• Last class: How to visualize a high-dim. distribution? (e.g. human genome)

\( X \in \mathbb{R}^d \) r. vector, \( \mathbb{E}X=0, \text{ Cov}(X)=EXX^T=\Sigma. \)

e.g. \( X \) = genome of a random person in the world

Eigenvectors \( v_i \) of \( \Sigma = "\text{principal components }" \) of \( X \)

- PCA: reduce dimension \( \mathbb{R}^d \to \mathbb{R}^2 \) by projecting \( X \)
  onto \( \text{span}\{v_1,v_2\} \)

- PCA assumes that we can compute the population cov. matrix

\[ \Sigma = \text{Cov}(X) = EXX^T \]

average over population

But we don’t have data of all population. We have:

• Finite sample \( X_1,\ldots,X_n \in \mathbb{R}^d \) iid copies of \( X \). \( \Rightarrow \) We approximate \( \Sigma \) by

\[ \Sigma_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^T \]

“Sample covariance matrix”

and hope that \( \text{PCA(sample)} = \text{PCA(population)} \), i.e.

\[ \lambda_i(\Sigma_n) \approx \lambda_i(\Sigma) \text{ and } v_i(\Sigma_n) \approx v_i(\Sigma). \]  (*)

• How large is \( n \)? \( n=O(\log d)? \) \( n=O(d)? \) \( n=O(d^2)? \) \( n=O(d^3)? \)

COVARIANCE ESTIMATION PROBLEM

• Our goal: \( n=O(d) \) suffices for (*)

• Plan

1. Approximate \( \Sigma_n \approx \Sigma \) in operator norm

2. Use perturbation theory to conclude (*)
By HW (operator norm for symmetric matrices),

$$\|\Sigma_n - \Sigma\| = \max_{v \in S^{d-1}} \frac{|v^T (\Sigma_n - \Sigma) v|}{v^T \Sigma_n v - v^T \Sigma v}$$

where $S = \text{unit sphere in } \mathbb{R}^d$

$$v^T \Sigma_n v = \sum_i \langle x_i, v \rangle^2$$

$$v^T \Sigma v = \sum_{i,j} \langle x_i, x_j \rangle \langle v_i, v_j \rangle$$

$$\Rightarrow \|\Sigma_n - \Sigma\| = \max_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (\langle x_i, v \rangle^2 - \langle x_i, v \rangle) \right|$$

$$(Z(v))_{v \in S^{d-1}}$$ Random process indexed by $v \in S^{d-1}$.

Compare to the Brownian motion (a.k.a. Wiener process) $$(B(t))_{t \in [0, \infty)}$$ indexed by time.

$$E \max_{t \leq T} |B(t)| \leq \sqrt{T}, \quad E \max_{v \in S^{d-1}} |Z(v)| \leq ?$$

Difficulty: a continuum of points in $S^{d-1}$. ⇒ Discretize.
**THE \( \varepsilon \)-NET METHOD**

**Prop (Finding a net)**
For \( \forall \varepsilon > 0 \), the unit sphere \( S^{d-1} \) has an \( \varepsilon \)-net \( x_1, \ldots, x_N \) with

\[
N \leq \left( \frac{2}{\varepsilon} + 1 \right)^d
\]

i.e. \( \forall x \in S^{d-1} \) is within dist \( \varepsilon \) from some \( x_i \):

\[
\forall x \in S^{d-1} \iff \exists i: \| x - x_i \|_2 \leq \varepsilon
\]

**Algorithm**
- Choose \( \forall x_1 \)
- Choose \( \forall x_2 \) at dist \( > \varepsilon \) from \( x_1 \)
- Choose \( \forall x_3 \) at dist \( > \varepsilon \) from \( \{ x_1, x_2 \} \)
- Choose \( \forall x_4 \) at dist \( > \varepsilon \) from \( \{ x_1, x_2, x_3 \} \)
- Stop whenever impossible

**Claim:** The \( (\varepsilon/2) \)-balls centered at \( x_i \) are disjoint

If not, \( \exists \ i \neq j, \exists y: \{ \| x_i - y \|_2 \leq \varepsilon/2 \}
\quad \iff \| x_j - y \|_2 \leq \varepsilon/2 \)

\[\triangle \Rightarrow \| x_i - x_j \|_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\]

But all \( x_i \) are \( \varepsilon \)-separated by construction.

- All these balls lie in the \( (1 + \varepsilon/2) \)-ball centered at \( 0 \) \( \Rightarrow \)

\[
\text{Vol}(B(1 + \varepsilon/2)) \geq N \cdot \text{Vol}(B(\varepsilon/2))
\]

\[\Rightarrow \quad N \leq \frac{\text{Vol}(B(1 + \varepsilon/2))}{\text{Vol}(B(\varepsilon/2))} = \left( \frac{1 + \varepsilon/2}{\varepsilon/2} \right)^d = \left( \frac{2}{\varepsilon} + 1 \right)^d.
\]

Remark: Covering = packing.
Prop (Computing the operator norm on a net)

Let $A$ be an $m \times n$ matrix, $N \subseteq S^{n-1}$ an $\varepsilon$-net. Then

$$\|A\| \leq \frac{1}{1-\varepsilon} \max_{x \in N} \|Ax\|$$

By def of operator norm, $\exists u \in S^{n-1}$:

$$\|Au\| = \|A\| \quad (\star)$$

By def of $\varepsilon$-net, $\exists x \in N$:

$$\|x-u\|_2 \leq \varepsilon.$$

$\Rightarrow \|Ax-Au\|_2 = \|A(x-u)\|_2 \leq \|A\| \cdot \|x-u\|_2 \quad (\text{def of operator norm})$

$$\leq \|A\| \cdot \varepsilon \quad (\star\star)$$

$\Rightarrow \|Ax\|_2 = \|Au - (Ax-Au)\|_2 \geq \|Au\|_2 - \|Ax-Au\|_2 \quad (\Delta \text{ineq.})$

$$\geq \|A\| - \|A\| \cdot \varepsilon \quad (\text{by } (\star) \text{ and } (\star\star))$$

$$= (1-\varepsilon) \|A\|.$$