Properties of MLE

Results based on some prob. theory

Law of large numbers

iid sample $X_1, \ldots, X_n$ with $E|X_1| < \infty$,

$$\bar{X} \xrightarrow{P} EX_1$$

$$\iff P(|\bar{X} - EX_1| > \varepsilon) \to 0 \text{ as } n \to \infty$$

Central Limit Theorem

iid $X_1, \ldots, X_n$ with $E|X_1| < \infty$ and $Var(X) = \sigma^2 < \infty$,

$$\sqrt{n}\,(\bar{X} - EX_1) \xrightarrow{d} N(0, \sigma^2)$$

$$\iff P(\sqrt{n}\,(\bar{X} - EX_1) \in (a, b]) \to \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

Slutsky's

If $A_n \xrightarrow{P} a$, $X_n \xrightarrow{d} X$ then $A_n X_n \xrightarrow{d} aX$

Properties to study; Consistency, asymptotic normality

and efficiency (next week)

note new notation instead of $\underset{\sim}{x}$

Setup $\underline{X}^n = (\underline{X}_1 \ldots \underline{X}_n)$ iid $p_{\theta_0}(x)$ where $p_{\theta_0}(x)$ model that belongs to model class $\mathcal{P} = \{p_\theta(x), \theta \in \Theta\}$

- We assume support of $p_\theta(x)$ does not depend on θ.

  (Example of violation $X \sim U[\theta, 1]$, $p_\theta(x) = \frac{1}{1-\theta} 1\{x \geq \theta\}$)

- We assume θ_0 is identified, meaning for $\theta_1 \neq \theta_2$ we have $p_{\theta_1}(x) \neq p_{\theta_2}(x)$ $\forall \theta \in \Theta$

- Likelihood $L(\theta | x^n) = \prod_{i=1}^{n} p_\theta(x_i)$ observed, function of θ for fixed $x^n = (x_1 \ldots x_n)$

- Estimate $\hat{\theta}(x^n) = \underset{\theta}{\text{argmax}}\ L(\theta | x^n)$

- <u>Estimator</u> $\hat{\theta}(\underline{X}^n)$ = function of r.v $\underline{X}^n = (\underline{X}_1 \ldots \underline{X}_n)$

  $= \underset{\theta}{\text{argmax}}\ L(\theta | \underline{X}^n)$

Why are MLEs good estimators?

Can often have
- consistency $\hat{\theta}(\underline{X}^n) \to^p \theta_0$
- asymptotic normality
  $\sqrt{n}\,(\hat{\theta}(\underline{X}^n) - \theta_0) \to^d N(0, V)$
- efficiency.
  V smallest variance among 'reasonable' classes of estimators

To show results, need some regularity conditions for p_θ.

Start with <u>Consistency</u>

- Not obvious that max of observed likelihood → max of expected

When is max of limit = limit of max?

Notation

$$L(\theta|\underline{X}^n) = \prod_{i=1}^n p_\theta(\underline{X}_i) \quad , \quad \ell(\theta|\underline{X}^n) = \log L(\theta|\underline{X}^n) = \sum_{i=1}^n \log p_\theta(\underline{X}_i)$$

<u>Normalized loglikelihood</u>

$$\boxed{\bar{\ell}(\theta|\underline{X}^n) = \frac{1}{n} \sum_{i=1}^n \log p_\theta(\underline{X}_i)}$$

depends on θ true density

- By WLLN, $\bar{\ell}(\theta|\underline{X}^n) \xrightarrow{P} E_{\theta_0}[\ell(\theta|\underline{X})] = \underbrace{\int \log p_\theta(x)\, p_{\theta_0}(x)\, dx}_{\text{denote by } \boxed{\ell_0(\theta)}}$

So, for large enough n we expect observed likelihood $\bar{\ell}(\theta)$ to be close to the expected $\ell_0(\theta)$.

We therefore expect the maximum of $\bar{\ell}$ to be close to the maximum of ℓ_0.

What is the maximum of $l_0(\theta)$?

<u>Information Inequality, Kullback-Leibler distance (KL)</u>

KL measures discrepancy between models f and g

Definition : $\boxed{KL(f,g) = E_f \log \frac{f(X)}{g(X)}} = \int \left(\log \frac{f(x)}{g(x)}\right) f(x)\,dx$

Now, $\boxed{KL(f,g) \geq 0}$ with equality if $f = g$ everywhere

Follows from Jensens inequality.

For convex $h(\,)$, $E(h(Z)) \geq h(E(Z))$

Here, use $h(\,) = -\log(\,)$

$$E_f \log \frac{f(X)}{g(X)} = E_f\left[-\log \frac{g(X)}{f(X)}\right] \geq -\log E_f\left[\frac{g(X)}{f(X)}\right]$$

$$= -\log \int \frac{g(x)}{f(x)} f(x)\,dx = -\log(1) = 0$$

Apply now to $f = p_{\theta_0}(x)$ and $g = p_\theta(x)$

$$\Rightarrow E_{\theta_0}\left[-\log \frac{p_\theta(X)}{p_{\theta_0}(X)}\right] \geq 0$$

$\Rightarrow l_0(\theta_0) \geq l_0(\theta)$ so θ_0 is the maximizer of $l_0(\theta)$.

We now need to show that the sequence of estimators $\hat{\theta}(\underline{X}^n)$ converges to θ_0 in probability
$n \to \infty$

We already have

- $\hat{\theta}(\underline{X}^n)$ is the maximizer of $\bar{l}(\theta)$
- θ_0 — " — " " of $l_0(\theta)$
- $\bar{l}(\theta) \xrightarrow{P} l_0(\theta)$ by WLLN
  $\forall \theta \in \Theta$

Conditions for lim max $\Leftrightarrow$ max lim is

- Compact parameter space Θ (or at least a condition on $\bar{l}(\theta)$ to not 'explode' for unbounded θ)

- <u>uniform convergence</u>

  That is,

  $$\sup_{\theta \in \Theta} |\bar{l}(\theta) - l_0(\theta)| \xrightarrow{P_{\theta_0}} 0$$

  $$\Leftrightarrow$$

  $$P_{\theta_0}\left(\sup_{\theta \in \Theta} |\bar{l}(\theta) - l_0(\theta)| > \varepsilon \right) \longrightarrow 0$$

- Assume also $p_\theta(x)$ continuous in $\theta \Rightarrow \bar{l}(\theta)$ continuous
- — " — $l_0(\theta)$ continuous in θ and
- θ_0 unique maximizer of $l_0(\theta)$.

Prove, given above conditions, that $\hat{\theta}(\underline{X}^n) \xrightarrow{p} \theta_0$.

<u>Idea</u>: Prove $\hat{\theta}(\underline{X}^n)$ guaranteed to be in neighborhood $(\theta_0 - \varepsilon, \theta_0 + \varepsilon) = \Theta(\varepsilon)$ for large enough n.

How guarantee this?

Need to make sure that $\underline{X}^n$ eventually such that $\bar{l}(\theta)$ is maximized in $\Theta(\varepsilon)$.

That is, prove that for large enough n, maximizing $\bar{l}(\theta)$ must lead to $\hat{\theta}(\underline{X}^n)$ in $\Theta(\varepsilon)$.

<u>Outline of proof structure</u>

We will consider 2 regions of Θ

$\Theta(\varepsilon)$ $\qquad$ $\Theta \cap \Theta(\varepsilon)^c$

As $n \to \infty$, can we be sure $\hat{\theta}(X^n)$ falls in $\Theta(\varepsilon)$ with probability $\to 1$?

Well, depends on how $\bar{l}(\theta)$ behaves in $\Theta(\varepsilon)$ and $\Theta \cap \Theta(\varepsilon)^c$.

Look in region $\theta \in \Theta \cap \Theta(\varepsilon)^c$.

Now, having assumed $l_0(\theta)$ continuous and $\Theta \cap \Theta(\varepsilon)^c$ is compact

$\longrightarrow$ $\underset{\theta \in \Theta \cap \Theta(\varepsilon)^c}{l_0(\theta)}$ attains a maximum at some θ^* in $\Theta \cap \Theta(\varepsilon)^c$

Since we assumed θ_0 is the unique maximizer of $l_0(\theta)$

we have $l_0(\theta^*) < l_0(\theta_0)$, or $\left(l_0(\theta_0) - l_0(\theta^*)\right) = \delta > 0$ (*)

for some value δ.

Now, define an event (sample outcome)

$$A(\bar{X}^n) = \left[\sup_{\theta \in \Theta \cap \Theta(\varepsilon)^c} |\bar{l}(\theta|\bar{X}^n) - l_0(\theta)| < \frac{\delta}{2}\right]$$

I.e. Consider only samples $\bar{X}^n$ where $\bar{l}(\theta|\bar{X}^n)$

is close to $l_0(\theta)$ for all $\theta \in \Theta \cap \Theta(\varepsilon)^c$.

For such $\bar{X}^n$, i.e. if $A(\bar{X}^n)$ is true,

$$\bar{l}(\theta) < l_0(\theta) + \frac{\delta}{2} \underbrace{\leq}_{\substack{\text{Since}\\ l_0(\theta)\\ \text{max @ } \theta^*\\ \text{if } \theta \text{ has to}\\ \text{be in } \Theta \cap \Theta(\varepsilon)^c}} l_0(\theta^*) + \frac{\delta}{2} \underbrace{=}_{\text{by (*)}} l_0(\theta_0) - \delta + \frac{\delta}{2} = l_0(\theta_0) - \frac{\delta}{2}$$

So, if $A(\bar{X}^n)$ is true, Then for $\theta \in \Theta \cap \Theta(\varepsilon)^c$ we have

$$\bar{l}(\theta) < l_0(\theta_0) - \frac{\delta}{2}$$

Look in region $\theta \in \Theta(\varepsilon)$.

Define a sample outcome, event,

$$B(\underline{X}^n) = \left[\sup_{\theta \in \Theta(\varepsilon)} |\bar{l}(\theta) - l_0(\theta)| < \frac{\delta}{2} \right]$$

[Question from class: why 2 events.
Consider this for fixed n. Since $\sup_{\theta \in \Theta(\varepsilon)}$ and $\sup_{\theta \in \Theta \cap \Theta(\varepsilon)^c}$ look at different values for θ, it is possible that for a fixed sample $\underline{X}^n$, $B(\underline{X}^n)$ might be true but not $A(\underline{X}^n)$]

Point is that as $n \to \infty$ both A and B are true

If $B(\underline{X}^n)$ is true,

$$\bar{l}(\theta) > l_0(\theta) - \frac{\delta}{2} \quad \forall \theta \in \Theta(\varepsilon)$$

In particular $\bar{l}(\theta_0) > l_0(\theta_0) - \frac{\delta}{2}$

For fixed n, $A(\underline{X}^n)$ and/or $B(\underline{X}^n)$ could be true / false and so we cannot really say where maximizer of $\bar{l}(\theta)$ will fall.
However, let $n \to \infty$ and we have

$$P_{\theta_0}\left(\sup_{\theta \in \Theta} |\bar{l}(\theta) - l_0(\theta)| > \varepsilon \right) \to 0$$

all values of θ

and so $P_{\theta_0}(A(\underline{X}^n) \cap B(\underline{X}^n)) \xrightarrow{n \to \infty} 1$

So eventually, both $A(\underline{X}^n)$ and $B(\underline{X}^n)$ are true

But if both $A(\underline{X}^n)$ and $B(\underline{X}^n)$ are true, then

in $\Theta(\varepsilon)$ $\bar{l}(\theta) > l_0(\theta) - \frac{\delta}{2}$ $\forall \theta \in \Theta(\varepsilon) \Rightarrow$ and $\bar{l}(\theta_0) > l_0(\theta_0) - \frac{\delta}{2}$

in $\Theta \cap \Theta(\varepsilon)^c$ $\bar{l}(\theta) < l_0(\theta_0) - \frac{\delta}{2}$

and since $\hat{\theta}(\underline{X}^n)$ is defined as the argmax of $\bar{l}(\theta)$

we will clearly have $\hat{\theta}(\underline{X}^n) \in \Theta(\varepsilon)$.

So $A(\underline{X}^n) \cap B(\underline{X}^n)$ true $\Rightarrow$ $\hat{\theta}(X^n) \in \Theta(\varepsilon)$

and so $P(A(\underline{X}^n) \cap B(\underline{X}^n)) = P(\hat{\theta}(X^n) \in \Theta(\varepsilon))$

$$\Big\downarrow {\scriptstyle n \to \infty}$$

$$1$$

$\rightarrow$ $\hat{\theta}(\underline{X}^n)$ is consistent.

So, <u>given</u> uniform convergence, we're all set. This can be more difficult to prove ...

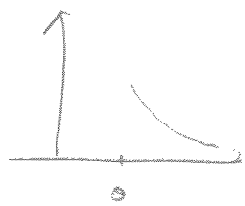
When support of $p_\theta(x)$ depends on θ (Example of breakdown of assumptions)

Here, we can usually get consistency, but not normality

Example

$X_i \sim$ iid shifted exponential

$$p_\theta(x) = e^{-(x-\theta)} I\{x \geq \theta\}$$



$$\Rightarrow L(\theta|\underline{x}) = \exp\left(-\sum_{i=1}^{n}(x_i-\theta)\right) I\{\min x_i \geq \theta\}$$

$\Rightarrow$ MLE is $\min(x_i) = X_{(1)} = \hat{\theta}$

Now, here $P_{\theta_0}(|X_{(1)}-\theta| > \varepsilon)$

$$= P_{\theta_0}(X_{(1)} > \theta_0+\varepsilon) + P_{\theta_0}(X_{(1)} < \theta_0-\varepsilon)$$

true for all $= \prod_{i=1}^{n} P_{\theta_0}(X_i > \theta_0+\varepsilon) = \exp(-n\varepsilon) \to 0$

so $\hat{\theta}$ is consistent

However, $\sqrt{n}(X_{(1)}-\theta_0)$ cannot be normal since by definition $X_{(1)}$ is always greater than θ_0.

Here,

$$P_{\theta_0}(n(X_{(1)}-\theta_0) \geq a) = P_{\theta_0}\left(X_{(1)} \geq \frac{a}{n}+\theta_0\right)$$

$$= \left(P_{\theta_0}\left(X_i \geq \frac{a}{n}+\theta_0\right)\right)^n = \exp(-a)$$

so $n(X_{(1)}-\theta_0) \xrightarrow{d} \mathrm{Exp}(1)$ (faster rate of convergence here than $\sqrt{n}$)