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

Lecture 3 – January 17

Lecturer: John Duchi *Scribe: John Cherian*

Warning: *these notes may contain factual errors*

Reading: VDV Chapter 5.1-5.2, 5.5; TPE Chapter 2.5

Outline of Lecture 2:

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

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

Claim 1. *If $r_n(T_n - \theta) \xrightarrow{d} T$, then if $\phi'(\theta)$ exists, $r_n(\phi(T_n) - \phi(\theta)) = r_n(\phi'(\theta)(T_n - \theta)) + o_p(r_n(T_n - \theta))$. Note that $r_n(T_n - \theta) \xrightarrow{p} 0$, i.e. $o_p(1)$, since the sequence is uniformly tight. By applying the Continuous Mapping Theorem and Slutsky's Lemma, we conclude that $r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)T$.*

Notation:

Definition 0.1. *Given distribution P on $\mathcal{X}$, function $f : \mathcal{X} \to \mathbb{R}^d$,*

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

If X_i, $i = 1, \ldots, n$, are observations, use P_n to denote the empirical distribution:

$$P_n f := \frac{1}{n} \sum_{i=1}^{n} f(x_i) \qquad \textit{for any function } f$$

Example 1: $P_n(A) = \frac{1}{n} \operatorname{card}\{i \in [n] : x_i \in A\}$ ♣

Setting: 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. $p_\theta = \frac{dP_\theta}{d\mu}$.

Idea: Our big idea for the day is that if we get enough observations from the same P_θ, we should be able to:

1. Identify θ
2. If $\ell_\theta(x) = \log p_\theta(x)$ is "smooth" enough, get explicit asymptotic normality results.

Definition 0.2.

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

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

Note: $\dot{\ell}_\theta \equiv \nabla \ell_\theta(x)$ *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.

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

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

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

Example 2 (Gaussian mean)**:** Let $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\theta, I)$ and $p_\theta(x) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|x-\theta\|_2^2}$. Then, $P_n \ell_n(\theta) = -\frac{1}{2n}\sum_{i=1}^n \|x_i - \theta\|_2^2 + \frac{d}{2}\log(2\pi)$. Note that this function is concave, so it has a unique global maximum.

Since $\nabla P_n \ell_n(\theta) = -\frac{1}{n}\sum_{i=1}^n (\theta - x_i) = 0$, $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n x_i$. And by the CLT, we know that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, I)$ ♣

Example 3 (2-sided exponential)**:** Let $X_i \overset{\text{iid}}{\sim} \mathit{Laplace}(\theta_0)$ and $p_\theta(x) = \frac{1}{2}e^{-|\theta - x|}$, i.e. $\log p_\theta(x) = -|\theta - x| - \log 2$. Then, $\hat{\theta}_n \in \operatorname{argmax}_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^n |\theta - x_i|$, i.e. $\hat{\theta}_n = \mathit{Median}(x_1, \ldots, x_n)$.

Note that even though the likelihood function is not smooth (the absolute value function has a discontinuity at 0), we are still able to derive asymptotic normality results for this estimator. These are, however, not presented in this lecture. ♣

In general, there are two relevant questions to ask about Maximum Likelihood and other estimators of θ_0:

1. Consistency: Does $\hat{\theta}_n \xrightarrow{p} \theta_0$ 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 $r_n(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$ and what is Z?

Consistency:

Definition 0.3 (Identifiability). *A model* $\{P_\theta\}_{\theta\in\Theta}$ *is* <u>*identifiable*</u> *if* $P_{\theta_1} \neq P_{\theta_2}$ *for all* $\theta_1, \theta_2 \in \Theta$ $(\theta_1 \neq \theta_2)$*.*

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

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 2 (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 We know by the Strong Law of Large Numbers that $P_n \ell_\theta(x) \xrightarrow{a.s.} P_{\theta_0} \ell_\theta(x)$ when $x_i \overset{\text{iid}}{\sim} P_{\theta_0}$.

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

We know that $D_{\mathrm{kl}}(P_{\theta_0} \| P_\theta) > 0$ <u>unless</u> $\theta \neq \theta_0$. So, for large enough n and finite Θ, we have that:

$$P_n\ell_{\theta_0}(x) - P_n\ell_\theta(x) > 0 \qquad\qquad \forall \theta \neq \theta_0$$

Therefore, $\hat{\theta}_n = \theta_0$ "eventually." $\square$

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

One sufficient condition often used for consistency results is a <u>uniform law</u>, i.e. for $x_i \overset{\text{iid}}{\sim} P$, we have that $\sup_{\theta\in\Theta} |P_n\ell_\theta - P\ell_\theta| \xrightarrow{p} 0$.

Example 4 (Uniform laws)**:** Suppose $\Theta = \mathbb{R}^d, x_i \overset{\text{iid}}{\sim} P$. Consider $P_n \langle \theta, x \rangle = \frac{1}{n}\sum_{i=1}^n \langle \theta, x_i \rangle$. Then, for each $\theta \in \Theta$, $P_n \langle \theta, x \rangle \xrightarrow{a.s.} \langle \theta, \mathbb{E}[x] \rangle$.

Does the uniform law, i.e. $\sup_{\theta\in\Theta} |\langle \theta, P_n x\rangle - \langle \theta, Px \rangle| \to 0$, hold? If $\operatorname{Cov} x \neq 0$, no. However, if Θ is compact, we will be able to prove this result.

Alternatively, $\hat{\theta}_n = \operatorname{argmax}_{\theta\in\Theta} P_n \langle x, \theta \rangle$ does satisfy $\left\langle \hat{\theta}_n, Px \right\rangle \to \sup_{\theta\in\Theta} \langle \theta, Px \rangle$. ♣

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

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:

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

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

3. $\nabla P_n \ell_{\hat{\theta}_n} = 0$.

4. Θ is a convex set.

Theorem 3. *Let $x_i \overset{\text{iid}}{\sim} P_{\theta_0}$ 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 We define the Fisher Information as $I_\theta := \mathbb{E}_\theta[\nabla\ell_\theta(x)\nabla\ell_\theta(x)^T] = \operatorname{Cov}_\theta \nabla\ell_\theta$.

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:

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

Similarly, given that $\nabla^2\ell_\theta = \nabla\left(\frac{\nabla p_\theta}{p_\theta}\right) = \frac{\nabla^2 p_\theta}{p_\theta} - \frac{\nabla p_\theta \nabla p_\theta^T}{p_\theta^2}$:

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

As a result:

$$\begin{aligned}
\mathbb{E}_\theta[\nabla^2\ell_\theta] &= -\mathbb{E}_\theta\left[\left(\frac{\nabla p_\theta}{p_\theta}\right)\left(\frac{\nabla p_\theta}{p_\theta}\right)^T\right] \\
&= -\operatorname*{Cov}_\theta(\nabla\ell_\theta(x)) = -I_\theta
\end{aligned}$$

Using what we have shown about the Fisher information, we now have a more compact representation of the asymptotic distribution described in the Theorem above.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, I_{\theta_0}^{-1} I_{\theta_0} I_{\theta_0}^{-1}) = \mathsf{N}(0, I_{\theta_0}^{-1}) \tag{3}$$

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.