Theoretical Statistics, STOR 655

Asymptotic Normality of MLE

Andrew Nobel

February 2023

Fisher Information

Fisher Information

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

(A) Parameter space $\Theta \subseteq \mathbb{R}^p$ is open

(B) Smoothness: $f(x|\theta)$ is 2x-continuously differentiable in θ for all $x \in \mathcal{X}$

(C) Integrability: For all $\theta \in \Theta$ and $1 \leq j \leq p$

$$\mathbb{E}_\theta\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta_j}\right)^2\right] < \infty$$

Fisher Information Matrix

Of interest: Derivatives of the log-likelihood. For $x \in \mathcal{X}$ and $\theta \in \Theta$ let

$$\psi(x, \theta) = \nabla_\theta \log f(x|\theta) = \left(\frac{\partial \log f(x|\theta)}{\partial \theta_1}, \ldots, \frac{\partial \log f(x|\theta)}{\partial \theta_p} \right)^t \in \mathbb{R}^p$$

$$\dot{\psi}(x, \theta) = \nabla_\theta^2 \log f(x|\theta) = \left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta_j \partial \theta_k} : 1 \leq j, k \leq p \right] \in \mathbb{R}^{p \times p}$$

Definition: The Fisher Information (FI) matrix of $\mathcal{P}$ at θ is

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

Note that $I(\theta)$ is non-negative definite

Fisher Information Matrix

Regularity conditions: Exchange of differentiation and integration

R1: $\frac{\partial}{\partial\theta_j} \int f(x|\theta)\, d\nu(x) \;=\; \int \frac{\partial}{\partial\theta_j} f(x|\theta)\, d\nu(x)$ for $1 \leq j \leq p$

R2: $\frac{\partial^2}{\partial\theta_j \partial\theta_k} \int f(x|\theta)\, d\nu(x) \;=\; \int \frac{\partial^2}{\partial\theta_j \partial\theta_k} f(x|\theta)\, d\nu(x)$ for $1 \leq j, k \leq p$

Note that $\int f(x|\theta)\, d\nu(x) = 1$ so

- R1 implies $\int \frac{\partial}{\partial\theta_j} f(x|\theta)\, d\nu(x) = 0$
- R2 implies $\int \frac{\partial^2}{\partial\theta_j \partial\theta_k} f(x|\theta)\, d\nu(x) = 0$

Alternate Expressions for the Fisher Information

Fact: Suppose that conditions (A) - (C) hold

1. If R1 holds then $\mathbb{E}_\theta \psi(X, \theta) = 0$ and $I(\theta) = \mathrm{Var}_\theta(\psi(X, \theta))$
2. If R1 and R2 hold then $I(\theta) = -\mathbb{E}_\theta(\dot{\psi}(X, \theta))$

Interpretation of the Fisher Information

Recall

- log-likelihood $\ell(\theta|x)$ = evidence for θ based on observation(s) x
- $\psi(x, \theta) = \nabla_\theta \ell(\theta|x)$ slope of log-likelihood at θ
- $\dot{\psi}(x, \theta) = \nabla_\theta^2 \ell(\theta|x)$ curvature of log-likelihood at θ

Suppose $X \sim f(x|\theta_0)$. Under the regularity conditions above

- Expected slope of log-likelihood at θ_0 is $\mathbb{E}_{\theta_0}\{\nabla_\theta \ell(\theta_0|X)\} = 0$
- Expected curvature of log-likelihood at θ_0 is $\mathbb{E}_{\theta_0}\{\nabla_\theta^2 \ell(\theta_0|X)\} = -I(\theta_0)$

Interpretation of the Fisher Information, cont.

Suppose $X \sim f(x|\theta_0)$. Taylor expansion of $\ell(\theta|x)$ around θ_0 gives

$$\begin{aligned} D(P_{\theta_0} : P_\theta) &= \int f(x|\theta_0) \log \frac{f(x|\theta_0)}{f(x|\theta)} = \mathbb{E}_{\theta_0} \left[\ell(\theta_0|X) - \ell(\theta|X) \right] \\ &\approx \mathbb{E}_{\theta_0} \left[(\theta - \theta_0)^t \nabla_\theta \ell(\theta_0|X) + \frac{1}{2} (\theta - \theta_0)^t \nabla_\theta^2 \ell(\theta_0|X) (\theta - \theta_0) \right] \\ &= \frac{1}{2} (\theta - \theta_0)^t I(\theta_0) (\theta - \theta_0) \end{aligned}$$

Upshot: When θ is close to θ_0, KL divergence between between P_{θ_0} and P_θ is determined by Fisher information $I(\theta_0)$

- $I(\theta_0)$ large $\Rightarrow$ more contrast between P_θ and P_{θ_0}
- $I(\theta_0)$ small $\Rightarrow$ less contrast between P_θ and P_{θ_0}

Fisher Information Examples

Poisson model: Family of densities wrt counting measure on $0, 1, 2, \ldots$.

$$\mathcal{P} = \left\{ f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!} : \theta > 0 \right\}$$

Fisher information of $\mathcal{P}$ at θ is given by $I(\theta) = \theta^{-1}$

Normal model: Family of densities wrt Lebesgue measure ν on $\mathbb{R}$

$$\mathcal{P} = \left\{ f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} : \mu \in \mathbb{R}, \sigma > 0 \right\}$$

Fisher information of family $\mathcal{P}$ at (μ, σ) is given by

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

Asymptotic Normality of the MLE

The Likelihood Equation

Let $X_1, X_2, \ldots \in \mathcal{X}$ be iid with $X_i \sim f(x|\theta_0) \in \{f(x|\theta) : \theta \in \Theta\}$

In searching for a maximum likelihood estimate it is natural to consider solutions $\hat{\theta}_n$ of the *likelihood equation*

$$\nabla_\theta \, \ell_n(\theta) \; = \; \sum_{i=1}^{n} \nabla_\theta \log f(X_i|\theta) \; = \; 0$$

Asymptotic Normality of MLE

Theorem: Assume $X_1, X_2, \ldots \in \mathcal{X}$ iid with $X_i \sim f(x|\theta_0) \in \mathcal{P}$ and that

1. A - C and R1 - R2 hold, and $I(\theta_0)$ is invertible (positive definite)
2. There exists $\delta_0 > 0$ and $K : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_{\theta_0} K(X) < \infty$ and
$$\max_{i,j} \sup_{\theta \in B(\theta_0, \delta_0)} |\dot{\psi}_{i,j}(x, \theta)| \;\leq\; K(x)$$
3. $P_\theta = P_{\theta_0}$ iff $\theta = \theta_0$

Then there exists a sequence $\hat{\theta}_n = \hat{\theta}_n(X_1^n)$ such that

1. $\nabla_\theta \ell_n(\hat{\theta}_n) = 0$ eventually almost surely
2. $\hat{\theta}_n \to \theta_0$ wp1
3. $n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}_p(0, I(\theta_0)^{-1})$

Asymptotic Normality of MLE

Key Consequence: Under the conditions of the theorem, if there exists a sequence of measurable estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots$ such that

1. $\dot{\ell}_n(\hat{\theta}_n) = 0$ with probability tending to one
2. $\hat{\theta}_n \to \theta_0$ wp1 (consistency)

then $n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}_p(0, I(\theta_0)^{-1})$.

In other words, any strongly consistent sequence of solutions to the likelihood equation is asymptotically normal

In general, a sequence of MLEs may *not* be consistent, even if it is a root of the likelihood equation

Non-Example: Uniform Distributions

Let $\mathcal{P} = \{P_\theta = U(0, \theta) : \theta > 0\}$ family of uniform distributions on the line

- $U(0, \theta)$ has density $f(x|\theta) = \theta^{-1}\mathbb{I}(0 \leq x \leq \theta)$
- First and second partials not well-defined or continuous
- MLE $\hat{\theta}_n(X_1^n) = \max(X_1, \ldots, X_n)$
- Can show $n(\hat{\theta}_n - \theta) \Rightarrow$ Exp(θ)
- Thus $n^{1/2}(\hat{\theta}_n - \theta) \to 0$ in probability