Stats 300b: Theory of Statistics **Winter 2018**

Lecture 3 – January 16

Lecturer: Yu Bai/John Duchi *Scribe: Shuangning Li, Theodor Misiakiewicz*

Warning: *these notes may contain factual errors*

Reading: VDV Chapter 5.1-5.6; ELST Chapter 7.1-7.3

Outline of Lecture 2:

1. Basic consistency and identifiability
2. Asymptotic Normality
   (a) Taylor expansions
   (b) Classical log-likelihood & asymptotic normality
   (c) Fisher Information

Recap of Delta Method Last lecture, we discussed the Delta Method (aka Taylor expansions). The basic idea was as follows:

Claim 1. *If $r_n(T_n - \theta) \stackrel{d}{\to} T$, and $\phi : \mathbb{R}^d \to \mathbb{R}^k$ is smooth, then $r_n(\phi(T_n) - \phi(\theta)) \to \phi'(\theta)T$, if $\phi'(\theta) \neq 0$.*

Idea of proof:

$$\begin{aligned} r_n(\phi(T_n) - \phi(\theta)) &= r_n(\phi'(\theta)(T_n - \theta) + o_p(T_n - \theta)) \\ &= r_n(\phi'(\theta)(T_n - \theta)) + o_p(r_n(T_n - \theta)) \\ &= r_n(\phi'(\theta)(T_n - \theta)) + o_p(1) \\ &\stackrel{d}{\to} \phi'(\theta)T. \end{aligned}$$

Notation: (from now on) Given distribution P on $\mathcal{X}$, function $f : \mathcal{X} \to \mathbb{R}^d$,

$$Pf := \int f dP = \int_{\mathcal{X}} f(x) dP(x) = \mathbb{E}_P[f(x)]$$

Example 1 (Empirical distributions)**:** Consider the observations $x_1, x_2, \ldots, x_n \in \mathcal{X}$. Let the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n 1_{x_i}$. For any set $A \subseteq \mathcal{X}$,

$$P_n(A) = \frac{1}{n}|\{i \in [n] : x_i \in A\}| = P_n 1_{\{x \in A\}}.$$

Hence for any function f, $P_n f = \frac{1}{n}\sum_{i=1}^n f(x_i)$. ♣

Taylor expansions

1. **Real-valued functions**

   For $f : \mathbb{R}^d \to \mathbb{R}$ differentiable at $x \in \mathbb{R}^d$,

   $$f(y) = f(x) + \nabla f(x)^T (y - x) + o(\|y - x\|). \text{ (Remainder version)}$$

   $$f(y) = f(x) + \nabla f(\tilde{x})^T (y - x). \text{ (Mean value version)}$$

   If f is twice differentiable,

   $$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + o(\|y - x\|^2). \text{ (Remainder version)}$$

   $$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(\tilde{x})(y - x). \text{ (Mean value version)}$$

2. **Vector-valued functions**

   Let $f : \mathbb{R}^d \to \mathbb{R}^k$, $f(x) = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{bmatrix}$. Define $Df(x) = \begin{bmatrix} \nabla f_1^T(x) \\ \nabla f_2^T(x) \\ \vdots \\ \nabla f_k^T(x) \end{bmatrix} \in \mathbb{R}^{k \times d}$ to be the Jacobian of f.

   Then,

   $$f(y) = f(x) + Df(x)(y - x) + o(\|y - x\|). \text{ (Remainder version)}$$

   But for mean value version, we don't necessarily have $\tilde{x}$ such that

   $$f(y) = f(x) + Df(\tilde{x})(y - x).$$

Example 2 (Failure of mean value version)**:** Let $f : \mathbb{R} \to \mathbb{R}^k$, $f(x) = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^k \end{bmatrix}$, then $Df(x) = \begin{bmatrix} 1 \\ 2x \\ \\ kx^{k-1} \end{bmatrix}$. Take $x = 0, y = 1$, then $f(y) - f(x) = \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$. Yet $Df(\tilde{x}) = \begin{bmatrix} 1 \\ 2\tilde{x} \\ \vdots \\ k\tilde{x}^{k-1} \end{bmatrix} \neq \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$. ♣

Example 3 (Quantitative continuity guarantees)**:** Recall the operator norm of A is

$$\|A\|_{op} = \sup_{\|u\|_2 = 1} \|Au\|_2,$$

this implied that $\|Ax\|_2 \leq \|A\|_{op}\|x\|_2$. For $f : \mathbb{R}^d \to \mathbb{R}^k$, differentiable, assume that Df is $L-$Lipschitz, i.e. $\|Df(x) - Df(y)\|_{op} \leq L\|x - y\|_2$. (Roughly, this means that $\|D^2 f(x)\| \leq L$.)

Claim 2. *We have*

$$f(y) = f(x) + Df(x)(y - x) + R(y - x),$$

where R is a remainder matrix (depending on x, y) that satisfy $\|R\|_{\mathrm{op}} \leq \frac{L}{2}\|y - x\|$ and $\|R(y - x)\| \leq \frac{L}{2}\|y - x\|^2$.

Proof Define $\phi_i(t) = f_i((1-t)x + ty)$, $\phi_i : [0,1] \to \mathbb{R}$. Note that $\phi_i(0) = f_i(x)$, $\phi_i(1) = f_i(y)$, and $\phi_i' = \big(\nabla f_i((1-t)x + ty)\big)^T (y-x)$. Then

$$Df((1-t)x+ty)(y-x) = \begin{bmatrix} \nabla f_1^T((1-t)x+ty) \\ \nabla f_2^T((1-t)x+ty) \\ \vdots \\ \nabla f_k^T((1-t)x+ty) \end{bmatrix} (y-x) = \begin{bmatrix} \phi_1'(t) \\ \phi_2'(t) \\ \vdots \\ \phi_k'(t) \end{bmatrix}.$$

Since $\phi_i(1) - \phi_i(0) = \int_0^1 \phi_1'(t)dt$,

$$\begin{aligned} f(y) - f(x) &= \int_0^1 Df((1-t)x+ty)(y-x)dt \\ &= \int_0^1 \big(Df((1-t)x+ty) - Df(x)\big)(y-x)dt + Df(x)(y-x). \end{aligned}$$

To bound the remainder term,

$$\begin{aligned} \| \int_0^1 \big(Df((1-t)x+ty) - Df(x)\big)(y-x)dt\| &\le \int_0^1 \|\big(Df((1-t)x+ty) - Df(x)\big)(y-x)\|dt \\ &\le \int_0^1 \|Df((1-t)x+ty) - Df(x)\|_{op}\|(y-x)\|dt \\ &\le \int_0^1 L\|t(y-x)\|\|(y-x)\|dt \\ &\le \int_0^1 Lt\|(y-x)\|^2 dt \\ &= \frac{L}{2}\|(y-x)\|^2. \end{aligned}$$

□

♣

Consistency and asymptotic distribution:

Setting:

1. We have some model family $\{P_\theta\}_{\theta\in\Theta}$ of distributions on $\mathcal{X}$, where $\Theta \subseteq \mathbb{R}^d$. Also, assume all P_θ have density p_θ with respect to base measure μ on $\mathcal{X}$, i.e. $p_\theta = \frac{dP_\theta}{d\mu}$.

2. We consider the log-likelihood of the distribution $\ell_\theta(x) = \log p_\theta(x)$, with

$$\nabla \ell_\theta(x) := \left[\frac{\partial}{\partial \theta_j} \log p_\theta(x)\right]_{j=1}^d \in \mathbb{R}^d$$

$$\nabla^2 \ell_\theta(x) := \left[\frac{\partial^2}{\partial \theta_i \theta_j} \log p_\theta(x)\right]_{i,j=1}^d \in \mathbb{R}^{d\times d}$$

For simplicity, we will denote: $\dot{\ell}_\theta \equiv \nabla \ell_\theta(x)$ and $\ddot{\ell}_\theta \equiv \nabla^2 \ell_\theta(x)$.

The gradient of the log-likelihood is often called the "score function." We will use this term to refer to $\nabla \ell_\theta(x)$ throughout future lectures.

3. Observe $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$ where θ_0 is unknown. Our goal is to estimate θ_0.

4. A standard estimator is to choose $\hat{\theta}_n$ to maximize the "likelihood," i.e. the probability of the data.

$$\hat{\theta}_n \in \operatorname*{argmax}_{\theta\in\Theta} P_n \ell_\theta(x)$$

Main questions:

1. Consistency: does $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to +\infty$?

2. Asymptotic distribution: does $r_n(\hat{\theta}_n \xrightarrow{p} \theta_0)$ converge in distribution ?

3. Optimality ? (in the next lecture)

Consistency:

Definition 0.1 (Identifiability)**.** *A model $\{P_\theta\}_{\theta\in\Theta}$ is identifiable if $P_{\theta_1} \neq P_{\theta_2}$ for all $\theta_1, \theta_2 \in \Theta$ ($\theta_1 \neq \theta_2$).*

Equivalently, $D_{\text{kl}}\left(P_{\theta_1} \| P_{\theta_2}\right) > 0$ when $\theta_1 \neq \theta_2$. Recall that $D_{\text{kl}}\left(P_{\theta_1} \| P_{\theta_2}\right) = \int \log \frac{dP_{\theta_1}}{dP_{\theta_2}} dP_{\theta_1}$.
Note that $P_{\theta_1} \neq P_{\theta_2}$ means that $\exists$ set $A \subseteq \mathcal{X}$ such that $P_{\theta_1}(A) \neq P_{\theta_2}(A)$.

Now that we have established what both identifiability and consistency mean, we can prove a basic result regarding the finite consistency of the Maximum Likelihood estimator (MLE).

Proposition 3 (Finite Θ consistency of MLE)**.** *Suppose $\{P_\theta\}_{\theta\in\Theta}$ is identifiable and* $\operatorname{card}\Theta < \infty$. *Then, if $\hat{\theta}_n := \operatorname{argmax}_{\theta\in\Theta} P_n \ell_\theta(x)$, $\hat{\theta}_n \xrightarrow{p} \theta_0$ when $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$.*

Proof of Proposition By the Strong Law of Large Numbers, we know that $P_n \ell_\theta(x) \xrightarrow{a.s.} P_{\theta_0} \ell_\theta(x)$ when $x_i \overset{\text{iid}}{\sim} P_{\theta_0}$.

$$\begin{aligned} P_{\theta_0}\ell_{\theta_0}(x) - P_{\theta_0}\ell_\theta(x) &= \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(x)}{p_\theta(x)}\right] \\ &= D_{\text{kl}}\left(P_{\theta_0} \| P_\theta\right) \end{aligned}$$

We know that $D_{\text{kl}}\left(P_{\theta_0} \| P_\theta\right) > 0$ unless $\theta = \theta_0$. Combining this remark with $P_n \ell_{\theta_0}(x) - P_n \ell_\theta(x) \xrightarrow{a.s.} D_{\text{kl}}\left(P_{\theta_0} \| P_\theta\right)$, we deduce that there exists $N(\theta)$ such that for all $n > N(\theta)$, we have $P_n \ell_{\theta_0}(x) - P_n \ell_\theta(x) > 0$ with probability 1.

It follows that for $n > \max_{\theta\in\Theta, \theta\neq\theta_0} N(\theta)$, we have $P_n \ell_{\theta_0}(x) > P_n \ell_\theta(x)$ for all $\theta \neq \theta_0$. Therefore $\hat{\theta}_n = \theta_0$ and we conclude that, for sufficiently large n and finite Θ, we have $\hat{\theta}_n = \theta_0$ "eventually." $\square$

Remark The above result can fail for Θ infinite even if Θ is countable.

Uniform law: One sufficient condition often used for consistency results is a uniform law, i.e. for $x_i \overset{\text{iid}}{\sim} P$, we have $\sup_{\theta\in\Theta} |P_n \ell_\theta - P\ell_\theta| \xrightarrow{p} 0$. In this case, if $P_{\theta_0}\ell_\theta < P_{\theta_0}\ell_{\theta_0} - 2\epsilon$ and $\sup_{\theta\in\Theta} |P_n \ell_\theta - P_{\theta_0}\ell_\theta| \leq \epsilon$, then $\hat{\theta}_n \neq \theta$. We will have:

$$\hat{\theta}_n \in \{\theta : P_{\theta_0}\ell_\theta \geq P_{\theta_0}\ell_{\theta_0} - 2\epsilon\}$$

Now, that we have established some basic definitions and results regarding the consistency of estimators, we turn our attention to understanding their asymptotic behavior.

Asymptotic Normality via Taylor Expansions:

Definition 0.2 (Operator norm)**.** $|||A|||_{\text{op}} := \sup_{\|u\|_2 \leq 1} \|Au\|_2$.
Note: $A \in \mathbb{R}^{k \times d}, u \in \mathbb{R}^d$ *and* $\|Ax\|_2 \leq |||A|||_{\text{op}} \|x\|_2$.

Before we do anything, we have to make several assumptions.

1. We have a "nice, smooth" model, i.e. the Hessian is Lipschitz-continuous. To be rigorous, the following must hold:

$$\left|\left|\left|\nabla^2 \ell_{\theta_1}(x) - \nabla^2 \ell_{\theta_2}(x)\right|\right|\right|_{\text{op}} \leq M(x) \|\theta_1 - \theta_2\|_2 \qquad \mathbb{E}_\theta[M^2(x)] < \infty$$

2. The MLE, $\hat{\theta}_n \in \text{argmax}_{\theta \in \Theta} P_n \ell_\theta(x)$, is consistent, i.e. $\hat{\theta}_n \xrightarrow{p} \theta_0$ under P_{θ_0}.

3. Θ is a convex set.

Theorem 4. *Let* $x_i \overset{\text{iid}}{\sim} P_{\theta_0}$, $\hat{\theta}_n$ *be the MLE (i.e.* $\nabla P_n \ell_{\hat{\theta}_n} = 0$*) and assume the conditions stated above. Then,* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, (P_{\theta_0}\nabla^2\ell_{\theta_0})^{-1} P_{\theta_0} \nabla\ell_{\theta_0} \nabla\ell_{\theta_0}^T (P_{\theta_0}\nabla^2\ell_{\theta_0})^{-1})$.

Remark Let us rewrite the asymptotic variance. Given that $\nabla^2 \ell_\theta = \nabla\left(\frac{\nabla p_\theta}{p_\theta}\right) = \frac{\nabla^2 p_\theta}{p_\theta} - \frac{\nabla p_\theta \nabla p_\theta^T}{p_\theta^2}$:

$$\mathbb{E}_\theta\left[\frac{\nabla^2 p_\theta}{p_\theta}\right] = \int \frac{\nabla^2 p_\theta}{p_\theta} p_\theta d\mu = \int \nabla^2 p_\theta d\mu = \nabla^2 \int p_\theta d\mu = 0$$

As a result:

$$\mathbb{E}_\theta[\nabla^2 \ell_\theta] = -\mathbb{E}_\theta\left[\left(\frac{\nabla p_\theta}{p_\theta}\right)\left(\frac{\nabla p_\theta}{p_\theta}\right)^T\right] = -\underset{\theta}{\text{Cov}}(\nabla \ell_\theta(x))$$

We define the <u>Fisher Information</u> as $I_\theta := \mathbb{E}_\theta[\nabla\ell_\theta(x)\nabla\ell_\theta(x)^T] = \text{Cov}_\theta \nabla\ell_\theta$ where the final equality holds because $\mathbb{E}_\theta[\nabla\ell_\theta(x)] = 0$ (θ maximizes $\mathbb{E}_\theta[\ell_\theta(x)]$). To show this, assume that we can swap $\nabla, \mathbb{E}$. Then, $\nabla\ell_\theta(x) = \nabla \log p_\theta(x) = \frac{\nabla p_\theta(x)}{p_\theta(x)}$. Using that result, we see that:

$$\mathbb{E}_\theta[\nabla \ell_\theta] = \mathbb{E}\left[\frac{\nabla p_\theta}{p_\theta}\right] = \int \frac{\nabla p_\theta}{p_\theta} p_\theta d\mu = \int \nabla p_\theta d\mu = \nabla \int p_\theta d\mu = \nabla(1) = 0.$$

We now have a more compact representation of the asymptotic distribution described in the Theorem above.

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

Consider $I_\theta = -\nabla^2 \mathbb{E}[\ell_\theta(x)]$. If the magnitude of the second derivative is "large," that implies that the log-likelihood is steep around the global maximum (making it "easy" to find). Alternatively, if the magnitude of $-\nabla^2 \mathbb{E}[\ell_\theta(x)]$ is "small," we do not have sufficient curvature to find the optimal θ.

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 $\widehat{\theta}_n = \text{argmax}_\theta\, P_n \ell_\theta(X)$, 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. □