# Why we like MLEs



- Consistency $\hat{\theta} \to^{p} \theta_0$
- Asymptotic normality $\sqrt{n}(\hat{\theta} - \theta_0) \to^{d} N(0, V_0^2)$ , $V_0^2 = \bar{I}(\theta_0)^{-1}$
- Asymptotic efficiency $V_0^2 = \bar{I}(\theta_0)^{-1}$

Consistency; we showed that the maximizer of the likelihood converges to the maximizer of the expected likelihood (Uniform conv. in prob of $\bar{\ell}$ to $\ell_0$)

Asympl Normality, we used the CLT applied to the score function and Slutsky's theorem (obs $\bar{I} \to$ expected $\bar{I}$) to conclude AN of $\hat{\theta}$

## Efficiency

Let's say we want to compare estimators $T^1(X^n)$ and $T^2(X^n)$

Let's say $V(T^1(X^n)) = \sigma_{n,1}^2(\theta) \to \sigma_1^2(\theta)$ and $V(T^2(X^n)) = \sigma_{n,2}^2(\theta) \to \sigma_2^2(\theta)$

The asymptotic efficiency of $T^1$ relative to $T^2$ is $\frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)} = e_{12}(\theta)$

Note, $e_{12}(\theta)$ may be $>1$ for some $\theta$ and $<1$ for others!

If $\sigma_1^2(\theta) = \frac{(g'(\theta))^2}{I(\theta)}$ we say that $T'(X^n)$ is asymptotically efficient.

This typically holds for MLEs <u>BUT</u> we can also get this result for other estimators

- UMVUEs (more later)

# Breakdown of assumptions

Problems we can encounter:

- Non-existence of MLE
- multiple solutions
- support of $p_\theta(x)$ depends on $\theta$

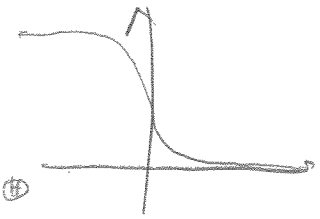
## Example of non-existence

(can happen for all $\underset{\sim}{X}$, or just some outcomes

Usually when $\Theta$ not compact or $\ell(\theta)$ not continuous in $\theta$

$X \sim Be\left(\frac{1}{1+e^{\theta}}\right)$

If observe $x=1 \Rightarrow L(\theta|1) = \frac{1}{1+e^{\theta}}$



$\Rightarrow$ max occurs on boundary of $\Theta$

$\Rightarrow$ max $\hat{\theta} = -\infty$ !

## Example of multiple solutions

Usually happens when $\theta$ can't be identified.

Example regression w. nonfull rank design matrix

$$y \sim N(X\theta, \sigma^2 I) \qquad \text{rank}(X) < p$$

$$(\theta_1, \ldots, \theta_p)$$

Sometimes, multiple solutions is a problem for small $n$ and goes away for large enough $n$.

Unless — # parameters increase with the sample size!

Example $\underset{\sim}{X} = (X_1, \dots, X_n)$ $\quad X_{ij} \sim N(\mu_i, \sigma^2)$ $\quad j = 1, 2$ (bivariate normal)

$n+1$ parameters $(\mu_1, \dots, \mu_n, \sigma^2) = \theta$

$$L(\theta | \underset{\sim}{X}) = \prod_{i=1}^{n} \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 \exp\left( -\frac{1}{2\sigma^2} \sum_{j=1}^{2} (X_{ij} - \mu_i)^2 \right)$$

$$\Rightarrow \hat{\mu}_i = \frac{1}{2}(X_{i1} + X_{i2}) \longrightarrow \text{no convergence here as } n \to \infty !$$

MLEs

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} (X_{ij} - \hat{\mu}_i)^2$$

$$\hat{\sigma}^2 \xrightarrow{P} \frac{\sigma^2}{2} \quad ! \qquad \text{since } \hat{\sigma}^2 = \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2$$

$$= \frac{1}{4n} \sum_{i=1}^{n} \left( \frac{X_{i1} - X_{i2}}{\sqrt{2}\sigma} \right)^2 2\sigma^2$$

$$\sim N(0,1)$$

$$= \frac{\sigma^2}{2} \cdot \frac{1}{n} \sum_{i=1}^{n} (N(0,1))^2$$

$$\chi^2_1 \text{ (mean 1)}$$

$$\to p\ 1$$

$$\to \frac{\sigma^2}{2} \quad !$$

When $\dim(\Theta)$ increases with $n$ we can have convergence to values other than $\theta_0$, or no convergence at all.

Example ctd

$X^n = (X_1, \ldots, X_n)$ $X_i$ iid with $X_i = (X_{i1}, \ldots, X_{in})$, $X_{ij} \sim N(\mu_i, \sigma^2)$, $j = 1, \ldots, n$

Again, have $n+1$ parameters to estimate.

$$L(\theta | X^n) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (X_{ij} - \mu_i)^2\right)$$

$$\Rightarrow \hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij} \quad i = 1, \ldots, n \quad \xrightarrow{p} \mu_i \text{ by WLLN}$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n^2} \sum_i \sum_j (X_{ij} - \hat{\mu}_i)^2 \quad \xrightarrow{p} \hat{\sigma}^2 \to \sigma^2$$

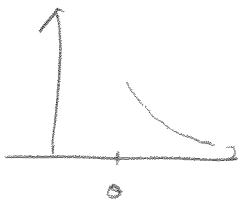
## When support of $p_\theta(x)$ depends on $\theta$ (Example of breakdown of assumptions)

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

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

since $n$ 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 $\theta_0$.

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

Before we discuss what happens to MLE procedure when we use the wrong model assumptions, we make a detour and discuss the general estimation problem.

---

Want to estimate $g(\theta)$ using estimator $T(X^n)$.
Is $T(X^n)$ a good estimator? "Good" means what?

First, what other type of estimators are there?

⊛ Method of Moments

Look @ $r$ first moments $E_\theta(X^j)$ $j = 1, \dots, r$

Given $f_\theta(x)$, $E_\theta(X^j) = m_j(\theta)$.

$m_j(\theta)$ can be estimated by the sample moment $\frac{1}{n}\sum_{i=1}^{n} X_i^j$

⇒ Equation system $m_j(\theta) = \overline{X^j}$ $j \geq 1$

⇒ solve for $\theta$.

⊖ Note, can have many choices for estimators of $\theta$ here.

⊕ Usually easy to work with and often consistent.

⊛ Via a loss function, e.g. least squares.

$$\min_\theta \sum (y_i - g_i(\theta))^2$$

(Many other variants of loss fcns to use, included penalized loss e.g. $\|y - g(\theta)\|^2 + \lambda \|g(\theta)\|^p$ etc.)

## Comparing estimates

Ok, so we have an estimator $T(X^n)$ for $g(\theta)$ (estimand)

Is $T(X^n)$ a 'good' estimator?

Commonly used criterion is the $\boxed{MSE(T,\theta) = E_\theta(T(X^n) - g(\theta))^2}$

$\Rightarrow$ Usually easy to work with in contrast with

eg. $E_\theta|T(X^n) - g(\theta)|$ or other norms.

Now, expanding the square we have

$$MSE(T,\theta) = Var_\theta T(X^n) + B^2(T,\theta)$$

$$\text{bias}^2 = (E_\theta(T(X^n)) - g(\theta))^2$$

Now, using $MSE(T,\theta)$ we want to find the best estimator.

Problem $MSE(T,\theta)$ usually depends on value of true $\theta$!

So, while an estimator $T'(X^n)$ may be better in terms of $MSE(T,\theta)$ than $T^*(X^n)$ for some $\theta$, it may be worse for others!

No uniformly optimal estimator!

Definition: inadmissable

If two estimators $S$ and $T$ are such that $MSE(T,\theta) \leq MSE(S,\theta)$ with strict inequality for some $\theta$, $T$ is better than $S$ and $S$ is called inadmissable

There are no best estimators though.

If $\theta = \theta_0$ fixed value, then $S(X^n) = \theta_0$ has $MSE(S, \theta_0) = 0$ and to beat it for any value $\theta_0 \Rightarrow T(X^n)$ must have $MSE(T, \theta) = 0$ everywhere → impossible task.

---

Problem is that we have put no restrictions on form of estimators.

⇒ What we tend to focus on is <u>classes of estimators</u> that can be compared using $MSE(\theta)$ (or some other risk function) and optimality is possible

**<u>Alt 1</u>**

Restrict to unbiased estimators

**<u>Alt 2</u>**

Use an <u>average</u> risk as your criteria

$\int MSE(T, \theta)\pi(\theta)\, d\theta$

(Bayes)

**<u>Alt 3</u>**

Use worst-case scenario as your criteria

$\max_{\theta} MSE(T, \theta)$

(minimax)

Let's focus on Alt 1, unbiased estimators.

Here $E_\theta(T(X^n)) = g(\theta)$ so $B(T,\theta) = 0$

and $MSE(T,\theta) = Var_\theta T(X^n)$

Now, within this class of estimators we say that $T$ is the uniformly minimum variance unbiased estimator (UMVUE) if $V(T(X^n)) \leq V(S(X^n))$ for any unbiased estimator $S$.

Which $T$ is optimal? Need the following:

Rao-Blackwell Theorem

If $T(X^n)$ is sufficient for $\theta$ [Remember: means $X \mid T(X)$ index of $\theta$]

and $S(X^n)$ is an estimator for $g(\theta)$;

By conditioning on $T(X^n)$ we can always get a better, or as good as $S$, estimator for $g(\theta)$.

Call this $T^*(X^n) = E(S(X^n) \mid T(X^n))$

Why? $\forall \theta \quad E_\theta(T^*(X) - g(\theta))^2 \leq E_\theta(S(X^n) - g(\theta))^2$

$MSE(T^*,\theta)$

with strict inequality only if $T^*(X^n) = S(X^n)$

Follows from $E_\theta[E(S(X^n) \mid T(X^n)] = E(S(X^n))$ so $B(T,\theta) = B(S,\theta)$

$= E(T^*(X^n))$

Now $Var(E(S \mid T)) \leq Var(S)$ since conditioning reduces variability $\Rightarrow E(E(S|T) - E(S|T))^2 = E(E(S|T) - E(S))^2 < E(S - E(S))^2$ #

## Completeness

$T$ is complete if $E_\theta(h(T)) = 0, \forall \theta$ is only possible if $h(T) = 0$

---

⇒ Main result, due to Lehmann & Scheffé

> If $T(X^n)$ is complete sufficient and $S(X^n)$ is an unbiased estimator of $g(\theta)$, then $T^*(X^n) = E(S(X^n) \mid T(X^n))$ is an UMVUE of $g(\theta)$. ($T^*$ is unique if $V(T^*) < \infty \ \forall \theta$)

Proof

From Rao-Blackwell we know $V(T^*) < V(S)$ for an unbiased estimator $S$

Now, we need to show it didn't matter what $S$ was so we can't improve on $T^*$ with another $S$.

But from completeness, if $T_1^* = h_1(T)$ and $T_2^* = h_2(T)$ are unbiased estimators of $g(\theta)$ (from $S_1$ and $S_2$), then

$$E(h_1(T) - h_2(T)) = E(h_1(T)) - E(h_2(T)) = 0 \quad \forall \theta$$

But since we assumed $T$ was complete ⇒ $h_1 = h_2$ must hold and it didn't matter what $S$ was.

---

Key to the UMVUE result is having a complete sufficient statistic. → usually true for exponential families.

- Use 1: Find unbiased estimator $S(X^n)$ and derive $E(S(X^n) \mid T(X^n))$
- Use 2: If we have such an $T(X^n)$, and we find an estimator $h(T(X^n))$ which is unbiased for $g(\theta)$ ⇒ $h(T(X^n))$ is UMVU since $E(h(T(X^n)) \mid T(X^n)) = h(T(X^n))$

## Exponential family

Let's assume we have $p_\theta(x)$ has support indep of $\theta$ and is 'nice' in term of exchange derivatives & integrals etc....

$\Rightarrow$ If $T^n(X^n)$ is an unbiased estimator for $g(\theta)$ and attains the lower bound in CR $\forall \theta \Rightarrow p_\theta$ is a one-parameter exp. family of form $p_\theta(x) = \exp\{\eta(\theta)T^n(x) - A(\theta) + c(x)\}$

$\Leftarrow$ If $p_\theta$ is a one-parameter exp. family given by $p_\theta(x) = \exp\{\eta(\theta)T(x) - A(\theta) + c(x)\}$ then (provided $\eta(\theta)$ has cts derivatives) $T(X^n)$ achieves the bound and is a UMVU for $E_\theta T(x)$

BUT $\rightarrow$ conditions may fail

(or)

an UMVU estimator for $g(\theta)$ may exist but not hit the bound (depends on $g(\cdot)$).

---

Some pros and cons of MLE vs UMVUE.

- Large $n$, usually doesn't matter, but MLEs tend to be easier to compute
- MLEs are invariant in sense that if $\hat\theta$ is a MLE of $\theta$ then $h(\hat\theta)$ is an MLE of $h(\theta)$.
  Unbiasedness property is not always so, i.e. $\hat\theta$ can be unbiased for $\theta$ but $g(\hat\theta)$ not for $g(\theta)$.
- If $p_\theta$ exp. family, both MLE and UMVUE functions of $T(X)$

## Examples

Poisson($\theta$) $\quad p_\theta(x) = \dfrac{e^{-\theta}\theta^x}{x!} \quad x = 0, 1, \ldots$

MLE of $\theta$ : $L(\theta|x^n) = \prod_{i=1}^n \dfrac{e^{-\theta}\theta^{x_i}}{x_i!} = \exp\{\underbrace{\log(\theta)}_{\eta(\theta)}\underbrace{\textstyle\sum x_i}_{T(x^n)} - n\theta + \textstyle\sum x_i!\}$

exp family

$\Rightarrow l(\theta|x^n) = \text{constant} - n\theta + \sum x_i \log(\theta)$

$\Rightarrow \dfrac{\partial}{\partial\theta} l(\theta|x^n) = -n + \dfrac{\sum x_i}{\theta} = 0 \quad \Rightarrow \hat\theta = \dfrac{\sum x_i}{n} = \bar x$

$\Rightarrow \dfrac{\partial^2}{\partial\theta^2} l(\theta|x^n) = -\dfrac{\sum x_i}{\theta^2} \Rightarrow \mathcal{FI}(\theta) = \dfrac{n\theta}{\theta^2} = \dfrac{n}{\theta}$

$V(\bar X) = \dfrac{V(X)}{n} = \dfrac{\theta}{n} = \mathcal{FI}(\theta)^{-1}$ hits bound.

Now consider $g(\theta) = \psi = e^{-\theta}$

$\log p_\psi(x) = (\log(-\log\psi))\textstyle\sum x_i + n\log\psi + \text{constant}$

$\dfrac{\partial}{\partial\psi} = \dfrac{n}{\psi} + \dfrac{\sum x_i}{\psi\log\psi} = 0 \Rightarrow \log\psi = -\bar x \Rightarrow \psi = e^{-\bar x}$

$= g(\hat\theta)$

What about UMVU? (sample size 1)

Consider $T(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{o/w} \end{cases}$, i.e. $T(X) = 1\{X=0\}$

$E(T(X)) = E(1\{X=0\}) = e^{-\theta}$, so $T(X)$ is an unbiased estimator for $e^{-\theta}$

$$V(T(X)) = E(T(X)^2) - e^{-2\theta}$$

$$= e^{-\theta}(1 - e^{-\theta})$$

Bound $\frac{(g'(\theta))^2}{I(\theta)} = \theta e^{-2\theta}$ $\Rightarrow$ $V(T(X)) \geq \theta e^{-2\theta}$

so does not hit the bound.

But $E(\mathbb{1}\{X=0\}|X) = \mathbb{1}\{X=0\}$ so $\mathbb{1}\{X=0\}$ is UMVU for $e^{-\theta}$.

Note, for mle $\hat{\theta} = e^{-X}$ is biased $\Rightarrow$ asymptotically unbiased

but $V(e^{-X}) < V(\mathbb{1}\{X=0\})$

(Try it at home, for $n=1, \ldots$)