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

Lecture 19 – March 14

Lecturer: John Duchi *Scribe: Steven Yadlowsky*

Warning: *these notes may contain factual errors*

Reading: Notes on course website (Contiguity and asymptotics)

1 Outline

- Recap
- Quadratic mean differentiability
  - Testing
  - Definitions and examples
- Local asymptotic normality
- Limiting Gaussian shifts

2 Recap

Measures Q_n are contiguous with respect to P_n (written $Q_n \triangleleft P_n$) if $Q_n(A_n) \to 0$ whenever $P_n(A_n) \to 0$.

Lemma 1. *Le Cam's 3rd Lemma If*

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

then $P_n \triangleleft \triangleright Q_n$ *and* $X_n \xrightarrow[Q_n]{d} \mathsf{N}(\mu + \tau, \Sigma)$.

Idea Asymptotically, we can change measure between P_n and Q_n, because $Q_n \triangleleft \triangleright P_n$ and $\log \frac{dQ_n}{dP_n} \to \mathsf{N}\left(-\frac{1}{2}\sigma^2, \sigma^2\right)$.

Goal Understand limiting behavior of experiments / random variables by this change of measure. Our motiviation will be via testing: When is testing point nulls versus point alternatives "appropriated" (which we will momentarily make precise) hard?

Recall

$$\|P-Q\|_{\text{TV}} = \sup_A |P(A)-Q(A)|,$$

$$d^2_{\text{hel}}(P,Q) = \frac{1}{2}\int \left(\sqrt{p}-\sqrt{q}\right)^2 \mathrm{d}\mu,$$

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

Now, consider the simple hypothesis testing problem of P_0 versus P_1, and the associated best error,

$$\inf_\psi P_0(\psi \neq 0) + P_1(\psi \neq 1) = 1 - \|P_0 - P_1\|_{\text{TV}} \ge 1 - \sqrt{2}d_{\text{hel}}(P_0, P_1).$$

Given sequences of tests $P_{0,n}$ versus $P_{1,n}$, we are interested in considering when the asymptotic error does not vanish,

$$\liminf_{n\to\infty} \inf_{\psi_n} P_{0,n}(\psi_n \neq 0) + P_{1,n}(\psi_n \neq 1) > 0,$$

which occurs whenever $1-\sqrt{2}d_{\text{hel}}(P_{1,n}, P_{0,n}) > 0$, or, put another way, when $d_{\text{hel}}(P_{1,n}, P_{0,n}) < \frac{1}{\sqrt{2}}$. Note that this bound may very well be loose– we can probably get tighter constant bounds on the Hellinger distance, but that is not our intention here. Instead, we are focusing on highlighting how Hellinger distance "plays nicely" with iid sampling (ie., product distributions). Specifically, consider the following:

$$\begin{aligned}
d^2_{\text{hel}}(P^n, Q^n) &= \frac{1}{2}\int \left(\sqrt{\mathrm{d}P^n} - \sqrt{\mathrm{d}P^n}\right)^2 \\
&= 1 - \int \sqrt{p(x_1)\dots p(x_n)}\sqrt{q(x_1)\dots q(x_n)}\,\mathrm{d}\mu \\
&= 1 - \left(\int \sqrt{p(x)}\sqrt{q(x)}\,\mathrm{d}\mu\right)^n = 1 - \left(1 - d^2_{\text{hel}}(P,Q)\right)^n.
\end{aligned}$$

In particular, given $d_{\text{hel}}(P,Q)$ we *know* $d_{\text{hel}}(P^n, Q^{)}$ of the product distributions. So, if we consider local alternatives P_0^n versus $P^n_{h/\sqrt{n}}$, then

$$\lim_{n\to\infty} d_{\text{hel}}(P_0^n, P^n_{h/\sqrt{n}}) < 1 \text{ if}$$

$$d^2_{\text{hel}}(P_0, P_{h/\sqrt{n}}) = O(\frac{1}{n}) \text{ as } n \to \infty,$$

because $(1-O(\frac{1}{n}))^n$ "$\to$" $\exp(-\text{something})$. With this in mind, our game plan is to understand when $d^2_{\text{hel}}(P_0, P_h/\sqrt{n}) = \frac{f(h)}{n} + o(\frac{1}{n})$.

3 Quadratic mean differentiability

Suppose $\{P_\theta\}_{\theta\in\Theta}$ is a smooth family of distributions (ie., ∇p_θ exists and is smooth). Then, using that

$$\sqrt{a+\delta} = \sqrt{a} + \frac{1}{2\sqrt{a}}\delta + O(\delta^2),$$

as $\delta \to 0$, we have that

$$\begin{aligned}
\sqrt{p_{\theta+h}} &= \sqrt{p_\theta + \nabla p_\theta^T h + O(\|h\|^2)} \\
&= \sqrt{p_\theta} + \frac{1}{2\sqrt{p_\theta}} \nabla p_\theta^T h + O(\|h\|^2) \\
&= \sqrt{p_\theta} + \frac{1}{2} \frac{\nabla p_\theta^T h}{p_\theta} \sqrt{p_\theta} + O(\|h\|^2) \\
&= \sqrt{p_\theta} + \frac{1}{2} \dot{\ell}_\theta^T h \sqrt{p_\theta} + O(\|h\|^2),
\end{aligned}$$

where $\dot{\ell}_\theta = \nabla \log p_\theta$ is the score function.

With this in mind, we define the following nice family of distributions for which the above expansion holds, almost definitionally. This will basically capture families of distributions that have nice behavior with the Hellinger distance.

Definition 3.1. *A family $\{P_\theta\}_{\theta\in\Theta}$ is quadratic mean differentiable (QMD) at $\theta \in \operatorname{int}\Theta$ if there is a score function $\dot{\ell}_\theta : \mathcal{X} \to \mathbb{R}^d$, so that*

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

as $h \to 0$.

Proposition 2. *(proved in the notes) $P_\theta \dot{\ell}_\theta = 0$ and $P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ is well-defined in a family which is QMD.*

Example 1: Exponential families. Let $p_\theta(x) = \exp(\theta^T T(x) - A(\theta))$, $A(\theta) = \log \int \exp(T(x)^T \theta)$. Then, $\{P_\theta\}$ is QMD with score $\dot{\ell}_\theta(x) = \nabla \log p_\theta(x) = T(x) - \nabla A(\theta) = T(x) - \mathbb{E}(T(x))$.
Proof (Sketch).

Without loss of generality, we can take $T(x) = x$ (we could do a change of measure with μ to make this precise).

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

Using that Lebesgue's dominated convergence theorem holds in an exponential family,

$$\begin{aligned}
\frac{1}{\|h\|^2} \int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \dot{\ell}_\theta \sqrt{p_\theta} \right) &= \int p_\theta O\left(\frac{\|h\|^4 + (h^T x)^4}{\|h\|^2} \right) \mathrm{d}\mu \\
&\to 0 \text{ as } h \to 0.
\end{aligned}$$

□

♣

Remark Fisher information $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ is exactly as before, $I_\theta = \mathrm{Cov}(T(X) - \nabla A(\theta))$.

The following heuristic (which is actually true, although unstated as such) is that if $\{P_\theta\}$ is QMD, then

$$\begin{aligned} d^2_{\mathrm{hel}}(P_{\theta+h}, P_\theta) \underbrace{=}_{\text{can be made rigorous}} & \frac{1}{2}\int \left(\frac{1}{2} h^T \dot{\ell}_t heta \sqrt{p_\theta}\right)^2 \mathrm{d}\mu + o(\|h\|^2) \\ &= \frac{1}{8}\int h^T \dot{\ell}_\theta \dot{\ell}_\theta^T h p_\theta \,\mathrm{d}\mu + o(\|h\|^2) \\ &= \frac{1}{8} h^T I_\theta h + o(\|h\|^2). \end{aligned}$$

For QMD families, we have

$$d^2_{\mathrm{hel}}(P_{\theta+h/\sqrt{n}}, P_\theta) = \frac{1}{8} h^T I_\theta h + o(\frac{1}{n}, \text{ and}$$

$$\lim_{n\to\infty} d^2_{\mathrm{hel}}(P^n_{\theta+h/\sqrt{n}}, P^n_\theta) = 1 - \exp(-\frac{1}{8} h^T I_\theta h) \in (0,1),$$

so that

$$\liminf_{n\to\infty} \inf_{\psi_n} P_\theta(\psi_n \neq 0) + P_{\theta+h/\sqrt{n}}(\psi_n \neq 1) > 0.$$

4 Local asymptotic normality

Definition 4.1. *A family $\{P_\theta\}_{\theta\in\Theta}$ is locally asymptotically normal (LAN) at $\theta \in \operatorname{int}\Theta$ if there exists a sequence $D_n \in \mathbb{R}^d$ and precision matrix $K \succeq 0$, such that for all $h \in \mathbb{R}^d$,*

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

where $\Delta_n \underset{P_{\theta,n}}{\overset{d}{\to}} \mathsf{N}(0, K)$, and $o_P(\|h\|)$ means converging in probability to 0 uniformly, if $\|h\|$ is bounded.

Remark

1. Basically, p_θ has a shifted quadratic expansion.
2. $h^T \Delta_n - \frac{1}{2} h^T K h \underset{P_{\theta,n}}{\overset{d}{\to}} \mathsf{N}(-\frac{1}{2} h^T K h, h^T K t)$, so this will imply contiguity.

Example 2: Gaussian shifts. Let $P_{h,n}$ be the distribution of $Y_i = h + \xi_i$, where $\xi_i \sim \mathsf{N}(0, \Sigma)$, and $i = 1, \dots n$. Then, calculations (omitted here) show that

$$\log \frac{\mathrm{d}P_{h/\sqrt{n},n}}{\mathrm{d}P_{0,n}} = \sqrt{n} h^T \Sigma^{-1} \bar{Y}_n - \frac{1}{2} h^T \Sigma^{-1} h,$$

so

$$\Delta_n = \sqrt{n}\Sigma^{-1}\bar{Y}_n \underset{P_{0,n}}{\overset{d}{\to}} \mathsf{N}(0, \Sigma^{-1}),$$

and $K = \Sigma^{-1}$ is the precision (or information) matrix. ♣

Example 3: QMD families. (see VdV ch 7 for details).

If $\{P_\theta\}$ is QMD, then

$$\log \frac{\mathrm{d}P_{h/\sqrt{n},n}}{\mathrm{d}P_{0,n}} = \left(\frac{1}{\sqrt{n}} \sum_{i=1} 6n\dot{\ell}(x_i) \right)^T h - \frac{1}{2} h^T I_\theta h + o_P(1).$$

So, QMD implies LAN with precision I_θ.

Proof (Sketch).

Let $P_n = P_{\theta + h/\sqrt{n}}$, and $P = P_\theta$.

$$\begin{aligned} \log \prod_{i=1}^n \frac{p_n}{p}(x_i) &= 2 \sum_{i=1}^n \log \sqrt{\frac{p_n}{p}(x_i)} \\ (*) &= 2 \sum_{i=1}^n \log \left(1 + \frac{1}{2} \underbrace{\left(2\sqrt{\frac{p_n}{p}}(x_i) - 2 \right)}_{W_{n,i}} \right) \\ &= 2 \sum_{i=1}^n \frac{1}{2} W_{n,i} - \frac{1}{8} W_{n,i}^2 + W_{n,i}^2 r(W_{n,i}), \end{aligned}$$

where $r(x) = O(|x|)$.

$$(*) = \underbrace{\sum_{i=1}^n W_{n,i}}_{(1)} - \underbrace{\frac{1}{4} \sum_{i=1}^n W_{n,i}^2}_{(2)} + o_P(1).$$

From here, we can use QMD to control (1) and (2), and if $g(x) = h^T \dot{\ell}_\theta(x)$, then

$$\begin{aligned} \mathrm{Var} \left(\sum_{i=1}^n W_{n,i} - \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i) \right) &\le n\mathbb{E} \left[(W_{n,1} - \frac{1}{\sqrt{n}} g(X_1))^2 \right] \\ &= no(\frac{1}{n}) = o(1), \end{aligned}$$

where the step to $o(\frac{1}{n})$ follows from the fact that the family is QMD.

Similar calculations show that in the end,

$$(*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i) - \frac{1}{2} h^T I_\theta h + o_P(1).$$

□

♣