MATH888: High-dimensional probability and statistics Fall 2021

Lecture 3 — September 13, 2021

Sebastien Roch, UW-Madison *Scribe: Weicong Lyu*

1 Overview

Last time we reviewed basic concepts in statistics and mostly focused on parametric models. This time we review some important facts in probability theory and then consider a more general context for comparing estimators.

2 Important Probability Facts

The slides for this section (`https://people.math.wisc.edu/~roch/hdps/roch-hdps-slides3.pdf`) provides a quick review of the Markov's inequality and the Chebyshev's inequality for bounding tails of distributions, monotonicity of $\mathcal{L}^p$ norms, the Schwarz inequality and modes of convergence. These materials are based on Williams [1], but the results can be found in any graduate-level probability textbook, e.g., `https://people.math.wisc.edu/~roch/grad-prob/`.

3 Basic Framework

Suppose we have a **family of distributions**, $\mathcal{P}$, over a measure space $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X}$ is the sample space and $\mathcal{B}$ is the associated σ-field. For example, $\mathcal{P}$ can be the family of all Gaussian distributions $\mathcal{N}(\mu, \sigma^2)$.

We are interested in estimating the **parameter**. In general, a parameter θ is a function from the family of distributions $\mathcal{P}$ to Θ, where $(\Theta, \mathcal{A})$ is the parameter space and $\mathcal{A}$ is the associated σ-field. θ can just be a parameter of a parametric family, or the mean and the covariance matrix for a more general collection of nonparametric distributions. For example, a parameter of the collections of Gaussian distributions is μ, the mean of the distribution.

Typically, the **data** we have are n independent and identically distributed (iid) samples $X_1, \ldots, X_n \sim \mathbb{P} \in \mathcal{P}$, where $\mathbb{P}$ is an unknown distribution from the family $\mathcal{P}$.

The **point estimator** $\hat{\theta}_n$ is a measurable function from the space of n data points $\mathcal{X}^n$ to the parameter space Θ.

A few remarks:

1. The sample space $\mathcal{X}$ is typically some measurable subset of $\mathbb{R}^p$. Continuous variables may be associated with Borel sets, while discrete variables may be associated with discrete topology.
2. $\theta(\mathbb{P})$ may not uniquely determines the distribution $\mathbb{P}$. In the Gaussian example, $\theta = \mu$, the mean of the distribution, does not uniquely determine the distribution $\mathbb{P}$; $\theta = (\mu, \sigma^2)$,

the mean and the variance of the distribution, uniquely determines $\mathbb{P}$. In general, for some nonparametric family of distributions we may only be interested in the mean, which does not uniquely determines the distribution.

3. $\mathbb{P}$ is the distribution of one sample. Let $\mathbb{P}^{(n)}$ be the joint distribution of $X_1, \ldots, X_n$ with corresponding expectation operator $\mathbb{E}^{(n)}$. Typically, we will have a sequence of point estimators $\hat{\theta}_n$ for each $n \geq 1$.

4 Loss Function

To compare estimators, we introduce a loss $\rho : \Theta \times \Theta \to \mathbb{R}_+$ which measures how close $\hat{\theta}_n$ is to $\theta(\mathbb{P})$. Typically, ρ is a semi-metric: ρ is symmetric, i.e., $\rho(a,b) = \rho(b,a)$, and ρ satisfies the triangle inequality.

Sometimes it is also useful to consider a non-negative and non-decreasing function Φ and look at the composition $\Phi \circ \rho$. For example, sometimes we want to work with the square of a metric, which is not necessarily a metric.

Later we will choose the loss function and compare the risks of two estimators.

Definition 1. *The risk is the expected loss*

$$\mathbb{E}^{(n)}[\Phi(\rho(\hat{\theta}_n, \theta(\mathbb{P})))],$$

where the expectation is taken over n iid samples from a fixed distribution $\mathbb{P}$.

The risk measures the expectation of how close the point estimator and the parameter are in the sense of loss. However, the risk depends on $\mathbb{P}$ and is hence hard to compare. In general, we want to define notions of optimality. One such notion is the minimax risk, a major quantity of interest this semester.

Definition 2. *The minimax risk is*

$$\mathcal{M}_n(\theta, \Phi \circ \rho) = \inf_{\hat{\theta}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{(n)}[\Phi \circ \rho(\hat{\theta}_n, \theta(\mathbb{P}))],$$

where the infimum is over all point estimators $\hat{\theta}_n$.

A minimax estimator, an estimator that achieves the minimax risk, is the best estimator in the sense that it minimizes the worst-case risk over all distributions in the family.

5 Mean Squared Error

Suppose that $\mathcal{X} \subset \mathbb{R}^p$ and $\Theta \subset \mathbb{R}^r$. The bias,

$$\text{bias}(\hat{\theta}_n, \theta) = \mathbb{E}^{(n)}(\hat{\theta}_n - \theta(\mathbb{P})),$$

measures how the estimator is centered around the true value. The bias itself does not tell us that much because even if the bias is zero, which implies that $\hat{\theta}_n$ is centered at $\theta(\mathbb{P})$, $\hat{\theta}_n$ can still have a large variance so that it is highly likely to be far from $\theta(\mathbb{P})$. Therefore, the variance of the estimator, $\text{Var}(\hat{\theta}_n)$, is also relevant. The mean squared error combines the two.

Definition 3. *The mean squared error (MSE) is*

$$\mathrm{MSE}(\hat{\theta}_n, \theta) = \mathbb{E}^{(n)}[\|\hat{\theta}_n - \theta(\mathbb{P})\|^2],$$

where $\| \cdot \|$ *is the Euclidean distance or the* L^2 *norm.*

The mean squared error is a particular notion of risk and very natural. The following lemma relates mean squared error to the bias and the variance.

Lemma 4 (Bias-Variance Decomposition)**.**

$$\mathrm{MSE}(\hat{\theta}_n, \theta) = \|\mathrm{bias}(\hat{\theta}_n, \theta)\|^2 + \sum_{i=1}^{r} \mathrm{Var}^{(n)}(\hat{\theta}_{n,i}),$$

where $\hat{\theta}_{n,i}$ *is the* i*th component of* $\hat{\theta}_n$*, and* $\mathrm{Var}^{(n)}$ *is the variance with respect to* $\mathbb{P}^n$*.*

According to the lemma, if the mean squared error is small, then both the bias and the variance are small.

References

[1] Williams, David. *Probability with Martingales.* Cambridge: Cambridge University Press, 1991.