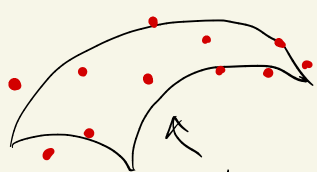
# LECTURE 23

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

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

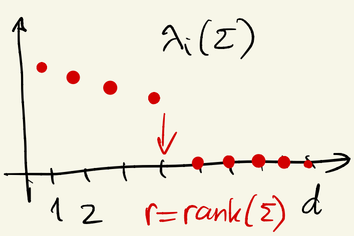
**Ex** • Human decisions are based on ~5-10 values
• Human faces have ~5-10 features we recognize.

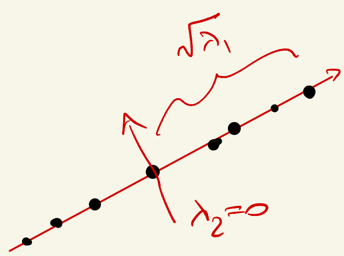
**LDP:**

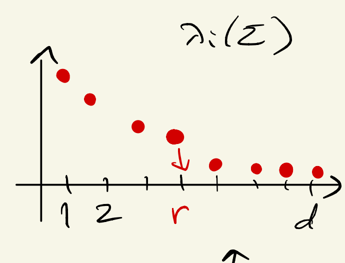



How do we find, parametrize $\mathcal{M}$?
How does our brain do it (in vision, thinking)?
How do we know it is true? From the eigenvalues of $\Sigma = \mathrm{Cov}(X) = \mathbb{E}XX^T$:

• If data $X \in \mathcal{M}$ ← linear subspace of dimension $r \ll n$
then $\Sigma$ has $\le r$ nonzero eigenvalues:
$\downarrow$



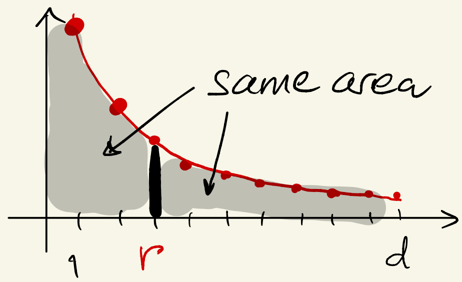




• If data $X$ lies close to $\mathcal{M}$, the eigs drop at $r$ (but not exactly to 0)

$r$ = "effective rank" of $\Sigma$
"effective dimension" of the data.

(Q) Mathematically, how to define the "effective rank" of $\Sigma$, = "effective dimension" of data?

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




Rigorously:

$$\boxed{\text{Def The effective rank of a PSD matrix } \Sigma \text{ is } \quad r(\Sigma) := \min\Big\{ r : \sum_{i \le r} \lambda_i \ge \sum_{i > r} \lambda_i \Big\}}$$

↑ eigs

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

Explains 50% of the variation of the data.

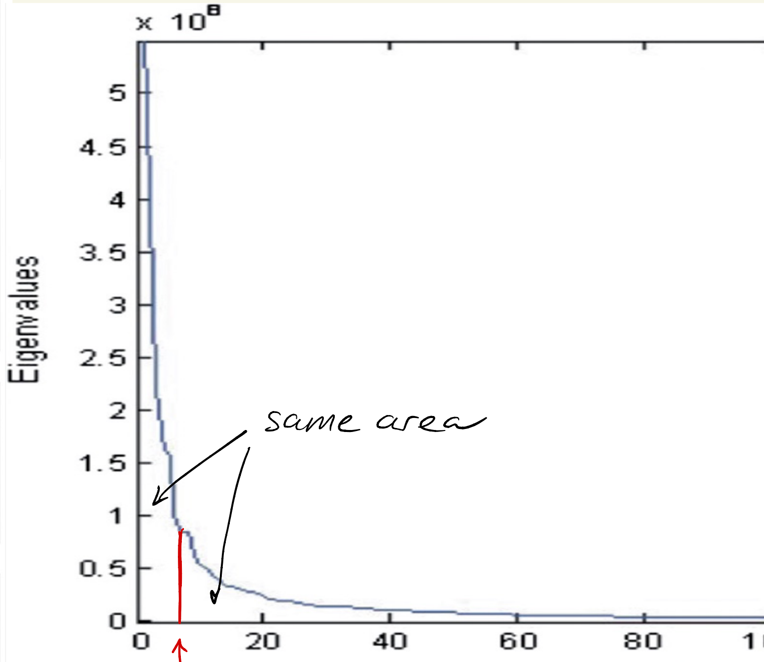
Ex (Eigenfaces) Data $X \sim \mathrm{Unif}\{\text{images of human faces}\}$

$X_i \in \mathbb{R}^{50{,}000 = d}$ $\qquad$ $\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$

$n = 2000$

Eigenvectors of $\Sigma$ = "eigenfaces"


eigenvalues of $\Sigma$




same area

$r \approx 6$ = 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.

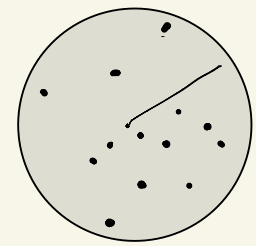
  We proved: a sample of size $n = O(d)$ suffices (lec...)

  We will prove: $n = O(r)$ suffices. ☺

remove bad outliers prior to applying

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

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



If $n \ge C \overset{\text{effective dim.}}{r} \log d$ then

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

sample covariance; $X_i$ = iid copies of $X$

Toward the proof:

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

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

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

$\sum_{i>r}\lambda_i \le \sum_{i\le r}\lambda_i$: def of $r$

To prove Thm, use matrix Bernstein inequality (last class):

> $\forall$ independent mean zero $d\times d$ symmetric random matrices $Z_i$ that satisfy $\|Z_i\| \le 1$ a.s., we have
>
> $$\Big\|\sum_{i=1}^{n} Z_i\Big\| \le C\sigma\sqrt{\log d} + CK\log d \quad \text{with prob} \ge 0.99$$
>
> where $\sigma^2 = \Big\|\sum_{i=1}^{n} \mathbb{E}Z_i^2\Big\|$.

## Proof of Thm

$$\|\Sigma_n - \Sigma\| = \Big\|\frac{1}{n}\sum_{i=1}^{n} \underbrace{(X_iX_i^T - \Sigma)}\Big\|$$

iid mean 0 random matrices $\Rightarrow$ use matrix Bernstein inequality:

$$\lesssim \frac{1}{n}\left(\sigma\sqrt{\log d} + K\log d\right) \quad \text{if } \|X_iX_i^T - \Sigma\| \le K \text{ a.s., where} \quad (*)$$

hides an absolute constant factor

$$\sigma^2 = \Big\|\sum_{i=1}^{n} \overbrace{\mathbb{E}(X_iX_i^T - \Sigma)^2}^{\text{all equal}}\Big\| = n\,\|\mathbb{E}(XX^T-\Sigma)^2\|$$

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

$$= \mathbb{E}X\underbrace{X^TX}_{= \|X\|_2^2}X^T - \underbrace{\Sigma^2}_{\succcurlyeq 0} \preccurlyeq 20r\|\Sigma\|\cdot \underbrace{\mathbb{E}XX^T}_{=\Sigma} = 20r\|\Sigma\|\Sigma$$

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

$$0 \preccurlyeq n\,\mathbb{E}(XX^T-\Sigma)^2 \preccurlyeq 20rn\|\Sigma\|\Sigma$$

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

$$\Rightarrow \sigma^2 = \|\cdots\| \le 20rn\|\Sigma\|^2$$

$K$: $\|XX^T - \Sigma\| \underset{\Delta\text{ inequality}}{\le} \underbrace{\|XX^T\|}_{\text{HW}} + \|\Sigma\| \le 21r\|\Sigma\|$

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

Substitute $\sigma^2$, $K$ into $(*)$ $\Rightarrow$

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

REMARKS

1. Logarithmic oversampling is needed in general **(HW)**
2. No distribution assumptions on $X$!