36-752: Advanced Probability **Spring 2018**

Lecture 22: April 24

Lecturer: Alessandro Rinaldo *Scribe: Ron Yurko*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

22.1 Efficient Likelihood Estimation and Testing

See the following for more in depth proofs and results:

- Chapter four of Wellner's notes
- Wellner's text on Empirical Process theory

22.1.1 Parametric Statistical Model

Let $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$, collection of probability measures on sample space $(\mathcal{X}, \mathcal{B}) = (\mathbb{R}^s, \mathcal{B}^s)$ indexed by a set $\Theta \subset \mathbb{R}^d$ parameter space

- $d =$ dimension of paramter space, Θ open subset
- Ex: $\theta = (\mu, \Sigma) \in \mathbb{R}^d \times C_{d,t} = \Theta$, where $C_{d,t}$ is the cone of PD $d \times d$ matrices.

P_θ to be $N(\mu, \Sigma), \mathcal{X} = \mathbb{R}^d$

22.1.2 Assumptions

- A0: Identifiability $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$
- A1: Support of $P_\theta = A \ \forall \ \theta$, (support of distribution is smallest closed set of S such that $P_A(S) = 1$)
- A2: $\exists$ σ-finite measure μ on sample space $(\mathcal{X}, \mathcal{B})$ such that $P_\theta << \mu \Rightarrow p_\theta = \frac{dP_\theta}{d\mu}$

We observe $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta_0}$ for some $\theta_0 \in \Theta$.

Notation $L_n(\theta) = \prod_{i=1}^n p_\theta(X_i)$ is the likelihood function, where $\ell_n(\theta) = \log L_n(\theta)$.

In HW5 we showed that under assumptions A0 - A2,

$$P_{\theta_0}(L_n(\theta_0) > L_n(\theta)) \to 1 \ \forall \theta \neq \theta_0, n \to \infty$$

(Used KL-divergence and LLN)

But because this actually makes no sense at all, we change the notation and replace P_{θ_0} by

$$\mathbb{P}\Big(\{\omega : L_n(\theta_0)(\omega) > L_n(\theta)(\omega)\}\Big)$$

meaning P_{θ_0} by

$$p^n_{\theta_0}\Big(\{(x_1,\dots,x_n) \in \mathbb{R}^{sn}, L_n(\theta_0, x_1,\dots,x_n) > L_n(\theta, x_1,\dots,x_n)\}\Big)$$

Definition 22.1 *The value $\hat{\theta}_n$ that maximizes $L_n(\theta)$ over Θ, if it exists and is unique, is the MLE of θ_0.*

$$\{\hat{\theta}_n\} = \{\theta^* : \theta^* = \underset{\theta\in\Theta}{sup} L_n(\theta)\}$$

the MLE is a singleton set.

But in many cases the MLE (1) may not exist and (2) may not be unique. Instead of $\hat{\theta}_n$, we may want to compute $\tilde{\theta}_n$, a root of the equation,

$$\dot{\ell}_n(\theta) = \nabla \ell_n(\theta) = 0$$

By the way $\Rightarrow$ MLE need not be consistent either! Neymann-Scott n independent pairs,

$$(X_n, Y_n) \sim N\Big(\begin{bmatrix}\mu_n\\ \mu_i\end{bmatrix}, \sigma^2 I_2\Big)$$

RHS: Unknown parameters $\mu_1, \dots \mu_n, \sigma$. Interested in estimating σ^2. Have simple estimator,

$$Z_i = X_i - Y_i \sim N(0, 2\sigma^2)$$
$$\frac{1}{2n}\sum_{i=1}^{n} Z_i^2 \sim \frac{\sigma^2}{n}\chi^2_n$$

which is unbiased and consistent!

Meanwhile the MLE of σ^2,

$$\frac{1}{4n}\sum_{i=1}^{n} Z_i^2 \xrightarrow{p} \frac{\sigma^2}{2}$$

is INCONSISTENT! Number of parameters is not fixed!!!

We move on to additional assumptions:

- A3: $\exists \Theta_0 \subset \Theta$ an open neighborhood of θ_0 such that
  - i $\log p_\theta(x)$ is twice continuously differentiable $a.e.[\mu]$
  - ii $\left|\frac{\partial^3 \log p_\theta(x)}{\partial\theta_i \partial\theta_j \partial\theta_k}\right| \leq M_{i,j,k}(x)\ \forall \theta \in \Theta$ where $M_{i,j,k}$ is such that $\mathbb{E}_{\theta_0}\Big[M_{i,j,k}(x)\Big]$ exists
- A4:
  - ii $\mathbb{E}_{\theta_0}[\dot{\ell}_j(\theta_0)] = 0, \dot{\ell}_j(\theta_0)$ is j^{th} coordinate of $\nabla\ell_n(\theta_0)$
  - iiii $\mathbb{E}_{\theta_0}[\dot{\ell}^2_j(\theta_0)] < \infty$
  - iiiii Let $I(\theta_0)$ be such that the i,j element is $\mathbb{E}_0[-\ddot{\ell}_{i,j}(\theta_0)] = \mathbb{E}_{\theta_0}[\dot{\ell}_j(\theta_0)\dot{\ell}_i(\theta_0)]$. Where $I(\theta_0)$ is the Fisher Information matrix assumed to be positive-definite and continuous function of θ in Θ_0.

(In practicality we are approximating and making assumptions that are close enough, not going to actually be able to verify all of these.)

Let $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}(\theta_0|X_i)$ and $\widetilde{I}(\theta_0) = I^{-1}(\theta_0)\dot{\ell}(\theta_0)$ so that

$$I^{-1}(\theta_0) = Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}(\theta_0, X_i)$$

called the efficient influence function!

Theorem 22.2 *Assume A0 - A4,*

i *With prob* $\to 1$, $\exists \tilde{\theta}_n$ *solution to likelihood equation* $\dot{\ell}_n(\theta) = 0$ *and* $\widetilde{\theta})n \xrightarrow{p} \theta_0$ *for some solution.*

ii $\tilde{\theta}_n$ *is asymptotic linear:*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = I^{-1}(\theta_0)Z_n + o_p(1)$$
$$\xrightarrow{D} N_d(0, I^{-1}(\theta_0))$$

which is the the Cramer-Rao lower bound.

iii $2log\tilde{\lambda}_n = 2log\frac{L_n(\tilde{\theta}_n)}{L_n(\theta_0)} \xrightarrow{D} \chi_d^2$, *likelihood ratio test*

<u>*Wald Test:*</u>

$$\sqrt{n}(\tilde{\theta}_n - \theta_0)^T \tilde{I}_n(\tilde{\theta}_n)\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{D} \chi_d^2$$

Where 3 ways of estimating Fisher Information matrix $\tilde{I}_n(\tilde{\theta}_n)$

- $I(\tilde{(\theta)}_n)$ *which we don't know how to compute*
- *Use* $\frac{1}{n}\sum_{i=1}^n \dot{\ell}(\tilde{\theta}_n, X_i)\dot{\ell}^T(\tilde{\theta}_n, X_i)$
- $-\frac{1}{n}\sum_{i=1}^n \ddot{\ell}(\theta_n, X_i)$

<u>*Rao Test:*</u> $R_n = Z_n^T \tilde{I}(\tilde{\theta}_n)Z)n \xrightarrow{D} \chi_d^2$

Proof.

i Existence and Consistency

Let $a > 0$ and $Q_a = \{\theta \in \Theta_0 : ||\theta - \theta_0|| = a\}$. We will show that for all a small enough,

$$P_{\theta_0}(\ell_n(\theta) < \ell_n(\theta_0) \forall \theta \in Q_d) \to 1$$

Use Taylor Series Expansion:

$$\frac{1}{n}(\ell_n(\theta) - \ell_n(\theta_0))$$
$$= \frac{1}{n}(\theta - \theta_0)^T \dot{\ell}_n(\theta_0) - \frac{1}{2}(\theta - \theta_0)^T(-\frac{1}{n}\ddot{\ell}_n(\theta_0))(\theta - \theta_0)$$
$$+\frac{1}{6n}\sum_i^d \sum_j^d \sum_k^d (\theta_i - \theta_i^0)(\theta_j - \theta_j^0)(\theta_k - \theta_k^0)\sum_i^n \gamma_{ijk}(X_i)M_{ijk}(X_i)$$

where $|\gamma_{ijk}(x_i)| \leq 1$.

Write this as:

$$= S_1 + S_2 + S_3$$

Next $S_1 \xrightarrow{p} 0$ by WLLN and Slutsky, $S_2 \xrightarrow{p} -\frac{1}{2}(\theta - \theta_0)I(\theta_0)(\theta - \theta_0)$ by the WLNN and Continuous Mapping Theorem where

$$(\theta - \theta_0)I(\theta_0)(\theta - \theta_0) \geq \lambda_{min}||\theta - \theta_0||^2 = \lambda_{min}a^2$$

where λ_{min} is the smallest eigenvalue of $I(\theta_i)$,

$$\inf_{x \neq 0} \frac{x^T A x}{x^T x} = \lambda_{min}(A)$$

$$x = (\theta - \theta_0)$$

$$A = I(\theta_0) S_3 \xrightarrow{p} \frac{1}{6} \sum_{i,j,k} (\theta_i - \theta_i^0)(\theta_j - \theta_j^0)(\theta_k - \theta_k^0)\mathbb{E}[\gamma_{ijk}(X_1)M_{ijk}(X_1)]$$

$$\leq \frac{1}{3}(da)^3 \sum_{ijk} m_{ijk}$$

See next lecture notes to see rest of proof.