L15: Principal Component Analysis

+ SVD

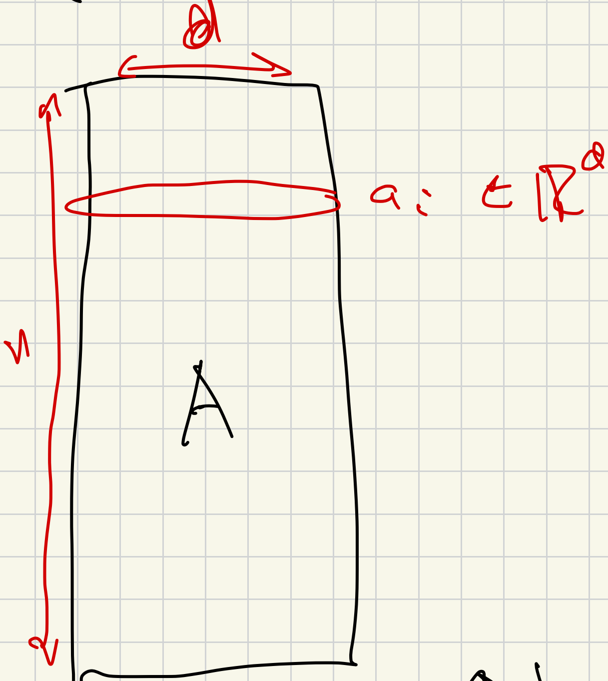
10.20.2025

Data Mining

Jeff M. Phillips

Input $A \in \mathbb{R}^{n \times d}$

$A = \{a_1, a_2, \dots a_n\} \subset \mathbb{R}^d$



dimension d is too big

Dimensionality Reduction

- each column is same units, comparable

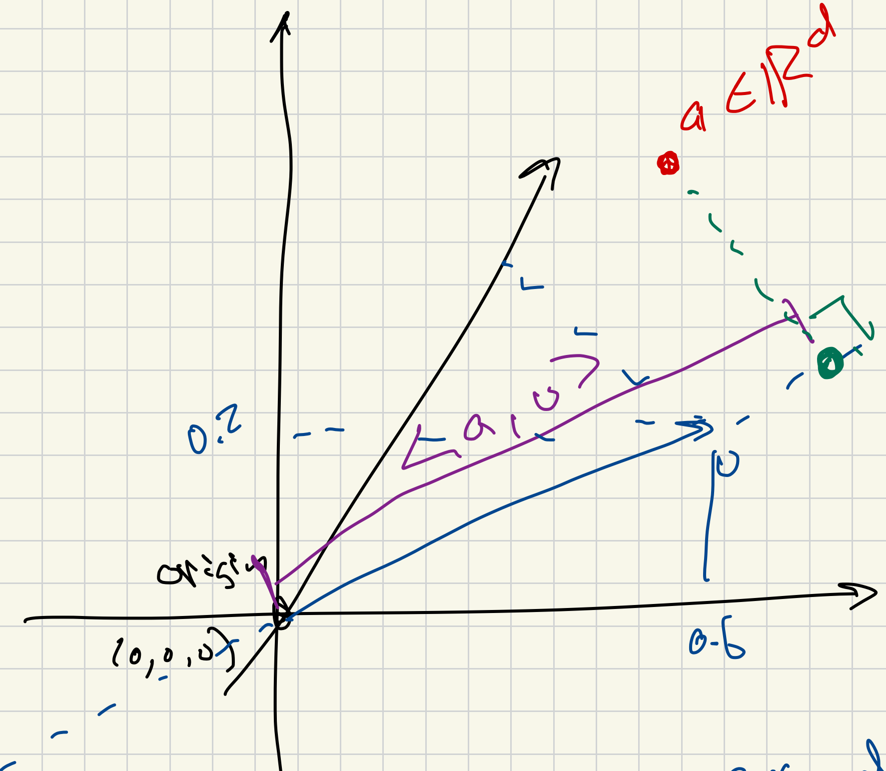
$$\|a_i - p\| = \sqrt{\sum_{j=1}^{d} (a_{ij} - p_j)^2}$$

Output: Mapping $\pi : \mathbb{R}^d \to \mathbb{R}^k$, $k \ll d$

$\pi(a_i) \in \mathbb{R}^d$

$\bar{\pi}(a_i) \in \mathbb{R}^k$

$$\langle a, u\rangle = \sum_{j=1}^{d} a_j u_j$$



$$\pi_u(a) = \underbrace{\langle a, u\rangle}_{\text{how far}} \cdot \underbrace{u}_{\text{where}} \in \mathbb{R}^d$$

u unit vector

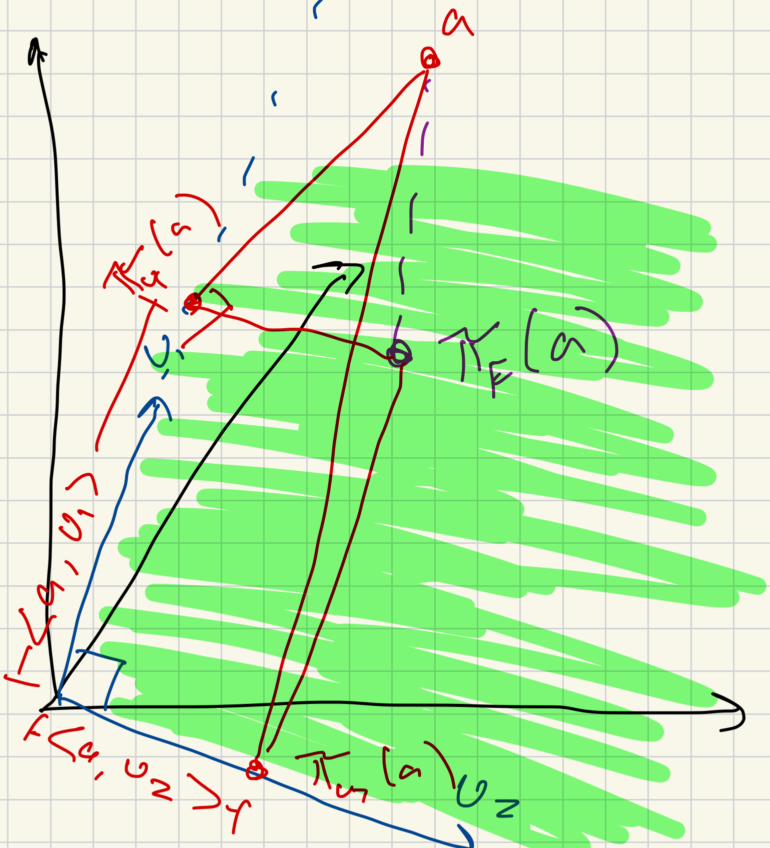
$u \in \mathbb{R}^d$

$\|u\| = 1$

ex $d = 3$

$u = (0.6, 0.2, 0.5)$

$$\pi_F = (\alpha_1, \alpha_2) \in \mathbb{R}^2$$
$$= (\langle v_1, a\rangle, \langle v_2, a\rangle)$$



$F = \begin{cases} v_1 & \text{unit vector} \in \mathbb{R}^d \\ v_2 & \text{unit vector} \in \mathbb{R}^d \end{cases}$

$\langle v_1, v_2\rangle = 0$

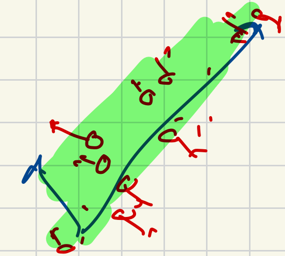
↳ orthogonal

$F = \{v_1, v_2, \dots v_k\}$

$\langle v_i, v_j\rangle = 0$

$$\pi_F(a) = \underbrace{\langle a, v_1\rangle}_{\alpha_1} v_1 + \underbrace{\langle a, v_2\rangle}_{\alpha_2} v_2$$
$$= \sum_{j=1}^{k} \underbrace{\langle a, v_j\rangle}_{\alpha_j} v_j$$

Sum of Squared Error   SSE



$$SSE(A, F) = \sum_{a_i \in A} \| a_i - \pi_F(a) \|^2$$

Goal: Find k-dim subspace $F = \{v_1, v_2, \dots v_k\}$

$\|v_j\| = 1$

$$F^* = \underset{F}{\arg\min}\; SSE(A, F)$$

$\langle v_j, v_{j'} \rangle = 0$

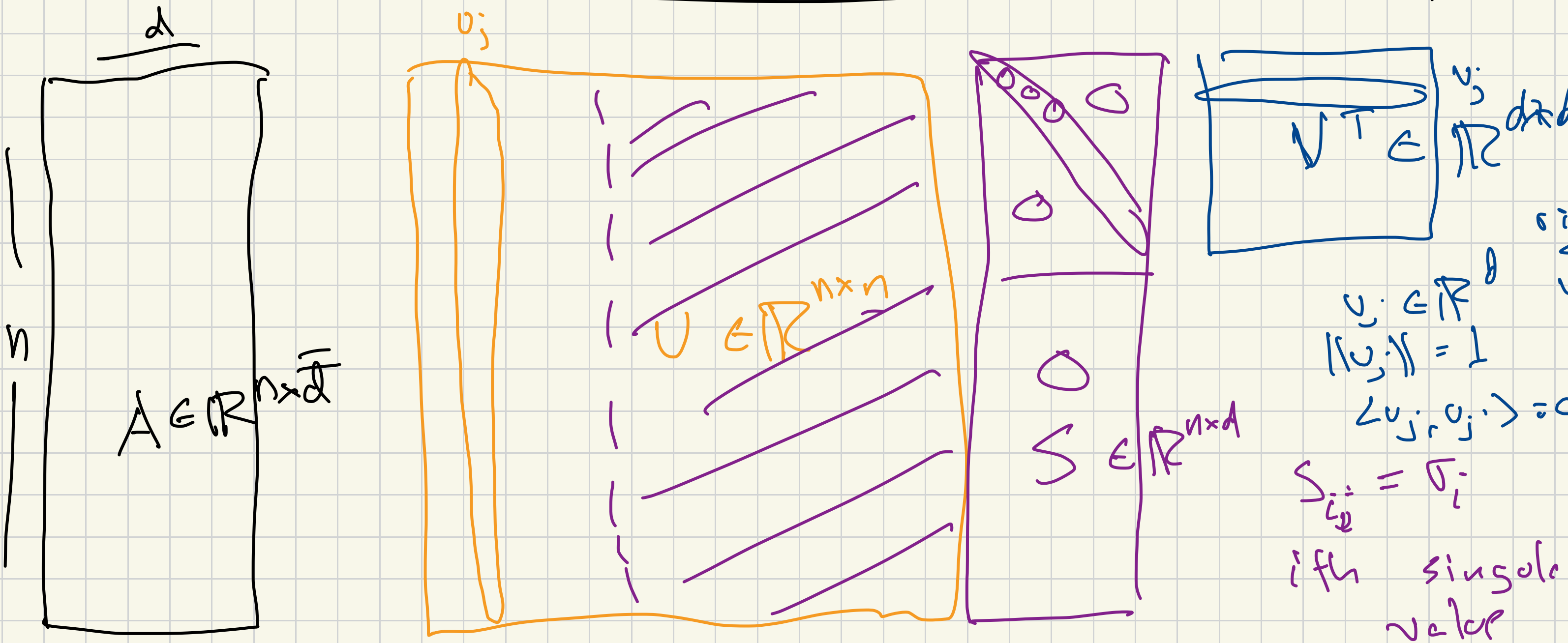
↳ solve   Singular Value Decomposition

SVD

Not goal for Principal Component Analysis

PCA

Singular Value Decomposition (SVD)



right sing. vectors

$v_j \in \mathbb{R}^d$

$\|v_j\| = 1$

$\langle v_j, v_{j'} \rangle = 0$

$S_{ii} = \sigma_i$

ith singular value

$$A = USV^T$$

left sing. vectors

$\|u_j\| = 1$

$\langle u_j, u_{j'} \rangle = 0$

$\sigma_i \geq \sigma_{i+1} \cdots \geq 0$

$Ax = (USV^T)x$

$A = USV^T = \sum_{j=1}^{d} \underbrace{u_j \sigma_j v_j^T}_{\sigma_j \underbrace{u_j v_j^T}_{\in \mathbb{R}^{n \times d}}}$

Low Rank k - approximation of A

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

$A_k \in \mathbb{R}^{n \times d} = \sum_{j=1}^{k} \sigma_j u_j v_j^T$

$F = \{v_1, v_2, \dots, v_k\}$

$$\sum_i \| a_i - \Pi_F(a_i) \|^2$$

$F = \{v_1, v_2 \ldots v_k\}$ top k Right sing vect

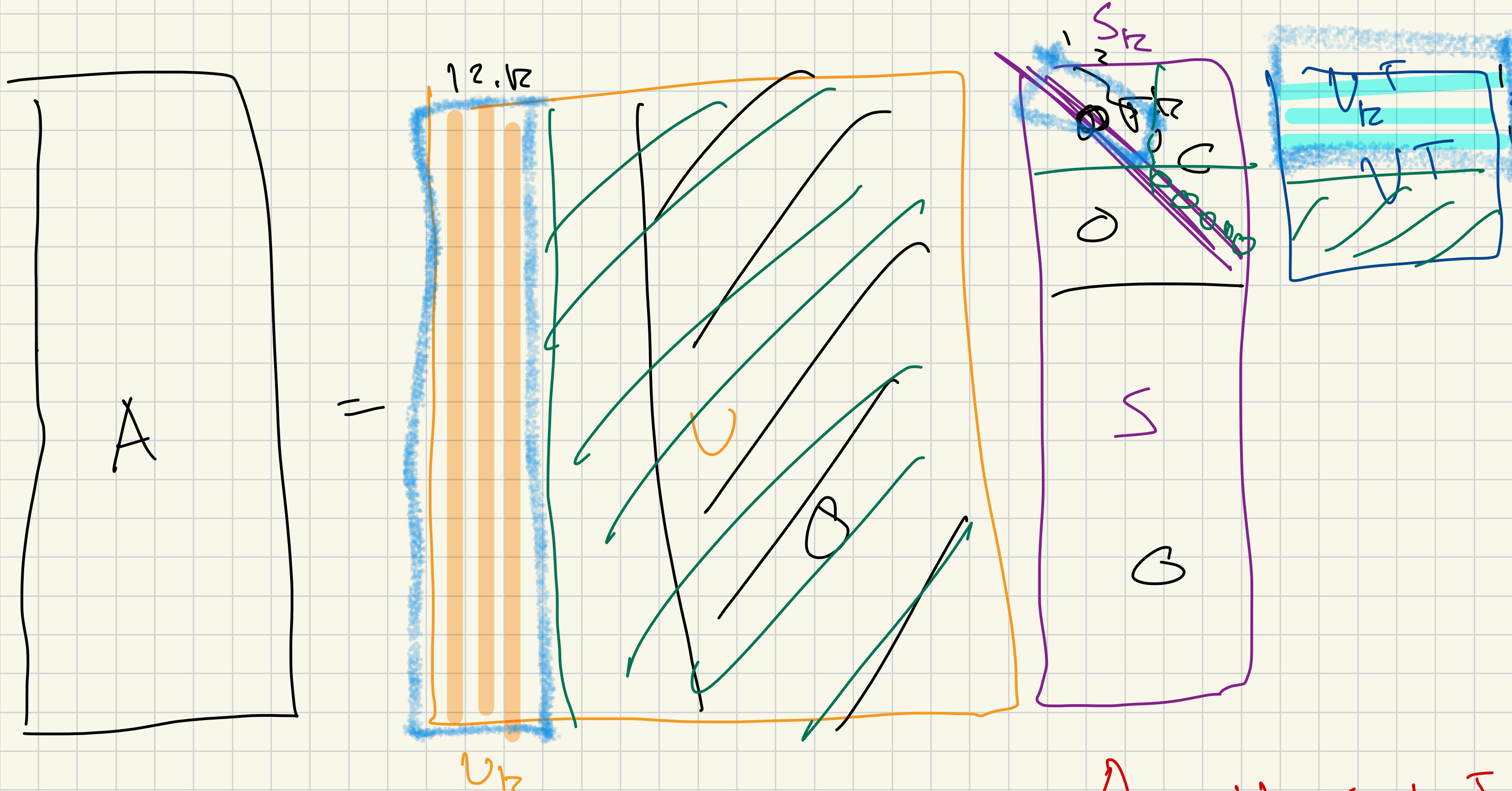
$$= \sum_i \Big\| \sum_{j=1}^{d} v_j \langle v_j, a_i \rangle - \sum_{j=1}^{k} v_j \langle v_j, a_i \rangle \Big\|^2$$

$$= \sum_i \Big\| \sum_{j=k+1}^{d} v_j \langle v_j, a_i \rangle \Big\|^2$$

Pythagorian

$$= \sum_i \sum_{j=k+1}^{d} \| v_j \langle v_j, a_i \rangle \|^2$$

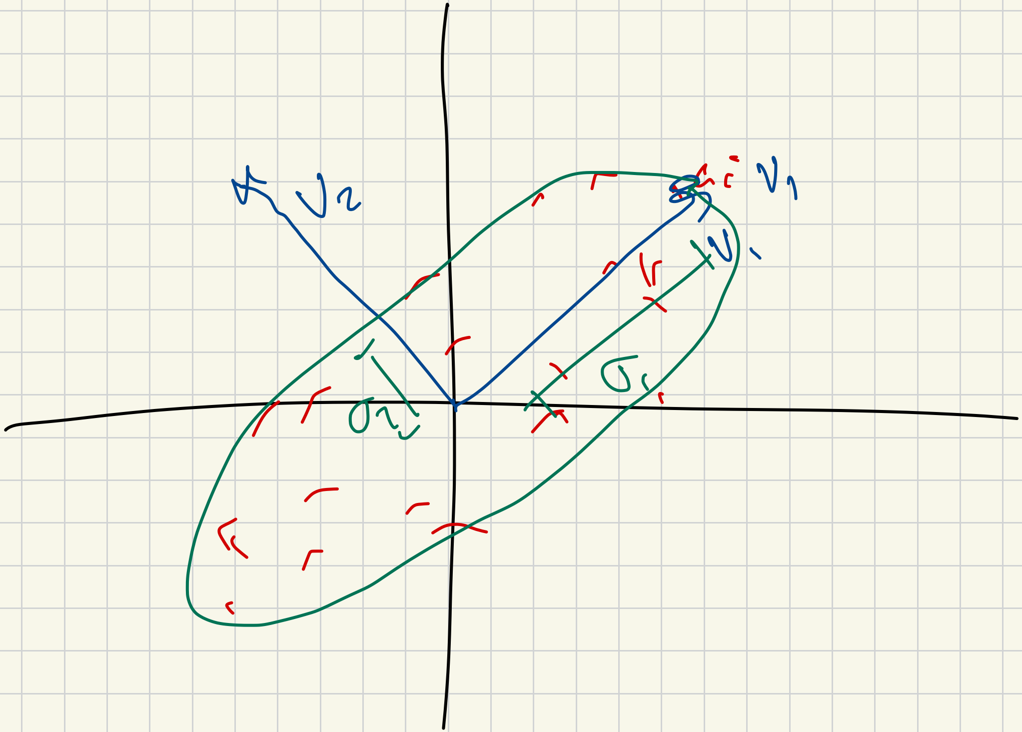
$$= \sum_{j=k+1}^{d} \boxed{\sum_{i=1} \langle v_j, a_i \rangle^2} = \sum_{j=k+1}^{d} \| A v_j \|^2 = \sum_{j=k+1}^{d} \sigma_j^2$$



$A_k = U_k S_k V_k^T$

best rank-k approx

equiv $S_{jj} = 0$ $j > k$

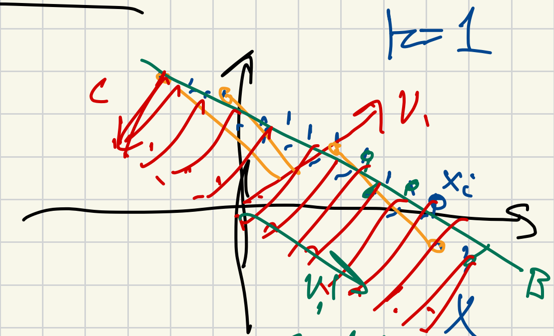


v_2
v_1
v_1
σ_1
σ_2

Principal Component Analysis PCA

$X = \{x_1, x_2 \dots x_n\} \subset \mathbb{R}^d$

$X \in \mathbb{R}^{n \times d}$



Goal: Find k-dim subspace B

minimize $\sum_{i=1}^{n} \| x_i - \pi_B(x_i) \|^2$

$svd(A)$

$\Downarrow$

V_1

PCA

1. center
2. SVD
3. return $v_1, \dots v_k$ $\longrightarrow$ principal vectors

Center $a_i = x_i - c$ $\quad c \in \mathbb{R}^d$

$c = \frac{1}{n} \sum_{i=1}^{n} x_i \leftarrow$ mean

$A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$