Theoretical Statistics, STOR 655

Information Inequality: Cramer-Rao Lower Bound

Andrew Nobel

February 2023

Information Inequality in One Dimension

Setting and Assumptions

Family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ of densities on $(\mathcal{X}, \mathcal{A})$ with base measure ν. Assume that

A1 Parameter space $\Theta \subseteq \mathbb{R}$ is open

A2 $\frac{\partial}{\partial\theta} f(x|\theta)$ exists for every $x \in \mathcal{X}$ and $\theta \in \Theta$, and for every $\theta \in \Theta$

$$\int \frac{\partial}{\partial\theta} f(x|\theta)\, d\nu \;=\; 0$$

A3 For each $\theta \in \Theta$ function $\psi(x, \theta) := \frac{\partial}{\partial\theta} \log f(x|\theta)$ well defined P_θ-a.s. and

$$I(\theta) \;=\; \mathbb{E}_\theta \left[\psi(X, \theta)^2\right] \;\in\; (0, \infty)$$

Note: A2 implies $\mathbb{E}_\theta \psi(X, \theta) = 0$ so that $I(\theta) = \mathrm{Var}_\theta(\psi(X, \theta))$

Information Inequality

Cramer-Rau Bound: Let $\hat{\theta} : \mathcal{X} \to \mathbb{R}$ be any statistic. Assume that A1 - A3 hold, and that for all $\theta \in \Theta$

1. $g(\theta) := \mathbb{E}_\theta[\hat{\theta}(X)] = \int_{\mathcal{X}} \hat{\theta}(x) f(x|\theta)\, d\nu$ exists
2. $g'(\theta) = \frac{\partial}{\partial \theta} g(\theta) = \int_{\mathcal{X}} \hat{\theta}(x) \frac{\partial}{\partial \theta} f(x|\theta)\, d\nu$

Then for each $\theta \in \Theta$

$$\mathrm{Var}_\theta(\hat{\theta}(X)) = \mathbb{E}_\theta(\hat{\theta}(X) - g(\theta))^2 \geq \frac{g'(\theta)^2}{I(\theta)}$$

Note: $\mathrm{Var}_\theta(\hat{\theta}(X)) =$ MSE of $\hat{\theta}(X)$ as estimate of $g(\theta)$

Interpretations of Information Inequality

Interpretation 1: Focus on function $g(\theta)$

- Wish to estimate $g(\theta)$, a parameter of interest
- CR gives lower bound on MSE of any unbiased estimate $\hat{\theta}$ of $g(\theta)$

Interpretation 2: Focus on statistic $\hat{\theta}$

- Regard $\hat{\theta}(X)$ as an unbiased estimate of $g(\theta) := \mathbb{E}_\theta[\hat{\theta}(X)]$
- CR gives lower bound $g'(\theta)^2/I(\theta)$ on MSE of $\hat{\theta}$

Interpretation 2': Focus on statistic $\hat{\theta}$

- Regard $\hat{\theta}(X)$ as *biased* estimate of θ with bias $b(\theta) = g(\theta) - \theta$
- CR gives lower bound $(1 + b'(\theta))^2/I(\theta)$ on the variance of $\hat{\theta}$

Fisher Information of an Independent Sample

Definition: Given $X_1, \ldots, X_n$ iid with $X_i \sim f(x|\theta) \in \mathcal{P}$ define the n-sample Fisher information

$$I_n(\theta) \;=\; \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X_1^n|\theta) \right)^2 \right]$$

Fact: Under the usual regularity conditions

1. $I_n(\theta) = n I(\theta)$
2. If $\hat{\theta}_n(X_1^n)$ has expectation $g(\theta) = \mathbb{E}_\theta[\hat{\theta}_n(X_1^n)]$ then

$$\mathsf{MSE}_\theta(\hat{\theta}(X_1^n)) \;\geq\; \frac{g'(\theta)^2}{n I(\theta)}$$

Gamma Family

Recall: For $\alpha, \beta > 0$ the Gam(α, β) distribution has density

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}}\, x^{\alpha-1}\, e^{-x/\beta}\, \mathbb{I}(x > 0)$$

Fact: If $X \sim$ Gam(α, β) then $\mathbb{E}X = \alpha\beta$ and $\mathrm{Var}(X) = \alpha\beta^2$

Example: Gamma Family with Fixed Shape Parameter

Consider: Model $\mathcal{P} = \{\text{Gam}(\alpha_0, \beta) : \beta > 0\}$ where $\alpha_0 > 0$ is fixed

1. Fisher Information $I(\beta) = \alpha_0/\beta^2$
2. If $\hat{\beta}(X_1^n)$ is unbiased for β then $\text{MSE}_\beta(\hat{\beta}(X_1^n)) \geq \beta^2/(n\alpha_0)$
3. Estimator $\hat{\beta}(X_1^n) = \overline{X}_n/\alpha_0$ achieves lower bound (UMVUE)
4. If $\hat{\beta}(X_1^n)$ is unbiased for $1/\beta$ then $\text{MSE}_\beta(\hat{\beta}(X_1^n)) \geq (n\beta^2\alpha_0)^{-1}$
5. Lower bound not achievable by any estimator

Example: Variance of Normal with Known Mean

Consider: Model $\mathcal{P} = \{\mathcal{N}(\mu_0, \sigma^2) : \sigma > 0\}$ with $\mu_0 \in \mathbb{R}$ known. Wish to estimate $g(\sigma) = \sigma^2$

- If $\hat{\sigma}_n$ unbiased for σ^2, CR gives $\mathrm{Var}_{\sigma^2}(\hat{\sigma}_n(X_1^n)) \geq 2\sigma^4/n$ which is achieved by unbiased estimator

$$\hat{\sigma}_n(X_1^n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$$

- However, it is easy to see that the biased estimator

$$\tilde{\sigma}_n(X_1^n) = \frac{1}{n+2}\sum_{i=1}^{n}(X_i - \mu_0)^2$$

  has mean squared error $2\sigma^4/(n+2)$, which is less than the CR bound

Moral: Biased estimators can outperform unbiased ones

Multivariate Information Inequality

Setting and Assumptions

Setting: Family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ of densities on $(\mathcal{X}, \mathcal{A})$ with base measure ν. Recall $\psi(x, \theta) := \nabla_\theta \log f(x|\theta)$. Assume that

A1. Parameter set $\Theta \subseteq \mathbb{R}^p$ is open

A2. $\nabla_\theta f(x|\theta)$ exists for every $x \in \mathcal{X}$ and $\theta \in \Theta$, and for every $\theta \in \Theta$

$$\int \nabla_\theta f(x|\theta)\, d\nu \;=\; 0$$

A3. For each $\theta \in \Theta$, $\psi(x, \theta)$ is well defined with P_θ-probability 1, and

$$I(\theta) \;=\; \mathbb{E}_\theta \left[\psi(X, \theta)\psi(X, \theta)^t\right] \quad \text{is invertible}$$

Note: A2 imples that $\mathbb{E}_\theta \psi(X, \theta) = 0$ for each θ, so $I(\theta) = \mathrm{Var}_\theta(\psi(X, \theta))$

Multivariate Information Inequality

Cramer-Rau Bound: Let $\hat{\theta} : \mathcal{X} \to \mathbb{R}^s$ be any statistic. Assume that for all θ

1. $g(\theta) := \mathbb{E}_\theta[\hat{\theta}(X)] = \int_{\mathcal{X}} \hat{\theta}(x) f(x|\theta) d\nu \in \mathbb{R}^s$ exists
2. $\dot{g}(\theta) = \int_{\mathcal{X}} \hat{\theta}(x) \nabla_\theta f(x|\theta)^t d\nu \in \mathbb{R}^{s \times p}$

Then for each $\theta \in \Theta$ we have the lower bound

$$\mathrm{Var}_\theta(\hat{\theta}(X)) \;\geq\; \dot{g}(\theta)\, I(\theta)^{-1}\, \dot{g}(\theta)^t$$

where $A \geq B$ means that $A - B \geq 0$

Multivariate Information Inequality

Cor: Let $X_1, \ldots, X_n$ be iid with $X_i \sim f(x|\theta)$. If $g(\theta) = \mathbb{E}\hat{\theta}(X_1^n)$ then

$$\mathrm{Var}_\theta(\hat{\theta}(X_1^n)) \;\geq\; n^{-1}\, \dot{g}(\theta)\, I(\theta)^{-1}\, \dot{g}(\theta)^t$$

Special case: If $g(\theta) = \theta$ this reduces to

$$\mathrm{Var}_\theta(\hat{\theta}(X_1^n)) \;\geq\; \frac{I(\theta)^{-1}}{n}$$

Nuisance Parameters

Dealing with Nuisance Parameters

Given: Family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^p$ open

Setting: Interested in inference about a subset $S \subset \{1, 2, \ldots, p\}$ of the components of θ with $|S| = s$. Consider restrictions

- For $\theta \in \mathbb{R}^p$ let $\theta_S = (\theta_j : j \in S) \in \mathbb{R}^s$
- For $A \in \mathbb{R}^{p \times p}$ let $A_S = \{a_{i,j} : i, j \in S\} \in \mathbb{R}^{s \times s}$

Task: Given $X_1, \ldots, X_n$ iid with $X_i \sim f(x|\theta)$ estimate $\theta_S \in \mathbb{R}^s$ using estimator $\tilde{\theta}_n : \mathcal{X}^n \to \mathbb{R}^s$

Terminology: The components θ_j with $j \notin S$ called *nuisance parameters*

Nuisance Parameters are Known

Case 1: Nuisance parameters θ_{S^c} known and fixed.

- Note that $\theta = (\theta_S, \theta_{S^c})$, and that $\psi(x,\theta)_j = 0$ if $j \in S^c$ as θ_{S^c} is fixed. Thus we may focus on $\psi_S(x,\theta)$.

- The FI of θ_S given the values of θ_{S^c} is

$$I(\theta_S|\theta_{S^c}) \;=\; \mathbb{E}_\theta\left[\psi_S(X,\theta)\psi_S(X,\theta)^t\right] \;=\; I(\theta)_S \;\in \mathbb{R}^{s\times s}$$

- Upshot: If $\tilde{\theta}_n(X_1^n)$ is unbiased for θ_S the CR bound gives

$$\mathrm{Var}_\theta(\tilde{\theta}_n(X_1^n)) \;\geq\; \frac{[I(\theta)_S]^{-1}}{n}$$

and similarly for estimates of $g(\theta_S)$

Nuisance Parameters are Unknown

Case 2: Nuisance parameters θ_{S^c} are unknown. Let $S = \{i_1, \ldots, i_s\}$

- Define $g : \mathbb{R}^p \to \mathbb{R}^s$ by $g(\theta) = \theta_S = (\theta_{i_1}, \ldots, \theta_{i_s})^t$
- Note that $\dot{g}(\theta) \in \mathbb{R}^{s \times p}$ has entries $\dot{g}(\theta)_{j,k} = \mathbb{I}(i_j = k)$
- If $\tilde{\theta}_n$ is unbiased for θ_S, then $\mathbb{E}_\theta \tilde{\theta}_n(X_1^n) = g(\theta)$, and CR bound gives

$$\mathrm{Var}_\theta(\tilde{\theta}_n(X_1^n)) \geq n^{-1} \dot{g}(\theta)\, I(\theta)^{-1}\, \dot{g}(\theta)^t = \frac{(I(\theta))_S^{-1}}{n}$$

Example: Univariate Normal Family

The normal family $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$ has Fisher Information

$$I(\mu, \sigma) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{bmatrix} \text{ and } I(\mu, \sigma)^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2/2 \end{bmatrix}$$

1. Estimating μ with σ^2 *known*

$$\mathrm{Var}_{\mu,\sigma}(\hat{\mu}_n(X_1^n)) \geq \frac{(I_{11}(\mu, \sigma))^{-1}}{n} = \frac{1}{n\sigma^2}$$

2. Estimating μ with σ^2 *unknown*

$$\mathrm{Var}_{\mu,\sigma}(\hat{\mu}_n(X_1^n)) \geq \frac{\left(I(\mu, \sigma)^{-1}\right)_{11}}{n} = \frac{1}{n\sigma^2}$$

Example: Univariate Gamma Family

The Gamma family $\mathcal{P} = \{\text{Gam}(\alpha, \beta) : \alpha, \beta > 0\}$ has Fisher Information

$$I(\alpha, \beta) = \begin{bmatrix} h(\alpha) & 1/\beta \\ 1/\beta & \alpha/\beta^2 \end{bmatrix} \text{ and } I(\mu, \sigma)^{-1} = \frac{\beta^2}{\alpha h(\alpha) - 1} \begin{bmatrix} \alpha/\beta^2 & -1/\beta \\ -1/\beta & h(\alpha) \end{bmatrix}$$

where $h(\alpha) = d^2 \log \Gamma(\alpha)/d\alpha^2$

1. Estimating β with α *known*

$$\text{Var}_{\alpha,\beta}(\hat{\beta}_n(X_1^n)) \geq \frac{\beta^2}{n\alpha} \text{ achieved by } \hat{\beta}_n(X_1^n) = \frac{\overline{X}_n}{\alpha}$$

2. Estimating β with α *unknown*

$$\text{Var}_{\alpha,\beta}(\hat{\beta}_n(X_1^n)) \geq \frac{\beta^2 h(\alpha)}{n(\alpha h(\alpha) - 1)} > \frac{\beta^2}{n\alpha}$$