Outline

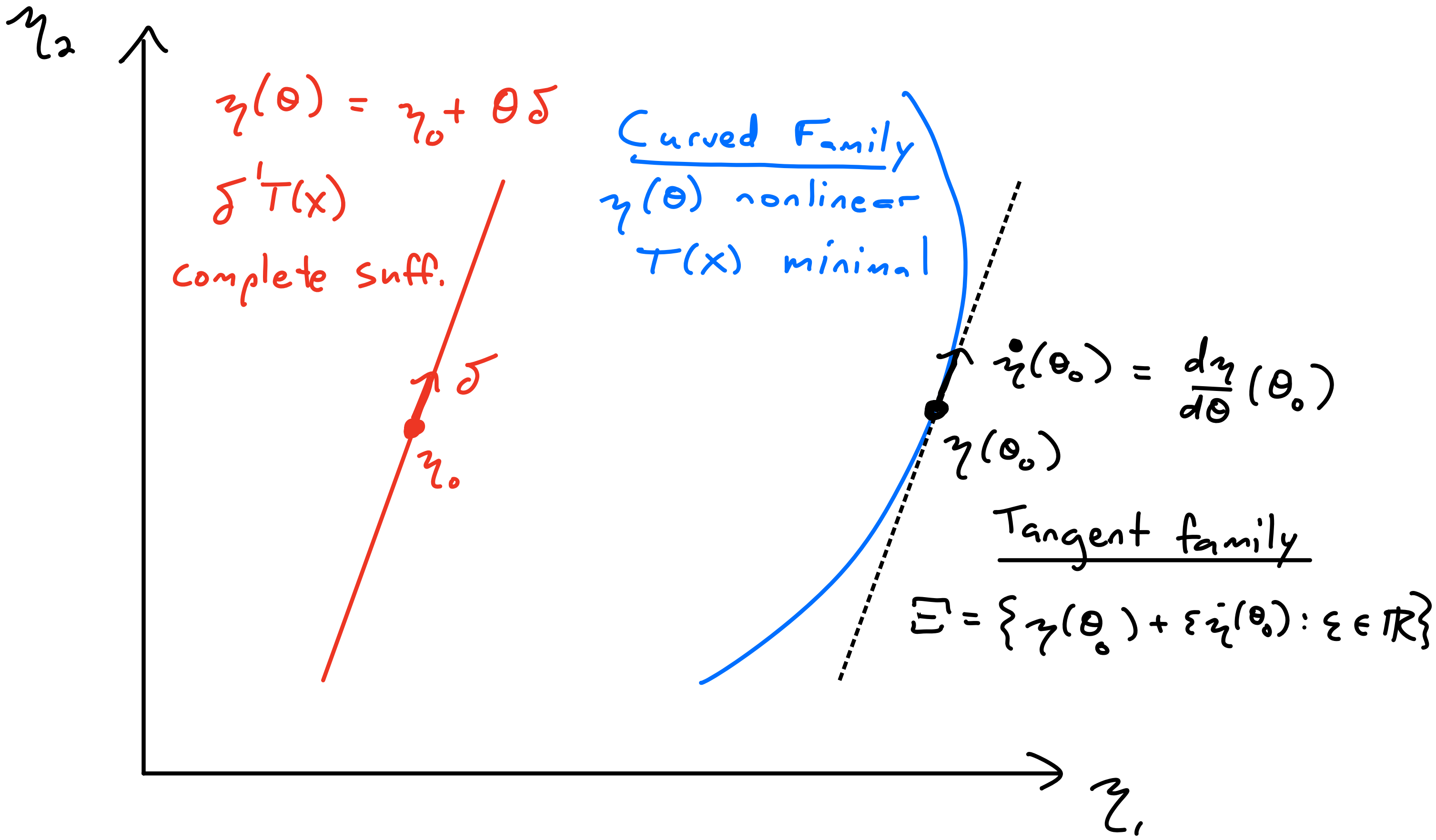
9/19/2023

1) Score function
2) Fisher information
3) Cramér-Rao Lower Bound
4) Examples

Motivation : Tangent family

$$p_\theta(x) = e^{\eta(\theta)'T(x) - A(\eta(\theta))}\, h(x) \qquad \eta: \mathbb{R} \to \mathbb{R}^2$$

$$\Xi = \{ \eta(\theta) : \theta \in \mathbb{R} \}$$



$$q_\varepsilon(x) = e^{(\eta(\theta_0) + \varepsilon\dot\eta(\theta_0))'T(x) - A(\cdots)}\, h(x)$$

$$= e^{\varepsilon \underbrace{\dot\eta(\theta_0)'(T(x) - \mathbb{E}_{\theta_0} T)}_{S_{\theta_0}(x)} - B(\varepsilon)}\, k(x)$$

Complete sufficient for tangent family at θ_0

Called <u>Score function</u>

Score function

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 complete 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)

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

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

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

Only true if these are the same value of θ!

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

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

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

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

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

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

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

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

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

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

1-param: $\quad \mathrm{Var}_\theta(\delta) \cdot \mathrm{Var}_\theta(\dot{\ell}(\theta;X)) \geq \mathrm{Cov}_\theta(\delta, \dot{\ell}(\theta;X))^2$

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

Multivariate: $\quad \theta \in \mathbb{R}^d, \quad g(\theta), \delta(X) \in \mathbb{R}$

$$\mathrm{Var}_\theta(\delta) \geq \nabla g(\theta)' J(\theta)^{-1} \nabla g(\theta)$$

Proof:

$$\begin{aligned} \mathrm{Var}_\theta(\delta) \cdot a' J(\theta) a &= \mathrm{Var}_\theta(\delta)\, \mathrm{Var}(a' \nabla \ell(\theta)) \\ &\geq \mathrm{Cov}_\theta(\delta, a' \nabla \ell(\theta))^2 \\ &= a' \nabla g \nabla g' a, \quad \text{for all } a \in \mathbb{R}^d \end{aligned}$$

$$\Rightarrow \mathrm{Var}_\theta(\delta) \geq \max_{a \neq 0} \frac{a' \nabla g \nabla g' a}{a' J(\theta) a} \overset{\text{Exercise}}{=} \nabla g' J(\theta)^{-1} \nabla g$$

$u = J(\theta)^{1/2} a \qquad \max_u \dfrac{u' J^{-1/2} \nabla g \nabla g' J^{-1/2} u}{u' u} \qquad u = J^{-1/2} \nabla g \qquad a = J^{-1} \nabla g$

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

Ex.: (i.i.d. sample)

$$X_1, \dots, X_n \overset{iid}{\sim} p_\theta^{(1)}(x) \qquad \theta \in \Theta \subseteq \mathbb{R}^d$$

p_θ "regular": common support, finite derivative wrt θ

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

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

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

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

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:

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

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

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

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

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

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

$$\leq 1$$

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

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

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

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

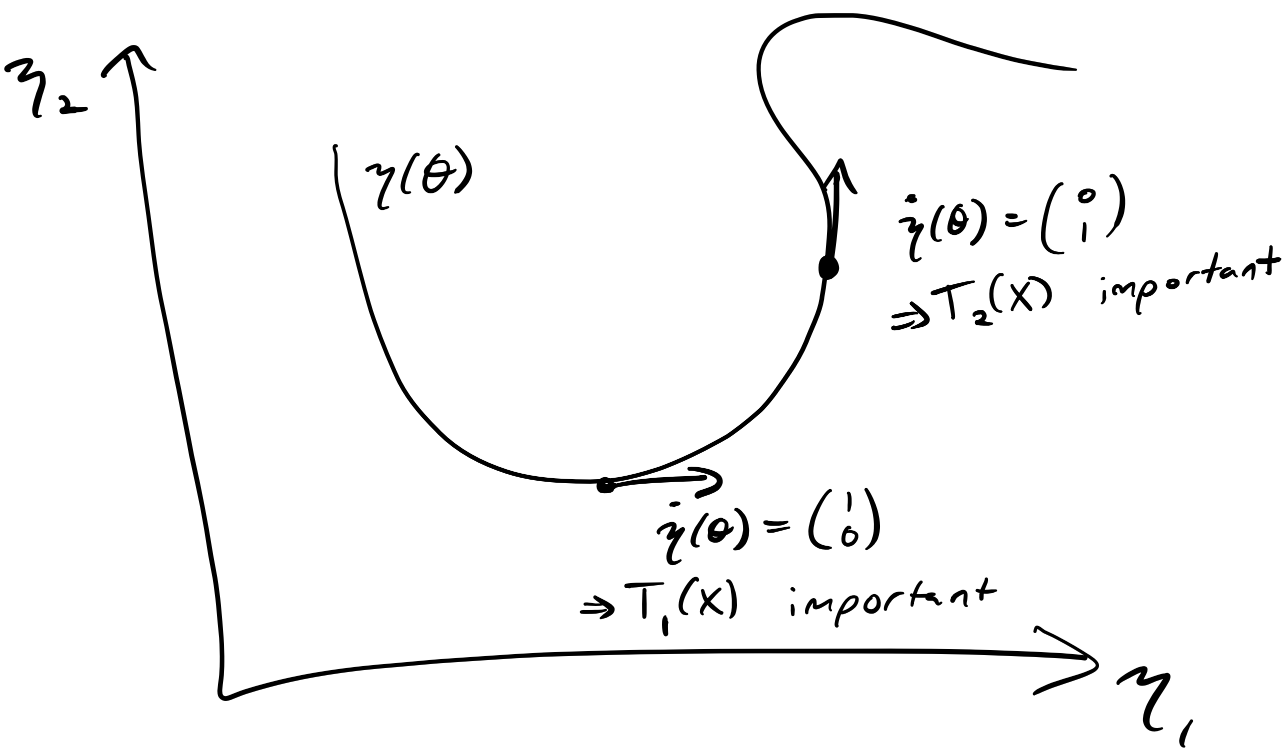
<u>Curved family</u>: $p_\theta(x) = e^{\eta(\theta)'T(x) - B(\theta)} h(x), \quad \theta \in \mathbb{R}$

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

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

$$\begin{aligned} \dot{\ell}(\theta; X) &= \dot{\eta}(\theta)' T(x) - \dot{\eta}(\theta)' \nabla_\eta A(\eta(\theta)) \\ &= \dot{\eta}(\theta)' \left(T(x) - \nabla_\eta A(\eta(\theta)) \right) \\ &= \dot{\eta}(\theta)' \left(T(X) - E_\theta T(X) \right) \end{aligned}$$

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



Fisher info as local metric

Kullback-Leibler Divergence

$$D_{KL}(p \,\|\, q) = \mathbb{E}_p\left[\log p(X) - \log q(X)\right]$$
$$= \int \log\left(\frac{p}{q}\right) p \, d\mu$$

Distance between two distributions

Parametric model

$$D_{KL}(\theta^* \,\|\, \theta) = D_{KL}(p_{\theta^*} \,\|\, p_\theta)$$
$$= \int \left(\ell(\theta^*) - \ell(\theta)\right) e^{\ell(\theta^*)} d\mu$$

Standard "distance" between two distributions

θ^* "real" distribution, function of θ

Maximized at $\theta = \theta^*$:

$$\frac{\partial}{\partial \theta_j} D_{KL}(\theta^* \| \theta) = -\int \frac{\partial \ell}{\partial \theta_j}(\theta)\, e^{\ell(\theta^*)} d\mu$$

$$= 0 \quad \text{at } \theta = \theta^*$$

$$\frac{\partial^2}{\partial \theta_j \partial \theta_k} D_{KL}(\theta^* \| \theta) = -\int \frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}(\theta)\, e^{\ell(\theta^*)} d\mu$$

$$= +J(\theta^*)_{jk} \quad \text{at } \theta = \theta^*$$

$d = 1$: