Lecture 2 Gaussian Estimation

Xuanxi Zhang

1 Lecture Topics

- Linear estimation with PCA and probabilistic interpretation.
- Efficient high-dimensional estimation.
- From Gaussian to Gaussian Mixtures: power of non-linear modeling.

Basic Setup

We observe data $X_1, \dots, X_n \in \mathbb{R}^d$ and we view them as i.i.d. samples from an underlying data distribution π (unknown).

2 First Moment Estimation

Given random vector $X \in \mathbb{R}^d$, its **mean**

$$m = \mathbb{E}[X] \in \mathbb{R}^d$$

and its **covariance**

$$\Sigma = \mathbb{E}\left[(X-m)(X-m)^T\right] \in \mathbb{R}^{d\times d}$$

where $\Sigma^T = \Sigma$ (symmetric matrix).

- m represents the center of mass.
- Σ specifies the spread along each direction of $\mathbb{R}^d$.In particular, Σ is positive semi-definite (psd).
- Recall that if X is a random vector of mean m and covariance Σ, then the random variable

  $$Z = \langle X-m, v\rangle, \quad v \in \mathbb{R}^d$$

  has mean 0 and variance $v^T\Sigma v$.

We can decompose X as:

$$X = \mathbb{E}[X] + (X - \mathbb{E}[X])$$

where

- $\mathbb{E}[X]$ is the deterministic part.
- $(X - \mathbb{E}[X])$ is a random component but has zero mean.

3 Dimensionality Reduction and Principal Component Analysis

Given a random variable in high dimension, we want reduce the number of variables, while preserving as much information as possible. It is quite intuitive to connect with normal dimensionality reduction techniques like low rank approximation or feature selection.

So consider a centered random vector $X \in \mathbb{R}^d$ with covariance $\Sigma = U\Lambda U^T$. We could change basis to the eigenvectors of Σ, which is the columns of U. Then we have

$$X = \sum_{j=1}^{d} \langle X, U_j\rangle U_j$$

Define $Y_j = \langle X, U_j\rangle$ as new random variables. Then we have

$$\mathbb{E}[Y_i Y_j] = \mathbb{E}(U_i^T X X^T U_j) = U_i^T \Sigma U_j = \lambda_j \delta_{ij}$$

which means Y_i are decorrelated and the variance of Y_i is λ_i.

3.1 PCA as an Optimal Linear Approximation

Go further to the ides of ”Dimensionality Reduction”, we can state it in the mathematic form. Given random variable X, we want to solve following optimization problem:

$$\min_{P\in\mathbb{R}^{d\times k},\ PP^T=I_k} \mathbb{E}\|(I-PP^T)X\|^2$$

The mathematical definition for a projector is : if a tranform $Q\in\mathbb{R}^{d\times d}$ satisfies $Q^2=Q$, then we call it a projector. Here given $P\in\mathbb{R}^{d\times k}$ has orthogonal columns, then PP^T and $I-PP^T$ are both projectors. The former one project X to the subspace spanned by the columns of P, while the latter one project X to the orthogonal complement of the subspace. Above optimization problem wants to minimizes the projection residual.

Some calculation

$$\mathbb{E}\|(I-PP^T)X\|^2 = \mathbb{E}\|X\|^2 - \mathbb{E}\|P^TX\|^2 = \mathrm{Tr}(\Sigma) - \mathrm{Tr}(P^T\Sigma P)$$

This is totally the same as matrix low rank approximation. Optimal solution is $P=(U_1,\dots,U_k)$.

3.2 Practical Importance

PCA is an **extremely useful visualization and summarization tool**, used in applications such as:

- Eigenfaces (face recognition),
- Gene expression analysis,
- Dimensionality reduction for large datasets.

4 Probabilistic Interpretation of PCA

Q:

1. What is the natural probabilistic model behind PCA?
2. What are the estimation properties and sample complexity of PCA?

4.1 From PCA to Probabilistic Modeling

PCA is only concerned with the **first two moments** of the distribution, meaning it considers only:

$$d+\frac{d(d+1)}{2}\text{ parameters.}$$

But there are infinitely many distributions with the first two moments **Analogy:** In numerical interpolation, we also have finite point constraints, but have to choose from infinite number of functions. We enforce a certain type of **smoothness** and choose the ”best” one from all possible functions.

4.2 Measuring the ”Smoothness” of Probability Distributions

Example 1: Suppose the input space X is a discrete set:

$$X=\{x_1,x_2,\dots,x_L\}.$$

The **uniform distribution** $P(X=x_i)=\frac{1}{L}$ is arguably the most ”regular.”

Example 2: Suppose the input space is the continuous interval:

$$X=[0,1].$$

Again, the **uniform measure** is the most regular.

Entropy Maximization. We can characterize these ”nice” distributions by **maximizing entropy**. For a discrete domain:

$$H(\pi)=-\sum_{i=1}^{L}\pi_i\log\pi_i.$$

Under appropriate assumptions (more on this in later lectures), for continuous distributions:

$$H(\pi) = -\int \log \frac{d\pi}{dx}(x)\,\pi(dx),$$

where $\pi \ll \text{Leb}$ (absolutely continuous with respect to the Lebesgue measure).

Entropy quantifies **uncertainty** and the amount of **information** revealed by observing a random event. It is a Fundamental quantity in Statistical mechanics, Information theory and so on.

4.3 Entropy-Based Characterization of Distributions

We can use entropy to **characterize distributions** under a given set of constraints. For example, solving:

$$\max_{\pi \in \mathcal{P}(X)} H(\pi)$$

subject to:

$$\mathbb{E}_{X\sim\pi}[\Phi(X)] = \Phi_0,$$

where $\Phi : X \to \mathbb{R}^K$ represents the **"sufficient statistics"**.

Important Remark: This constrained optimization problem is **not always well-defined**; its feasibility depends on:

- The domain X.
- The constraints imposed by Φ.

Example 1: If $\Phi \equiv 0$ (no constraint), and

$$X = [0,1],$$

then the solution is:

$$\pi^* = \text{Uniform}([0,1]).$$

Example 2: If $\Phi \equiv 0$ and

$$X = \mathbb{R},$$

then there is **no solution**, as the entropy **blows up**.

Example 3:

Maximum Entropy with Polynomial Constraints. Consider the case where:

$$\Phi(x) = (x, xx^T)$$

representing all polynomials of degree 1 and 2, with $X = \mathbb{R}$.

We focus on feasible constraints and rewrite:

$$\Phi(x) = (x, (x-m)^2),$$

with given moments:

$$\Phi_0 = (m, \sigma^2).$$

$$\max_{p:\mathbb{R}\to\mathbb{R}} -\int p(x)\log p(x)\,dx$$

subject to:

$$\begin{cases} p(x) \geq 0, \quad \forall x, \\ \int p(x)\,dx = 1, \\ \int xp(x)\,dx = m, \\ \int (x-m)^2 p(x)\,dx = \sigma^2. \end{cases}$$

- This is a **constrained optimization problem** in infinite dimensions.
- First, observe that the function:

  $$[0,1] \ni t \mapsto -t\log t$$

  is concave.

- Additionally, the constraints are **linear** in p.
- We can characterize the maximizer by analyzing its **first-order critical points**.

We define the Lagrangian function:

$$\mathcal{L}(p) = -\int p(x)\log p(x)\,dx + \lambda\left(\int p(x)dx - 1\right) + \beta\left(\int p(x)x\,dx - m\right) + \gamma\left(\int p(x)(x-m)^2dx - \sigma^2\right).$$

Taking the derivative with respect to $p(x)$:

$$\frac{\delta\mathcal{L}}{\delta p(x)} = -\log p(x) + 1 + \lambda + \beta x + \gamma(x-m)^2 = 0.$$

Solving for $p(x)$

$$\log p(x) = a + bx - cx^2.$$

$$\Rightarrow p(x) = Ce^{-bx-cx^2},$$

where C, b, c are chosen such that the constraints are satisfied.
Conclusion: Gaussian Distribution

$$\pi = \mathcal{N}(m, \sigma^2)$$

is the **Gaussian distribution** with mean m and variance σ^2.

- The same calculation in d dimensions shows that the **maximum entropy distribution** with given mean and covariance is the **multivariate Gaussian**.
- This is known as the **"maximally non-committal"** distribution (Jaynes), based on the principle of maximum entropy (MaxEnt).
- **Re-interpretation of PCA:** When we only model the first two moments, we are implicitly assuming an underlying Gaussian model.

5 Statistical Aspect

Now we consider the **empirical version** of the problem, where we observe $X_1, \ldots, X_n$ rather than an abstract random variable X which has mean m and variance Σ. We have empirical mean

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Assume, without loss of generality, that $\mathbb{E}(\hat{m}) = m = 0$. And empirical covariance:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T.$$

Fact: $\hat{\Sigma}$ is also symmetric and positive semi-definite (psd), and $\mathbb{E}[\hat{\Sigma}] = \Sigma$.
By the **law of large numbers**, we know that as $n \to \infty$:

$$\hat{m} \to m, \quad \hat{\Sigma} \to \Sigma \quad \text{(almost surely)}.$$

5.1 Accuracy of Covariance Estimator

Question: How good is this estimator of the covariance? In other words, for a desired expected relative accuracy

$$\mathbb{E}\|\Sigma - \hat{\Sigma}\|_{\text{op}} \leq \varepsilon,$$

how many samples n are required?

Theorem 1 (Vershynin)**.** Assume $X \in \mathbb{R}^d$ is a random vector such that:

$$\|X\|_2 \leq K\sqrt{\mathbb{E}\|X\|_2^2}$$

for some constant K. Then, for any $\varepsilon > 0$:

$$\frac{\mathbb{E}\|\hat{\Sigma} - \Sigma\|}{\|\Sigma\|} \leq \varepsilon$$

whenever:

$$n \approx \varepsilon^{-2} d \log d.$$

Proof: Recitation Tomorrow.

Consequence: PCA is Not Cursed by Dimension

- We can estimate eigenvalues with **small relative error**.
- When eigenvalues are sufficiently spaced, we can also estimate principal components with small relative error, as stated in the **Davis-Kahan theorem**.