Local Asymptotics and Optimality

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- Motivation with testing
- Quadratic mean differentiability and local asymptotic normality
- Asymptotically most powerful tests
- Limiting Gaussian experiments
- Local asymptotic minimax theorems

Reading:

- van der Vaart, *Asymptotic Statistics* Chs. 6–8
- Lehmann & Romano, *Testing Statistical Hypothesis* Ch. 12.3, 13.1–13.3

Recapitulation

- Measures Q_n are contiguous w.r.t. P_n, $Q_n \lhd P_n$, if $Q_n(A_n) \to 0$ whenever $P_n(A_n) \to 0$
- Le Cam's third lemma states that

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

  implies $X_n \xrightarrow{d}_{Q_n} \mathcal{N}(\mu + \tau, \Sigma)$
- asymptotic change of measure from $P_n \lhd\rhd Q_n$ as $\log \frac{dQ_n}{dP_n}$ has mean $-\frac{1}{2}\sigma^2$

Goal: understand limits of random experiments to get optimality

Testing motivation

idea: look at optimal pairs of tests and parameterize them

some distances on distributions:

$$\|P - Q\|_{\mathsf{TV}} := \sup_A |P(A) - Q(A)| = \frac{1}{2}\int |p - q| d\mu$$

$$d_{\mathsf{hel}}^2(P, Q) := \frac{1}{2}\int (\sqrt{dP} - \sqrt{dQ})^2 = \frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2 d\mu$$

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

and optimal test error

$$\inf_{\psi} \{P_0(\psi(X) \neq 0) + P_1(\psi(X) \neq 1)\} = 1 - \|P_0 - P_1\|_{\mathsf{TV}}$$

Asymptotics in pairwise tests

Lemma (Asymptotically non-trivial testing)

For any sequence of distributions $P_{0,n}$ vs. $P_{1,n}$, we have

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

if and only if

$$\limsup_n d_{\mathrm{hel}}(P_{0,n}, P_{1,n}) < 1.$$

Why Hellinger distances? they work well with i.i.d. sampling:

$$d^2_{\mathrm{hel}}(P^n, Q^n) = 1 - \left(1 - d^2_{\mathrm{hel}}(P, Q)\right)^n$$

- tests asymptotically non-trivial when $d^2_{\mathrm{hel}}(P_{0,n}, P_{1,n}) \asymp \frac{1}{n}$

Quadratic mean differentiability

- expectation: if $\{p_\theta\}$ is "smooth" family of densities,

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

- using $\sqrt{p_\theta} \in L^2(P_0)$ can make this hold in mean square sense

Definition

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

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \dot{\ell}_\theta \sqrt{p_\theta} \right)^2 d\mu = o(\|h\|^2) \quad \text{as } h \to 0.$$

Existence of information and Hellinger distance

Proposition

If $\{P_\theta\}$ is QMD at θ with score $\dot{\ell}_\theta$, then

- $P_\theta \dot{\ell}_\theta = 0$ *and the Fisher information* $I_\theta := P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ *exists*
- *the Hellinger distance is* $d_{\text{hel}}^2(P_{\theta+h}, P_\theta) = \frac{1}{8} h^T I_\theta h + o(\|h\|^2)$

Quadratic mean differentiability is typical

- typical case: p_θ is a μ-probability density in a neighborhood of θ_0
- elements of $I_\theta = \int \frac{\dot{p}_\theta}{p_\theta} \frac{\dot{p}_\theta^T}{p_\theta} p_\theta d\mu$ are continuous in θ

Lemma

Under above conditions, $\{P_\theta\}$ is QMD near θ_0

Exponential families are QMD

Example (Exponential families)

Let $p_\theta(x) = \exp(\theta^T x - A(\theta))$, $A(\theta) = \log \int e^{\theta^T x} d\mu(x)$. Then $\{P_\theta\}$ is QMD with score

$$\dot{\ell}_\theta(x) = \nabla \log p_\theta(x) = x - \nabla A(\theta) = x - \mathbb{E}_\theta[X]$$

Local asymptotic normality

idea: for "nice" families, log-likelihood ratios should look locally quadratic (and give a CLT)

Definition (LAN families)

A family $\{P_{\theta,n}\}_{\theta\in\Theta}$ is *locally asymptotically normal* (LAN) at $\theta \in \operatorname{int}\Theta$ if there exists a sequence of random variables $\Delta_n \in \mathbb{R}^d$ and *information* (or precision) matrix $K \succeq 0$ such that

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

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

Gaussian shift families

Example (Gaussian shifts)

Let $P_{h,n}$ be distributions

$$Y_i = h + \xi_i, \quad \xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma), \ i = 1, \ldots, n.$$

Then

$$\log \frac{dP_{h/\sqrt{n},n}}{dP_{0,n}}(Y_1^n) = \sqrt{n} h^T \Sigma^{-1} \overline{Y}_n - \frac{1}{2} h^T \Sigma^{-1} h$$

Quadratic mean differentiable families

Proposition (QMD families are LAN)

If $\{P_\theta\}$ *is QMD at* θ *with score* $\dot{\ell}_\theta$ *and* $P_n = P^n_{\theta+h/\sqrt{n}}$, $P = P^n_\theta$,

$$\log \frac{dP_n}{dP}(X_1^n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_\theta(X_i)^T h - \frac{1}{2} h^T I_\theta h + o_P(1)$$

Optimal testing in a LAN family

- testing $H_0 : P_{0,n}$ vs. $H_1 : P_{h/\sqrt{n},n}$ as $n \to \infty$
- Neyman-Pearson (likelihood ratio) test is optimal: for $L_n := \frac{dP_{h/\sqrt{n},n}}{dP_{0,n}}$

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

- limits and alternatives:

$$(\log L_n, \log L_n) \xrightarrow[P_{0,n}]{d} \mathcal{N}\left(\begin{bmatrix} -\frac{1}{2}h^T Kh \\ -\frac{1}{2}h^T Kh \end{bmatrix}, \begin{bmatrix} h^T Kh & h^T Kh \\ h^T Kh & h^T Kh \end{bmatrix} \right)$$

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

Levels and power for the Neyman-Pearson test

some observations on $\phi_{n,h}$:

$$\alpha = \mathbb{E}_{P_0}[\phi_{n,h}] = P_0(\log L_n > c_{n,h}) + o(1)$$
$$= \mathbb{P}\left(\mathcal{N}\left(-\frac{1}{2}h^T Kh, h^T Kh\right) > c_{n,h}\right) + o(1)$$

- direct computation of thresholds $c_{n,h}$

$$c_{n,h} = (1-\alpha) \text{ quantile of } \mathcal{N}\left(-\frac{1}{2}h^T Kh, h^T Kh\right) + o(1)$$
$$= -\frac{1}{2}h^T Kh + z_{1-\alpha}\sqrt{h^T Kh} + o(\|h\|^2)$$

Observation (Neyman-Pearson power under local alternatives)

Under the above conditions, the power of $\phi_{n,h}$ is

$$\mathbb{E}_{h/\sqrt{n}}[\phi_{n,h}] \to 1 - \Phi\left(z_{1-\alpha} - \sqrt{h^T Kh}\right) = \Phi\left(z_\alpha + \sqrt{h^T Kh}\right)$$

Asymptotically optimal tests

Definition

A sequence $\{\phi_n\}$ of tests of θ_0 against θ_n is *asymptotically most powerful* (AMP) if

i. $\limsup_n \mathbb{E}_{\theta_0}[\phi_n] \le \alpha$

ii. for any sequence of tests ψ_n with $\limsup_n \mathbb{E}_{\theta_0}[\psi_n] \le \alpha$,

$$\limsup_n \{\mathbb{E}_{\theta_n}[\psi_n] - \mathbb{E}_{\theta_n}[\phi_n]\} \le 0.$$

Theorem

Let $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}}$ be LAN at θ_0. Then $\phi_n = \phi_n(X_1^n)$, $X_i \overset{\text{iid}}{\sim} P_\theta$ is AMP against local alternatives at level α iff $\mathbb{E}_{\theta_0}[\phi_n] \to \alpha$ and

$$\limsup_n \mathbb{E}_{\theta_0 + h/\sqrt{n}}[\phi_n] = 1 - \Phi(z_{1-\alpha} - h\sqrt{K}) = \Phi(z_\alpha + h\sqrt{K}).$$

Estimation lower bounds

idea: if we can show everything is Gaussian in the limit, we can get estimation lower bounds

Example

In model $X \sim \mathcal{N}(\theta, \Sigma)$, $\theta \in \mathbb{R}^d$ the minimax ℓ_2^2 risk is

$$\inf_{\widehat{\theta}} \sup_{\theta} \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|_2^2] = \mathbb{E}_\theta[\|X - \theta\|_2^2] = \mathrm{tr}(\Sigma).$$

Local asymptotic normality and sufficiency

- locally asymptotically normal family $\{P_{\theta,n}\}_{\theta\in\Theta}$ with

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

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

- Le Cam's third lemma:

$$\Delta_n \xrightarrow[P_{\theta+h/\sqrt{n},n}]{d} \mathcal{N}(Kh,K) \quad \text{i.e.} \quad Z_n := K^{-1}\Delta_n \xrightarrow[P_{\theta+h/\sqrt{n},n}]{d} \mathcal{N}(h,K^{-1})$$

idea: asymptotically, Δ_n should be sufficient for h

Heuristics: limiting Gaussianity

goal: show "local" experiments $P_{h/\sqrt{n},n}$ look like Gaussian shifts
heuristic: estimate h in a Bayesian model

$$h \sim \underbrace{\mathcal{N}(0, \Gamma)}_{=:\pi} \quad \text{and} \quad X^n \sim P_{h/\sqrt{n},n}$$

- posterior on h is approximately

$$\pi(h \mid X^n)$$
$$\propto \exp\left(-\frac{1}{2}(h - (K + \Gamma^{-1})^{-1}\Delta_n)^T (K + \Gamma^{-1})(h - (K + \Gamma^{-1})^{-1}\Delta_n)\right)$$

Notation for asymptotic Gaussian Posteriors

For $K \succeq 0$, $\Gamma \succ 0$ define

$$\mathsf{G}_{K,\Gamma}(\cdot \mid z) = \mathcal{N}\left((K + \Gamma^{-1})^{-1} K z, (K + \Gamma^{-1})^{-1}\right)$$

- posterior of $h \mid z$ in model

$$h \sim \mathcal{N}(0, \Gamma), \quad Z \mid h \sim \mathcal{N}(h, K^{-1})$$

- idea: for $Z_n := K^{-1}\Delta_n$, $h \mid Z_n$ should be almost $\mathsf{G}_{K,\Gamma}(\cdot \mid Z_n)$

Asymptotic Gaussian Posteriors

- prior $\pi^{\Gamma,c}$ is $\mathcal{N}(0,\Gamma)$ truncated to $\{h \in \mathbb{R}^d : \|h\| \le c\}$
- model:

$$h \sim \pi^{\Gamma,c}, \quad X^n \mid h \sim P_{h/\sqrt{n},n}, \quad \pi^{\Gamma,c}(\cdot \mid X^n) := \text{posterior on } h \mid X^n$$

- marginal $\overline{P}_n(\cdot) = \int P_{h/\sqrt{n},n}(\cdot) d\pi^{\Gamma,c}(h)$
- define $Z_n := K^{-1}\Delta_n(X^n) = K^{-1}\Delta_n$

Theorem (Le Cam)

Let above conditions hold. Then for all $\epsilon > 0$, there exist $C < \infty$ and $N < \infty$ such that $c \ge C$ and $n \ge N$ imply

$$\int \left\| \mathsf{G}_{K,\Gamma}(\cdot \mid Z_n(x^n)) - \pi^{\Gamma,c}(\cdot \mid x^n) \right\|_{\mathrm{TV}} d\overline{P}_n(x^n) \le \epsilon.$$

Remarks

- for LAN families, the *true* posterior under truncated Gaussian prior is (on average) Gaussian conditional on $Z_n = K^{-1}\Delta_n$
- other notions in which limits must be Gaussian

Theorem (van der Vaart Thm. 7.10)

Let $\{P_{\theta,n}\}$ be LAN at θ with information I_θ. If T_n converge in distribution under $P_{\theta+h/\sqrt{n},n}$ for each h, then

$$T_n \underset{P_{\theta+h/\sqrt{n},n}}{\overset{d}{\longrightarrow}} T$$

where T is a (randomized) statistic in $\{\mathcal{N}(h, I_\theta^{-1})\}_{h\in\mathbb{R}^d}$

Local Minimax Theorems

insight: we can reduce everything to estimation in Gaussian shift experiments $\mathcal{N}(h, K^{-1})$

Definition (Quasi-convexity)

A function $L : \mathbb{R}^d \to \mathbb{R}$ is *quasi-convex* if for each $\alpha \in \mathbb{R}$, the sublevel sets $\{x : L(x) \leq \alpha\}$ are convex

Anderson's Lemma

Lemma (Anderson)

Let L be symmetric and quasi-convex, $A \in \mathbb{R}^d \times \mathbb{R}^k$, and $X \sim \mathcal{N}(\mu, \Sigma)$. Then

$$\inf_v \mathbb{E}[L(AX - v)] = \mathbb{E}[L(A(X - \mu))] = \mathbb{E}[L(A\Sigma^{1/2}W)]$$

for $W \sim \mathcal{N}(0, I)$

The local asymptotic minimax theorem

Theorem

Let $L : \mathbb{R}^d \to \mathbb{R}$ be quasi-convex, symmetric, and bounded, and $\{P_{\theta,n}\}$ be LAN at θ_0 with precision (information) $K \succ 0$. Then

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \inf_{\widehat{\theta}_n} \sup_{\|\theta-\theta_0\|\le \frac{c}{\sqrt{n}}} \mathbb{E}_{P_{\theta,n}}\left[L(\sqrt{n}(\widehat{\theta}_n(X^n)-\theta))\right] \ge \mathbb{E}[L(K^{-1/2}W)], \quad W \sim \mathcal{N}(0,I).$$

Local asymptotic minimax theorem for QMD families

Corollary

Let $\{P_\theta\}$ be QMD at θ_0 with Fisher information I_{θ_0} and $\pi_{c,n}$ be $\mathcal{N}(\theta_0, \frac{b(c)}{n} I)$, where $b(c) = \sqrt{c}$, truncated to $\|\theta - \theta_0\| \leq c/\sqrt{n}$. Then

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \inf_{\widehat{\theta}_n} \int \mathbb{E}_{P_\theta^n}\left[L(\sqrt{n}(\widehat{\theta}_n - \theta))\right] d\pi_{c,n}(\theta) \geq \mathbb{E}[L(Z)]$$

for $Z \sim \mathcal{N}(0, I_{\theta_0}^{-1})$.

Proof of local asymptotic minimax theorem

- w.l.o.g. take $L(z) \in [0, 1]$ and rescale to perturbations $\{h : \|h\| \leq c\}$

Completing the proof: substitute in posteriors

- posterior $\pi(\cdot \mid x^n)$ on h similar to $\mathsf{G}_{K,\Gamma}(\cdot \mid z_n(x^n))$

Extensions and Corollaries

- differentiable functions: estimating $\psi(\theta)$ for a smooth function ψ of θ

- non-parametric scenarios: we wish to estimate $\theta(P) \in \mathbb{R}^d$ for a "smooth" function θ
   i. fix P_0, construct sub-models

$$dP_h \propto (1 + hg)_+ \, dP_0$$

   for function $g \in L^2(P_0)$, $P_0 g = 0$
   ii. evaluate derivatives

$$\lim_{h \downarrow 0} \frac{\theta(P_h) - \theta(P_0)}{h}$$