**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**: $\mathbb{R}^d$ → $\mathbb{R}^r$  
  data $\in$ low-dim. manifold $\mathcal{M}$

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 $\Sigma = \text{Cov}(x) = EXX^T$:

- If data $X \in \mathcal{M}$ - linear subspace of dimension $r \leq n$ then $\Sigma$ has $\leq r$ nonzero eigenvalues:

  $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$

  \[ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \]

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

  $r = \text{"effective rank" of } \Sigma$

  \[ r = \text{"effective dimension" of the data.} \]
Mathematically, how to define the “effective rank” of $\Sigma$, or “effective dimension” of data?

Find $r$ that splits the area into two equal halves.

Rigorously:

**Def.** The effective rank of a PSD matrix $\Sigma$ is

$$ r(\Sigma) = \min \{ r : \frac{\sum_{i \leq r} \lambda_i}{\sum_{i \geq r} \lambda_i} \geq \frac{\sum_{i \geq r+1} \lambda_i}{\sum_{i \leq r} \lambda_i} \} $$

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

Explains 50% of the variation of the data.

**Ex. (Eigenfaces)** Data $X \sim \text{Unif}\{\text{images of human faces}\}$
Eigenvectors of $\Sigma$ — "eigenfaces" — eigenvalues of $\Sigma$.

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 $r$ ($\leq d$).

- Example: Covariance estimation $\Rightarrow$ PCA.
  We proved: a sample of size $n = O(d)$ suffices (e.g.,)
  We will prove: $n = O(r)$ suffices.
Theorem 1

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

$$\|X\|_2^2 \leq 10 \|E\|_2^2 \text{ a.s.}$$

If $n \geq C r \log d$ then

$$\|\Sigma_n - \Sigma\| \leq 0.1 \|\Sigma\|$$

**Effective dim.**  $\frac{1}{n} \sum_i X_i X_i^T$  
Sample covariance, $X_i = \text{iid copies of } X$

**Toward the proof:**

Lemma

$$E \|X\|_2^2 \leq 2 r \|\Sigma\|$$

$$E X^T X = E \text{ tr}(X^T X) = E \text{ tr}(XX^T) = \text{ tr} E [XX^T] = \text{ tr}(\Sigma)$$

Cyclic property of trace  
Linesarity

$$= \sum_{i=1}^d \lambda_i \quad \sum_{i \leq r} \lambda_i + \sum_{i > r} \lambda_i \leq 2 \sum_{i \leq r} \lambda_i \leq 2 r \|\Sigma\|.$$  

Eigenvalues of $\Sigma$
Proof of Thm

Use Matrix Bernstein inequality (lec 30, November 11):

\[
\Sigma_n - \Sigma = \frac{1}{n} \sum_{i=1}^{n} (X_i X_i^T - \Sigma) \] . Use Matrix Bernstein Ineq. (lec 30, Nov. 11) \[
\leq \frac{1}{n} \left( 3\sqrt{\log d} + K \log d \right) \] where \( \sigma^2 = \left\| \sum_{i=1}^{n} E(X_i X_i^T - \Sigma)^2 \right\| \), (*)

\[
\sigma^2 = n \left\| E(X X^T - \Sigma)^2 \right\|
\]

\[
0 \leq E(x x^T)^2 - E(x x^T) \Sigma - \Sigma E(x x^T) + \Sigma^2
\]

\[
= E X X^T x x^T - \Sigma^2 \leq 20r \| \Sigma \|^2 : E X X^T = 20r \| \Sigma \|^2
\]

\[
\| x x^T - \Sigma \| \leq 20r n \| \Sigma \|^2
\]

\[
K \quad \| x x^T - \Sigma \| \leq \| x x^T \|^2 + \| \Sigma \| \leq 21r \| \Sigma \|^2
\]

Substitute \( \sigma^2, K \) into M.B.I. (* p.4) \[
\| \Sigma_n - \Sigma \| \leq \frac{1}{n} \left( \sqrt{n} \| \Sigma \|^2 \log d + r \| \Sigma \|^2 \log d \right) = \left( \sqrt{\log d \frac{n}{n}} + \frac{r \log d}{n} \right) \| \Sigma \|
\]

REMARKS

1. Logarithmic oversampling is needed in general \( (K) \)

2. No distribution assumptions on \( X \)!