

# ECE 532 - lecture 12 - more SVD and PCA

①

$$\text{Recall } A = [U_1 U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 \Sigma_1 V_1^T$$

orthogonal      |      orthogonal  
 diagonal  $\Sigma_1$   
 and positive.

$$\text{we also saw } \|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sigma_1.$$

which  $x$  achieves this maximum? It's  $v_1$  (first right singular vec.)

$$Av_1 = U_1 \Sigma_1 V_1^T v_1 = U_1 \Sigma_1 e_1 = U_1 \sigma_1 e_1 = \sigma_1 u_1$$

$$\text{so } \frac{\|Av_1\|}{\|v_1\|} = \frac{\sigma_1 \|u_1\|}{\|v_1\|} = \sigma_1 \quad (\text{since } u_1, v_1 \text{ are normalized}).$$

In general, we have:

$A v_i = \sigma_i u_i \quad \text{and} \quad A^T u_i = \sigma_i v_i \quad \forall i.$

This is related to notion of eigenvalues:

$$A^T A v_i = A^T (\sigma_i u_i) = \sigma_i^2 v_i \Rightarrow \sigma_i^2 \text{ is eigenvalue of } A^T A \text{ with eigenvector } v_i$$

$$A A^T u_i = A (\sigma_i v_i) = \sigma_i^2 u_i \Rightarrow \sigma_i^2 \text{ is eigenvalue of } A A^T \text{ with eigenvector } u_i$$

(this is one way to compute SVD).

(2)

Directly, we have:  $A^T A = (U \Sigma V^T)^T (U \Sigma V^T)$

$$= V \underbrace{\Sigma^T U^T U \Sigma V^T}_{I}$$

eigenvalue  
decomposition!  $\Rightarrow = V \underbrace{\Sigma^T \Sigma}_{\Sigma} V^T$

$$\hookrightarrow (\Sigma_1 \ 0)^\top (\Sigma_1 \ 0) = (\sigma_1 \ \dots \ \sigma_r \ 0).$$

similarly,  $A A^T = U \Sigma^T \Sigma U^T$ .

★ if  $Q$  is symmetric and positive semidefinite,  $\lambda_i = \sigma_i$   
(eigenvalues are the same as singular values).

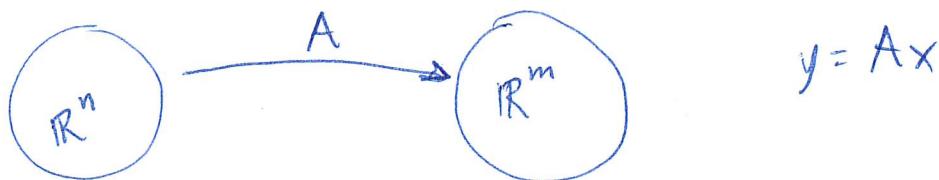
★ in general, a square matrix  $B$

→ might have complex eigenvalues

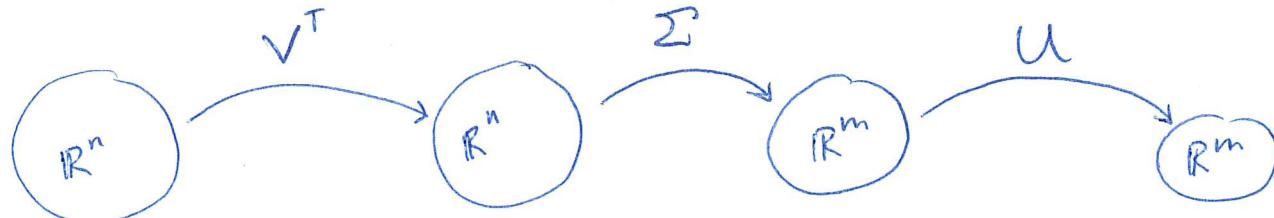
→ might not have orthogonal eigenvectors ( $P D P^{-1}$  instead of  $P D P^T$ )  
→ might not be diagonalizable at all! (Jordan form)  $P D P^T$ )

whereas every matrix (even non-square) has an SVD  
and real positive singular values.

Geometric interpretation of SVD,

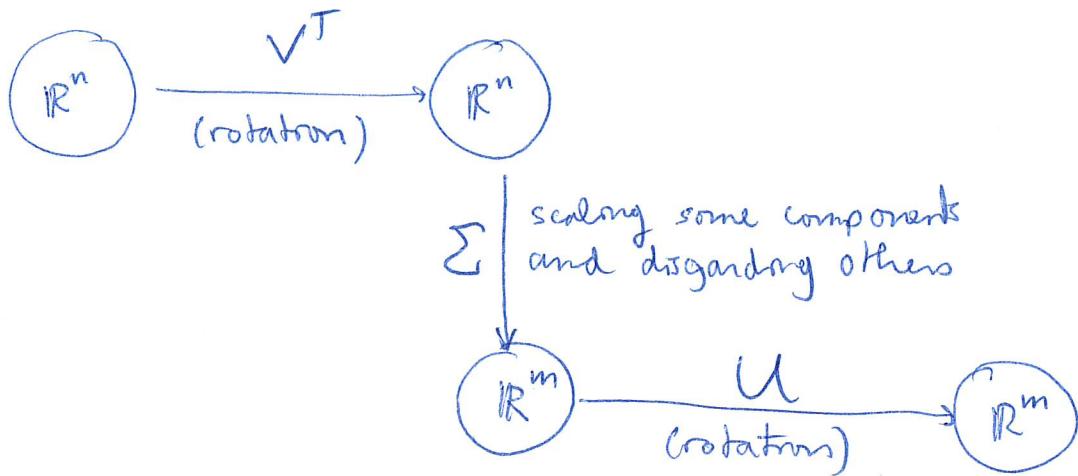
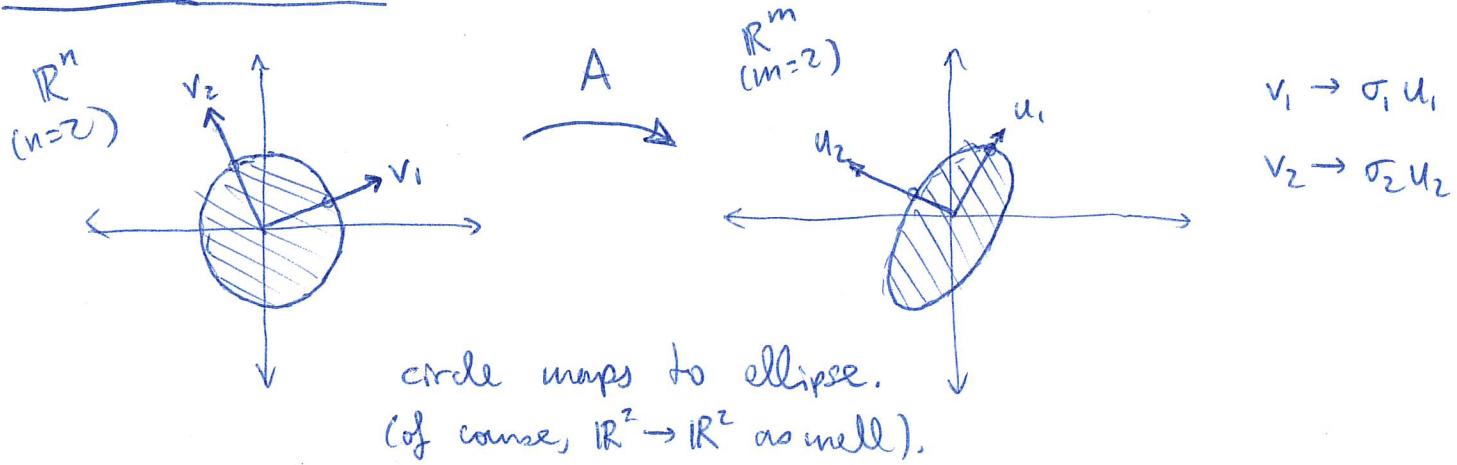
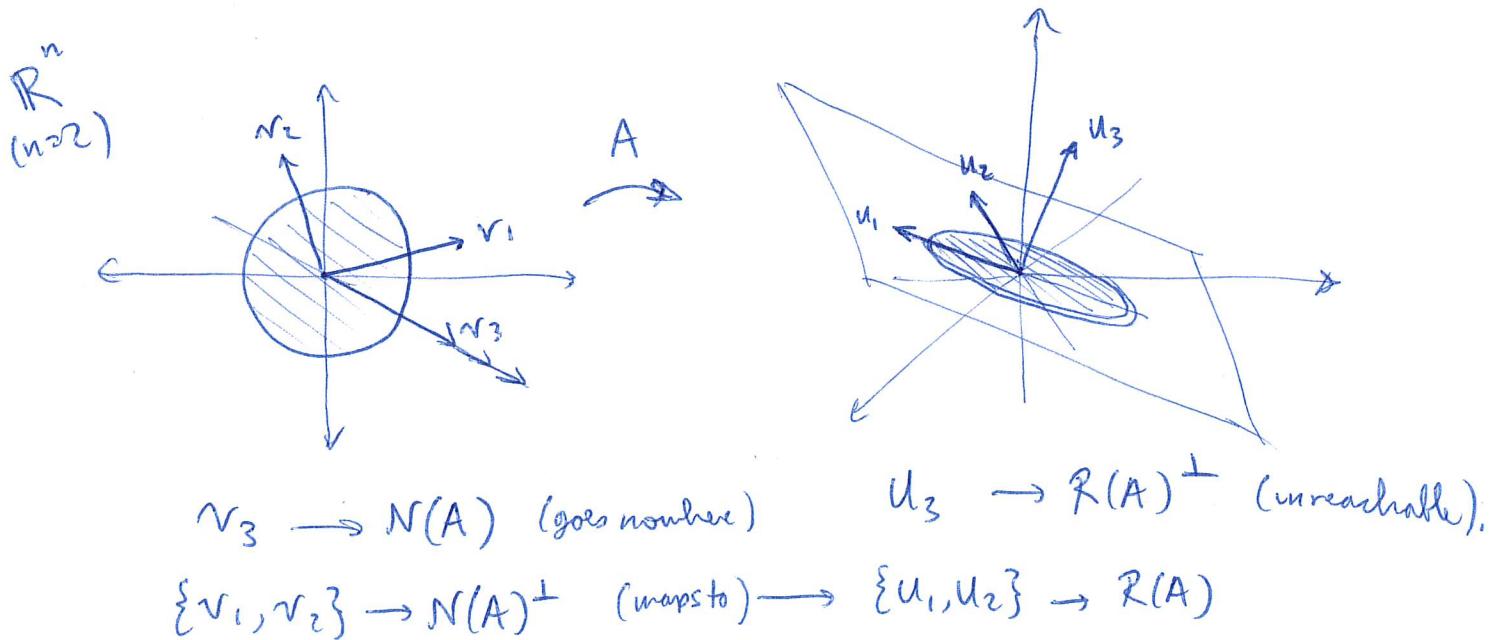


instead;



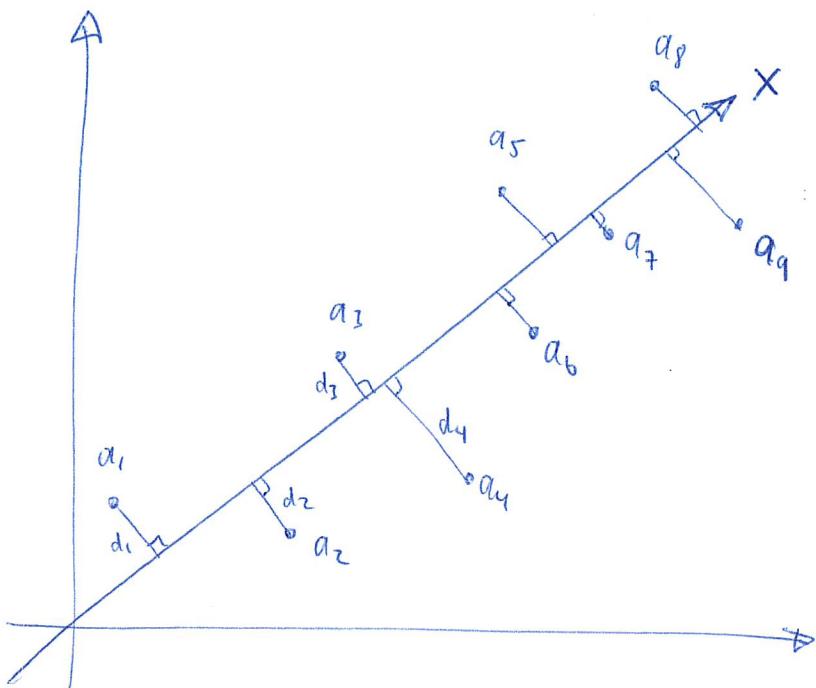
$$y = U \Sigma V^T x$$

(3)

in 2D  $\rightarrow$  2Din 2D  $\rightarrow$  3D

(4)

## Finding the closest line to a set of points. Example



$$a_i \in \mathbb{R}^d \quad i=1, \dots, n$$

are points and we want to minimize

$$\sum_{i=1}^n d_i^2$$

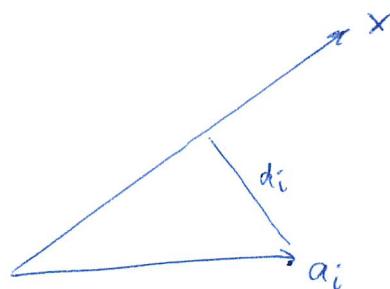
(sum of squares of distances to the line).

★ This is different from regression! Here, our line is coordinate-independent; rotating the points also rotates the line. In regression, this is not the case.

★ why not  $\sum_{i=1}^n d_i$  (sum of distances) or  $\max_i d_i$  (max dist?)

it's a choice! we will see these other types of distance measurements later in class.

in this scenario,  $d_i^2 = \|a_i - \text{proj}_x a_i\|^2$



(5)

$$\text{recall } \text{proj}_x a = \frac{x^T a}{x^T x} x$$

Note: we can interchange matrix-vector-scalar multiplication!

$$\text{e.g. } \underbrace{x^T a}_{\substack{\text{scalar} \\ \text{vector}}} x = \underbrace{xx^T a}_{\substack{\text{scalar} \\ \text{vector}}} = \underbrace{xx^T a}_{\substack{\text{matrix} \\ \text{vector}}} \quad (*)$$

$$\text{also, } (a^T x)^2 = \underbrace{(a^T x)(a^T x)}_{\text{scalar-scalar multiplication.}} = a^T (x x^T) a = \underbrace{x^T (a a^T) x}_{\text{quadratic forms.}}$$

$$\begin{aligned}
 d_i^2 &= \|a_i - \text{proj}_x a_i\|^2 \\
 &= \|a_i - \frac{x^T a_i}{x^T x} x\|^2 \quad \text{use trick (*)} \\
 &= \|a_i - \frac{1}{x^T x} (x x^T) a_i\|^2 \quad \text{factor } a_i \text{ from the right.} \\
 &= \left\| \left( I - \frac{1}{x^T x} (x x^T) \right) a_i \right\|^2 \quad \text{use fact that } \|v\|^2 = v^T v. \\
 &= a_i^T \left( I - \frac{1}{x^T x} (x x^T) \right) \left( I - \frac{1}{x^T x} (x x^T) \right) a_i \\
 &= a_i^T \left( I - \frac{1}{x^T x} (x x^T) - \frac{1}{x^T x} (x x^T) + \underbrace{\frac{1}{(x^T x)^2} x x^T x x^T}_{\text{scalar.}} \right) a_i \\
 &= a_i^T \left( I - \frac{1}{x^T x} x x^T \right) a_i
 \end{aligned}$$

(6)

(cont'd).

$$\begin{aligned}
 d_i^2 &= a_i^T (I - \frac{1}{x^T x} x x^T) a_i \\
 &= a_i^T a_i - \frac{1}{x^T x} a_i^T x x^T a_i \quad \text{using trick on} \\
 &\quad \text{prior page.} \\
 &= a_i^T a_i - \frac{1}{x^T x} x^T (a_i a_i^T) x.
 \end{aligned}$$

Now minimizing  $d_i^2$  is equivalent to maximizing  $\frac{1}{x^T x} x^T (a_i a_i^T) x$  because of the negative sign and because  $a_i^T a_i$  is constant.

$$\begin{aligned}
 &\max_{x \neq 0} \sum_{i=1}^n \frac{1}{x^T x} x^T (a_i a_i^T) x \\
 &= \max_{x \neq 0} \frac{1}{x^T x} x^T \left( \sum_{i=1}^n a_i a_i^T \right) x \\
 &= \max_{x \neq 0} \frac{x^T A^T A x}{x^T x} \\
 &= \max_{x \neq 0} \frac{\|A x\|^2}{\|x\|^2} \quad X_{opt} = v_1, \text{ the first} \\
 &\quad \text{right singular vector!} \\
 &= \sigma_1^2
 \end{aligned}$$

If  $A = \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$

then  $A^T A = \sum_{i=1}^n a_i a_i^T$

so the best line is  $x = v_1$  where  $v_1$  is first col. of  $V$   
 in the SVD  $A = U \Sigma V^T$  and  $A = \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix}$  is the data matrix

In general, the best approximation to a set of points by a  $k$ -dimensional subspace can be found in a similar fashion. Here, instead of choosing a direction  $x$ , we choose an orthonormal basis  $\{w_1, w_2, \dots, w_k\}$ . Replace  $\text{proj}_x a_i$  with  $\text{proj}_W a_i$  where  $W = [w_1, w_2, \dots, w_k]$ .

$$\begin{aligned} \sum_{i=1}^n d_i^2 &= \sum_{i=1}^n \|a_i - \text{proj}_W a_i\|^2 \\ &= \sum_{i=1}^n \|a_i - Ww^T a_i\|^2 \\ &= \sum_{i=1}^n a_i^T (I - Ww^T)^2 a_i \\ &= \sum_{i=1}^n a_i^T (I - Ww^T) a_i \\ &= \left( \sum_{i=1}^n a_i^T a_i \right) - \sum_{i=1}^n a_i^T Ww^T a_i \\ &= \left( \sum_{i=1}^n a_i^T a_i \right) - \text{trace} \left[ W^T \underbrace{\left( \sum_{i=1}^n a_i a_i^T \right)}_{A^T A} W \right]. \end{aligned}$$

This is called PCA ("principal component analysis")

$$\begin{aligned} x^T x &= \text{trace}(xx^T) \\ \text{so } (Wa)^T (Wa) &= \text{tr}(Wa a^T W^T) \end{aligned}$$

$$\text{tr}(x^T x) = \|x\|_F^2$$

so minimizing  $\sum_{i=1}^n d_i^2$  is equivalent to maximizing  $\text{trace}(W^T A^T A W)$

$\Rightarrow$  maximize  $\|AW\|_F^2$  if we let  $A = U, \Sigma, V^T$   
 $W \in \mathbb{R}^{n \times k}$ ,  
orthogonal

$$\|AW\|_F^2 = \|U, \Sigma, V^T W\|_F^2 = \|\Sigma, V^T W\|_F^2$$

maximized when  $W = [v_1, v_2, \dots, v_k]$ .

i.e.  $\text{Span}\{v_1, \dots, v_k\}$  is best  $k$ -dimensional subspace.  
these are called the principal components.