Stats 300b: Theory of Statistics **Winter 2017**

Lecture 4 – January 19

Lecturer: John Duchi *Scribe: Xiaotong Suo*

Warning: *these notes may contain factual errors*

Reading: VDV Chapter 3

Outline of the lecture:

I Asymptotic Normality & Fisher information

(a) Basic Asymptotic Normality result

(b) Fisher information

i. Definitions, Examples

ii. Information Inequality (Cramer Rao Bound)

1 The basic normality result

As in the previous lecture, we assume as always that we have a model family $\{P_\theta\}_{\theta\in\Theta}$, each distribution P_θ having density p_θ with respect to some base measure μ on $\mathcal{X}$. We also use our usual notation that $\ell_\theta(x) := \log p_\theta(x)$ is the log-likelihood. In order to get our asymptotic normality results, we require a number of conditions on the smoothness of the log-likelihood so as to perform appropriate Taylor expansions. Recall briefly that if a function $f : \mathbb{R}^d \to \mathbb{R}$ that is k-times continuously differentiable, then

$$f(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v + \cdots + \mathrm{Rem}(x+v)[v^{\otimes k}],$$

where $v^{\otimes k}$ indicates the k-th order tensor of v, i.e. the tensor in $\mathbb{R}^{n^k}$ indexed by $[v^{\otimes k}]_{i_1,\ldots,i_k} = v_{i_1} \cdots v_{i_k}$, and Rem is a remainder function such that $\mathrm{Rem}(x+v)$ acts linearly on the argument $v^{\otimes k}$ and $\mathrm{Rem}(x+v) \to 0$ as $v \to 0$. In some instances, we may say stronger things, such as if the $(k-1)$th derivative is Lipschitz. To keep things concrete, suppose $\nabla^2 f$ is Lipschitz, meaning that $\left|\!\left|\!\left| \nabla^2 f(x) - \nabla^2 f(y) \right|\!\right|\!\right|_{\mathrm{op}} \le M \|x-y\|$ for some $M < \infty$. In this case, we may take the remainder term to satisfy $|\!|\!| \mathrm{Rem}(x+v) |\!|\!|_{\mathrm{op}} \le M \|v\|$.

With these preliminaries out of the way, we begin with the major theorem we would like to prove, which is that so long as the log likelihood $\ell_\theta(x) := \log p_\theta(x)$ is suitably smooth and that the MLE $\widehat{\theta}_n$ is consistent, then $\widehat{\theta}_n$ is asymptotically normal.

Theorem 1. *Let $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$ where $\theta_0 \in \operatorname{int}\Theta$. Assume that $\ell_\theta(x) = \log p_\theta(x)$ is smooth enough that $\mathbb{E}_{\theta_0}[\nabla \ell_{\theta_0} \nabla \ell_{\theta_0}^T]$ exists and that the Hessian $\nabla^2 \ell_\theta(x)$ is $M(x)$-Lipschitz in θ, that is,*

$$\left|\!\left|\!\left| \nabla^2 \ell_{\theta_1}(x) - \nabla^2 \ell_{\theta_2}(x) \right|\!\right|\!\right|_{\mathrm{op}} \le M \|\theta_1 - \theta_2\|,$$

where $\mathbb{E}_{\theta_0}[M(X)^2] < \infty$. Assume additionally that the MLE $\widehat{\theta}_n$ is consistent, $\widehat{\theta}_n \xrightarrow{p} \theta_0$.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0}^{-1})$$

where $I_\theta = \mathbb{E}_\theta[\nabla \ell_\theta \nabla \ell_\theta^T]$ is the Fisher information.

Proof Let $\widehat{r}(x) \in \mathbb{R}^{d\times d}$ be the remainder matrix in Taylor expansion of the gradients of the individual log likelihood terms around θ_0 guaranteed by Taylor's theorem (which certainly depends on $\widehat{\theta}_n - \theta_0$), that is,

$$\nabla \ell_{\widehat{\theta}_n}(x) = \nabla \ell_{\theta_0}(x) + \nabla^2 \ell_{\theta_0}(x)(\widehat{\theta}_n - \theta_0) + \widehat{r}(x)(\widehat{\theta}_n - \theta_0),$$

where by Taylor's theorem $\|\widehat{r}(x)\|_{\text{op}} \le M(x)\|\widehat{\theta}_n - \theta_0\|$. Writing this out using the empirical distribution and that $\widehat{\theta}_n = \operatorname{argmax}_\theta P_n \ell_\theta(X)$, we have

$$\nabla P_n \ell_{\widehat{\theta}_n} = 0 = P_n \nabla \ell_{\theta_0} + P_n \nabla^2 \ell_{\theta_0}(\widehat{\theta}_n - \theta_0) + P_n \widehat{r}(X)(\widehat{\theta}_n - \theta_0). \tag{1}$$

But of course, expanding the term $P_n \widehat{r}(X) \in \mathbb{R}^{d\times d}$, we find that

$$P_n \widehat{r}(X) = \frac{1}{n}\sum_{i=1}^n \widehat{r}(X_i) \quad \text{and} \quad \|P_n \widehat{r}\|_{\text{op}} \le \underbrace{\frac{1}{n}\sum_{i=1}^n M(X_i)}_{\stackrel{a.s.}{\to} \mathbb{E}_{\theta_0}[M(X)]} \underbrace{\|\widehat{\theta}_n - \theta_0\|}_{\stackrel{p}{\to} 0} = o_P(1).$$

In particular, revisiting expression (1), we have

$$\begin{aligned} 0 &= P_n \nabla \ell_{\theta_0} + P_n \nabla^2 \ell_{\theta_0}(\widehat{\theta}_n - \theta_0) + o_P(1)(\widehat{\theta}_n - \theta_0). \\ &= P_n \nabla \ell_{\theta_0} + \left(P_{\theta_0} \nabla^2 \ell_{\theta_0} + (P_n - P_{\theta_0})\nabla^2 \ell_{\theta_0} + o_P(1)\right)(\widehat{\theta}_n - \theta_0). \end{aligned}$$

The strong law of large numbers guarantees that $(P_n - P_{\theta_0})\nabla^2 \ell_{\theta_0} = o_P(1)$, and multiplying each side by $\sqrt{n}$ yields

$$\sqrt{n}(P_{\theta_0}\nabla^2 \ell_{\theta_0} + o_P(1))(\widehat{\theta}_n - \theta_0) = -\sqrt{n} P_n \nabla \ell_{\theta_0}.$$

Applying Slutsky's theorem gives the result: indeed, we have $T_n = \sqrt{n} P_n \nabla \ell_{\theta_0}$ satisfies $T_n \stackrel{d}{\to} \mathsf{N}(0, I_{\theta_0})$ by the central limit theorem, and noting that $P_{\theta_0}\nabla^2 \ell_{\theta_0} + o_P(1)$ is eventually invertible gives

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \stackrel{d}{\to} \mathsf{N}(0, (P_{\theta_0}\nabla^2 \ell_{\theta_0})^{-1} I_{\theta_0} (P_{\theta_0}\nabla^2 \ell_{\theta_0})^{-1})$$

as desired. $\square$

2 Fisher Information

Definition 2.1. *For a model family $\{P_\theta\}, \theta \in \Theta$ on $\mathcal{X}$. The fisher information is $I_\theta = I(\theta) = \mathbb{E}_\theta[\nabla l_{\theta_0} \nabla l_{\theta_0}^T] = \mathrm{Cov}_\theta(\nabla l_\theta)$. When ∇ and $\mathbb{E}$ are interchangable, then*

$$I_\theta = -\mathbb{E}[\nabla^2 \log P_\theta(x)]$$

Example 1: Normal location family. $\{\mathsf{N}(\theta, \sigma^2)\}_{\theta\in\mathbb{R}}$, where θ is unknown,

$$\frac{\partial}{\partial\theta}\log P_\theta(x) = \frac{\theta - x}{\sigma^2}.$$

Thus,

$$\mathbb{E}[(\frac{\partial}{\partial\theta}\log P_\theta(x))^2] = \frac{\mathrm{Var}(X)}{\sigma^4} = \frac{1}{\sigma^2}$$

Heuristically speaking, if $\sigma^2 \to 0$, then it's easy to estimate the mean. If $\sigma^2 \to 0$, then it's hard to estimate θ because heavy tails. So fisher information roughly tells us how easy or hard to estimate a parameter. ♣

Remark What if we care about $\tau = h(\theta)$ instead of θ? Then inverse function theorem yields:

$$\frac{\partial}{\partial \tau} h^{-1}(\tau)(h(\theta)) = \frac{1}{h'(h^{-1}(\tau))} = \frac{1}{h'(\theta)}$$

Therefore, we have

$$I(\tau) = I(h(\theta)) = \frac{I(\theta)}{h'(\theta)^2}$$

when $h'(\theta) \neq 0$. We can see this using the chain rule:

$$\begin{aligned}\frac{\partial}{\partial \tau} \log P_{h^{-1}(\tau)} &= \frac{\partial}{\partial \tau} \log P_\theta \\ &= \frac{\partial \log P_\theta}{\partial \theta} \frac{\partial \theta}{\partial \tau} \\ &= \frac{\partial \log P_\theta}{\partial \theta} \frac{\partial h^{-1}(\tau)}{\partial \tau}\end{aligned}$$

Example 2: Normal location $h(\theta) = \theta^2$. $h'(\theta) = 2\theta$, so

$$I(\theta^2) = \frac{1}{4\theta^2} I(\theta) = \frac{1}{4\theta^2\sigma^2}$$

In particular, as $\theta \to 0$, $I(\theta) \to \infty$. Suppose $\theta = 0$, let $\hat{\theta}_n = (\frac{1}{n}\sum_{i=1}^n x_i)^2$, then

$$n(\frac{1}{n}\sum_{i=1}^n x_i)^2 = (\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i)^2 \xrightarrow{d} z^2$$

where $z \sim N(0, \sigma^2)$. Therefore, we have an order of n convergence. In this case, our estimator converges faster than $\sqrt{n}$. So heuristically speaking, if we have a higher fisher information, our estimator is somehow better. ♣

Additivity Property of Fisher information If $x_1 \sim P_\theta$, $x_2 \sim Q_\theta$, x_1, x_2 independent, then $I_{x_1,x_2}(\theta) = I_{x_1}(\theta) + I_{x_2}(\theta)$.
Proof Since x_1 and x_2 are independent,

$$\mathrm{Cov}(\nabla \log P_\theta(x_1) + \nabla \log q_\theta(x_2)) = \mathrm{Cov}(\nabla \log P_\theta(x_1)) + \mathrm{Cov}(\nabla \log q_\theta(x_2)) = I_1 + I_2$$

□

Corollary 2. *If $x_i \overset{\text{iid}}{\sim} p_\theta$, $I(\theta) = Info(x_i)$, then $I_n(\theta) = nI(\theta)$.*

Information Inequality We start with proving covariance "lower bound".

For any decision procedure $\delta : \mathcal{X} \to \mathbb{R}$ and any function $\psi : \mathcal{X} \to \mathbb{R}$, we have

$$\mathrm{Var}(\delta) \geq \frac{\mathrm{Cov}(\delta, \psi)^2}{\mathrm{Var}(\psi)}$$

Proof The proof uses Cauchy Schwarz.

$$\begin{aligned}\mathrm{Cov}(\delta, \psi) &= \mathbb{E}[(\delta - \mathbb{E}\delta)(\psi - \mathbb{E}\psi)] \\ &\leq \{\mathbb{E}[(\delta - \mathbb{E}\delta)^2]\}^{1/2}\{\mathbb{E}[(\delta - \mathbb{E}\delta)^2]\}^{1/2} \\ &= \{\mathrm{Var}(\delta)\}^{1/2}\{\mathrm{Var}(\psi)\}^{1/2}\end{aligned}$$

Therefore, we have

$$\mathrm{Var}(\delta) \geq \frac{\mathrm{Cov}(\delta, \psi)^2}{\mathrm{Var}(\psi)}$$

□

Theorem 3. *(1 dimensional information inequality). Assume that* $\mathbb{E}_\theta[\delta] = g(\theta)$ *ids differentiable at* θ *and* P_θ *is regular enough that* $\dot{P}_\theta = \frac{\partial}{\partial\theta}P_\theta$, $\int P_\theta d\mu = \frac{\partial}{\partial\theta}\int P_\theta d\mu = 0$, *and* $\int \delta(x)P_\theta\dot{(}x)d\mu = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[\delta] = g'(\theta)$. *Then*

$$\mathrm{Var}_{\theta_0}(\delta) \geq \frac{(g'(\theta_0))^2}{I(\theta_0)}$$

Why do we care about this? Suppose that $\mathbb{E}_\theta[\delta] = 0$ and $g(\theta) = \theta$. Therefore, $g'(\theta) = 1$ and

$$\mathbb{E}_\theta((\delta - \theta)^2) \geq \frac{1}{I(\theta)}$$

In 1 dimension, any unbiased estimator has $MSE \geq \frac{1}{I(\theta)}$. However, this result is not true when our estimator is biased.

Proof Take $\psi(x) = \frac{\partial}{\partial\theta}\log P_\theta(x) = \frac{\partial}{\partial\theta}l_\theta(x) = \frac{\dot{P}_\theta(x)}{P_\theta(x)}$. By covariance inequality,

$$\begin{aligned}\underset{\theta}{\mathrm{Cov}}(\delta, \psi) &= \mathbb{E}_\theta[(\delta - g(\theta))(\psi - E[\psi])] \\ &= \mathbb{E}_\theta[\delta, \psi] \\ &= \mathbb{E}_\theta[\delta\frac{\dot{P}_\theta(x)}{P_\theta(x)}] \\ &= \int \delta(x)\dot{P}_\theta(x)\frac{P_\theta(x)}{P_\theta(x)}d\mu(x) \\ &= \frac{\partial}{\partial\theta}\int \delta P_\theta d\mu \\ &= \frac{\partial}{\partial\theta}\mathbb{E}_\theta[\delta] = g'(\theta)\end{aligned}$$

Therefore, we have

$$\mathrm{Var}(\delta) \geq \frac{\mathrm{Cov}(\delta, \psi)^2}{\mathrm{Var}(\psi)} = \frac{g'(\theta)}{I(\theta)}$$

□

Remark This result is unsatisfying in two senses:

1. Tied to mean square error (MSE)
2. Requires unbiasedness

We will cover a major theorem later in the class, which is better than this result. Roughly speaking, we will show that

$$\mathbb{E}_\theta[L(\sqrt{n}(\widehat{\theta}_n - \theta))] \geq \mathbb{E}[L(Z)], \quad Z \sim \mathsf{N}(0, I(\theta)^{-1}), \quad \forall \widehat{\theta}_n$$

and symmetric quasi-convex $L : \mathbb{R}^d \to \mathbb{R}_+$. (In fact, this holds for *almost* all θ, though not necessarily for all θ.) More precisely, the following result holds true. Let $h \in \mathbb{R}^d$ and $\theta_0 \in \Theta$ be arbitrary, where $\Theta \subset \mathbb{R}^d$ is open. In addition, assume that the distributions P_θ have log-likelihoods smooth enough that the conditions of Theorem **??** are satisfied. Then for any sequence of estimators $\widehat{\theta}_n : \mathcal{X}^n \to \Theta$ and any quasi-convex symmetric loss $L : \mathbb{R}^d \to \mathbb{R}_+$,

$$\liminf_{c\to\infty} \liminf_{n} \sup_{\|h\|\leq c} \mathbb{E}_{\theta_0 + h/\sqrt{n}} \left[L(\sqrt{n}(\widehat{\theta}_n - (\theta_0 + h/\sqrt{n})))\right] \geq \mathbb{E}[L(Z)], \quad Z \sim \mathsf{N}(0, I_{\theta_0}^{-1}).$$

That is, under perturbations of the true parameter θ_0 by amounts shrinking as $1/\sqrt{n}$, we have a locally difficult estimation problem. (Here $\mathbb{E}_\theta$ denotes expectation taken w.r.t. i.i.d. sampling under P_θ.)

Multidimensional Information Inequality We now generalize the result to $\theta \in \mathbb{R}^d$.

Lemma 4. *Let $\delta : \mathcal{X} \to \mathbb{R}$, $\psi : \mathcal{X} \to \mathbb{R}^d$, $\mathbb{E}_\theta(\psi) = 0$. Define $\gamma = [\mathrm{Cov}(\delta, \psi_j]_{j=1}^d$, $C = \mathrm{Cov}_\theta(\psi) = \mathbb{E}_\theta[\psi\psi^T] = 0$. Then*

$$\mathrm{Var}(\delta) \geq \gamma^T C^{-1} \gamma.$$

Proof Let $v \in \mathbb{R}^d$ be arbitrary. By 1 dimensional covariance inequality,

$$\begin{aligned}\mathrm{Var}(\delta) &\geq \frac{\mathrm{Cov}(\delta, v^T\psi)^2}{\mathrm{Var}(v^T\psi)} \\ &= \frac{(\gamma^T v)^2}{v^T C v} = \gamma^T C^{-1} \gamma\end{aligned}$$

The last inequality uses the following fact:

Fact 5. *If $A > 0$, then*

$$\sup_v \frac{(v^T u)^2}{v^T A v} = u^T A^{-1} u$$

Proof We first use Cauchy Schwartz.

$$(v^T u)^2 = (A^{1/2}v)^T (A^{-1/2}u)^2 \leq \|A^{1/2}v\|_2^2 \|A^{-1/2}u\|^2 = v^T A v u^T A^{-1} u$$

Therefore, for all v,

$$\frac{(v^T u)^2}{v^T A v} \leq u^T A^{-1} u$$

Now, we take $v = A^{-1}u$ to achieve this upper bound. $\triangledown$

$\square$

Theorem 6. *Let $g(\theta) = \mathbb{E}_\theta[\delta] \in \mathbb{R}^d$, with lots of regularity. Then we have*

$$\mathrm{Var}_\theta(\delta) \geq \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)$$

where $I(\theta) = \mathbb{E}_\theta[\nabla l_\theta \nabla l_\theta^T]$.

Proof Let $\psi = \nabla l_\theta(x)$ in covariance lower bound. $\mathbb{E}_\theta[\psi] = 0$, $\mathrm{Cov}(\delta, \psi) = \mathbb{E}[\delta \nabla l_\theta] = \nabla \mathbb{E}_\theta[\delta] = \frac{\nabla \delta(\theta)}{\gamma}$. $\square$

Corollary 7. *If $\hat{\theta} : x \to \Theta \in \mathbb{R}^d$ is unbiased,*

$$\mathbb{E}_\theta[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] \geq I(\theta)^{-1}$$

Proof Take $v \in \mathbb{R}^d$, $\delta(x) = v^T \hat{\theta}$. Then

$$\begin{aligned} \mathbb{E}(\delta) &= v^T \theta = g(\theta) \\ \implies \nabla g(\theta) &= v \end{aligned}$$

So

$$\begin{aligned} \mathbb{E}[(v^T(\hat{\theta} - \theta))^2] &\geq v^T I(\theta)^{-1} v \\ &= \mathbb{E}[v^T (\hat{\theta} - \theta)(\hat{\theta} - \theta)^T v] \end{aligned}$$

$\square$