Stat210A: Theoretical Statistics **Lecture Date: September 23th, 2014**

Information Inequality

Lecturer: Michael I. Jordan *Scribe: Weiqiao Han*

1 Variance Bound and Information

Last time, we showed

$$\mathrm{Var}_\theta(\delta) \geq \frac{\mathrm{Cov}_\theta^2(\delta,\psi)}{\mathrm{Var}_\theta(\psi)}. \tag{1}$$

Let $\mathbb{P}$ be dominated, i.e., density exists. Consider $\theta, \theta+\Delta$. First, we want to show $E_{\theta+\Delta}\delta - E_\theta\delta$ is covariance.

Define

$$L(X) = \begin{cases} \frac{p_{\theta+\Delta}(X)}{p_\theta(X)} & p_\theta(X) > 0 \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

$E_{\theta+\Delta}h(x) = \int hp_{\theta+\Delta}d\mu = \int hLp_\theta d\mu = E_\theta h(X)L(X)$.

For $h=1$, $E_\theta L(X) = 1$.

For $h=\delta$, $E_{\theta+\Delta}\delta = E_\theta L\delta$.

Define $\psi(X) = L(X) - 1$. Then $E_\theta\psi(X) = 0$.

So $E_{\theta+\Delta}\delta - E_\theta\delta = E_\theta L\delta - E_\theta\delta = E_\theta\psi\delta = \mathrm{Cov}(\psi,\delta)$.

Now suppose δ is unbiased, i.e., $E_\theta\delta = g(\theta)$. Then $\mathrm{Cov}_\theta(\delta,\psi) = g(\theta+\Delta) - g(\theta)$.

$\mathrm{Var}_\theta\,\psi = E_\theta(\frac{p_{\theta+\Delta}(X)}{p_\theta(X)} - 1)^2 \Rightarrow \mathrm{Var}_\theta(\delta) \geq \frac{(g(\theta+\Delta)-g(\theta))^2}{E_\theta(\frac{p_{\theta+\Delta}(X)}{p_\theta(X)}-1)^2}$. This is called the *Hammersley-Chapman-Robbins inequality*, the longest name in statistics.

The right hand side is large when $p_{\theta+\Delta}(X)$ and $p_\theta(X)$ are not every distinct. The numerator says it's easy to estimate smooth things.

Heuristically,

$$\mathrm{Var}_\theta(\theta) \geq \frac{(\frac{g(\theta+\Delta)-g(\theta)}{\Delta})^2}{E_\theta(\frac{p_{\theta+\Delta}(X)-p_\theta(X)}{p_\theta(X)}/\Delta)} \to \frac{(g'(\theta))^2}{E_\theta(\frac{\partial\log p_\theta(X)}{\partial\theta})^2} \text{ as } \Delta\to 0 \tag{3}$$

The denominator $E_\theta(\frac{\partial\log p_\theta(X)}{\partial\theta})^2$ is called *Fisher Information.*

Less heuristically, let $g(\theta) = E_\theta\delta$. Take $\psi = \frac{\partial\log p_\theta}{\partial\theta}$. Assuming Regularity,

$$g'(\theta) = \frac{\partial}{\partial\theta}\int\delta p_\theta d\mu = \int\delta\frac{\partial}{\partial\theta}p_\theta d\mu = \int\delta\frac{\partial\log p_\theta}{\partial\theta}p_\theta d\mu = \int\delta\psi p_\theta d\mu \Rightarrow E_\theta\delta\psi = g'(\theta). \tag{4}$$

When $\delta = 1$, $E_\theta \psi = 0$. So $\mathrm{Cov}_\theta(\delta, \psi) = E_\theta \delta \psi = g'(\theta)$.

Use Cauchy-Schwarz, $\mathrm{Var}_\theta(\theta) \geq \frac{(g'(\theta))^2}{\mathrm{Var}_\theta(\psi)}$, where $\mathrm{Var}_\theta(\psi) = E_\theta(\frac{\partial \log p_\theta(X)}{\partial \theta})^2 = I(\theta)$, which is the definition of Fisher Information, and we can write the above inequality as $\mathrm{Var}_\theta(\theta) \geq \frac{(g'(\theta))^2}{I(\theta)}$.

With more regularity (e.g., domain convex), $1 = \int p_\theta d\mu \Rightarrow 0 = \int \frac{\partial^2 p_\theta}{\partial \theta^2} d\mu = \int \frac{\partial}{\partial \theta}(\frac{\partial}{\partial \theta p_\theta}) d\mu = \int \frac{\partial}{\partial \theta}(\frac{\partial \log p_\theta}{p_\theta} p_\theta) d\mu = \int \frac{\partial^2 \log p_\theta}{\partial \theta^2} p_\theta d\mu + \int (\frac{\partial \log p_\theta}{\partial \theta})^2 p_\theta d\mu \Rightarrow I(\theta) = -E_\theta(\frac{\partial^2 \log p_\theta}{\partial \theta^2})$.

$I(\theta)$ represents the average curvature of likelihood. Low curvature makes problem hard, and vice versa.

Fisher Information is not invariant under reparametrization, but information inequality is invariant under reparametrization.

To see this, let $h : \Xi \to \Omega, h(\xi) = \theta$. Let $Q_\xi = P_{h(\xi)}, \tilde{g}(\xi) = g(h(\xi)), q_\xi(x) = P_{h(\xi)}(x)$.

$$\tilde{I}(\xi) = \tilde{E_\xi}(\frac{\partial \log q_\xi(X)}{\partial \xi})^2 = \tilde{E_\xi}(\frac{\partial \log p_{h(\xi)}(X)}{\partial \xi})^2 = \tilde{E_\xi}(\frac{\partial \log p_\theta(X)}{\partial \theta}\frac{d\theta}{d\xi})^2 = (h'(\xi))^2 \tilde{E_\xi}(\frac{\partial \log p_\theta(X)}{\partial \theta})^2 \tag{5}$$

$$= (h'(\xi))^2 E_\theta(\frac{\partial \log p_\theta(X)}{\partial \theta})^2 = (h'(\xi))^2 I(\theta) \tag{6}$$

In contrast, $\tilde{g}'(\xi) = h'(\xi) g'(\theta) \Rightarrow \frac{(\tilde{g}'(\xi))^2}{I(\theta)} = \frac{(\tilde{g}'(\xi))^2}{I(\theta)}$.