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 Earliest Starting Point for Probabilistic Modeling?

Rather than specifying the full distribution, we can first focus on the first moments: 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 $v^T\Sigma v$.

A Naïve but Useful Decomposition

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

where

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

Next Step: Exploiting Covariance. We now proceed with the next step of this decomposition by exploiting covariance.

3 Principal Component Analysis (Pearson, 1901)

- Given a centered random vector $X \in \mathbb{R}^d$, can we decompose it as a sum:

$$X = \sum_{j=1}^{d} Y_j V_j$$

  such that Y_j are decorrelated, i.e.,

$$\mathrm{Cov}(Y_i, Y_j) = 0 \quad \text{if } i \neq j?$$

  where V_j form an orthogonal basis.
- How do we find such a decomposition?
- Why is it useful?

3.1 Largest Direction of Variance

Question: What is the largest direction of variance?

Recall: The variance of $\langle X - \mathbb{E}[X], v\rangle$ is given by $v^T \Sigma v$. The solution $v_{\max}$ is the eigenvector corresponding to the largest eigenvalue of Σ.

$$v_{\max} = \arg\max_{\|v\|_2=1} \; v^T \Sigma v.$$

Using a Lagrange multiplier λ:

$$L(v) = v^T \Sigma v + \lambda(\|v\|^2 - 1)$$

and setting $\nabla L(v) = 0$ leads to the eigenvalue equation.

Since Σ is positive semidefinite, we have:

$$\Sigma = U \Lambda U^T, \quad \text{with } \lambda_i \geq 0.$$

Consequence: Denote U_j as the j-th column of U. The random variables

$$Y_j = \langle X, U_j \rangle$$

are decorrelated. Their variances decrease as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

$$Y = U^T X, \quad \mathbb{E}[YY^T] = \mathbb{E}[U^T X X^T U] = U^T U \Lambda U^T U = \Lambda.$$

3.2 PCA as an Optimal Linear Approximation

Another consequence of PCA is that it provides an **optimal linear approximation** or **dimensionality reduction** with respect to mean-squared error:

$$\min_{P \text{ ortho. projection in dim } k} \mathbb{E}\|X - PP^T X\|^2$$

where the term $PP^T X$ represents the k-dimensional "features." P is ortho. projection in dim k means $P \in \mathbb{R}^{d\times k}$ and $P^T P = I_k$.

Error decomposition

$$\mathbb{E}\|(I - PP^T)X\|^2 = \mathbb{E}\|X\|^2 - \mathbb{E}\|P^T X\|^2 = \sigma^2 - \mathrm{Tr}(P^T \Sigma P) = \sigma^2 - \sum_{i=1}^{d} \lambda_i \|P^T U_i\|^2.$$

Optimal solution is $P = (U_1, \ldots, U_k)$.

3.3 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:

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

- In contrast, a **probability model** defined over $\mathbb{R}^d$ is infinite-dimensional, as it requires **all** moments to fully describe the distribution.
- **Analogy:** Consider an analogy where we attempt to fit a function using a finite set of points. Some moments of the distribution are known—how do we interpoleta them? We need more constraints or regularization.
- **Regularization:** enforce a certain type of **smoothness**.
- The goal is to find the **"smoothest"** distribution that agrees with some given moments.

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.

Question: Is there a variational principle at play?

Answer: 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 and Uncertainty Entropy quantifies **uncertainty** and the amount of **information** revealed by observing a random event.

Fundamental Quantity in:

- Statistical mechanics
- Information theory
- ...

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:

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

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 Estimating Principal Components

- So far, we have studied how to extract principal components from the covariance Σ of a random vector.
- In practice, we observe $x_1, \dots, x_n \in \mathbb{R}^d$ as i.i.d. samples.
- **Empirical version?**

Empirical Mean

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

Assume, without loss of generality, that $m = 0$.

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).

$$\mathbb{E}[\hat{\Sigma}] = \mathbb{E}[XX^T] = \Sigma.$$

By the **law of large numbers**, we know that as $n \to \infty$:

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

Accuracy of Covariance Estimator

Question: How good is this estimator of the covariance?
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**.