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

Lecture 19 – March 13

Lecturer: John Duchi *Scribe: Saied Mehdian*

Warning: *these notes may contain factual errors*

Reading: TSH 12.3, 13.1-13.3; VdV 6, 7.1-7.3

Outline: Asymptotic Testing and Optimality

- Hellinger distance
- Quadratic Mean Differentiability
- Local Asymptotic normality
- Asymptotically most powerful tests

1 Recap

Measures Q_n are contiguous to P_n, $Q_n \lhd P_n$, if $Q_n(A_n) \to 0$ when $P_n(A_n) \to 0$. An important consequence is Le Cam's third lemma:

$$(X_n, \log \frac{dQ_n}{dP_n}) \xrightarrow[P_n]{d} \mathcal{N}\left(\begin{bmatrix} \mu \\ \frac{-1}{2}\sigma^2 \end{bmatrix}, \begin{bmatrix} \Sigma & \tau^T \\ \tau & \sigma^2 \end{bmatrix}\right)$$

then $Q_n \lhd\rhd P_n$ and $X_n \to \mathcal{N}(\mu + \tau, \Sigma)$.

Idea: We would like to asymptotically change measures using $Q_n \lhd\rhd P_n$ because $\log \frac{dQ_n}{dP_n} \xrightarrow[P_n]{d} z \sim \mathcal{N}(\frac{-1}{2}\sigma^2, \sigma^2)$ with $\mathbb{E}[e^z] = 1$, so e^z is densty like quantity.

Today, we use this to get changes of measure and prove optimal power of tests.
As starting point, we look at optimal test and related distances between distributions.

Recall: $\|P - Q\|_{TV} = \sup_A |P(A) - Q(A)| = \frac{1}{2}\int |p - q| d\mu$ and $d^2_{Hel}(P, Q) = \frac{1}{2}\int (\sqrt{p} - \sqrt{p})^2 d\mu$ where $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$. Then:

$$d^2_{Hel}(P, Q) \le \|P - Q\|_{TV} \le d_{Hel}(P, Q)\sqrt{2 - d^2_{Hel}(P, Q)}$$

and

$$\inf_{\psi: \mathcal{X} \to \{0,1\}} P_0(\psi(x) \neq 0) + P_1(\psi(x) \neq 1) = 1 - \|P_0 - P_1\|_{TV}.$$

Consider sequences $P_{0,n}$ and $P_{1,n}$. Then

$$\liminf_n \inf_\psi P_{0,n}(\psi(x) \neq 0) + P_{1,n}(\psi(x) \neq 1) > 0$$

if and only if

$$\limsup_{n\to\infty} d_{Hel}(P_{0,n}, P_{1,n}) < 1$$

Why should we use Hellinger distance? Ans: It plays nicely with iid sampling and Taylor expansions.

Note that

$$d^2_{Hel}(P^n, Q^n) = 1 - \int \sqrt{p(x_1)\dots p(x_n)}\sqrt{q(x_1)\dots q(x_n)}d\mu$$
$$=1 - (\int \sqrt{dP}\sqrt{dQ})^n = 1 - (1 - d^2_{Hel}(P,Q))^n$$

If we think about local alternatives, say P_0 vs $P_{\frac{h}{\sqrt{n}}}$, if testing is right difficulty, we want $d^2_{Hel}(P_0, P_{\frac{h}{\sqrt{n}}}) \approx \frac{\|h\|^2}{n} \asymp \frac{1}{n}$.

2 Quadratic Mean Differentiability (Le Cam 1970's)

Suppose $\{p_\theta\}_{\theta\in\Theta}$ is a "nice" family of densitites. Then using $\sqrt{a+\delta} = \sqrt{a} + \frac{\delta}{2\sqrt{a}} + o(\delta^2)$, we expect $p_{\theta+h} = p_\theta + \dot{p}_\theta^T h + O(\|h\|^2)$ and

$$\sqrt{p_{\theta+h}} = \sqrt{p_\theta} + \frac{\dot{p_\theta}^T h}{2\sqrt{p_\theta}} + O(\|h\|^2)$$
$$= \sqrt{p_\theta} + \frac{1}{2}\dot{\ell}_\theta h\sqrt{p_\theta} + O(\|h\|^2)$$

where $\dot{\ell}_\theta = \frac{\partial}{\partial\theta}\log p_\theta = \frac{\dot{p}_\theta}{p_\theta}$.

Nice fact: $\sqrt{p_\theta}$ is in L^2 for any density and always $\int(\sqrt{p_\theta})^2 = \int p_\theta = 1$.
(There is a paper by David Pollard in 1997 Festschrift for Le Cam exploring this)
Derivatives in mean squared error allow us to get interesting results.

Definition 2.1. *Quadratic Mean Differentiability (QMD)*
A family $\{P_\theta\}_{\theta\in\Theta}$ is QMD at $\theta \in int\Theta$ if there exists a score function $\dot{\ell} : X \to \mathbb{R}^d$ such that

$$\int(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^T\dot{\ell}\sqrt{p_\theta})^2 d\mu = o(\|h\|^2)$$

as $h \to 0$.

Proposition (Notes on website or reading): For a QMD family, $P_\theta\dot{\ell}_\theta = 0$ and $I_\theta := P_\theta\dot{\ell}_\theta\dot{\ell}_\theta^T$ exists.

Example 1: Let $\{P_\theta\}_{\theta\in\Theta}$ be an exponential family.

$$p_\theta = \frac{dP_\theta}{d\mu} = exp(\theta^T T(x) - A(\theta)), \quad A(\theta) = \log\int e^{\theta^T T(x)}d\mu(x)$$

Then $\{P_\theta\}_{\theta\in\Theta}$ is QMD with score function $\dot{\ell}_\theta(x) = T(x) - \mathbb{E}_\theta[T(x)] = T(x) - \nabla A(\theta) = \nabla_\theta \log p_\theta(x)$.

Sketch of Proof (VdV Lemma 7.6, which actually proves that if $\dot{\ell}_\theta = \frac{\partial}{\partial\theta}\log p_\theta$ and the usual Fisher information is continuous in θ, i.e., $\mathbb{E}_\theta[(\frac{\partial}{\partial\theta}p_\theta)^2/p_\theta^2]$ is continuous then family is QMD with score $\dot{\ell}_\theta$).
For us, let $T(x) = x$ without loss of generality.

$$\begin{aligned}
&\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^T(X - \nabla A(\theta))\sqrt{p_\theta(x)} \\
=&exp\left(\frac{1}{2}(\theta^T - A(\theta))\right)\left[exp(\frac{1}{2}(h^T x - \frac{1}{2}[A(\theta+h) - A(\theta)])) - 1 - \frac{1}{2}h^T(x - \nabla A(\theta))\right] \\
=&\sqrt{p_\theta}\left[1 + \frac{1}{2}h^T x - \frac{1}{2}(A(\theta+h) - A(\theta)) - 1 - \frac{1}{2}h^T(x - \nabla A(\theta))\right] \\
+&\sqrt{p_\theta}\left[exp(\frac{1}{2}(h^T x - \frac{1}{2}[A(\theta+h) - A(\theta)])) - 1 - \frac{1}{2}h^T x + \frac{1}{2}(A(\theta+h) - A(\theta))\right] \\
=&\sqrt{p_\theta}\left[\frac{1}{4}h^T\nabla^2 A(\theta + \tilde{h})h\right] + \sqrt{p_\theta}O(\|h\|^2)
\end{aligned}$$

Thus

$$\frac{1}{\|h\|^2}\int(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^T\dot{\ell}\sqrt{p_\theta})^2 d\mu = \int p_\theta O(\frac{\|h\|^2}{\|h\|})d\mu \to 0$$

as $\|h\| \to 0$ by dominated convergence. □

♣

Remark How about Hellinger distance? Let $\{P_\theta\}_{\theta\in\Theta}$ be QMD at θ.

$$\begin{aligned}
&d^2_{Hel}(P_\theta, P_{\theta+h}) = \frac{1}{2}\int(\sqrt{p_{\theta+h}} - \sqrt{p_\theta})^2 d\mu = \frac{1}{2}\int(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^T\dot{\ell}_\theta\sqrt{p_\theta} + \frac{1}{2}h^T\dot{\ell}_\theta\sqrt{p_\theta})^2 d\mu \\
=&\frac{1}{2}\int(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^T\dot{\ell}_\theta\sqrt{p_\theta})^2 d\mu + \frac{1}{8}\int h^T\dot{\ell}_\theta\dot{\ell}_\theta^T h p_\theta d\mu + \frac{1}{2}\int(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^T\dot{\ell}_\theta\sqrt{p_\theta})h^T\dot{\ell}_\theta\sqrt{p_\theta} \\
=&o(\|h\|^2) + \frac{1}{8}\int h^T\dot{\ell}_\theta\dot{\ell}_\theta^T h p_\theta d\mu + \sqrt{o(\|h\|^2)O(\|h\|^2)} = o(\|h\|^2) + \frac{1}{8}\int h^T\dot{\ell}_\theta\dot{\ell}_\theta^T h p_\theta d\mu \\
=&\frac{1}{8}h^T I_\theta h + o(\|h\|^2)
\end{aligned}$$

So, Fisher Information determines distances (locally, as $h \to 0$) and so it should appear in asymptotic optimality of tests.

3 Local Asymptotic Normality

Hope/Idea: In "nice" families, we might have limiting normality or CLT to use Guassanity for optimality.

Definition 3.1. *A family $\{P_{\theta,n}\}_{\theta\in\Theta}$, $n \in \mathbb{N}$ is locally asymptotically normal (LAN) at $\theta \in int\Theta$ if there exists a sequence Δ_n (random) and matrix $K \succeq 0$ (information precision) such that for all $h \in \mathbb{R}^d$,*

$$\log\frac{dP_{\theta+\frac{h}{\sqrt{n}},n}}{dP_{\theta,n}} = h^T\Delta_n - \frac{1}{2}h^T K h + o_{P_{\theta,n}}(\|h\|)$$

where $\Delta_n \xrightarrow[P_{\theta,n}]{d} \mathcal{N}(0, K)$

Remark $h^T\Delta_n - \frac{1}{2}h^TKh \xrightarrow[P_{\theta,n}]{d} \mathcal{N}(-\frac{1}{2}h^TKh, h^TKh)$. So, we will be able to use continguity/Le Cam's third lemma to get limits under $Q_n := P_{\theta+\frac{h}{\sqrt{n}}}$.

Example 2: Gaussian shift family
Let $P_{h,n}$ be distributions of $Y_i = \frac{1}{\sqrt{n}}h + \xi_i$ for $\xi_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$

$$\log \frac{dP_{\frac{h}{\sqrt{n}},n}}{dP_{0,n}}(Y_{1:n}) = \sqrt{n}h^T\Sigma^{-1}\bar{Y}_n - \frac{1}{2}h^T\Sigma^{-1}h$$

where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Certainly, $\sqrt{n}\bar{Y}_n \xrightarrow[P_{0,n}]{d} \mathcal{N}(0, \Sigma^{-1})$. So, $\Delta_n = \sqrt{n}\Sigma^{-1}\bar{Y}_n \xrightarrow[P_{0,n}]{d} \mathcal{N}(0, \Sigma^{-1})$ and the family is LAN with precision/information Σ^{-1}. ♣

Example 3: Quadratic Mean Differentiable family
Proposition: If $\{P_\theta\}$ is QMD and $P_n := P_\theta + \frac{h}{\sqrt{n}}$, $P = P_\theta^n$, then

$$\log \frac{dP_n}{dP}(X_1, \ldots, X_n) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_\theta(x_i)\right)^T h - \frac{1}{2}h^T I_\theta h + o_p(1)$$

So QMD property implies LAN property with $\Delta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_\theta(x_i)$ such that

$$\Delta_n \xrightarrow[P_{\theta,n}^n]{d} \mathcal{N}(0, I_\theta), \quad I_\theta = \mathbb{E}[\dot{\ell}_\theta \dot{\ell}_\theta^T]$$

♣

Let's consider testing in a LAN family. We will show how to get asymptotically optimal powers for each level $\alpha \in (0, 1)$.

Suppose $\{P_{\theta,n}\}$ is LAN, and consider tests of $p_{0,n}$ vs $p_{\frac{h}{\sqrt{n}},n}$. We consider simple hypothesis tests. So optimal test (by Neyman-Pearson) is $L_n = \frac{dP_{h/\sqrt{n},n}}{dP_{0,n}}$.

$$\phi_{n,h} = \begin{cases} 1 & if \log L_n > c_{n,h} \\ \gamma_{n,h} & if \log L_n = c_{n,h} \\ 0 & if \log L_n < c_{n,h} \end{cases}$$

for some $\gamma_{n,h}$ and $c_{n,h}$.
We know that $\log L_n = h^T\Delta_n - \frac{1}{2}h^TKh + o_{P_{0,n}}(1)$. So,

$$(\log L_n, \log L_n) \xrightarrow[P_n]{d} \mathcal{N}\left(\begin{bmatrix} \frac{-1}{2}h^TKh \\ \frac{-1}{2}h^TKh \end{bmatrix}, \begin{bmatrix} h^TKh & h^TKh \\ h^TKh & h^TKh \end{bmatrix}\right)$$

So, by Le Cam's third lemma, we know that

$$\log L_n \xrightarrow[P_{h/\sqrt{n},n}]{d} \mathcal{N}(\frac{1}{2}h^TKh, h^TKh)$$

Consider asymptotically level-α tests $\phi_{n,h}$ (optimal by Neyman-Pearson).

$$\lim_{n\to\infty} \mathbb{E}_{P_0}[\phi_{n,h}] = \alpha = P_0(\log L_n > c_{n,h}) + o(1)$$

So if $w_n \sim \mathcal{N}(\frac{-1}{2}h^T Kh, h^T Kh)$, then

$$P_0(\log L_n > c_{n,h}) = P(w_n > c_{n,h}) + o(1) \text{ and } c_{n,h} = \frac{1}{2}h^T Kh + z_{1-\alpha}\sqrt{h^T Kh} + o(1)$$

as $n \to \infty$, where $z_{1-\alpha}$ = the $(1-\alpha)$-quantile of $\mathcal{N}(0,1)$.
Thus, the powers of $\phi_{n,h}$ must satisfy $\mathbb{E}_{\frac{h}{\sqrt{n}}}[\phi_{n,h}] \to 1 - \Phi(z_{1-\alpha} - \sqrt{h^T Kh})$.

Definition 3.2. *For testing θ_0 against alternatives θ_n, sequence $\{\phi_n\}$ is asymptotically most powerful (AMP) at level α if $\limsup_{n\to\infty} \mathbb{E}_{\theta_0}[\phi_n] \leq \alpha$ and for all tests $\{\psi_n\}$ satisfying $\limsup_{n\to\infty} \mathbb{E}_{\theta_0}[\psi_n] \leq \alpha$, we have $\limsup_{n\to\infty} \mathbb{E}_{\theta_n}[\psi_n] - \mathbb{E}_{\theta_n}[\phi_n] \leq 0$.*

Theorem 1. *Let $\{P_\theta\}$ be LAN, $\theta \in \mathbb{R}$, $\theta_0 \in \Theta$. Then $\phi_n = \phi_n(X_1, \ldots, X_n)$ is AMP at level α if and only if $\mathbb{E}_{\theta_0}[\phi_n] \to \alpha$ and, for all $h \in \mathbb{R}$, $\lim_{n\to\infty} \mathbb{E}_{\frac{h}{\sqrt{n}}}[\phi_n] = 1 - \Phi(z_{1-\alpha} - h\sqrt{K})$*