36-789 - High Dimensional Statistsics II	**Fall 2015**

Lecture 8: Nov 19

Lecturer: Alessandro Rinaldo *Scribes: Kirthevasan Kandasamy*

8.1 Equaliser Rule (Cont'd)

Example 8.1 (Example of Equaliser Rule) $X \sim \mathrm{Bin}(n, \theta)$ *where* $\theta \sim \mathrm{Beta}(\alpha, \beta)$. *The risk is* $R(\theta, \hat{\theta}) = \mathbb{E}_\theta[(\theta - \hat{\theta})^2]$. *Given an observation* x, *the posterior* $\Theta | X = x$ *is* $\mathrm{Beta}(\alpha + x, \beta + n - x)$. *The Bayes' rule in* L_2 *error is the posterior expectation,*

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

The risk of $\hat{\theta}(\pi)$ *is* $R(\hat{\theta}(\pi), \theta) = \alpha^2/(\alpha + \beta + n)^2$. *When* $\alpha = \beta = \sqrt{n}/2$, *then* $R(\hat{\theta}(\pi), \theta) = 1/4(1 + \sqrt{n})^2$ *is a constant. Therefore,*

$$\hat{\theta} = \frac{1}{1 + \sqrt{n}}\frac{1}{2} + \frac{\sqrt{n}}{1 + \sqrt{n}}\frac{x}{n}$$

is minimax.

8.2 Minimax Hypothesis Testing

8.2.1 Parametric Models

Let $\mathcal{P} = \{p_\theta; \theta \in \Theta\}$, $\Theta \in \mathbb{R}^d$ and $X = \{X_1, X_2, \ldots, X_n\} \in \mathcal{X}$ be drawn i.i.d from $p_\theta \in \mathcal{P}$. Let $\Theta_0, \Theta_1 \subset \Theta$. In hypothesis testing, our goal is to determine whether $H_0 : \theta \in \Theta_0$ or $H_1 : \theta \in \Theta_1$ upon observing $X \sim p_\theta$.

We say Θ_i is *simple* if it is a singleton and *composite* otherwise.

A test function is a function $\psi : \mathcal{X} \to \{0, 1\}$. We accept H_0 if $\psi(X) = 0$ and reject H_0 if $\psi(X) = 1$. If ψ takes values in $[0, 1]$ then it is a randomised test and we reject H_0 with probability $\psi(X)$.

If $\theta \in \Theta_0$, then $\mathbb{E}_\theta[\psi(X)]$ is the type I error - the probability of false rejection.
If $\theta \in \Theta_1$, then $\mathbb{E}_\theta[1 - \psi(X)]$ is the type II error - the probability of false acceptance.

8.2.2 Two Definitions of Risk for simple vs simple tests: $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$

1. We define the risk at $\{\theta_0, \theta_1\}$ as $R(\psi, \theta) = c_0\mathbb{E}_\theta[\psi(X)]\mathbb{1}(\theta = \theta_0) + \mathbb{E}_\theta[1 - \psi(X)]\mathbb{1}(\theta = \theta_1)$. The overall risk is then, $R(\psi, \theta) = R(\psi, \theta_0) + R(\psi, \theta_1)$. If $c_0 = 1$ then this is just the sum of the type I and type II errors.

2. Let $\Psi_\alpha = \{\psi; \mathbb{E}_{\theta_0}[\psi(X)] \leq \alpha\}$ be the set of all tests with type I error at most α. Define the risk of any $\psi \in \Psi_\alpha$ as $R_\alpha(\psi) = \mathbb{E}_{\theta_1}[1 - \psi(X)]$.

Theorem 8.2 *In both cases above, the optimal test is of the form*

$$\psi(X) = \begin{cases} 1 & \textit{if } \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0} \geq c \\ 0 & \textit{otherwise} \end{cases}$$

where $\frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0}$ *is the likelihood ratio. For the first definition we set* $c = c_0$ *and for the latter we set* $c = c_\alpha$ *such that* $\mathbb{E}_{\theta_0}[\psi(X)] = \alpha$.

8.2.3 Two Definitions of Risk for composite vs composite tests

Similar to above, we define

1. $R(\psi, \Theta_0, \Theta_1) = \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\psi(X)] + \sup_{\theta \in \Theta_1} \mathbb{E}_\theta[1 - \psi(X)]$.
2. $R_\alpha(\psi, \Theta_0, \Theta_1) = \sup_{\theta \in \Theta_1} \mathbb{E}_\theta[1 - \psi(X)]$ where $\psi \in \Psi_\alpha(\Theta_0) = \{\psi \, : \, \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\psi(X)] \leq \alpha\}$.

Definition 8.3 (Minimax Test) *A test* ψ^m *is minimax if*

1. $R(\psi^m, \Theta_0, \Theta_1) = \inf_\psi R(\psi, \Theta_0, \Theta_1)$*, for the first definition above.*
2. $R(\psi^m, \Theta_0, \Theta_1) = \inf_{\psi \in \Psi_\alpha(\Theta_0)} R_\alpha(\psi, \Theta_0, \Theta_1)$*, for the second definition.*

Bayesian Testing: Let $\mathcal{P}_0 = \{p_\theta : \theta \in \Theta_0\}$, $\mathcal{P}_1 = \{p_\theta; \theta \in \Theta_1\}$. Under minimal regularity conditions,

1. The minimax risk for the first definition above is,
$$\sup\{1 - d_{TV}(p_0, p_1) \, : \, p_0 \in \mathrm{conv}(\mathcal{P}_0), p_1 \in \mathrm{conv}(\mathcal{P}_1)\}$$
where $p \in \mathrm{conv}(\mathcal{P}_i) \implies p = \int_{\Theta_i} p_\theta \mathrm{d}\pi(\theta)$ where π is some prior on Θ_i (hence the name "Bayesian" testing ??). The supremum is always realised by some pair of priors on Θ_0, Θ_1.
2. The minimax risk is,
$$\sup \left\{ \inf_{\psi \in \Psi_\alpha(\mathcal{P}_0)} \mathbb{E}_{p_1}[1 - \psi(X)]; p_0 \in \mathrm{conv}(\mathcal{P}_0), p_1 \in \mathrm{conv}(P_1) \right\}.$$
This is realised by a pair of least favourable priors on Θ_0, Θ_1.

To obtain a lower bound on a minimax risk, put priors π_0, π_1 on Θ_0 and Θ_1 and consider the testing problem p_0 vs p_1,

$$p_0 = \int_{\Theta_0} p_\theta \mathrm{d}\pi_0(\theta), \qquad p_1 = \int_{\Theta_1} p_\theta \mathrm{d}\pi_1(\theta).$$

For this simplified problem, the Neyman Pearson result tells us that the minimax risk is attained by the test $\psi(X) = 1$ if $\mathrm{d}p_1/\mathrm{d}p_0 \geq c$ and 0 otherwise. As to how tight the lower bound is depends on how well we choose the priors π_0 and π_1.

In high dimensional literature, we usually treat $\Theta_0 = \{\theta_0\}$ and take Θ_1 to be composite. The risk is $R(\psi, \Theta_0, \Theta_1) = \mathbb{E}_{\theta_0}[\psi(X)] + \sup_{\theta \in \Theta_1} \mathbb{E}_\theta[1 - \psi(X)]$.

Let $n, d \to \infty$ and $\Theta_1 = \Theta_1(n, d)$. A test ψ is said to be *asymptotically powerful* if $\lim_{n \to \infty} R(\psi, \Theta_0(n, d), \Theta_1(n, d)) = 0$ and *powerless* if $\lim_{n \to \infty} R(\psi, \Theta_0(n, d), \Theta_1(n, d)) \geq 1$

8.2.3.1 Some Examples

1. $H_0 = \mathcal{N}_n(0, 1)$ and $H_1 = \mathcal{N}_n(\theta, I)$, $\theta \neq 0$. We formulate H_1 as $\{\mathcal{N}(\theta, I); \|\theta\| \geq \gamma_{n,d}\}$ for some $\gamma_{n,d}$.
   Q: At what rate can $\gamma_{n,d}$ go to zero while keeping the test asymptotically powerful.

2. $Y = X\beta + \epsilon$, $\epsilon = \{\epsilon_1, \ldots, \epsilon_n\} \sim \mathcal{N}_n(0, I)$ and $X_{n,d}$ is of rank $\min(n, d)$. We want to test for $H_0 = (\beta = \mathbf{0})$ vs $H_1 = (\beta \neq \mathbf{0})$.
   A natural test in this set up is to compute $\|PY\|^2$ where P is the column span projection of X. Under the null, $\|PY\| \sim \chi^2_{\min(n,d)}$ and centrality parameter $\|X\beta\|^2$. A test that rejects for large values of $\|PY\|^2$ is asymptotically powerless if $\|X\beta\|^2/\sqrt{min(n, d)} \to 0$.

8.2.4 Lower Bounds for Hypothesis Testing

General Strategy: Put a prior π on Θ_1. The minimax risk is lower bounded by

$$\inf_{\psi} \left(\mathbb{E}_{\theta_0}[\psi(X)] + \int_{\Theta_1} \mathbb{E}_{\theta_1}[(1 - \psi(X))] \mathrm{d}\pi(\theta) \right)$$

and this infimum is achieved by the likelihood ratio test. Denoting $p_1 = \int_{\Theta_1} p_\theta \mathrm{d}\pi$ the test is given by,

$$\psi(X) = \begin{cases} 1 & \text{if } \frac{\mathrm{d}p_1}{p_0} > 1 \\ 0 & \text{otherwise} \end{cases}$$

Denote the likelihood ratio by $L(x) = \mathrm{d}p_1(x)/\mathrm{d}p_0$. Then, we know $1 - d_{TV}(p_1, p_0) = 1 - \frac{1}{2}\mathbb{E}_{\theta_0}[|L(x) - 1|] \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_{\theta_0}[(L(x)^2 - 1)]}$.

For instance, in the regression example above $L(Y) = \mathbb{E}_\pi[\exp(Y^\top X\beta - \frac{\|X\beta\|^2}{2})]$. If $\mathbb{E}[L(x)^2] \to 1$, then the test is asymptotically powerless.