Outline

1) Log-likelihood and score
2) Fisher information
3) Cramér-Rao Lower Bound
4) Hammersley-Chapman-Robbins Ineq.

Log-likelihood, score

Assume $\mathcal{P}$ has densities p_θ wrt μ, $\Theta \subseteq \mathbb{R}^d$

Common support: $\{x : p_\theta(x) > 0\}$ same $\forall \theta$

Recall $\ell(\theta; X) = \log p_\theta(X)$,

Thought of as random function of θ

Def The score is $\nabla \ell(\theta; X)$; plays a key role in many areas of statistics, esp. asymptotics.

Can think of as "local sufficient statistic":

$$p_{\theta_0+\eta}(x) = e^{\ell(\theta_0+\eta\,;\,x)}$$

$$\approx e^{\eta' \nabla \ell(\theta_0; x)} p_{\theta_0}(x) \qquad \text{for } \eta \approx 0$$

Differential identities: (assuming enough regularity)

$$1 = \int_{\mathcal{X}} e^{\ell(\theta; x)} d\mu(x)$$

$$\frac{\partial}{\partial \theta_j} \Rightarrow \quad 0 = \int \frac{\partial}{\partial \theta_j} \ell(\theta; x)\, e^{\ell(\theta; x)} d\mu(x)$$

$$\Rightarrow \quad \mathbb{E}_\theta \left[\nabla \ell(\theta; X) \right] = 0$$

Only true if these are the same value of θ!

$$\frac{\partial}{\partial\theta_k} \Rightarrow \quad 0 = \int \left(\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_k} + \frac{\partial \ell}{\partial\theta_j} \cdot \frac{\partial \ell}{\partial\theta_k} \right) e^{\ell} \, d\mu$$

$$= \mathbb{E}_\theta\left[\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_k}\right] + \mathbb{E}_\theta\left[\frac{\partial \ell}{\partial\theta_j} \frac{\partial \ell}{\partial\theta_k}\right]$$

$$\Longrightarrow \quad \underbrace{\mathrm{Var}_\theta[\nabla \ell(\theta; X)]}_{= J(\theta)} = \mathbb{E}_\theta[-\nabla^2 \ell(\theta; X)]$$

same θ same θ

Called "<u>Fisher Information</u>"

[It is possible to extend this definition to certain cases where ℓ is not even differentiable, e.g. Laplace location family, but for our purposes we can just assume "sufficient regularity."]

Try with another statistic $\delta(X)$, let

$$g(\theta) = \mathbb{E}_\theta[\delta(X)] \quad \text{("unbiased estimator")}$$

$$g(\theta) = \int \delta e^{\ell} \, d\mu$$

$$\Rightarrow \quad \nabla g(\theta) = \int \delta \nabla \ell \, e^{\ell} \, d\mu = \mathbb{E}_\theta[\delta(X) \nabla \ell(\theta; X)]$$

$$= \mathrm{Cov}_\theta(\delta(X), \nabla \ell(\theta; X))$$

why? Since $\mathbb{E}\nabla\ell = 0$

Combining these results with Cauchy-Schwarz gives us the <u>Cramér-Rao Lower Bound</u> or <u>Information Lower Bound</u>:

<u>1-param:</u> $\text{Var}_\theta(\delta) \cdot \text{Var}_\theta(\dot{\ell}(\theta;X)) \geq \text{Cov}_\theta(\delta, \dot{\ell}(\theta;X))^2$

$$\Rightarrow \quad \text{Var}_\theta(\delta) = \dot{g}(\theta)^2 / J(\theta)$$

<u>$\theta \in \mathbb{R}^d$, $g(\theta) \in \mathbb{R}$:</u> $\text{Var}_\theta(\delta) \geq \nabla g(\theta)' J(\theta)^{-1} \nabla g(\theta)$

<u>Interp</u>: If $g(\theta)$ is estimand, no unbiased estimator can have smaller verience than $\nabla g(\theta)' J(\theta)^{-1} \nabla g(\theta)$

<u>Ex.</u>: (i.i.d. sample)

$$X_1, \ldots, X_n \overset{iid}{\sim} p_\theta^{(1)}(x) \qquad \theta \in \Theta$$

$$X \sim p_\theta(x) = \prod_i p_\theta^{(1)}(x_i)$$

Let

$$\ell_1(\theta; x_i) = \log p_\theta^{(1)}(x_i)$$

$$\ell(\theta; x) = \sum_i \ell_1(\theta; x_i)$$

$$\begin{aligned} J(\theta) &= \text{Var}_\theta(\nabla \ell(\theta; X)) \\ &= \text{Var}_\theta\Big(\sum_i \nabla \ell_1(\theta; X_i)\Big) \\ &= n J_1(\theta) \end{aligned}$$

where $J_1(\theta)$ is Fisher info in single observation

$\Rightarrow$ Lower bound scales like n^{-1} ($SD \asymp n^{-1/2}$ for "regular" families)

Efficiency

CRLB is not nec. attainable.

We define the efficiency of an unbiased estimator as:

$$\text{eff}_\theta(\delta) = \frac{\text{CRLB}}{\text{Var}_\theta(\delta)} \left(= \frac{1/J(\theta)}{\text{Var}_\theta(\delta)} \quad \text{if } g(\theta) = \theta \in \mathbb{R} \right)$$

$$\text{eff}_\theta(\delta) \leq 1$$

We say $\delta(X)$ is efficient if $\text{eff}_\theta(\delta) = 1 \ \forall \theta$

Depends on $\text{Corr}_\theta(\delta(X), \nabla \ell(\theta; X))$:

$$\text{eff}_\theta(\delta) = \frac{\text{Cov}_\theta^2(\delta(X), \dot{\ell}(\theta; X))}{\text{Var}_\theta(\delta) \cdot \text{Var}_\theta(\dot{\ell}(\theta))}$$

$$= \text{Corr}_\theta^2(\delta, \dot{\ell}(\theta))$$

$$\leq 1$$

$\delta(X)$ is efficient $\Leftrightarrow$ $\text{Corr}_\theta^2(\delta, \dot{\ell}(\theta)) = 1 \ \forall \theta$

Rarely achieved in finite samples but we can approach it asymptotically as $n \to \infty$

Hammersley-Chapman-Robbins Ineq.

CRLB requires differentiation under integral

Can make more general statement if we replace $\nabla \ell(\theta;x)$ with finite-difference:

$$\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1 = e^{\ell(\theta+\varepsilon;x) - \ell(\theta;x)} - 1$$

$$\left(\approx \varepsilon' \nabla \ell(\theta;x) \quad \text{small } \varepsilon \right)$$

$$\mathbb{E}_\theta\left[\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right] = \int \left(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right) p_\theta \, d\mu = 1 - 1 = 0$$

(assuming common support, or $P_{\theta+\varepsilon} \ll P_\theta$)

$$\begin{aligned}
\mathrm{Cov}_\theta\left(\delta, \frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right) &= \int \delta \left(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right) p_\theta \, d\mu \\
&= \mathbb{E}_{\theta+\varepsilon}(\delta) - \mathbb{E}_\theta(\delta) \\
&= g(\theta+\varepsilon) - g(\theta)
\end{aligned}$$

$$\Rightarrow \mathrm{Var}_\theta(\delta) \geq \frac{(g(\theta+\varepsilon) - g(\theta))^2}{\mathbb{E}_\theta\left[\left(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right)^2\right]}$$

CRLB follows from $\varepsilon \to 0$, but $\sup_\varepsilon$ gives better bound

Ex. Exponential Families

$$p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x)$$

$$\ell(\eta; x) = \eta' T(x) - A(\eta) + \log h(x)$$

$$\nabla \ell(\eta; X) = T(x) - \nabla A(\eta)$$

$$= T(X) - \mathbb{E}_\eta T(X)$$

$$\mathrm{Var}_\eta(\nabla \ell(\eta)) = \mathrm{Var}_\eta(T(X)) = \nabla^2 A(\eta)$$

$$\nabla^2 \ell(\eta; X) = -\nabla^2 A(\eta)$$

$$\mathbb{E}_\eta[-\nabla^2 \ell(\eta; X)] = \nabla^2 A(\eta) \quad \checkmark$$

So any unbiased est. of η has

$$\mathrm{Var}_\eta(\delta) \geq \nabla^2 A(\eta)^{-1}$$

Curved family: $p_\theta(x) = e^{\eta(\theta)'T(x) - B(\theta)} h(x), \quad \theta \in \mathbb{R}$

$B(\theta) = A(\eta(\theta))$

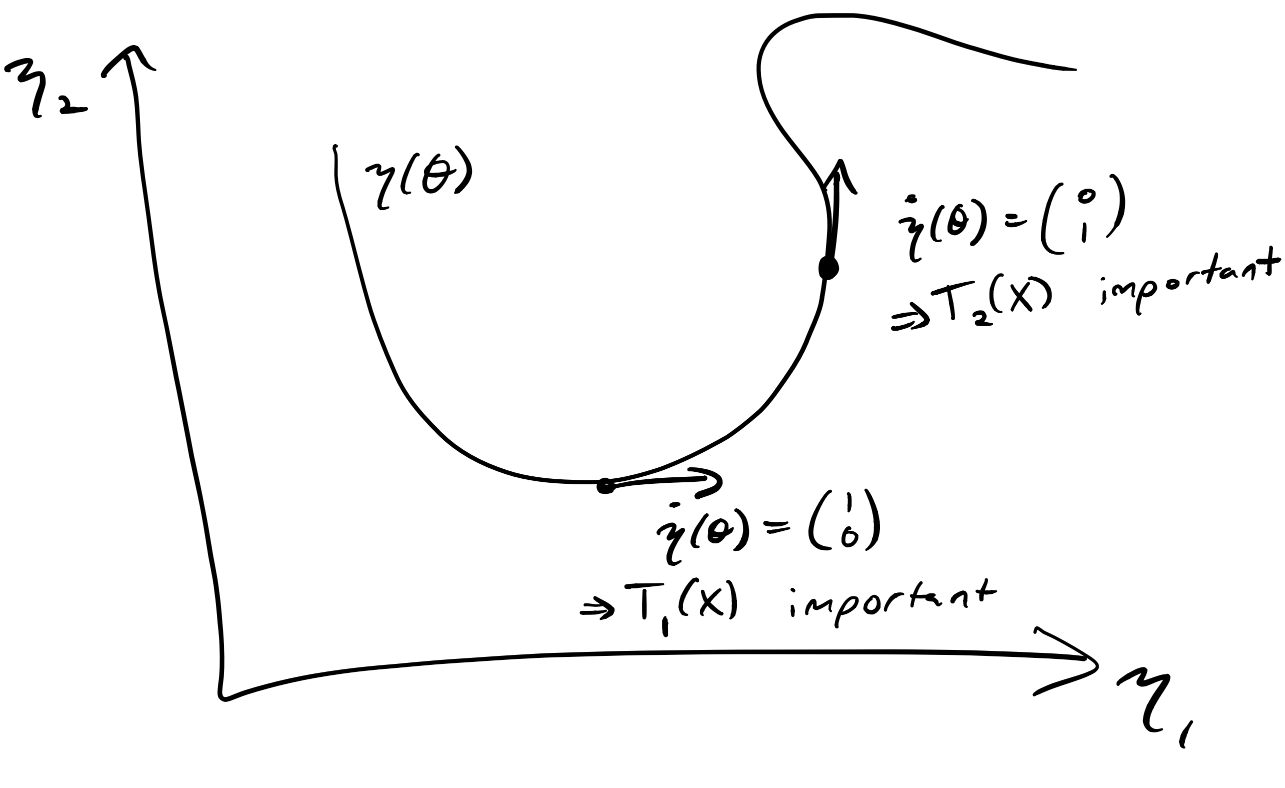
$$\ell(\theta; X) = \eta(\theta)' T(x) - B(\theta) + \log h(x)$$

$$\dot{\ell}(\theta; X) = \dot{\eta}(\theta)' T(x) - \dot{\eta}(\theta)' \nabla_\eta A(\eta(\theta))$$

$$= \dot{\eta}(\theta)' \left(T(x) - \nabla_\eta A(\eta(\theta)) \right)$$

$$= \dot{\eta}(\theta)' \left(T(X) - E_\theta T(X) \right)$$

$\Longrightarrow$ $\dot{\eta}(\theta)' T(X)$ is "locally complete suff. stat."



Doubts about unbiasedness

The UMVUE might be very inefficient, or inadmissible, or just dumb, in cases where another approach makes much more sense.

Ex. $X \sim \text{Bin}(1000, \theta)$

Estimate $g(\theta) = \mathbb{P}_\theta(X \geq 500)$

UMVUE is $1\{X \geq 500\}$ (why?)

$\Rightarrow$ $X = 500$? Conclude $g(\theta) = 100\%$

$X = 499$? Conclude $g(\theta) = 0\%$

This is not epistemically reasonable!!

Could do much better with e.g. MLE or a Bayes estimator.

In fact, our theorem should make us suspicious of UMVUE's: every idiotic function of T is a UMVUE (of its own expectation)

Gaussian Sequence Model

$X_i \overset{iid}{\sim} N(\mu_i, 1) \qquad i=1,\dots,d \qquad$ indep.

or $\quad X \sim N_d(\mu, I_d) \qquad \mu \in \mathbb{R}^d$, estimate $\rho^2 = \|\mu\|^2$

X is complete sufficient

$$\mathbb{E}_\mu \|X\|^2 = \mathbb{E}_0\left[\|\mu + X\|^2\right]$$

$$= \|\mu\|^2 + \mathbb{E}_0\|X\|^2 + \underbrace{\cancel{2\mathbb{E}_0[\mu' X]}}_{0}$$

$$= \|\mu\|^2 + d$$

$$\Rightarrow \quad \delta(x) = \|X\|^2 - d$$

If $\mu = 0$, $\quad \delta(x) < 0 \quad$ about half the time!

$$(\|X\|^2 - d)_+ = \max(0, \|X\|^2 - d)$$

strictly dominates UMVU

Gets worse: Ex 4.7 in Keener

$X \sim$ Truncated Poisson(θ)

$$p_\theta(x) = \frac{\theta^x e^{-\theta}}{x!\,(1-e^{-\theta})} \qquad x = 1, 2, \ldots \quad \theta > 0$$

Estimate $g(\theta) = e^{-\theta}$ (mass lost to truncation)

Keener shows UMVUE is $\delta(X) = (-1)^{X+1}$

$$\frac{e^{-\theta}}{1-e^{-\theta}}\left(\theta - \frac{\theta^2}{2} + \frac{\theta^3}{3!} - \cdots\right) = \frac{e^{-\theta}}{\cancel{1-e^{-\theta}}}\left[\cancel{1 - \left(1 + (-\theta) + \frac{(-\theta)^2}{2!} + \cdots\right)}\right]$$

[Idiotic but we cannot improve using any unbiased estimator

Sometimes insisting on unbiasedness leads us to absurd results.

Unbiasedness has bad reputation, but other methods have their problems too.]