setting.
Ex. 5 exam scores of 2000 students
$d = 5$, $n = 2000$.
Due to the advent of computers and sensor technology, now we often have $d \gg n$, the "neo-classical" setting.
Ex. The face database: $d = 128^2$, $n = 143$.

PCA & SVD

Note Title

Recall the centered data matrix

$\tilde{X} := [\tilde{x}_1 \cdots \tilde{x}_n] \in \mathbb{R}^{d \times n}$

$\tilde{x}_j := x_j - \bar{x}$, $\bar{x} := \frac{1}{n} \sum_{j=1}^{n} x_j$,

and the sample covariance matrix

$S := \frac{1}{n} \tilde{X} \tilde{X}^T$

Then, PCA is nothing but the eigendecomposition of S

$S = \Phi \Lambda \Phi^T$, $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$

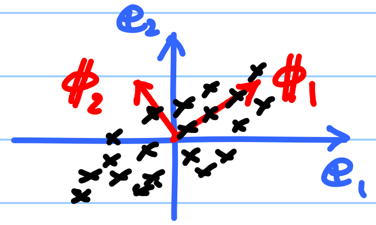
$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$.

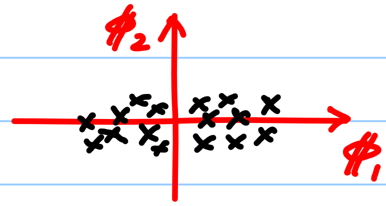
$\Phi := [\phi_1 \cdots \phi_d] \in \mathbb{R}^{d \times d}$ is an ortho. matrix, and $\{\phi_1, \cdots, \phi_d\}$ form an ONB of $\mathbb{R}^d$.

$\phi_j^T \tilde{X}$ is said to be the jth principal components of $\tilde{X}$.

These are nothing but the expansion coefficients of $\tilde{X}$ w.r.t. the ONB vector ϕ_j.

If $\tilde{X}$ forms a "cigar" shape, then $\phi_j^T \tilde{X}$ are the coordinate values of $\tilde{X}$ under the rotated axes





- Hence viewing the given dataset under the principal axes $\phi_1, \phi_2, \cdots,$ provides us better interpretations of the data than viewing them under the original axes $e_1, e_2, \cdots$.
- PCA is also often used as a tool to do dimension reduction and feature extraction by keeping only top k PCA coordinates where $k \ll d$, i.e.,

$$\Phi_k := [\phi_1 \cdots \phi_k] \in \mathbb{R}^{d \times k}$$

$$\mathbb{R}^d \ni \tilde{x}_j \longmapsto \underbrace{\Phi_k^T \tilde{x}_j}_{\text{top } k \text{ PCA coordinates or top } k \text{ Principal components of } \tilde{x}_j.} \in \mathbb{R}^k$$

Note that using these top k principal components, we can approximate the original data x_j by

$$x_j \approx \bar{x} + \Phi_k \Phi_k^T \tilde{x}_j$$

Of course the approximation gets better and better as k increases. In fact, if $k = d$, then x_j is recovered exactly (within machine ε).

Now we'll face the problem when we compute the eigendecomposition of $S = \Phi \Lambda \Phi^T$:

(1) If d is large, we cannot compute this eigendecomposition because we cannot hold $\Phi \in \mathbb{R}^{d \times d}$ in computer memory, and its computational cost is $O(d^3)$, i.e., too expensive to compute.

(2) Fortunately, we often do not need all d eigenvectors, most likely, only first k eigenvectors $k \ll d$.

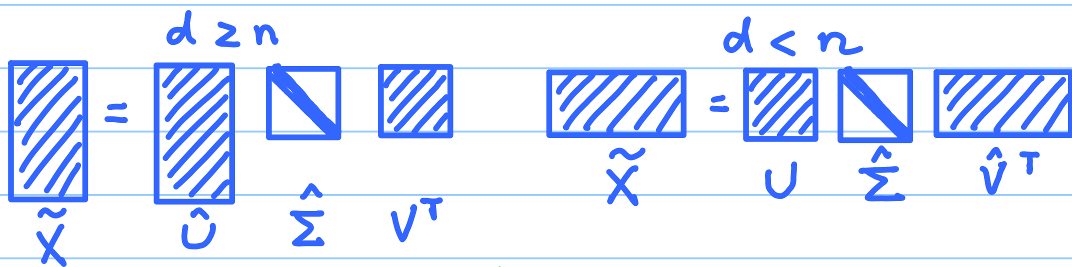
(3) Moreover if $d > n$, then $\text{rank}(S) = n-1$ if $\mathbb{x}_j$'s are linearly indep. So, after the first $n-1$ eigenvectors are useless!

Why? $S = \frac{1}{n} \tilde{X} \tilde{X}^T = \frac{1}{n} \{ \underbrace{\tilde{\mathbb{x}}_1 \tilde{\mathbb{x}}_1^T}_{\text{rank } 1} + \cdots + \underbrace{\tilde{\mathbb{x}}_n \tilde{\mathbb{x}}_n^T}_{\text{rank } 1} \}$

So looks like $\text{rank}(S) = n$.
But since $\tilde{\mathbb{x}}_1 + \cdots + \tilde{\mathbb{x}}_n = 0$ because the mean $\bar{\mathbb{x}}$ is subtracted from each data vector $\mathbb{x}_j$ (i.e., $\tilde{\mathbb{x}}_j = \mathbb{x}_j - \bar{\mathbb{x}}$)
Hence, S loses 1 rank.
So, $\text{rank}(S) = n-1$.

Now, let's consider the reduced SVD of $\tilde{X}$:

$$\tilde{X} = \hat{U} \hat{\Sigma} V^T$$



Just consider the "neo-classical" setting, i.e., $d \geq n$ (e.g., the face image database)

Then consider the sample covariance matrix S using the above SVD:

$$S = \frac{1}{n} \tilde{X} \tilde{X}^T = \frac{1}{n} \hat{U} \hat{\Sigma} \underline{V^T V} \hat{\Sigma}^T \hat{U}^T$$

$$= \frac{1}{n} \hat{U} \hat{\Sigma} \hat{\Sigma}^T \hat{U}^T = \frac{1}{n} \hat{U} \hat{\Sigma}^2 \hat{U}^T$$

Now $\hat{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_{n-1}, \underline{\underline{0}})$ if $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are linearly indep.

So, $\hat{\Sigma}^2 = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_{n-1}^2, 0)$.

Finally, S can be written as

$$S = \underline{\hat{U}} \underbrace{\left(\frac{1}{n} \hat{\Sigma}^2\right)}_{= \mathrm{diag}(\sigma_1^2/n, \cdots, \sigma_{n-1}^2/n, 0)} \underline{\hat{U}^T}$$

columns are orthonormal.

Comparing this with the eigendecomposition

$S = \Phi \Lambda \Phi^T$, we can conclude that

$$\begin{cases} \Phi(:, 1:n) = \hat{U} \\ \Lambda(1:n, 1:n) = \frac{1}{n} \hat{\Sigma}^2 = \mathrm{diag}(\sigma_1^2/n, \cdots, \sigma_{n-1}^2/n, 0) \end{cases}$$

In fact, only the $1:n-1$ portion is useful since $\sigma_n = 0$.

Hence, we should use the reduced SVD of $\tilde{X}$ (not S) for computing PCA!!
Do not use the eigendecomposition of S unless d is small.

<u>Note</u>: $\tilde{X} V = \hat{U} \hat{\Sigma} V^T V = \hat{U} \hat{\Sigma}$

$$= [\sigma_1 u_1, \cdots \sigma_{n-1} u_{n-1}, 0]$$

$$= [\tilde{X} v_1, \cdots, \tilde{X} v_n]$$

So, $u_j = \frac{1}{\sigma_j} \tilde{X} v_j$, $j = 1, \cdots, n-1$.

In other words, each principal axis u_j is just a linear combination of the (centered) input vectors $\tilde{x}_1, \cdots, \tilde{x}_n$!

Now let's do MATLAB experiments using the face image database consisting of 143 faces each of which has $128 \times 128 = 16384$ pixels,
i.e., $d = 16384$, $n = 143$.