MATH 312
LECTURE 19

: Principal Components Analysis.

Summary of SVD and pseudoinverse: A $m\times n$, $\operatorname{rank}(A)=r$

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

$$\underset{\substack{m\times m\\ \text{orthogonal}\\ \text{matrix}}}{U} = \underbrace{\left[\underbrace{\vec{u}_1 \dots \vec{u}_r}_{C(A)} \ \underbrace{\vec{u}_{r+1} \dots \vec{u}_m}_{N(A^T)}\right]}_{\substack{\text{orthonormal eigenvectors}\\ \text{of } AA^T}} \qquad V = \underbrace{\left[\underbrace{\vec{v}_1 \dots \vec{v}_r}_{C(A^T)} \ \underbrace{\vec{v}_{r+1} \dots \vec{v}_n}_{N(A)}\right]}_{\substack{\text{orthonormal eigenvectors}\\ \text{of } A^TA}} \ \substack{n\times n\\ \text{orthogonal}\\ \text{matrix}}$$

The order is important: $\vec{u}_i, \vec{v}_i$ follow the order of the singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$.

- Reduced SVD: $\quad U_r = \left[\vec{u}_1 \dots \vec{u}_r\right] \quad V_r = \left[\vec{v}_1 \dots \vec{v}_r\right]$

$A = U_r \Sigma_r V_r^T$, where $\quad \Sigma_r = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$ $r\times r$ diagonal (recall that Σ is $m\times n$)

or

$A = \sigma_1 \vec{u}_1 \vec{v}_1^T + \dots + \sigma_r \vec{u}_r \vec{v}_r^T$.

• Pseudoinverse: $A^+ = V\Sigma^+ U^T$

1) A^+A is the projection matrix onto $C(A^T)$

$$A^+A = V\Sigma^+ U^T U\Sigma V^T = V\Sigma^+\Sigma V^T = \vec{v}_1\vec{v}_1^T + \dots + \vec{v}_r\vec{v}_r^T$$

(recall that the $\vec{v}$'s are orthonormal)

2) AA^+ is the projection matrix onto $C(A)$

$$AA^+ = U\Sigma V^T V\Sigma^+ U^T = U\Sigma\Sigma^+ U^T = \vec{u}_1\vec{u}_1^T + \dots + \vec{u}_r\vec{u}_r^T$$

(the $\vec{u}$'s are orthonormal)

Therefore:

1) If a system has infinite solutions, i.e.

$A\vec{x} = \vec{b} \rightsquigarrow \vec{x} = \vec{x}_p + \vec{x}_n$ with $\vec{x}_p$ any particular solution (not necessarily in row space), $\vec{x}_n \in N(A)$

the shortest length solution (the only one living in the row space) is the projection of $\vec{x}$ onto the row space:

$$\vec{x}_r = A^+A\vec{x} = A^+A\vec{x}_p + \underbrace{A^+\cancel{A}\vec{x}_n}_{\vec{0}} = A^+\vec{b}.$$

(2) If $A\vec{x} = \vec{b}$ no sol. $\rightarrow$ Solve $A\vec{x} = P_{C(A)}\vec{b} = AA^+\vec{b} \Rightarrow \vec{x} = A^+AA^+\vec{b} = A^+\vec{b}$)
(shortest length least squares sol.

Example: Find the shortest length solution,

$$\left.\begin{aligned} x + y + z &= 1 \\ y - z &= 1 \end{aligned}\right\}$$

Sol:

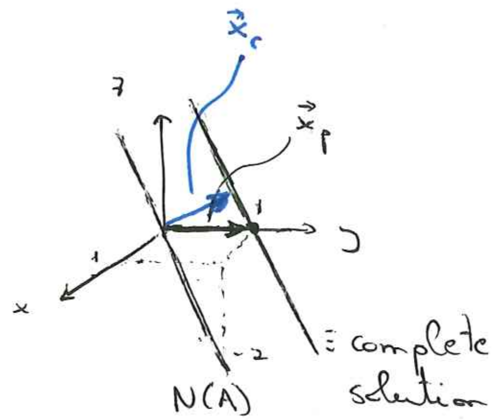
Let's do it as if we didn't know about the pseudoinverse.

$A\vec{x} = \vec{b}$ with $A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}$, $\vec{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$,

$$\left[\begin{array}{ccc|c} \boxed{1} & 1 & 1 & 1 \\ 0 & \boxed{1} & -1 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} \boxed{1} & 0 & 2 & 0 \\ 0 & \boxed{1} & \underset{\uparrow \text{ free}}{-1} & 1 \end{array}\right] \qquad \begin{aligned} x_3 &= \alpha \\ x_2 &= 1 + \alpha \\ x_1 &= -2\alpha \end{aligned}$$

That is,

$$\vec{x} = \underbrace{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}}_{\vec{x}_p} + \alpha \underbrace{\begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}}_{\text{basis for } N(A)} \quad \overset{\text{geometrically}}{\rightsquigarrow}$$



We can see in the picture solutions different than $\vec{x}_p$ with shorter length. Indeed,

$\vec{x}_p = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ is not in the row space, as it is not orthogonal to the nullspace.

If we find the projection matrix onto $C(A^T)$ then the shortest length solution (which we denote $\vec{x}_r$ because it is in the row space) is

$$\vec{x}_r = P_{C(A^T)}(\vec{x}_p + \vec{x}_n) = P_{C(A^T)}\vec{x}_p + \underbrace{\cancel{P_{C(A^T)}\vec{x}_n}}_{\vec{0}\ \text{(as } C(A^T) \perp N(A))}.$$

Notice that we are lucky this time: the rows of A are orthogonal, so

$$P_{C(A^T)} = \frac{\vec{a}_1\vec{a}_1^T}{\|\vec{a}_1\|^2} + \frac{\vec{a}_2\vec{a}_2^T}{\|\vec{a}_2\|^2}, \text{ and so,}$$

$$\vec{x}_r = \vec{a}_1 \frac{1}{3}\vec{a}_1^T\vec{x}_p + \vec{a}_2\frac{1}{2}\vec{a}_2^T\vec{x}_p = \frac{1}{3}\begin{bmatrix}1\\1\\1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0\\1\\-1\end{bmatrix} = \begin{bmatrix}1/3\\5/6\\-1/6\end{bmatrix}$$

$$\left(\text{Check: } \begin{bmatrix}1/3 & 5/6 & -1/6\end{bmatrix}\begin{bmatrix}-2\\1\\1\end{bmatrix} = 0\right).$$

- Let's do it now using the pseudoinverse:

$\vec{x}_r = A^+\vec{b}$, where

$A^+ = V\Sigma^+ U^T \longrightarrow$ we need U, Σ, V

1) $\underline{\Sigma}$: A is $2\times3 \rightarrow AA^T$ 2×2

$$AA^T = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$ (A has orthogonal rows)

$$\Downarrow$$

$$\lambda_1 = 3,\ \lambda_2 = 2 \rightarrow \sigma_1 = \sqrt{3},\ \sigma_2 = \sqrt{2}$$

$$\Sigma = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix}$$

2) $\underline{\underline{V}}$

$$A^TA = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}$$

$$A^TA - 3I = \begin{bmatrix} -2 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} -2 & 1 & 1 \\ 0 & -1/2 & 1/2 \\ 0 & 1/2 & -1/2 \end{bmatrix} \rightarrow \vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\frac{1}{\sqrt{3}}$$

$$A^TA - 2I = \begin{bmatrix} -1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} -1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \rightarrow \vec{v}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}\frac{1}{\sqrt{2}}$$

$$A^TA \rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix} \rightarrow \vec{v}_3 = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}\frac{1}{\sqrt{6}}$$

$$V = \begin{bmatrix} 1/\sqrt{3} & 0 & -2/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \end{bmatrix}$$

3) $\underline{\underline{U}}$

$$\vec{u}_1 = \frac{A\vec{v}_1}{\sigma_1} = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\frac{1}{\sqrt{3}} = \frac{1}{3}\begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$\vec{u}_2 = \frac{A\vec{v}_2}{\sigma_2} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}\frac{1}{\sqrt{2}} = \frac{1}{2}\begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Check: $U\begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix}\begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ -2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}$ //

Thus,

$$A^+ = \begin{bmatrix} 1/\sqrt{3} & 0 & -2/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \end{bmatrix}\begin{bmatrix} 1/\sqrt{3} & 0 \\ 0 & 1/\sqrt{2} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 1/3 & 0 \\ 1/3 & 1/2 \\ 1/3 & -1/2 \end{bmatrix}$$

$$AA^+ = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\substack{\text{projection onto} \\ C(A) = \mathbb{R}^2}}, \quad A^+A = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 5/6 & -1/6 \\ 1/3 & -1/6 & 5/6 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 2 & 2 & 2 \\ 2 & 5 & -1 \\ 2 & -1 & 5 \end{bmatrix}$$

Finally, $\vec{x}_r = \begin{bmatrix} 1/3 & 0 \\ 1/3 & 1/2 \\ 1/3 & -1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 5/6 \\ -1/6 \end{bmatrix}$.

III.) Principal Components Analysis.

Now we will have a bunch of data points, which we write in a matrix:

$$X = \begin{bmatrix} \vec{x}_1 & \cdots & \vec{x}_n \\ & & \end{bmatrix}, \quad \vec{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{im} \end{bmatrix} \in \mathbb{R}^m$$

→ Each row corresponds to a variable.

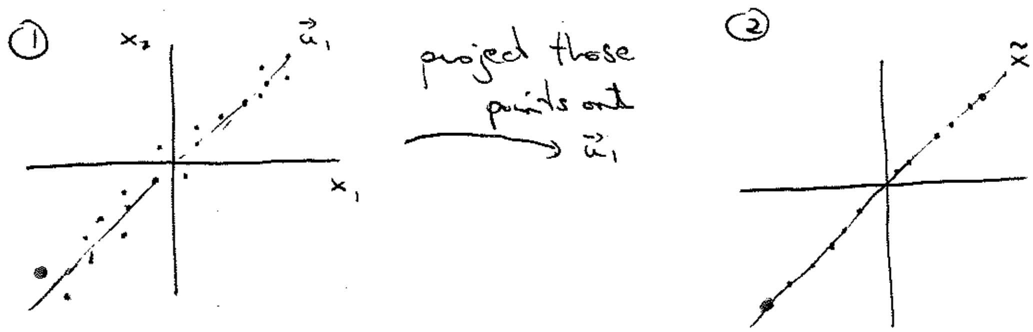
(High dimensional data)

points, samples, measurements

Examples: The cols. can be all the students at Penn, while the rows might be their height, age, major, ...

- The goal is to be able to distinguish or classify the points with much less than m variables, without losing much information.

Imagine we have points in $\mathbb{R}^2$ $(m=2)$ that look as follows:



In ①, each point is characterized by two variables (x_1, x_2).

In ②, we lose some information (red and green points cannot be distinguish in ②), however, most of them are still split apart.

→ Most variance among data remains in ②.

→ In ②, we only need one variable $(\tilde{x})$ = the distance along $\vec{u}_1$.

↳ each original point $\vec{x}$ (cols. of X) are now identified by their projection onto $\vec{u}_1$.

- What is the relation with SVD?

Write $X = \left[\vec{x}_1 \ \ldots \ \vec{x}_n \right] = \sigma_1 \vec{u}_1 \vec{v}_1^T + \ldots + \sigma_r \vec{u}_r \vec{v}_r^T$.

Notice that: $A\vec{v}_i = \sigma_i \vec{u}_i \Rightarrow \underbrace{A^T A \vec{v}_i}_{= \sigma_i^2 \vec{v}_i} = \sigma_i A^T \vec{u}_i \Rightarrow \left\| \cancel{A} \, \sigma_i \vec{v}_i^T = \vec{u}_i^T A \right\|$

so $X = \vec{u}_1(\vec{u}_1^T X) + \dots + \vec{u}_r(\vec{u}_r^T X)$, where

$$\vec{u}_i(\vec{u}_i^T X) = \left[\begin{array}{c|c|c} \vec{u}_i\vec{u}_i^T\vec{x}_1 & \dots & \vec{u}_i\vec{u}_i^T\vec{x}_n \end{array}\right]$$

projection of the data points $\vec{x}_1, \dots, \vec{x}_n$ onto $\vec{u}_i$.

In our 2d example,

$$X = \underbrace{\vec{u}_1(\vec{u}_1^T X)}_{} + \vec{u}_2(\vec{u}_2^T X).$$

this gives picture ② (each column gives where the original points go onto the line spanned by $\vec{u}_1$).

<u>Remark</u>: $\vec{u}_1$ gives the line ~~of best~~ through the origin of best fit.

$\vec{u}_1, \vec{u}_2$ gives the plane through the origin of best fit. (*)

⋮

→ The $\vec{u}$'s are called principal components. They provide a new basis (new variables) to represent our data.

(*) We have to substract the mean to each row to obtain good results.

Example: Database of 216 patients, 121 with ovarian cancer and 95 without it.

Each patient identified through 4000 genes.

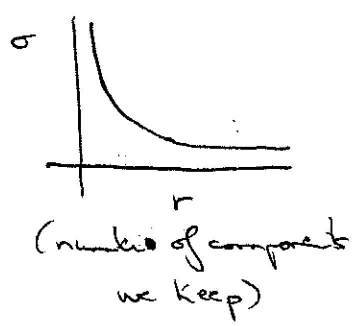
$$\vec{X} = \begin{bmatrix} \vec{x}_1 & \cdots & \vec{x}_{216} \end{bmatrix} \Big\} \text{ 4000 rows}$$

We want to see if there are some common factors that predict the ovarian cancer.

It seems likely that many of the genes are correlated.

We perform the SVD ⇝ we might obtain a lot of $\vec{u}$'s. (up to 4000!)

How many are important? ⇝ We can look the weights σ^2. (graph)



Let's plot the patients in the basis given by $\{\vec{u}_1, \vec{u}_2, \vec{u}_3\}$, that, we project each $\vec{x}_i$ into the subspace spanned by ↑

vector of 4000 components

$$\vec{\tilde{x}}_i = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \vec{u}_1(\vec{u}_1^T \vec{x}_i) + \vec{u}_2(\vec{u}_2^T \vec{x}_i) + \vec{u}_3(\vec{u}_3^T \vec{x}_i) \Rightarrow$$

$$\Rightarrow \vec{\tilde{x}}_i \Big|_{\{\vec{u}_1, \vec{u}_2, \vec{u}_3\}} = \begin{bmatrix} \vec{u}_1^T \vec{x}_i \\ \vec{u}_2^T \vec{x}_i \\ \vec{u}_3^T \vec{x}_i \end{bmatrix}$$

This can be plotted in a 3d graph!

(see Matlab ...)

<u>Example:</u> Eigenfaces

Database 1600 pictures → 25 people × 64 pictures/person.

Each picture is 192·168 pixels.
$= 32256$

$$X = \begin{bmatrix} \vec{x}_1 & \dots & \vec{x}_{1600} \\ & & \\ & & \end{bmatrix} \Big\} \text{ 32256 rows (pixels)}$$

each image is stored as a column vector.

· We do the SVD of X − mean of rows.

↳ The $\vec{u}$'s are called "eigenfaces" → they keep most common features in human faces.

↳ They provide a basis for the space of "faces".

A new face ("test face") can be projected into the eigenspace given by the $\vec{u}$'s (building using X, the "training set").

↳ Now we can see if it is close to one of the faces in X or not → face recognition.

$$\text{testface} = \begin{bmatrix} \\ \\ \\ \end{bmatrix} \rightsquigarrow \underbrace{U_r U_r^T \,\text{testface}}_{\substack{\text{approximation of testface} \\ \text{using first } r \text{ eigenfaces.}}} \qquad \left(U_r = \begin{bmatrix} \vec{u}_1 & \dots & \vec{u}_r \end{bmatrix} \right.$$