MATH888: High-dimensional probability and statistics Fall 2021

Lecture 4 — September 15, 2021

Sebastien Roch, UW-Madison *Scribe: Subhojyoti Mukherjee*

1 Overview

In the last lecture we reviewed some important facts in probability theory and then considered a more general context for comparing estimators. In this lecture we will redefine a few notations and study the Bias-variance decomposition and Cramer-Rao bound.

2 Notations

We make slight adjustments to notations in this lecture. We simplify them and redefine as follows:

1. We have a family of distributions $\mathcal{P}$ over a sample space $\mathcal{X}$. We will assume everything is measurable.
2. Parameter $\theta : \mathcal{P} \to \Theta$, where Θ is the parameter space.
3. We are provided with the data: n i.i.d $\mathcal{X}$-valued $X_1, X_2, \ldots, X_n \sim P \in \mathcal{P}$
4. We define the estimator: $\hat{\theta}^n : \mathcal{X}^n \to \Theta$

Remark 1. *All events and random variables are over a Probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $\mathbb{P}$ is the underlying measure of the joint probability space. P is the distribution if one of the samples $X_1, X_2, \ldots, X_n$. The expectation and variance of Z is denoted by $\mathbb{E}[Z]$ and $\mathrm{Var}(Z)$ respectively. If need to specify the distribution of samples P, we use a sub-script. $\mathbb{P}_P$, $\mathbb{E}_P$, Var_P. The $\mathbb{P}_p$ is defined on an event, the $\mathbb{E}_P$, Var_P are defined over some function of X_i's.*

Remark 2. *We often drop the superscript of $\hat{\theta}^n$. That is we will write $\hat{\theta}$.*

Definition 3. *We define the following definitions which are taken from [Pan16]*

(a) The bias of $\hat{\theta}$ is $bias_P(\hat{theta}, \theta) = \mathbb{E}_P[\hat{\theta} - \theta]$. For short we will write $\mathrm{bias}(\hat{\theta})$. Bias is positive if you over-estimate and negative if you under-estimate. It is not a metric as such.

(b) The mean squared error (MSE) of $\hat{\theta}$ is

$$\mathrm{MSE}_P(\hat{\theta}, \theta) = \mathbb{E}_P(\|\hat{\theta} - \theta\|^2).$$

For short we will write $\mathrm{MSE}_P(\hat{\theta})$.

3 Bias-Variance Decomposition [Pan16], (Lemma 3.4)

We introduce the notation $\hat{\theta} = (\hat{\theta}, \ldots, \hat{\theta_p}) \in \mathbb{R}^p$ is a p-dimensional vector. Then we have

$$\text{MSE}(\hat{\theta}) = \| \text{bias}(\hat{\theta})\|_2^2 + \sum_{k=1}^{p} \text{Var}(\hat{\theta}).$$

Proof. The proof is for $p = 1$. We have that

$$\begin{aligned}\mathbb{E}(\hat{\theta}, \theta) &= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2 \\ &= \underbrace{\mathbb{E}[(\mathbb{E}(\hat{\theta}) - \theta)^2]}_{\text{bias}(\hat{\theta})} + \underbrace{\mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2}_{\text{Var}(\hat{\theta},\theta)} + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\underbrace{\mathbb{E}(\hat{\theta}) - \theta)}_{\text{Term (a)}}]\end{aligned}$$

where the term (a) is deterministic and $\mathbb{E}(\hat{\theta}) - \theta = 0$. Hence we get the result. □

Definition 4. *We say that $\hat{\theta}^{(n)}$, $n \geq 1$ is consistent if $\hat{\theta}^n \to \theta$ as $n \xrightarrow{P} \infty$, where $\xrightarrow{P}$ means convergence in probability.*

Lemma 5. *We can show that $\mathbb{P}[\|\hat{\theta} - \theta\|_2^2 > \epsilon] \leq \dfrac{\text{MSE}(\hat{\theta})}{\epsilon^2}$. This implies consistency if*

$$\text{MSE}(\hat{\theta}^n) \to 0.$$

Proof. Let $Z = \|\hat{\theta} - \theta\|_2^2$. So Z is a random variable and positive. Using Markov's inequality we can show that

$$\mathbb{P}(\|\hat{\theta} - \theta\|_2^2 > \epsilon^2) \leq \frac{\mathbb{E}(\|\hat{\theta} - \theta\|_2^2)}{\epsilon^2}.$$

Note that $\text{MSE}(\hat{\theta}) = \|\hat{\theta} - \theta\|_2^2$. Hence the claim of the lemma follows. □

Definition: We define an unbiased estimator if $\text{bias}_P(\hat{\theta}, \theta) = 0, \forall \theta, P$.

4 Cramer-Rao Bound (Special case)

We now prove the Cramer-Rao bound for the special case when dimension $p = 1$. For the multi-dimensional general case see Thm 6.6 in [LC06].

First we define the following setting required for this proof:

(a) $\mathcal{X}$ is finite.

(b) $\Theta \subseteq \mathbb{R}$ is open

(c) $\mathcal{P} = \{P(\cdot; \theta) : \theta \in \Theta\}$ where $\mathbb{P}(X = x) = P(x; \theta)$

(d) $P(x; \theta)$ is continuously differentiable for all x wr.t. θ

(e) $P(x;\theta) > 0$.

Theorem 6. *Let $\hat{\theta}$ be an unbiased estimator of $g(\theta)$ where $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable. Then*

$$Var(\hat{\theta}) \geq \frac{(g'(\theta))^2}{n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log P(X_1;\theta)\right)^2\right]}$$

where $\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log P(X_1;\theta)\right)^2\right] = I(\theta)$ is the fisher information matrix.

Remark: The $Var(\hat{\theta})$ is same as MSE($\hat{\theta}$) as the bias($\hat{\theta}$) is 0.

Proof. We will prove this Theorem 6 in the next lecture. □

References

[LC06] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[Pan16] Victor M Panaretos. Statistics for mathematicians. *Compact Textbook in Mathematics. Birkhäuser/Springer*, 142:9–15, 2016.