Stats 300b: Theory of Statistics **Winter 2019**

Lecture 3 – January 15

Lecturer: John Duchi *Scribe: Nian Si*

Warning: *these notes may contain factual errors*

Reading: VDV Chapter 3-4

Outline of Lecture 2:

1. Basic consistency and identifiability
2. Asymptotic Normality Results
   (a) Taylor expansions & Fisher Information
   (b) Moment method (not covered)

1 Recap: Delta method (Taylor expansions)

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

If $r_n(T_n - \theta) \xrightarrow{d} T$, then $r_n(\phi(T_n) - \phi(\theta)) = r_n(\phi'(\theta)(T_n - \theta)) + o_p(1) \xrightarrow{d} \phi'(\theta)T$.

2 Today: Consistency and Asymptotic Normality

Idea: Often log-likelihoods of models are smooth enough to permit Taylor Taylor approximations, So we can apply Delta method and CLTs to understand estimators.

Notation and Setting: Model family $\{P_\theta\}_{\theta\in\Theta}$ of distributions on space $\mathcal{X}$ and $\Theta \in \mathbb{R}^d$. Let log-likelihood of model P_θ with density p_θ be $l_\theta(x) := \log p_\theta(x)$.

Definition 2.1. *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: If X_i, $i = 1, \ldots, n$, are observations, we use P_n to denote the empirical distribution, i.e, $P_n := \frac{1}{n}\sum_{i=1}^n I_{X_i}$. So, $P_n(A) = \frac{1}{n}|\{i \in [n] : x_i \in A\}|$ and $P_n f = \frac{1}{n}\sum_{i=1}^n f(x_i)$.
♣

Definition 2.2.

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

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

Note: $\dot{\ell}_\theta(x) \equiv \nabla \ell_\theta(x)$ and $\ddot{\ell}_\theta(x) \equiv \nabla^2 \ell_\theta(x)$.

Problem Today: Observe $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$ but θ_0 is unknown. Our goal is to estimate θ_0.

A standard estimation is Maximum likelihood:

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\, P_n \ell_\theta(x).$$

Three important questions:

1. Consistency: Does $\hat{\theta}_n \overset{p}{\to} \theta_0$ as $n \to \infty$?
2. What is the asymptotic distribution and the rate of convergence of $\hat{\theta}_n$ to θ_0, i.e. for what $r_n \to \infty$, does $r_n(\hat{\theta}_n - \theta_0) \overset{d}{\to} Z$ and what is Z?
3. Optimality?

We will talk briefly about (1), and more about (2).

2.1 Consistency

Definition 2.3 (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, $D_{\text{kl}}(P_{\theta_1} \| P_{\theta_2}) > 0$ when $\theta_1 \neq \theta_2$. Recall that $D_{\text{kl}}(P_{\theta_1} \| P_{\theta_2}) = \int \log \frac{dP_{\theta_1}}{dP_{\theta_2}} dP_{\theta_1}$.

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 1 (Basic consistency for finite Θ)**.** *Suppose $\{P_\theta\}_{\theta\in\Theta}$ is identifiable and $|\Theta| < \infty$. Then, if $\hat{\theta}_n := \operatorname{argmax}_{\theta\in\Theta} P_n \ell_\theta(x)$ and $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$, $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$.*

Proof of Proposition We know by the Strong Law of Large Numbers that $P_n \ell_\theta(x) \overset{a.s.}{\to} P_{\theta_0}\ell_\theta(x)$ when $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$. Then,

$$P_{\theta_0}\ell_{\theta_0}(x) - P_{\theta_0}\ell_\theta(x) = \mathbb{E}_{\theta_0}\left[\log\left(\frac{p_{\theta_0}(x)}{p_\theta(x)}\right)\right] = D_{\text{kl}}(P_{\theta_0} \| P_\theta) > 0$$

for $\theta \neq \theta_0$. So, eventually we have that $P_n \ell_{\theta_0}(x) > P_n \ell_\theta(x)$ for all $\theta \neq \theta_0$. □

Remark Sometimes, the above result can fail when $|\Theta| = \infty$ even if the model is identified.

One sufficient condition often used for consistency results is a uniform law of large numerbs,: $\sup_{\theta\in\Theta} |P_n\ell_\theta - P\ell_\theta| \overset{p}{\to} 0$.

2.2 Asymptotic Normality and Taylor Expansions:

Definition 2.4 (Operator norm)**.**

$$\|A\|_{op} := \sup_{\|u\|_2 \leq 1} \|Au\|_2 = \sup_{\|u\|_2\leq 1, \|v\|_2 \leq 1} u^T A v.$$

Note: $\|Ax\| \leq \|A\|_{op} \|x\|$.

Assume we have a nice smooth model family. Specifically, we assume

1. $\mathbb{E}_{\theta_0}\left[\nabla \ell_{\theta_0} \nabla \ell_{\theta_0}^T\right]$ exists.

2. Lipschitz smoothness condition on second derivatives:

$$\left\|\nabla^2 \ell_{\theta_1}(x) - \nabla^2 \ell_{\theta_2}(x)\right\|_{op} \leq M(x) \left\|\theta_1 - \theta_2\right\|_2$$

   for θ_1 and θ_2 near θ_0 and $\mathbb{E}_\theta[M^2(x)] < \infty$.

Note: Taylor expansions can be a little tricker. If $f : \mathbb{R}^d \to \mathbb{R}^d$, let $Df(\theta) := [\nabla f_1(\theta), \ldots, \nabla f_d(\theta)]^d \in \mathbb{R}^{d\times d}$. Then $\|Df(\theta) - Df(\theta')\|_{op} \leq M(x) \left\|\theta_1 - \theta_2\right\|_2$ implies

$$f(\theta) = f(\theta_0) + \left(Df(\theta_0) + E_\theta\right)(\theta - \theta_0),$$

when $E_\theta \in \mathbb{R}^{d\times d}$ and $\|E_\theta\| \leq M \left\|\theta_1 - \theta_2\right\|_2$.

NOT mean-value-like results. We do **NOT** get that for some $\tilde{\theta}$ between θ, θ_0,

$$f(\theta) = f(\theta_0) + \left(Df(\tilde{\theta})\right)(\theta - \theta_0).$$

Theorem 2. *Let $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$ and assume the consistency $\hat{\theta}_n \overset{p}{\to} \theta_0$ and $P_n \nabla \ell(\hat{\theta}_n) = 0$ and the conditions stated above hold. Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} \mathsf{N}(0, (P_{\theta_0}\nabla^2 \ell_{\theta_0})^{-1} P_{\theta_0}(\nabla \ell_{\theta_0} \nabla \ell_{\theta_0}^T)(P_{\theta_0}\nabla^2 \ell_{\theta_0})^{-1}).$$

Intuition: If Hessian $P_{\theta_0}\nabla^2$ is "big" then lots of curvature makes estimation easier; on the other hand, if it is small, then little curvature makes estimation hard.

"Simplifying" Remarks: Usually, we can swap ∇(differentiation) and $\int$(expectation).

Then,

$$\nabla^2 \ell_\theta = \nabla\left(\frac{\nabla p_\theta}{p_\theta}\right) = \frac{\nabla^2 p_\theta}{p_\theta} - \frac{\nabla p_\theta \nabla p_\theta^T}{p_\theta^2}.$$

If $\nabla \mathbb{E} = \mathbb{E}\nabla$,

$$\mathbb{E}_{\theta_0}\left[\frac{\nabla^2 p_{\theta_0}}{p_{\theta_0}}\right] = \int \frac{\nabla^2 p_{\theta_0}}{p_{\theta_0}} p_{\theta_0} d\mu = \int \nabla^2 p_{\theta_0} d\mu = \nabla^2 \int p_{\theta_0} d\mu = \nabla^2 1 = 0.$$

So,

$$P_{\theta_0}\nabla^2 \ell_{\theta_0}(x) = -P_{\theta_0}(\nabla \ell_{\theta_0} \nabla \ell_{\theta_0}^T) = -I_{\theta_0} = \text{Fisher Information}.$$

Consequence: substitute Fish information into our asymptotic covariance.

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \overset{d}{\to} N(0, I_{\theta_0}^{-1}).$$

Intuition: If information matrix is large, I_{θ_0} is "large", problem is easier. Slope or score function $\nabla \ell_\theta$ is large, which means it is easy to find $\mathbb{E}_{\theta_0}(\nabla \ell_\theta) = 0$.

Proof of Theorem Taylor expansions + CLTs + Slutsky

Let $E_{\hat{\theta}_n} \in \mathbb{R}^{d\times d}$ 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 $\hat{\theta}_n - \theta_0$), that is,

$$\nabla \ell_{\hat{\theta}_n}(x) = \nabla \ell_{\theta_0}(x) + \left(\nabla^2 \ell_{\theta_0}(x) + E_{\hat{\theta}_n}(x)\right)(\hat{\theta}_n - \theta_0)),$$

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

$$0 = \nabla P_n \ell_{\hat{\theta}_n} = P_n \nabla \ell_{\theta_0} + P_n \left(\nabla^2 \ell_{\theta_0} + E_{\hat{\theta}_n}\right)(\hat{\theta}_n - \theta_0). \tag{3}$$

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

$$P_n E_{\hat{\theta}_n}(X) = \frac{1}{n}\sum_{i=1}^n E_{\hat{\theta}_n}(X_i) \leq \underbrace{\frac{1}{n}\sum_{i=1}^n M(X_i)}_{\xrightarrow{a.s.} \mathbb{E}_{\theta_0}[M(X)]} \underbrace{\|\hat{\theta}_n - \theta_0\|}_{\xrightarrow{p} 0} = o_P(1).$$

In particular, revisiting expression (3), we have

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

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

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

Applying Slutsky's theorem gives the result: indeed, we have $T_n = \sqrt{n} P_n \nabla \ell_{\theta_0}$ satisfies $T_n \xrightarrow{d} \mathsf{N}(0, P_{\theta_0}(\nabla \ell_{\theta_0} \nabla \ell_{\theta_0}^T))$ by the central limit theorem, and noting that $P_{\theta_0}\nabla^2 \ell_{\theta_0} + o_P(1)$ is eventually invertible gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} (P_{\theta_0}\nabla^2_{\theta_0})^{-1}\mathsf{N}(0, P_{\theta_0}(\nabla \ell_{\theta_0} \nabla \ell_{\theta_0}^T))$$

as desired. □

Remark If the model is not a true model, but we still have $\theta_0 = \text{argmax}_{\theta\in\Theta} \mathbb{E}\left[\log p_\theta(x)\right]$ and $\nabla_\theta \mathbb{E}\left[\log p_{\theta_0}(x)\right] = 0$, then proof is completely identical, once we have consistence $\hat{\theta}_n \xrightarrow{p} \theta_0$.