EECS 16B
Spring 2022
Lecture 23
4/12/2022

LECTURE 23

- complete min. norm solution
- low rank approximation
- Principal Component Analysis (PCA)

Last lecture: $A\vec{x}=\vec{y}$, $A \in \mathbb{R}^{m \times n}$

$\vec{x}=A^{\dagger}\vec{y}$, where $A^{\dagger}$ was defined using SVD of A, is:

- unique solution of $A\vec{x}=\vec{y}$ when A is square and invertible ($m=n=r$) because $A^{\dagger}=A^{-1}$ in that case
- LS solution when A is tall ($m>n$). If full column rank also ($n=r$), then
$$A^{\dagger}=(A^TA)^{-1}A^T,$$
which recovers LS solution studied before.
- Min. norm solution when A is wide ($n>m$) and infinitely many solutions exist. If full row rank ($m=r$),
$$A^{\dagger}=A^T(AA^T)^{-1}.$$

Thus,

$$\boxed{\vec{x}_{MN}=A^T(AA^T)^{-1}\vec{y}.}$$

Example: Controllability

$$\underbrace{\vec{x}_{target}-A^{\ell}\vec{x}[0]}_{\vec{y}} = \underbrace{[A^{\ell-1}B, \ldots, AB, B]}_{C_\ell} \begin{bmatrix} u[0] \\ \vdots \\ u[\ell-1] \end{bmatrix}$$

Solution exists when system $\ell \geq$ state dimension by controllability.

From boxed eq'n above for Min Norm Solution:

$$\begin{bmatrix} u[0] \\ \vdots \\ u[l-1] \end{bmatrix}_{MN} = C_l^T (C_l C_l^T)^{-1} (\vec{x}_{target} - A^l \vec{x}[0]).$$

For the vehicle control example (Lecture 19):

$$A = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \quad B = \frac{\Delta}{RM} \begin{bmatrix} \frac{1}{2}\Delta \\ 1 \end{bmatrix}$$

$$AB = \frac{\Delta}{RM} \begin{bmatrix} \frac{3}{2}\Delta \\ 1 \end{bmatrix}, \quad A^2B = A \cdot AB = \frac{\Delta}{RM} \begin{bmatrix} \frac{5}{2}\Delta \\ 1 \end{bmatrix}$$

$$\vdots$$

$$A^{l-1}B = \frac{\Delta}{RM} \begin{bmatrix} (l-\frac{1}{2})\Delta \\ 1 \end{bmatrix}$$

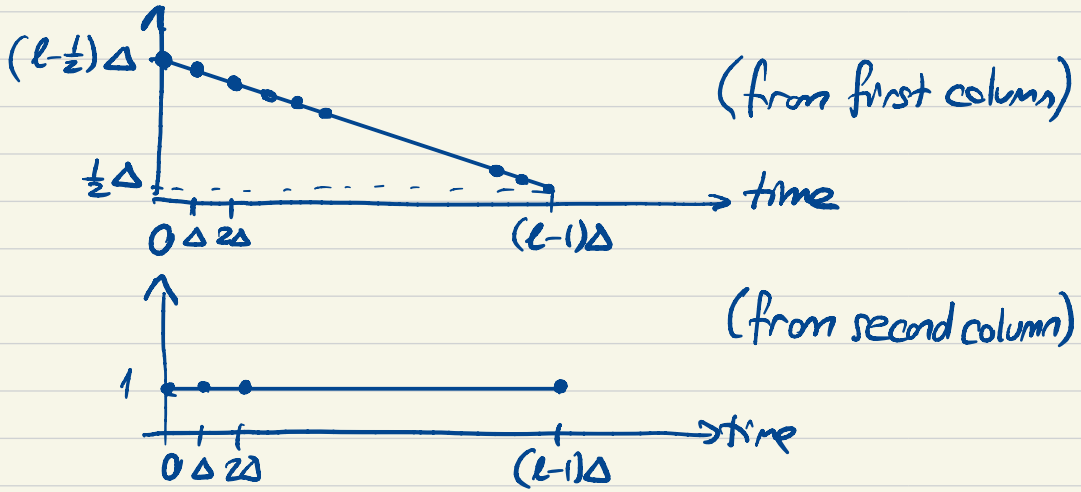
$$C_l = \frac{\Delta}{RM} \begin{bmatrix} (l-\frac{1}{2})\Delta & \cdots & \frac{5}{2}\Delta & \frac{3}{2}\Delta & \frac{1}{2}\Delta \\ 1 & & 1 & 1 & 1 \end{bmatrix}$$

$\vec{x}[0] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\vec{x}_{target} = \begin{bmatrix} 1000 \\ 0 \end{bmatrix}$, $\Delta = 0.1\,s$, $RM = 5000\,kgm$

$$\begin{bmatrix} u[0] \\ \vdots \\ u[l-1] \end{bmatrix}_{MN} = \underbrace{\frac{\Delta}{RM}}_{2 \cdot 10^{-5}} \underbrace{\begin{bmatrix} (l-\frac{1}{2})\Delta & 1 \\ \vdots & \vdots \\ \frac{3}{2}\Delta & 1 \\ \frac{1}{2}\Delta & 1 \end{bmatrix}}_{C_l^T} \underbrace{(C_l C_l^T)^{-1} \begin{bmatrix} 1000 \\ 0 \end{bmatrix}}_{\begin{bmatrix} * \\ * \end{bmatrix}}$$

Thus, min. norm solution is a linear combination of the columns of C_ℓ^T (rows of C_ℓ).

Therefore, min. norm control sequence is a linear combination of two sequences:



Solution shown in Lecture 19 is indeed a weighted sum of these two.

Low-rank Approximation

Given a high rank matrix $A \in \mathbb{R}^{m \times n}$ with

$$r \approx \min\{m, n\},$$

find an approximation with rank $\ell \ll \min\{m, n\}$.

SVD suggests the following heuristic:

$$A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T = \underbrace{\sum_{i=1}^{\ell} \sigma_i \vec{u}_i \vec{v}_i^T}_{=: A_\ell,\ \text{serves as rank-}\ell\ \text{approximation}} + \underbrace{\sum_{i=\ell+1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T}_{\text{discard, bc } \sigma_{\ell+1}, \ldots \sigma_r \text{ smaller than } \sigma_1, \ldots \sigma_\ell}$$

Note:

- When $\ell \ll \min\{m, n\}$, A_ℓ has far fewer entries to store than A.

Ex: $m = n = 1000$, $\ell = 10 \Rightarrow$ A has 10^6 entries
u_i, v_i of dimension 1000 each, so 2000 total per outer product. $\ell = 10$ terms: $20{,}000 \ll 10^6$.

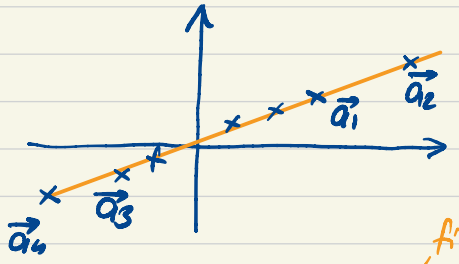
- In many data sets a few singular values are dominant and the rest are much smaller. If A_ℓ captures dominant ones, $A - A_\ell$ is small.

- Eckart-Young Theorem (Note 17) states that the SVD truncation above is more than a heuristic: A_ℓ gives the least possible deviation from A that is possible with a rank-ℓ matrix. More precisely,

A_ℓ above solves: $\min_{B \in \mathbb{R}^{m \times n}} \|A - B\|_F$ ← Frobenius norm
such that $\text{rank}(B) = \ell$.

Principal Component Analysis (PCA)

Suppose A has $m=2$ rows and many more columns ($n \gg 2$). If $r=1$, what does a scatter plot of columns of A look like?



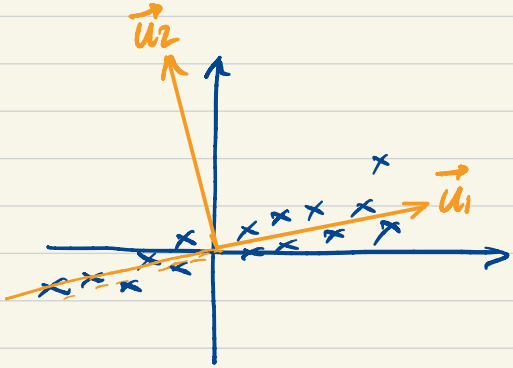
$A = \sigma_1 \vec{u}_1 \vec{v}_1^T$ $\quad \vec{a}_1 = \sigma_1 v_1(1) \vec{u}_1$, $\vec{a}_2 = \sigma_1 v_1(2) \vec{u}_1$

(first entry of $\vec{v}_1$; second entry of $\vec{v}_1$)

What if $r=2$, but $\sigma_1 \gg \sigma_2$?

$A = \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T$

Column i of A, $\vec{a}_i = \sigma_1 v_1(i) \vec{u}_1 + \sigma_2 v_2(i) \vec{u}_2$



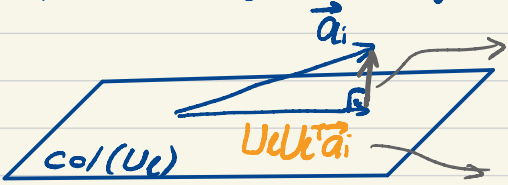
$v_1(i)$ and $v_2(i)$ are generally in the same range since $\|\vec{v}_1\| = \|\vec{v}_2\| = 1$, but since $\sigma_1 \gg \sigma_2$ we expect $\sigma_1 v_1(i)$ to be bigger than $\sigma_2 v_2(i)$ for most columns $i \in \{1, 2, \ldots, n\}$. Therefore, most columns will be slanted towards $\vec{u}_1$.

Generalizing this example, for a matrix A with ℓ dominant singular values, we expect the scatter plot of columns to cluster around the subspace spanned by:

$$\vec{u}_1, \ldots \vec{u}_\ell$$

or, equivalently, the column space of $U_\ell := [\vec{u}_1 \cdots \vec{u}_\ell]$.

We can find the closeness of the ith column $\vec{a}_i$ to this subspace using the projection:



$U_\ell U_\ell^T$ is a projection matrix since U_ℓ has orthonormal columns (see last lecture)

Thus, distance of $\vec{a}_i$ to Col(U_ℓ):

$$\| \vec{a}_i - U_\ell U_\ell^T \vec{a}_i \|.$$

An aggregate measure of closeness of all columns $\vec{a}_1, \ldots \vec{a}_n$ can be obtained from sum of squared distances:

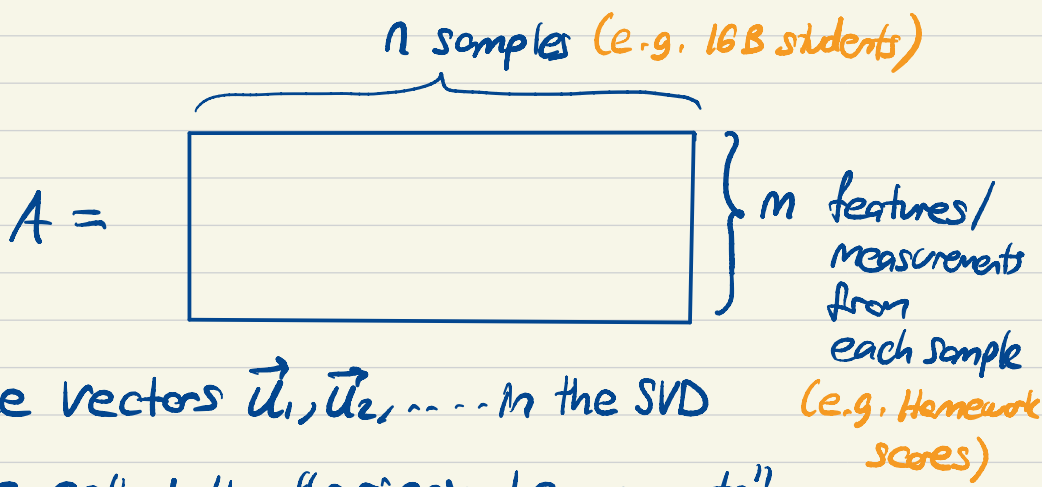
$$\sum_{i=1}^{n} \| \vec{a}_i - U_\ell U_\ell^T \vec{a}_i \|^2$$

Theorem 4 in Note 17 says the subspace spanned by $\vec{u}_1, \ldots \vec{u}_\ell$ minimizes the sum of squared distances among all ℓ-dimensional subspaces: if we choose another matrix W with ℓ orthonormal columns:

$$\sum_{i=1}^{n} \| \vec{a}_i - W W^T \vec{a}_i \|^2 \geq \sum_{i=1}^{n} \| \vec{a}_i - U_\ell U_\ell^T \vec{a}_i \|^2.$$

closeness to another subspace spanned by columns of W

PCA: When A is a collection of data



the vectors $\vec{u}_1, \vec{u}_2, \ldots$ in the SVD are called the "principal components."

Thus, PCA/SVD shows dominant directions in data sets. The vectors $\vec{u}_1, \vec{u}_2 \ldots$ corresponding to dominant singular values give a lower dimensional representation of the data.

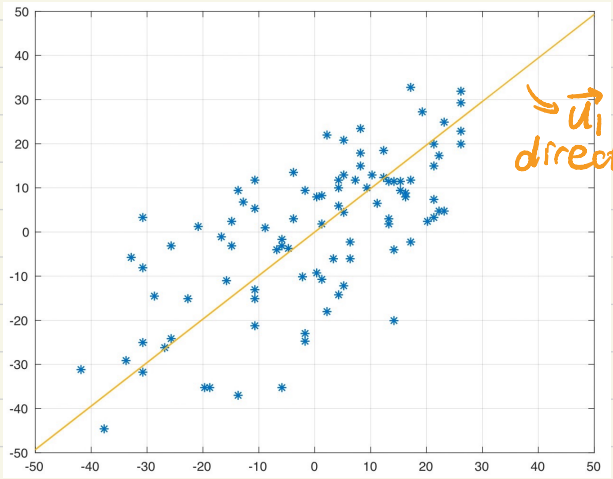
Recall: $\vec{u}_1, \vec{u}_2, \ldots$ are evectors of the $m \times m$ matrix AA^T.

In statistics $\frac{1}{n-1} AA^T$ is called the covariance matrix and the principal components are obtained from its evectors. It is assumed that the

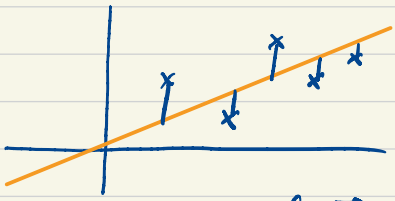
average of each row is subtracted from that row so $\sum_{j=1}^{n} a_{ij} = 0$, $i = 1, \ldots m$.

This ensures that the center of the column vectors is 0, i.e. the data is centered around the origin.

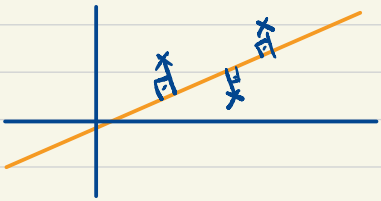
Ex: 2×94 matrix A containing two midterm scores of 94 students. Each column is a student and each row a midterm. The average of each midterm subtracted from corresponding row to center the data at the origin:



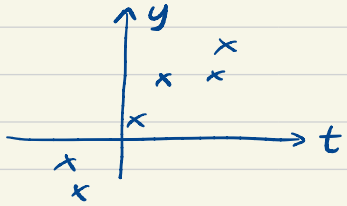
Fitting a subspace to represent data points resembles Least Squares (LS), but in LS fit we minimize the squared sum of <u>vertical</u> distances:



whereas the subspace of $\vec{u}_1, \vec{u}_2, \ldots$ discussed above minimizes squared sum of <u>perpendicular</u> distances:



<u>Example</u>: Suppose we have data points $(t_1, y_1), (t_2, y_2), \ldots (t_n, y_n)$ and want to fit a line



$y = \alpha t$. We can choose α as the LS solution to $\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\vec{y}} = \underbrace{\begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}}_{\vec{t}} \alpha \Rightarrow \alpha_{LS} = (\vec{t}^T\vec{t})^{-1}\vec{t}^T\vec{y} = \frac{\vec{t}^T\vec{y}}{\vec{t}^T\vec{t}}$.

Alternatively we can form the data matrix

$$A = \begin{bmatrix} t_1 & t_2 & \cdots & t_n \\ y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} \vec{t}^T \\ \vec{y}^T \end{bmatrix}$$

whose columns are the data points and find $\vec{u}_1$ from SVD which gives a direction that closely fits data. In this case $\vec{u}_1$ is an evector of $AA^T = \begin{bmatrix} \vec{t}^T \\ \vec{y}^T \end{bmatrix} [\vec{t}\ \vec{y}] = \begin{bmatrix} \vec{t}^T\vec{t} & \vec{t}^T\vec{y} \\ \vec{y}^T\vec{t} & \vec{y}^T\vec{y} \end{bmatrix}$, corresponding to its largest evalue.

Take for example $\vec{t} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$, $\vec{y} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$. Then

$$\alpha_{LS} = \frac{\vec{t}^T\vec{y}}{\vec{t}^T\vec{t}} = \frac{3}{2}$$

. On the other hand

$AA^T = \begin{bmatrix} 2 & 3 \\ 3 & 6 \end{bmatrix}$ and you can show $\vec{u}_1 = \begin{bmatrix} 0.4719 \\ 0.8817 \end{bmatrix}$.

$$\text{slope} = \frac{0.8817}{0.4719}$$

$$= 1.8685$$

$$\neq \alpha_{LS}$$