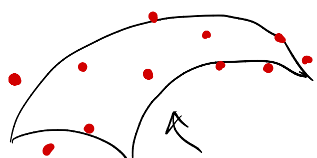
LECTURE 33

LD Paradigm: High-dimensional data has low-dimensional structure.

This allows us to visualize, think about data, world;
lift the curse of high dimensionality.

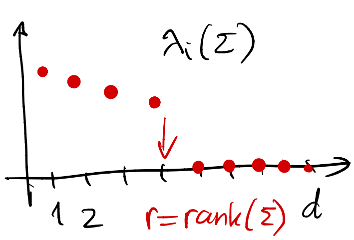
Ex • Human decisions are based on ~5–10 values.
• Human faces have ~ 5–10 features we recognize.

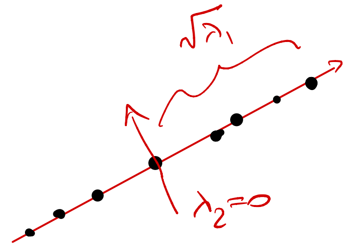
LDP:

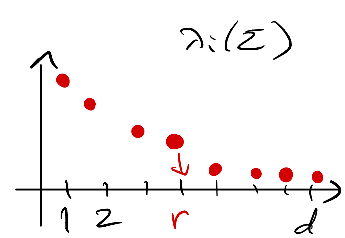


How do we find, parametrize $\mathcal{M}$?
How does our brain do it (in vision, thinking)?
How do we know it is true? From the eigenvalues of $\Sigma = \mathrm{Cov}(X) = \mathbb{E}XX^T$:

• If data $X \in \mathcal{M} \leftarrow$ linear subspace of dimension $r \leq n$
then Σ has $\leq r$ nonzero eigenvalues:
$\downarrow$







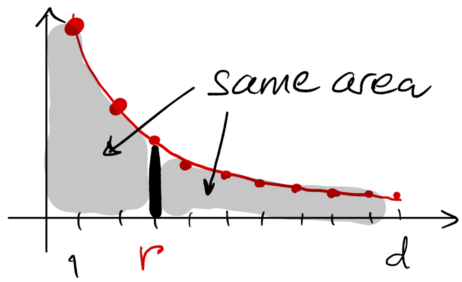
$\uparrow$

• If data X lies close to $\mathcal{M}$, the eigs drop at r (but not exactly to 0)

$r =$ "effective rank" of Σ
"effective dimension" of the data.

Ⓠ Mathematically, how to define the "effective rank" of Σ,
$=$ "effective dimension" of data ?

Ans Find r that splits the area into two equal halves:



Rigorously:

Def The effective rank of a PSD matrix Σ is

$$r(\Sigma) := \min\Big\{ r : \sum_{i \le r} \lambda_i \ge \sum_{i > r} \lambda_i \Big\}$$

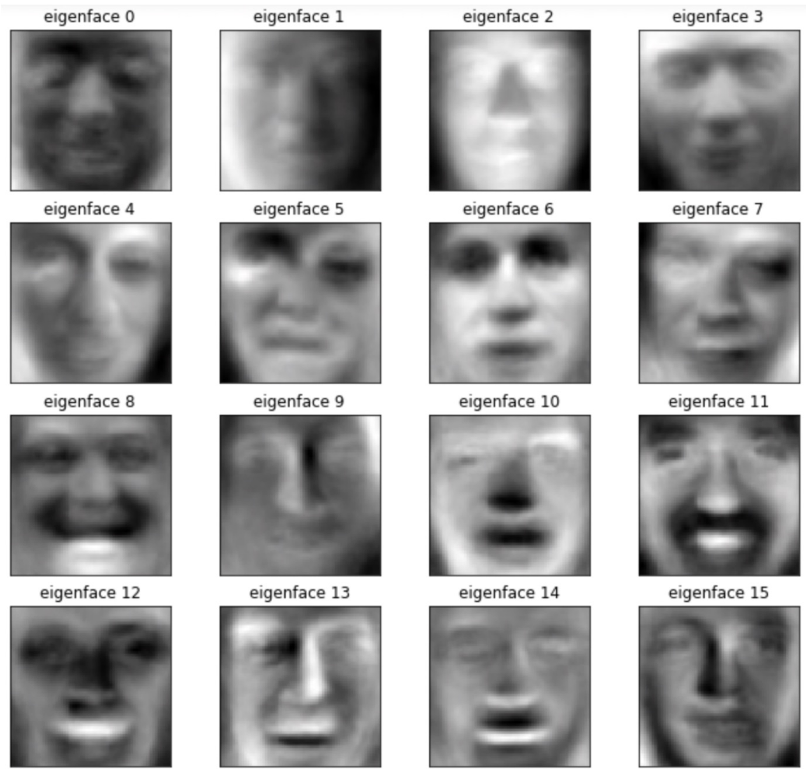
↑ eigs

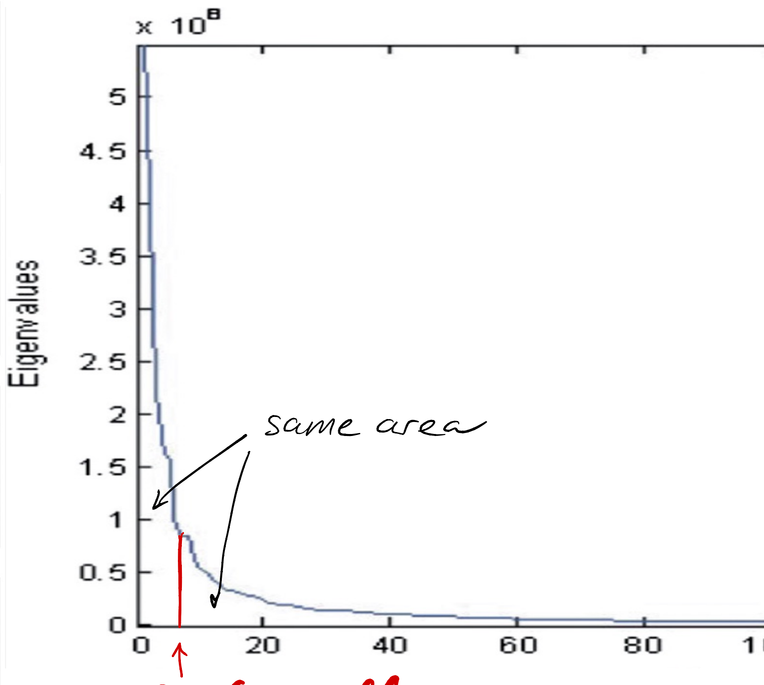
$=$ effective dimension of X if $\Sigma = \mathrm{Cov}(X)$.

Explains 50% of the variation of the data.

Ex (Eigenfaces) Data $X \sim \mathrm{Unif}\{\text{images of human faces}\}$

Eigenvectors of Σ = "eigenfaces" eigenvalues of Σ





$r \approx 6$ = effective dimension

6 facial features explain 50% of variability

- LD Paradigm $\Rightarrow$ in all ML results, the dimension d of the data can be replaced by the effective dimension r $(\ll d)$.
- Example: Covariance estimation $\Rightarrow$ PCA.
  We proved: a sample of size $n = O(d)$ suffices (lec...)
  We will prove: $n = O(r)$ suffices. ☺

remove bad outliers prior to applying

THM Let X be a random vector in $\mathbb{R}^d$ such that

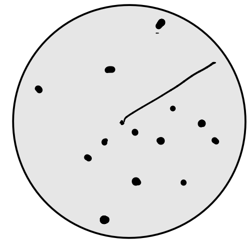
$$\|X\|_2^2 \le 10\, \mathbb{E}\|X\|_2^2 \quad \text{a.s.}$$

If $n \ge C\, r \log d$ then

$$\|\underbrace{\Sigma_n}_{=\frac{1}{n}\sum_1^n X_i X_i^T} - \underbrace{\Sigma}_{=\mathbb{E}XX^T}\| \le 0.1\, \|\Sigma\|$$

effective dim.

sample covariance; X_i = iid copies of X



Toward the proof:

Lem $\mathbb{E}\|X\|_2^2 \le 2r\|\Sigma\|$

$$\mathbb{E}\|X\|_2^2 = \mathbb{E}X^TX = \mathbb{E}\,\mathrm{tr}(\underbrace{X^TX}_{1\times 1 \text{ matrix}}) \overset{\text{cyclic property of trace}}{=} \mathbb{E}\,\mathrm{tr}(XX^T) \underset{\text{linearity}}{=} \mathrm{tr}\,\mathbb{E}[XX^T] = \mathrm{tr}(\Sigma)$$

$$= \sum_{i=1}^d \underbrace{\lambda_i}_{\text{eigs of }\Sigma} = \underbrace{\sum_{i\le r}\lambda_i}_{\ge} + \underbrace{\sum_{i>r}\lambda_i}_{\text{def of } r} \le 2\sum_{i\le r}\underbrace{\lambda_i}_{\le \lambda_1 = \|\Sigma\|} \le 2r\|\Sigma\|.$$

Proof of Thm Use Matrix Bernstein inequality (Lec 30, November 11):

$$\Sigma_n - \Sigma = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(X_iX_i^T - \Sigma)}_{\text{iid mean 0 random matrices}}.$$

Use Matrix Bernstein ineq. (Lec 30, Nov. 11) $\Rightarrow$

$$\underset{\uparrow\ \text{hides an absolute constant factor}}{\lesssim} \frac{1}{n}\left(\sigma\sqrt{\log d} + K\log d\right) \text{ where } \sigma^2 = \Big\| \sum_{i=1}^{n} \underbrace{\mathbb{E}(X_iX_i^T - \Sigma)^2}_{\text{all equal}} \Big\|, \quad (*)$$

$$\|XX^T - \Sigma\| \le K.$$

all equal $\Rightarrow$

σ^2 $= n\|\underbrace{\mathbb{E}(XX^T - \Sigma)^2}\|$

$$0 \preceq \mathbb{E}(XX^T)^2 - \underbrace{\mathbb{E}(XX^T)}_{=\Sigma}\Sigma - \Sigma\underbrace{\mathbb{E}(XX^T)}_{=\Sigma} + \Sigma^2$$

$$= \mathbb{E}X\underbrace{X^TX}_{\|X\|_2^2 \le 10\,\mathbb{E}\|X\|_2^2 \le 20r\|\Sigma\|}X^T - \underbrace{\Sigma^2}_{\succeq 0} \preceq 20r\|\Sigma\|\cdot \underbrace{\mathbb{E}XX^T}_{=\Sigma} = 20r\|\Sigma\|\Sigma$$

$\|X\|_2^2 \underset{\uparrow\ \text{assumption}}{\le} 10\,\mathbb{E}\|X\|_2^2 \underset{\uparrow\ \text{Lem}}{\le} 20r\|\Sigma\|$

$$\Rightarrow \sigma^2 \le \boxed{20rn\|\Sigma\|^2}.$$

since $(0 \preceq A \preceq B \Rightarrow \|A\| \le \|B\|)$ **HW**

K $\quad \|XX^T - \Sigma\| \underset{\uparrow\ \Delta\text{ inequality}}{\le} \|\underbrace{XX^T}\| + \|\Sigma\| \le \boxed{21r\|\Sigma\|}$

$\|XX^T\| = \|X\|_2^2 \le 20r\|\Sigma\|$, as above

Substitute σ^2, K into M.B.I. (* p. 4) $\Rightarrow$

$$\|\Sigma_n - \Sigma\| \lesssim \frac{1}{n}\left(\sqrt{rn\|\Sigma\|^2\log d} + r\|\Sigma\|\log d\right) = \underbrace{\left(\sqrt{\frac{r\log d}{n}} + \frac{r\log d}{n}\right)}_{\le\ 0.01 \text{ if } n > Cr\log d.}\|\Sigma\|$$

REMARKS 1. Logarithmic oversampling is needed in general **(HW)**

2. No distribution assumptions on X!