FoDA L21

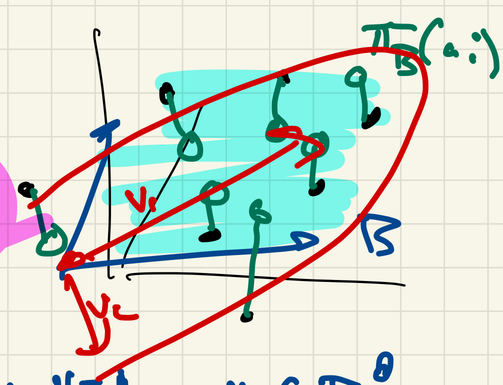
Principal Component Analysis (PCA) & MultiDimensional Scaling (MDS)

Nov 8, 2022

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

Goal $\underset{B \text{ rank } k}{\arg\min} \sum_{i=1}^{n} \| a_i - \pi_B(a_i) \|^2$

B contains 0

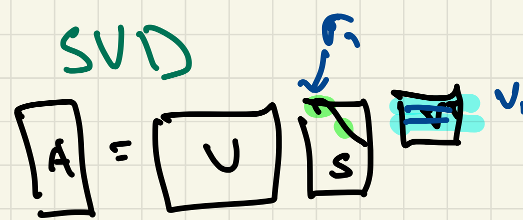


$B: \quad V_B = \{v_1, v_2 \dots v_k\} \subset \mathbb{R}^d \qquad \|v_i\| = 1 \qquad v_i \in \mathbb{R}^d$

$\langle v_j, v_i \rangle = 0$

Solve for V_B through SVD

$A = U S V^T$



how much info

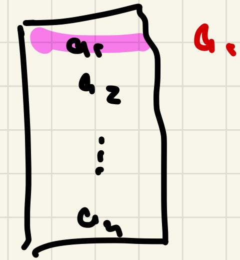
$\sigma_1 \quad v_1$

$\sigma_2 \quad v_2$

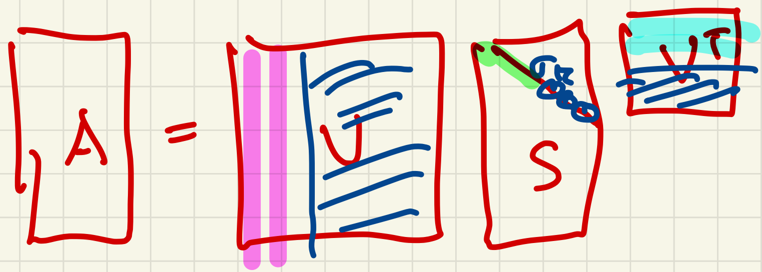
$\sigma_3 \quad v_3$

orthogonal basis

$A = \{a_1 \ \dots \ a_n\} \in \mathbb{R}^d$

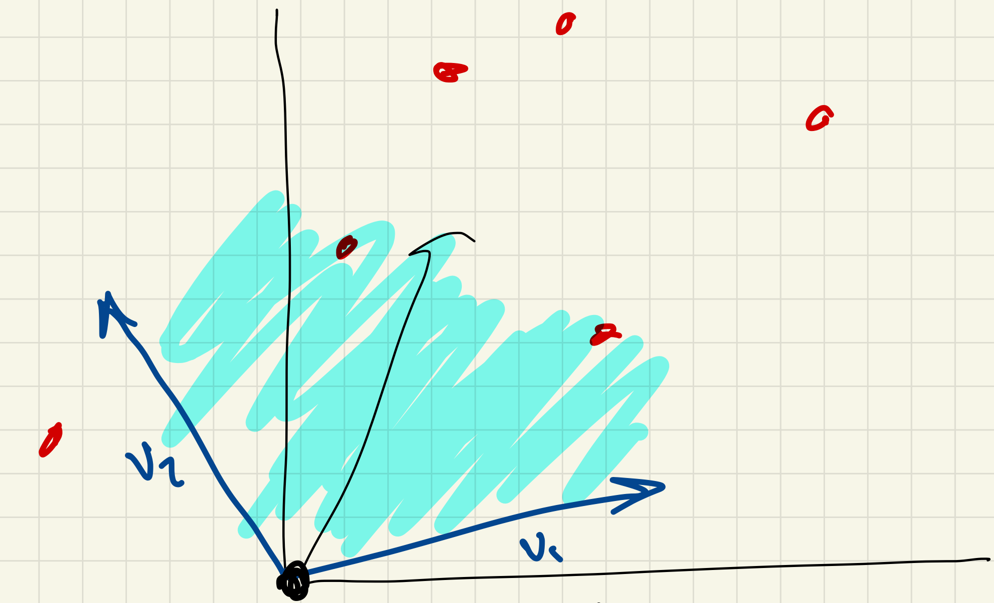


$A \in \mathbb{R}^{n \times d}$

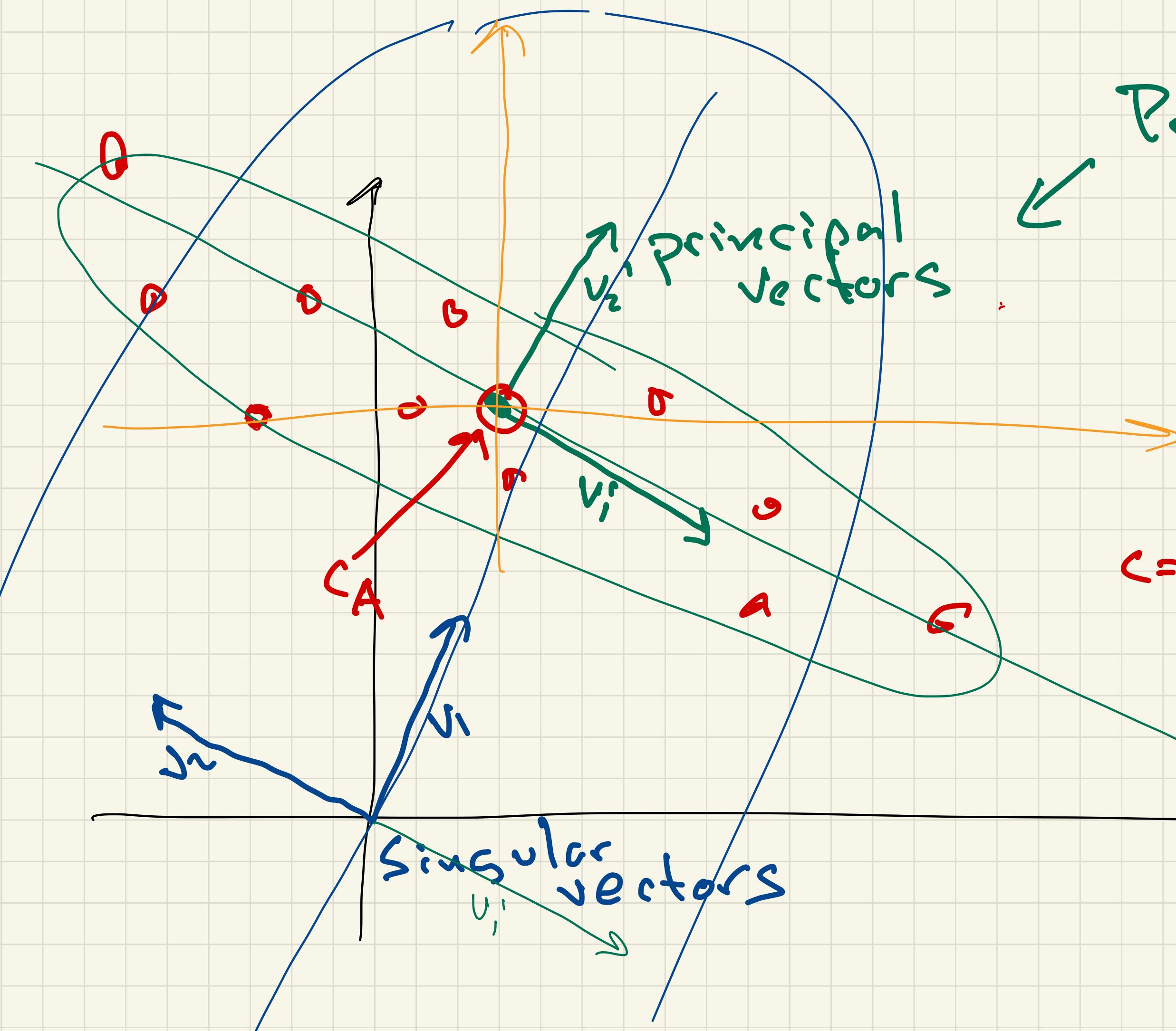


$$A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T \quad \in \mathbb{R}^{n \times d}$$

$$\|A - A_k\|_{2 \text{ or } F}$$



origin $= (0, 0 \ldots 0)$



Principal Component Analysis (PCA)

$$c = (\bar{a}_1, \ldots, \bar{a}_d)$$

$$\bar{a}_j = \frac{1}{n} \sum_{i=1}^{n} A_{ij}$$

average of jth coord.

PCA

0. Input $A \in \{a_1, \ldots a_n\} \in \mathbb{R}^d$

1. Find center $c_A = (\bar{a}_1, \ldots \bar{a}_d) = \frac{1}{n} \sum_{i=1}^{n} a_i$

2. For all $a_i \in A$ } centering make $c_A = 0$

$$\tilde{a}_i = a_i - c_A$$

$$\tilde{A} = \{\tilde{a}_1, \ldots \tilde{a}_n\}$$

3. $U, S, V^T = \text{svd}(\tilde{A})$

① $a_i \to b_i \in \mathbb{R}^k$

$$b_i = (\langle v_1, a_i \rangle, \ldots, \langle v_k, a_i \rangle)$$

4. Principal component : $v_1, \ldots v_k$ ← right sing. vectors of $\tilde{A}$

also report c_A

② $\hat{a}_i = c_A + \sum_{j=1}^{k} v_j \langle v_j, a_i \rangle \in \mathbb{R}^d$

Centering Matrix

$$C_n = I_n - \frac{1}{n}\underbrace{\mathbb{1}\mathbb{1}^T}$$

$$\underbrace{\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}}_{n \times n} \; \frac{1}{n}\underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{n \times n} = \underbrace{\begin{bmatrix} 1-1/n & & -\frac{1}{n} \\ & 1-1/n & \\ -\frac{1}{n} & & \ddots \end{bmatrix}}_{n \times n}$$

$$\bar{A} = C_n A = \left(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right)A = A - \frac{1}{n}\mathbb{1}\mathbb{1}^T A$$

$$\text{PCA} = \text{SVD}(\text{centered}(A))$$

$$= \text{SVD}(C_n A)$$

Multi Dimensional Scaling (MDS)

Input n objects $= \Omega$ & $x_1, \ldots, x_n$

distance function $d : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$

Goal: Find $a_1, \ldots, a_n \in \mathbb{R}^k$ $k \ll n$

$d(x_i, x_j) \approx \| a_i - a_j \|$

input distance matrix $D \in \mathbb{R}^{n \times n}$

$D^{(2)} = \{ D_{ij}^{(2)} = d(x_i, x_j)^2 \}$ $D_{ij} = d(x_i, x_j)$

classical MDS

Input $D \in \mathbb{R}^{n \times n}$ distance matrix

1. Double centering $M = -\frac{1}{2} C_n D^{(2)} C_n$

2. eigendecomposition

$$[L, V] = \text{eig}(M) \qquad M = V L V^T$$

3. Return $Q = V_k \underbrace{L_k^{1/2}}_{\text{sing. values}}$

$$\|q_i - q_j\| \approx D_{ij}$$

$$Q = \{q_1, q_2, \ldots q_n\} \in \mathbb{R}^k$$

Why does cMDS work?

$A = \{a_1, \dots, a_n\}$

D_{ij} think comes from $\|a_i - a_j\|$

$a_i, a_j \in \mathbb{R}^d$

$(a-b)^2 = a^2 + b^2 - 2ab$

$$\|a_i - a_j\|^2 = \|a_i\|^2 + \|a_j\|^2 - 2\langle a_i, a_j \rangle$$

$$\boxed{\langle a_i, a_j \rangle} = \frac{1}{2}\left(\|a_i\|^2 + \|a_j\|^2 - \boxed{\|a_i - a_j\|^2}\right)$$

assume $a_1 = 0 \in \mathbb{R}^d$

$\|a_i\| = \|a_i - a_1\|$

$\|a_i\| = D_{i1}$

$$(AA^T)_{ij} = \langle a_i, a_j \rangle$$

$A_k \in \mathbb{R}^k = V_k L_k^{1/2}$

$$\mathrm{eig}(AA^T) = [L, V]$$

L left sing vect (A)

sq. sing values A