L20: Principal Component Analysis & Multi Dimensional Scaling

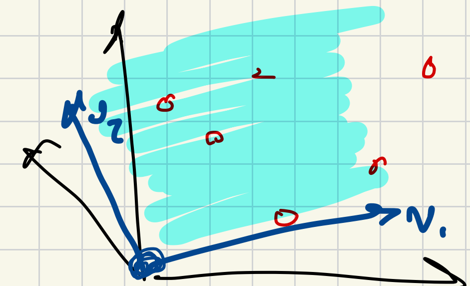
Mar 25, 2026

FoDA

Jeff M. Phillips

<u>Input</u> $A = \{a_1, \dots, a_n\} \subset \mathbb{R}^d$ $\Longleftarrow$ all coords have same units & meaning

<u>Goal</u> $\displaystyle \underset{\substack{F \text{ rank } k \\ (\text{contain } 0)}}{\arg\min} \sum_{i=1}^{n} \| a_i - \pi_F(a_i) \|^2$ $\Longleftarrow$ remove constraint



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

$\|v_i\| = 1$

$\langle v_i, v_j \rangle = 0$

Solve for V_F using SVD

$A = USV^T$

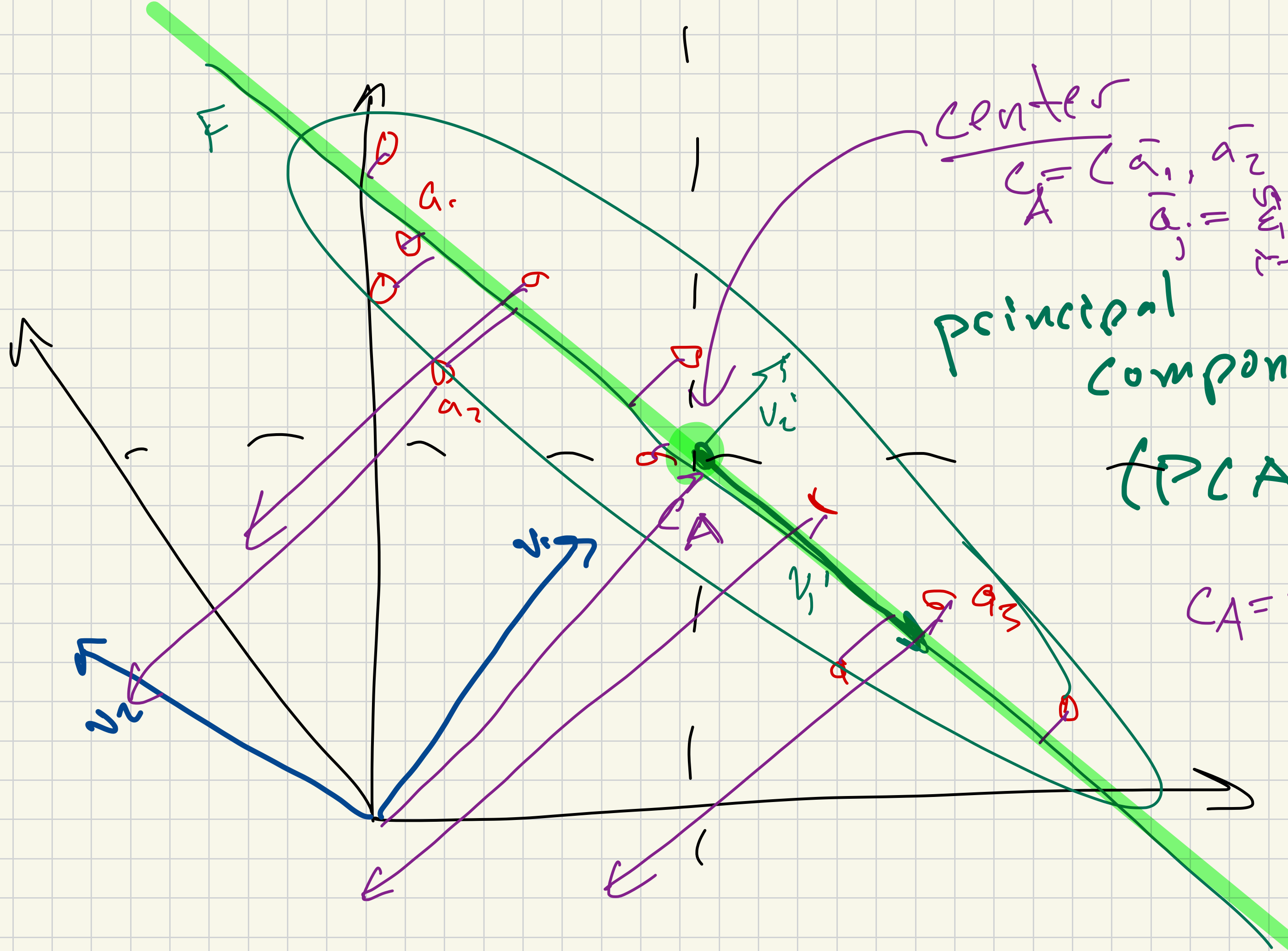
$\boxed{A} = \boxed{U}\,\boxed{S}\,\boxed{V^T}$

best rank k

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

$\Longrightarrow$ or project to k-dim

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



center

$c_A = (\bar{a}_1, \bar{a}_2, \dots, \bar{a}_d)$

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

principal components (PCA)

$CA = \frac{1}{n} \sum_{i=1}^{n} a_i$

PCA

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

1. Find center $C_A = (\bar{a}_1, \bar{a}_2 \dots \bar{a}_d) = \frac{1}{n} \sum_{i=1}^{n} a_i$

2. For all $a_i \in A$

   $\tilde{a}_i = a_i - C_A$ $\quad \tilde{A} = \{\tilde{a}_1, \dots \tilde{a}_n\}$ centering

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

4. Principal Components : $v_1, v_2, \dots v_k$

   ① $a_i \to b_i \in \mathbb{R}^k$

   $b_i = (\langle a_i, v_1 \rangle, \langle a_i, v_2 \rangle \dots \langle a_i, v_k \rangle)$

   ② $a_i \to \hat{a}_i \in \mathbb{R}^d$

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

Centering Matrix

$$C_n = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T \qquad \mathbb{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\left(I_n = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} - \frac{1}{n}\left(\mathbb{1}\mathbb{1}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\right) = \begin{bmatrix} 1-1/n & & -\frac{1}{n} \\ & 1-1/n & \\ -\frac{1}{n} & \ddots & \\ & & 1-1/n \end{bmatrix} \right)$$

centering

$$\tilde{A} = C_n A$$

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

$$= SVD(C_n A)$$

MultiDimensional Scaling (MDS)

Input n objects $x_1, x_2 \ldots x_n \in \Omega$

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

Goal Find $a_1, a_2 \ldots a_n \in \mathbb{R}^k$ $k << n$

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

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

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

2. eigendecomposition

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

3. Return $Q = V_k L_k^{1/2} \in \mathbb{R}^{n \times k}$ $\quad \|q_i - q_j\| \approx D_{ij}$

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

why does cMDS work

$$D_{ij} = \|a_i - a_j\| \quad \text{for some } \{a_1, a_2, \dots, a_n\} \in \mathbb{R}^d$$

$$D^{(2)}_{ij} \qquad (a-b)^2 = a^2 + b^2 - 2ab$$

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

cancel w/ centering

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

$$M_2 = (A A^T)_{ij} = \langle a_i, a_j\rangle$$