FoDA : Principal Component Analysis (PCA)

L21

centering, PCA, MDS

Dimensionality Reduction

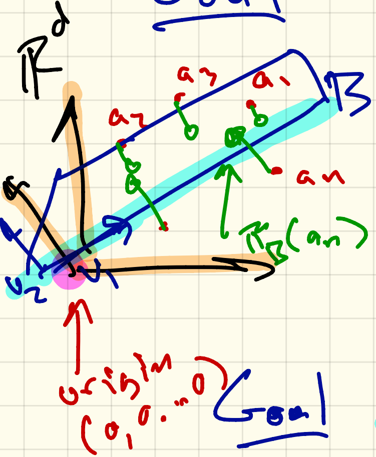
Input $A \subset \mathbb{R}^d$

$$a_1, a_2 \dots a_n \in \mathbb{R}^d$$

$$d(a_i, a_j) = \|a_i - a_j\|_2$$

Goal Low-dimensional Representation

$a_i \mapsto b_i$



$$B \subset \mathbb{R}^k \qquad b_1, b_2 \dots b_n$$

B represents k-dim subspace of $\mathbb{R}^d$

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

Goal minimize V_B $\quad \|\pi_B(A) - A\|^2 = \sum_{i=1}^{n} \|\overbrace{\pi_B(a_i) - a_i}^{r_i}\|^2$

Alternative Problem

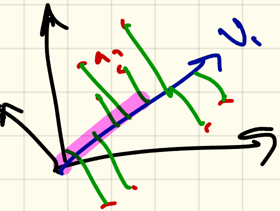
data points

features

	f_1	f_2	f_3	f_4	... f_n
x_1	7.1	blue			
x_2	8.3	green			
⋮	⋮				
x_n	9.4	black			

blue	green	... black
f_{21}	f_{22}	... f_{27}
1	0	0
0	1	0
0	0	0
0	0	0
0	0	1

can't mix
numerical
categorical
in unsupervised



High Dimensional Data

$a_1, a_2, \dots, a_n \in \mathbb{R}^d$

How do I sort?

↳ The first dimension v_1 provides the best 1-dim ordering

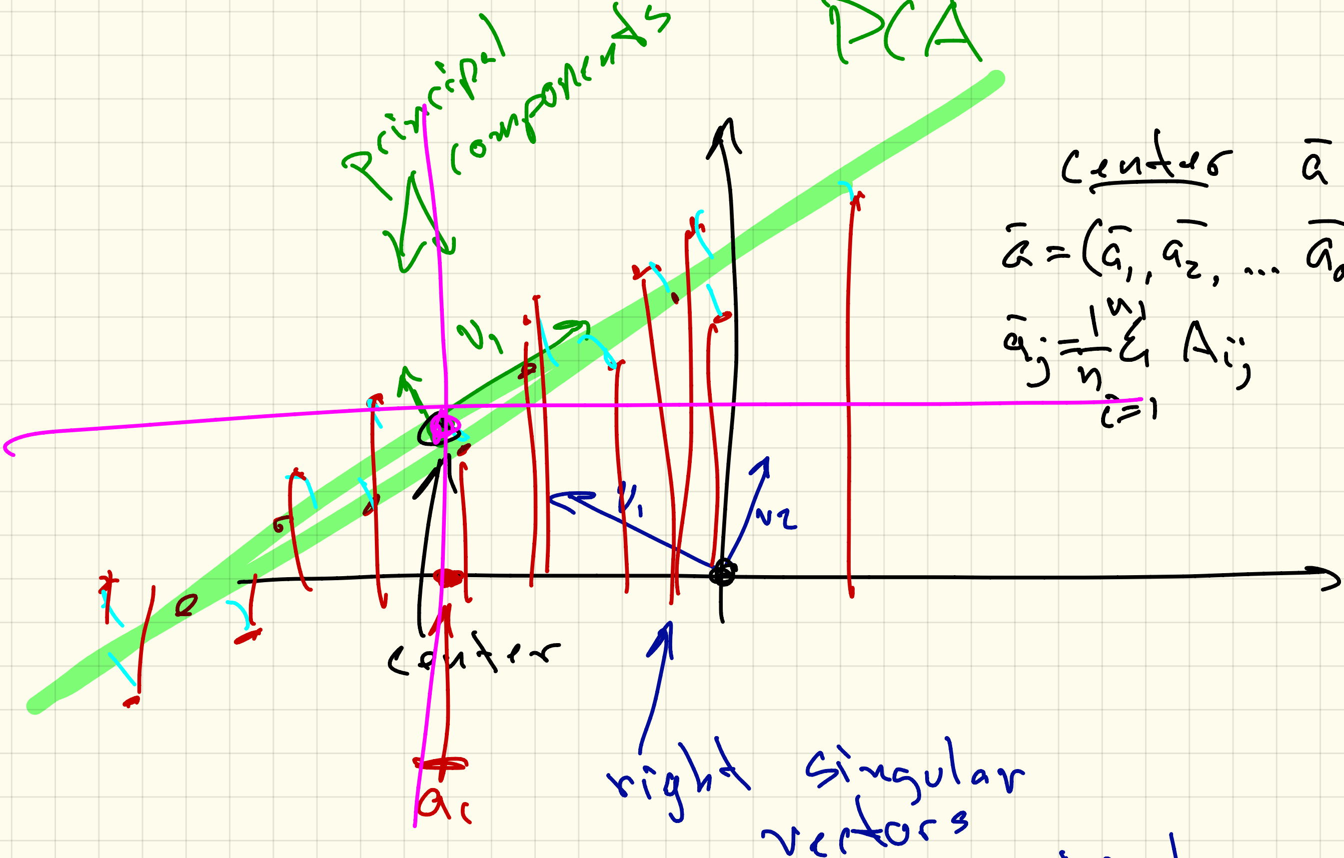
$\{\langle v_1, a_1\rangle, \langle v_1, a_2\rangle, \dots \langle v_1, a_n\rangle\}$

PCA

center $\bar{a} \in \mathbb{R}^d$

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

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



Not Good!

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

Two ways

1. compute center $\bar{a} = (\bar{a}_1, \dots, \bar{a}_d)$

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

$$\tilde{A}_{ij} = A_{ij} - \bar{a}_j$$

2. centering matrix

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

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

$$- \begin{bmatrix} 1/n & 1/n & \dots & 1/n \\ \vdots & & & \\ 1/n & & \dots & 1/n \end{bmatrix}$$

$$A - \frac{1}{n}[\mathbb{1}\mathbb{1}^T]A = \tilde{A}_{ij} = C_n A$$

Principal Component Analysis

Input $A \in \mathbb{R}^{n \times d}$, param $k < d < n$

Output $B : V_B = \{v_1 .. v_k\}$ minimizes $\sum_{i=1}^{n} \| \pi_B(a_i) - a_i \|^2$

1. $\tilde{A} = C_n A$ ← centering

2. $USV^T = \text{svd}(\tilde{A})$

3. $V_B = \{v_1, v_2, \ldots v_k\}$

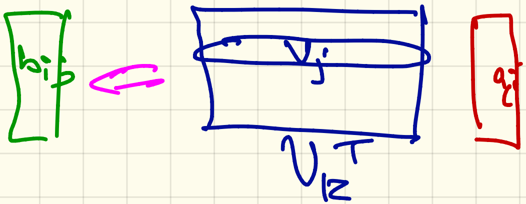
$B = \{b_1, \ldots b_n\}$

$B = V_k^T A \subset \mathbb{R}^k$

If I want k-dimensional pts

$$B = \{b_1, \dots, b_n\} \in \mathbb{R}^k$$

$$B = V_k^T A = b_i = V_k^T a_i$$



$$b_{ij} = \langle v_j, a_i \rangle$$

If I want d-dim pts in k-dim subspace

1. Centering $\bar{a} \in \mathbb{R}^d$

2. SVD USV^T

$$\tilde{A}_k = U_k S_k V_k^T \in \mathbb{R}^{n \times d}$$

$$[A_k]_{ij} = [\bar{A}_k]_{ij} + \bar{a}_j$$

undo

SVD
best subspace
though origin

PCA
best subspace
(not necessarily
w/ origin)

Variance

v_1 highest variance subspace

R.V. X

$$\mathrm{Var}[X] = E\left[\left(X - E[X]\right)^2\right]$$

$$\Pr[x] = \frac{1}{|X|}$$

Sample expected value $E[x] = \frac{1}{|X|} \sum_{x \in X} x \Pr[x]$

$$= \frac{1}{|X|} \sum_{x \in X} x$$

centering makes 0

sample variance

$$E_x\left[\left(X - E_x[x]\right)^2\right]$$

$$X = \{\langle v, a_i \rangle, \langle v, a_2 \rangle \dots\}$$

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left(\langle v, a_i \rangle - \frac{1}{n} \sum_{i=1}^{n} \langle v, a_i \rangle\right)\right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\langle v, a_i \rangle^2\right)$$

MultiDimensional Scaling (MDS)

Input (Ia) set n objects and a distance d

(Ib) $D \in \mathbb{R}^{n \times n}$ $\quad D_{ij} = d(x_i, x_j)$

Goal: Embedding of n objects in $\mathbb{R}^k$ as $Q = \{q_1 \dots q_n\}$

so $\|q_i - q_j\| \approx D_{ij} = d(x_i, x_j)$

Classical MDS

$D^{(2)} \approx -AA^T$

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

Input $D \in \mathbb{R}^{n \times n}$, k

1. Square $D^{(2)}$: $D^{(2)}_{ij} = D_{ij}^2$
2. $M = -\frac{1}{2} C_n D^{(2)} C_n$ ← double centering
3. $[L, V] = \text{eig}(M)$
4. Return $Q = V_k L_k^{1/2}$

top k eigenvectors

$\sqrt{\ }$ eigenvalues $\approx$ sing. values

Why does Classical MDS work?

Assume $\exists$ $a_1 \dots a_n$ $\quad A = \begin{bmatrix} a_1 \\ a_2 \\ a_n \end{bmatrix} \in \mathbb{R}^{n \times p}$ $\quad \|a_i - a_j\|$, $d(a_i, a_j) = D_{ij}$

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

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

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

$$[AA^T]_{ij} = \langle a_i, a_j \rangle = \frac{1}{2}\left(\|a_i\|^2 + \|a_j\|^2 - D_{ij}^2\right)$$

assume $a_1 = \text{Origin } (0, 0 \dots 0)$

$$[AA^T]_{ij} = \frac{1}{2}\left(D_{i1}^2 + D_{j1}^2 - D_{ij}^2\right)$$