FoDA

L18 : Dimensionality Reduction

SVD

Dimensionality Reduction

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

$A = \{a_1, a_2, \ldots, a_n\}$
$a_i \in \mathbb{R}^d$

- too high for vis
- runtime depend on d dimensionality d
- statical : signal vs. noise
  k-dim (d-k)-dim

Data as matrix

$A \in \mathbb{R}^{n \times d} \rightarrow$ SVD $\rightarrow$ map to each $f(a_i) \rightarrow b_i \in \mathbb{R}^k$

$d = 10{,}000 \rightarrow k = 2, 3, 10, 100$

Unsupervised Learning

Input A not (X, y)

no labels

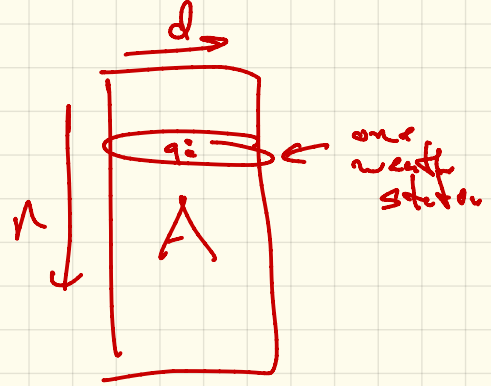
↳ Goal: Simpler Representation of Data.

---

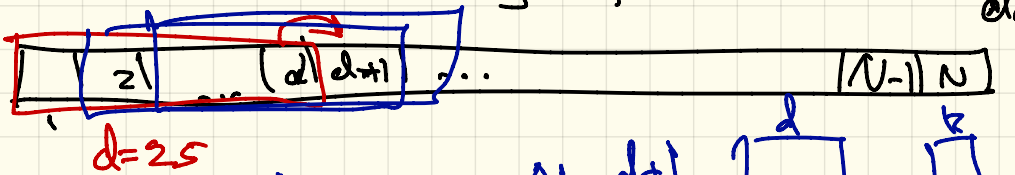
optimize over $A \to B$

Data Matrix

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

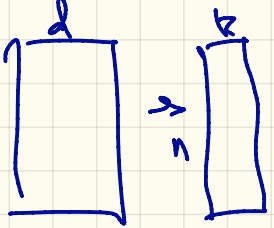


- n weather stations measure temp at d time points
- n users who rate d movies
- stock prices: closing price for N days



$N \rightarrow n = N - d + 1$

$d = 25$



All coordinates have the same units!

$$\|q_i - q_j\| = \text{Euclidean Dist}$$

$$= \sqrt{\sum_{\ell=1}^{d} (q_{i\ell} - q_{j\ell})^2}$$

$q_{i\ell}$: ℓth coord.

adding coordinates

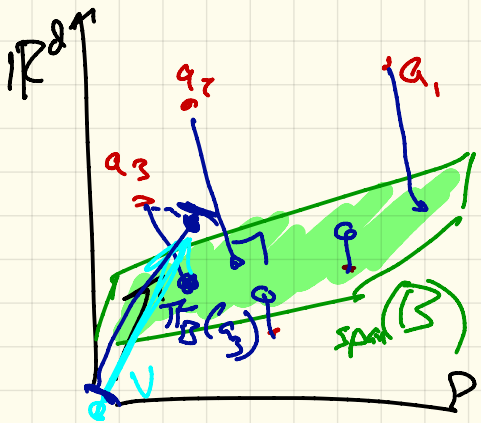
Projections

Goal: $A \to B \subset \mathbb{R}^d$

$rank(B) = k$

Projecting $a_i \in \mathbb{R}^d$ to span (B)

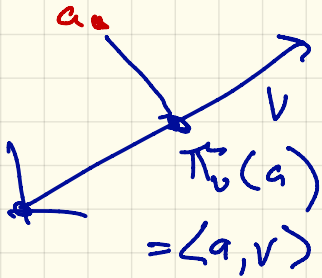
$\pi_B(a_i)$



B be 1-dimensional $v \in \mathbb{R}^d$

$\|v\| = 1$

$\pi_v(a_i) = \langle v, a_i \rangle v$



k-dimensional subspace B

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

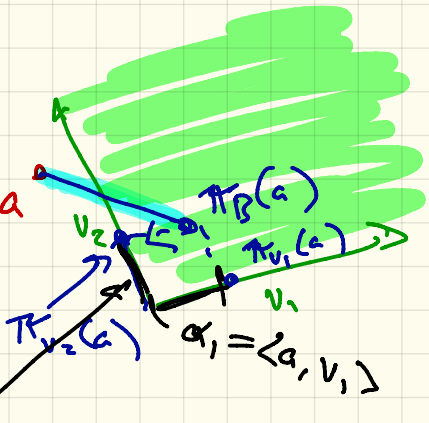
- $\|v_i\| = 1$
- pair v_i, v_j : $\langle v_i, v_j \rangle = 0$

- For any $x \in B$: $x = \sum_{j=1}^{k} \alpha_j v_j$ $\quad \alpha_j = \langle x, v_j \rangle$

$$\pi_B(a) = \sum_{j=1}^{k} \langle v_j, a \rangle v_j$$

vectors in $\mathbb{R}^d$

also $\rightarrow$ $(\alpha_1, \alpha_2, \dots \alpha_k) \in \mathbb{R}^k$



Sum of Squared Errors

$$SSE(A, B) = \sum_{a_i \in A} \| a_i - \pi_B(a_i) \|^2$$

Euclidean dist

Goal: best k-dimensional subspace B
(represented by $V_B = \{v_1, v_2 \dots v_k\}$)
to minimize

$$B^* = \arg\min_B \left(SSE(A, B) \right)$$

SVD $\Rightarrow$ solve for B s.t. B contains Origin $(0, 0 \dots 0)$

PCA $\Rightarrow$ has no origin restriction.
Principal Component Analysis.

Singular Value Decomposition (SVD)

Fortran code LAPACK

Input $A \in \mathbb{R}^{n \times d}$ (ex, $d < n$)

output 3 matrices

$U \in \mathbb{R}^{n \times n}$

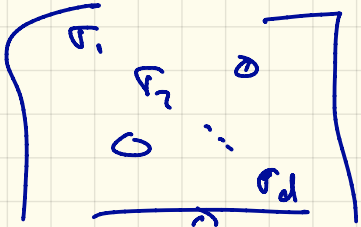
$S \in \mathbb{R}^{n \times d}$

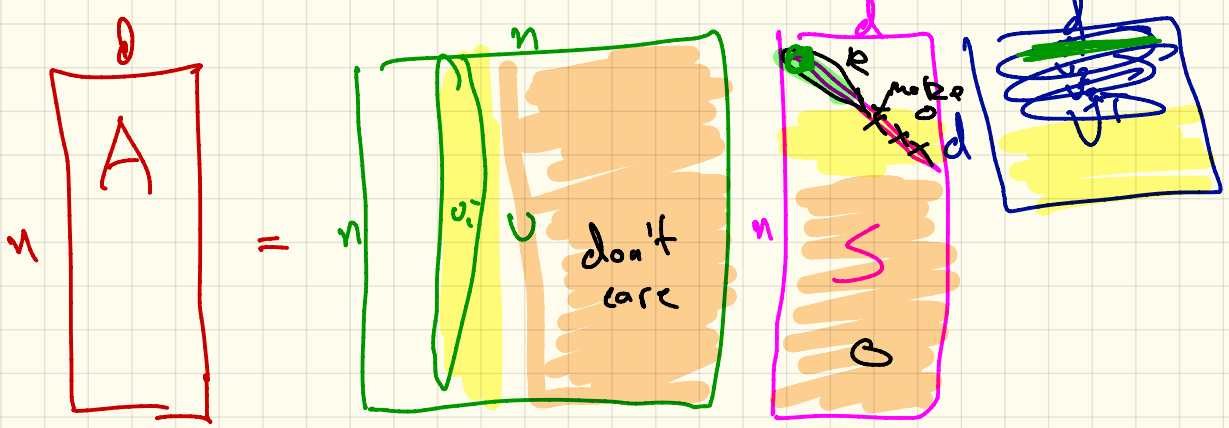
$V \in \mathbb{R}^{d \times d}$

$A = U S V^T$

orthogonal

diagonal $S = \text{diag}(\sigma_1, \sigma_2, \ldots)$





$\|v_i\| = 1$

$\langle v_j, v_i \rangle = 0$

$\sigma_1 > \sigma_2 > \cdots > \sigma_k$

$v_1 \quad v_2$

Best subspace $V_B = \{v_1, v_2, \ldots v_k\}$