36-789 Topics in High Dimensional Statistics II **Fall 2015**

Lecture 9: November 24

Lecturer: Alessandro Rinaldo *Scribes: Yining Wang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

In this lecture we review the equalizer rule for exact minimax estimation and then proceed to minimax hypothesis testing (also known as minimax detection). Finally we consider a high-dimensional detection example, where we want to decide whether there is signal in the underlying model.

9.1 The equalizer rule

Suppose Θ is the parameter space and let $d : \Theta \times \Theta \to \mathbb{R}^+$ be a specific loss function (e.g., the ℓ_2 loss $d(\theta, \theta') = \|\theta - \theta'\|_2^2$). The *risk* of an estimator $\hat{\theta}$ is defined as $\mathbb{E}_\theta[d(\hat{\theta}, \theta)]$, where the expectation is taken over i.i.d. sampled from the underlying distribution parameterized by the true parameter θ. Let π be a prior distribution over the parameter space Θ. The *Bayes risk* of an estimator $\hat{\theta}$ with respect to prior π is defined as

$$R(\hat{\theta}, \pi) = \int_\Theta \mathbb{E}_\theta[d(\hat{\theta}, \theta)]\mathrm{d}\pi(\theta).$$

The *posterior risk* of an estimator $\hat{\theta}$ with respect to prior π and data X is defined as

$$r(\hat{\theta}|X) = \mathbb{E}_{\theta\sim\pi}[d(\hat{\theta}, \theta)|X].$$

A simple observation is that $R(\hat{\theta}, \pi)$ can also be expressed as an integration over posterior risk of $\hat{\theta}$, as shown below:

$$R(\hat{\theta}, \pi) = \int_{\mathcal{X}} r(\hat{\theta}|X) d\mu_X(X). \tag{9.1}$$

The *Bayes rule* estimator with respect to prior π is the estimator $\hat{\theta}$ that minimizes the posterior risk $r(\hat{\theta}|X)$ at every X. It is known that when ℓ_2 loss is used, the Bayes rule is the posterior mean $\mathbb{E}[\theta|X]$.

The *equalizer rule* asserts that an estimator is *minimax* if it is the Bayes rule with respect to some prior π and achieves constant risk for all underlying parameter θ. More specifically, we have the following proposition:

Proposition 9.1 (The equalizer rule) *Let $\hat{\theta}(\pi)$ be the Bayes rule with respect to some prior π. If*

$$\mathbb{E}_\theta[d(\hat{\theta}(\pi), \theta)] = C, \quad \forall \theta \in \Theta$$

for some constant C, then $\hat{\theta}(\pi)$ is minimax:

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta[d(\hat{\theta}(\pi), \theta)] = \inf_{\hat{\theta}} \sup_\theta \mathbb{E}_\theta[d(\hat{\theta}, \theta)].$$

Example: Binomial distribution . Suppose $X \sim B(n, \theta)$ for $\theta \in \Theta = [0, 1]$. Consider the Beta prior $\theta \sim \text{Beta}(\alpha, \beta)$, The posterior distribution of θ conditioned on X is then

$$\theta | X = x \sim \text{Beta}(\alpha + x, \beta + n - x).$$

Under the ℓ_2 loss function $d(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, the Bayes rule is the posterior mean:

$$\hat{\theta}(\pi) = \frac{\alpha + x}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}.$$

Taking $\alpha = \beta = \sqrt{n}/2$, we have

$$R(\hat{\theta}(\pi), \theta) = \frac{1}{4(1 + \sqrt{n})^2}, \quad \forall \theta \in \Theta,$$

which is a constant function with respect to the underlying parameter θ. Subsequently, by the equalizer rule, we claim that the minimax estimator for θ is

$$\hat{\theta} = \frac{1}{1 + \sqrt{n}} \cdot \frac{1}{2} + \frac{\sqrt{n}}{1 + \sqrt{n}} \cdot \frac{X}{n}.$$

9.2 General hypothesis testing theory

Consider distribution class $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ for some parameter space $\Theta \subseteq \mathbb{R}^d$. Suppose $X_1, \cdots, X_n \overset{i.i.d.}{\sim} p_\theta$ for some $\theta \in \Theta$. We want to test:

$$H_0 : \theta \in \Theta_0; \qquad H_1 : \theta \in \Theta_1;$$

for some $\Theta_0, \Theta_1 \subseteq \Theta$. Conventionally we also assume that $\Theta_0 \cap \Theta_1 = \emptyset$. We call a hypothesis testing problem *simple* if each one of Θ_0, Θ_1 only has one parameter; that is, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. A test function ψ is a function from $\mathcal{X}$ to $\{0, 1\}$ such that

$$\psi(X) = \begin{cases} 1, & \text{reject } H_0; \\ 0, & \text{fail to reject } H_0. \end{cases}$$

The *type-I error* of a testing function ψ is defined as $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \psi$, while the *type-II error* if defined as $\sup_{\theta \in \Theta_1} (1 - \mathbb{E}_\theta \psi)$.

From now on we shall consider the simple hypothesis testing case, where $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$ for some distinct $\theta_0, \theta_1 \in \Theta$. There are two standard ways of defining the "risk" of the simple hypothesis testing problem:

1. The risk of ψ is defined as
$$R(\psi) = R(\psi, \theta_0) + R(\psi, \theta_1),$$
where
$$R(\psi, \theta) = c_0 \mathbb{E}_\theta \psi \cdot 1[\theta = \theta_0] + (1 - \mathbb{E}_\theta \psi) \cdot 1[\theta = \theta_1]$$
for some constant $c_0 > 0$.

2. Neyman-Pearson's approach ("bi-criteria"). First define
$$\Psi_\alpha = \{\psi : \mathbb{E}_{\theta_0} \psi \leq \alpha\}$$
to be all tests that have type-I error controlled by some constant $\alpha \in (0, 1)$. The risk of a test $\psi \in \Psi_\alpha$ is then defined as
$$R_\alpha(\psi) = \mathbb{E}_{\theta_1}[1 - \psi].$$

The following lemma (usually referred to as *Neyman-Pearson lemma*) asserts that the optimal test for both risk formulations are likelihood ratio tests.

Lemma 9.2 (Neyman-Pearson) *For both risk formulations the optimal test ψ^* takes the form*

$$\psi^*(X) = \begin{cases} 1, & \text{if } p_1(X)/p_0(X) \geq c; \\ 0, & \text{otherwise.} \end{cases}$$

Here we assumed that $p_1(X)/p_0(X) = c$ with probability zero. Note that for risk formulation 1, set $c = c_0$ and for risk formulation 2, set c such that $\mathbb{E}_{\theta_0}\psi^* = \alpha$.

The two risk formulations can also be generalized to the composite hypothesis case:

1. For the first formulation, define

$$R(\psi, \Theta_0, \Theta_1) = c_0 \cdot \sup_{\theta \in \Theta_0} \mathbb{E}_\theta \psi + \sup_{\theta \in \Theta_1} \mathbb{E}_\theta [1 - \psi]$$

   for some constant $c_0 > 0$.

2. For the second formulation with constant $\alpha \in (0, 1)$, define

$$R_\alpha(\psi, \Theta_0, \Theta_1) = \sup_{\theta \in \Theta_1} \mathbb{E}_\theta [1 - \psi]$$

   for those $\psi \in \Psi_\alpha = \{\psi : \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\psi] \leq \alpha\}$.

9.3 Minimax test (minimax detection)

Definition 9.3 (minimax test) *A test ψ^M is* minimax optimal *if*

$$R(\psi^M, \Theta_0, \Theta_1) = \inf_\psi R(\psi, \Theta_0, \Theta_1)$$

for the first formulation or

$$R_\alpha(\psi^M, \Theta_0, \Theta_1) = \inf_{\psi \in \Psi_\alpha} R_\alpha(\psi, \Theta_0, \Theta_1).$$

Under regularity conditions, one can show that

1. The minimax risk of $R(\psi, \Theta_0, \Theta_1)$ is

$$\sup_{p_0, p_1} \{1 - \|p_0 - p_1\|_1; p_0 \in \text{conv}(\mathcal{P}_0), p_1 \in \text{conv}(\mathcal{P}_1)\}$$

   and the minimax test is achieved by a Bayes test. Here $\text{conv}(\cdot)$ is the convex hull of a distribution class.

2. The minimax risk of $R_\alpha(\psi, \Theta_0, \Theta_1)$ is

$$\sup_{p_0, p_1} \left\{ \inf_{\psi \in \Psi_\alpha} \mathbb{E}_{p_1}[1 - \psi]; p_0 \in \text{conv}(\mathcal{P}_0), p_1 \in \text{conv}(\mathcal{P}_1) \right\}.$$

We next consider an example of high-dimensional minimax detection. Consider $\Theta_0 = \{\theta_0\}$ and $\Theta_1 \subseteq \mathbb{R}^d$. Typically we assume $\theta_0 = 0$ is the zero vector and $\Theta_1(n, d)$ changes with n (the number of samples) and d (the number of variables). The risk of a testing function is defined as

$$R(\psi, \Theta_0, \Theta_1) = \mathbb{E}_{\theta_0}\psi + \sup_{\theta \in \Theta_1} \mathbb{E}_{\theta_1}[1 - \psi],$$

or under a Bayesian formulation

$$R(\psi, \Theta_0, \Theta_1) = \mathbb{E}_{\theta_0}\psi + \int_{\Theta_1} \mathbb{E}_{\theta_1}[1 - \psi]\mathrm{d}\pi(\theta_1)$$

with respect to some prior distribution π over Θ_1. Under the high-dimensional testing scenario, we usually adopt the following definition of *asymptotic power* to quantify the power of a test ψ:

Definition 9.4 (asymptotic power) *A test ψ is* asymptotically powerful *if*

$$\lim_{n \to \infty} R(\psi, \Theta_0(n, d), \Theta_1(n, d)) = 0.$$

On the other hand, ψ is asymptotically powerless *if*

$$\liminf_{n \to \infty} R(\psi, \Theta_0(n, d), \Theta_1(n, d)) = 1.$$

Let's now consider the example of high-dimensional normal mean testing problem. We have

$$H_0 : \mathcal{N}(0, I), \qquad H_1 : \mathcal{N}(\theta, I),$$

where $\theta \in \Theta_1(n, d) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \geq r_{n,d}\}$. The goal is to find the fastest rate of $r_{n,d}$ going to zero while still making the test asymptotically powerful. As a perhaps simpler example, consider low-dimensional linear regression

$$H_0 : \beta = 0, \qquad H_1 : \beta \neq 0.$$

For fixed design $X \in \mathbb{R}^{n \times d}$, a natural test is to consider $\|XX^\dagger y\|_2^2$. Under H_0 we have

$$\|XX^\dagger y\|_2^2 \sim \chi^2_{\min(n,d)}.$$

Therefore, the test is powerless if

$$\frac{\|X\beta\|_2^2}{\sqrt{\min(n, d)}} \to 0.$$