MATH888: High-dimensional probability and statistics Fall 2021

Lecture 4 — September 15, 2021

Sebastien Roch, UW-Madison *Scribe: Ajinkya Kokandakar*

1 Overview

The previous lecture reviewed important concepts in classical probability theory, and introduced the loss function and the risk function that constitute the basic framework for the comparison of estimators. Intuitively, the idea behind the loss function is to measure how close the estimator is to the true (but unknown) value of the target *parameter* by endowing the parameter space with a semi-metric.

The main topic of this lecture is the Cramér-Rao Bound, which is a lower bound for the variance of unbiased estimators for a target parameter.

2 Basic Framework and Notation : A review

We suppose that there is a abstract probability space $(\Omega, \mathcal{F})$ which is the source of all randomness. However, we are only concerned with *samples* $X_1, X_2, \ldots, X_n$ that take values in the *sample space* denoted $\mathcal{X}$. In measure theoretic terms, a *sample* X is a random variable $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{B})$. For our purposes, $\mathcal{X} \subseteq \mathbb{R}^p$. The samples are drawn from a fixed but unknown distribution on the sample space, and our goal is to estimate functions of this distribution. With this as our objective, we use the following framework for the discussion in this lecture:

- *Family of distributions* – There is a family of distributions $\mathcal{P}$ over the sample space, which contains the true distribution P, i.e. $P \in \mathcal{P}$.
- *Parameter* – We are interested in estimating a function of the distribution $\theta : \mathcal{P} \to \Theta$ where Θ is the *parameter space*. In parametric families, we are typically interested in estimating a parameter of the distribution. For example, we might consider the family of normal distributions parameterized by mean μ and variance σ^2 and we are interested in making inferences about the mean μ. Notice that $\mu = \mathbb{E}_{(\mu,\sigma^2)}(X) = \int_{\mathbb{R}} x \, dP_{(\mu,\sigma^2)}(x) = \mu(P_{(\mu,\sigma^2)})$, and is thus a function of the underlying (true) probability measure.
- *Data* – The data consists of a sequence of n i.i.d samples, $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P \in \mathcal{P}$
- *Estimator* – An estimator $\hat{\theta}^{(n)} = \hat{\theta}^{(n)}(X_1, X_2, \ldots, X_n)$ is a function that maps the data to the parameter space i.e. $\hat{\theta}^{(n)} : \mathcal{X}^n \to \Theta$. An estimator is understood to be a function of the sample and hence we only write $\hat{\theta}^{(n)}$. Moreover, we often drop the superscript in the estimator $\hat{\theta}^{(n)}$ and simply use $\hat{\theta}$ instead.

Remark 1. *The expectation and variance of a random variable Z are denoted as $\mathbb{E}[Z]$ and $\mathrm{var}(Z)$. If we need to specify the distribution of the samples from P, we use $\mathbb{P}_P(\cdot), \mathbb{E}_P(\cdot)$ and $\mathrm{var}_P(\cdot)$ for the probability, expected value and variance under the distribution P respectively.*

2.1 Mean squared error and the Bias-Variance tradeoff

Recall the definition of the bias of an estimator from the previous lecture:

Definition 2. *The bias of an estimator is defined as:*

$$\text{bias}_P(\hat{\theta}, \theta) = \mathbb{E}_P[\hat{\theta} - \theta(P)].$$

When the distribution P and parameter θ are obvious from the context, we write $\text{bias}(\hat{\theta})$ for ease of notation. Positive bias indicates that the estimator provides an over-estimate the target parameter *on average*. All else equal, we typically prefer to have zero bias, which motivates the following definition.

Definition 3 (Unbiasedness)**.** *An estimator $\hat{\theta}$ is said to be unbiased if* $\text{bias}(\hat{\theta}, \theta) = 0$ *for all $P \in \mathcal{P}$.*

It is important to note that an unbiased estimator must have zero bias for *all* possible distributions, for the following reason: for any given estimator we can usually find *some* distribution under which the estimator has zero bias. For example, for some $c \in \mathbb{R}$, the constant estimator $\hat{\theta} = c$ has zero bias for any P with $\theta(P) = c$, but this is hardly a useful estimator for any P that leads to a different value of θ.

Definition 4 (MSE, Panaretos [1] Def 3.3)**.** *The mean squared error (MSE) is defined as:*

$$\text{MSE}_P(\hat{\theta}) = \mathbb{E}[\|\hat{\theta} - \theta(P)\|^2]$$

where $\|\cdot\|$ is the L^2 norm.

For brevity we use $\text{MSE}(\hat{\theta})$ when the distribution and the parameter are unambiguous. It is clear that the MSE satisfies the definition of a risk function given in the previous lecture.

The following lemma expresses the the MSE of an estimator in terms of its bias and variance.

Lemma 5 ([1], Lemma 3.4)**.** *Let $\hat{\theta} \in \mathbb{R}^p$ be an estimator for a p-dimensional parameter θ, and $\hat{\theta}_k \in \mathbb{R}$ denote the k-th component of θ, i.e. $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)^\top$, then the following always holds:*

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

This result is called the *bias-variance decomposition*.

Proof. We only prove the bias-variance decomposition for the case $p = 1$. By the definition of MSE,

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + \mathbb{E}[(\mathbb{E}[\hat{\theta}] - \theta)^2] + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)] \\
&= \text{var}(\hat{\theta}) + \mathbb{E}[\text{bias}(\hat{\theta})^2] + 2(\mathbb{E}[\hat{\theta}] - \theta)\underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])]}_{=0} \\
&= \text{bias}(\hat{\theta})^2 + \text{var}(\hat{\theta})
\end{aligned}$$

which concludes the proof.

□

We define another impotant notion in classical statistics in the following definition.

Definition 6 (Consistency)**.** *A sequence of estimators* $\{\hat{\theta}^{(n)}\}$ *is* ***consistent*** *if* $\hat{\theta}^{(n)} \xrightarrow{i.p.} \theta$ *as* $n \to \infty$ *for all* $P \in \mathcal{P}$*.*

Lemma 7. *For any* $\varepsilon > 0$*,*

$$\mathbb{P}(\|\hat{\theta} - \theta\| > \varepsilon) \leq \frac{\mathrm{MSE}(\hat{\theta})}{\varepsilon^2}$$

Proof. Let $Z = \|\hat{\theta} - \theta\|^2$, then using Markov's inequality with Z and $c = \varepsilon^2$ we get:

$$\mathbb{P}(\|\hat{\theta} - \theta\| > \varepsilon) = \mathbb{P}(\|\hat{\theta} - \theta\|^2 > \varepsilon^2) \leq \frac{\mathbb{E}[Z]}{\varepsilon^2}$$

which is the required inequality. □

The following corollary is a direct consequence of the above lemma.

Corollary 8. *A sequence of estimators* $\hat{\theta}^{(n)}$ *is consistent if* $\mathrm{MSE}(\hat{\theta}^{(n)}) \to 0$ *as* $n \to \infty$*.*

We can reduce the mean squared error by either reducing the bias, the variance or both. It is easy to see that the mean squared error of an unbiased estimator is the variance itself. Therefore, when comparing among the class of unbiased estimators we only need to care about the variance of the estimators. This motivates the Cramér-Rao bound in the next section.

3 Cramér-Rao Bound

The Cramér-Rao bound is a lower bound for the variance of any unbiased estimators of a parameter of interest under certain regularity conditions. A lower bound is useful because it provides a benchmark to compare estimators against. If an estimator achieves the lower bound, it is optimal in the sense that no other unbiased estimator can have a lower variance. An estimator that achieves the lower bound is therefore "efficient" in the sense that it has the least possible variance.

In this section we consider the special case of the Cramér-Rao bound for random variables that can take only finitely many (real) values. For a more general result that applies to random variables in $\mathbb{R}^p$, see *Theorem 6.6* in Lehmann and Casella [2].

For the theorem that follows consider the following setup:

- The *sample space* $\mathcal{X}$ is finite, i.e. $|\mathcal{X}| < \infty$.
- The *parameter space* $\Theta \subseteq \mathbb{R}$ is an open set in the real line
- The family of distribution consists of discrete distributions indexed by θ such that

  $$\mathcal{P} = \{\ p(\ \cdot\ ; \theta)\ :\ \theta \in \Theta\ \}$$

  where $p(x, \theta) > 0$ and $p(x, \theta)$ is continuously differentiable in θ for all $x \in \mathcal{X}$.

Suppose we have a sample $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} p(\cdot;\theta)$ as described above, then we have the following theorem:

Theorem 9. *If $\hat{\vartheta}^{(n)},\ n \in \mathbb{N}$ is a sequence of unbiased estimators for $g(\theta)$ i.e. $\mathbb{E}[\hat{\vartheta}^{(n)}] = g(\theta)$ for all $n \in \mathbb{N}$, where $g : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function, then:*

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

where $\mathbb{E}\left[\frac{\partial}{\partial\theta}\log p(X_1;\theta)\right]^2 =: I(\theta)$ is called the Fisher Information for θ.

To understand the lower bound, note that:

$$I(\theta) := \mathbb{E}\left[\frac{\partial}{\partial\theta}\log p(x;\theta)\right] = \mathbb{E}\left[\frac{\frac{\partial p(x;\theta)}{\partial\theta}}{p(x;\theta)}\right]$$

Thus the Fisher information quantifies the expected relative rate of change of the likelihood in response to a small change in θ. If this quantity is larger, it is easier to distinguish the likelihood functions that induce different values of θ, i.e. it is easier to distinguish between different value of θ. It is in this sense, that $I(\theta)$ captures "information" about the parameter θ, and therefore the name "information" matrix. The larger the information matrix, the easier it is to discern between different parameter values and the more information we have about the parameter. Notice that a larger Fisher information value corresponds to a lower Cramér-Rao bound. A note of caution is due at this point: the Cramér-Rao bound is *not* a sharp lower bound, which means that there may not exist any unbiased estimator that achieves this lower bound.

References

[1] Panaretos, V. M. (2016). *Statistics for Mathematicians: A Rigorous First Course.* Birkhauser/Springer.

[2] Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation.* Springer.