Basics of asymptotic normality in estimation

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- Empirical process notation
- Consistency
- Asymptotic normality and Taylor expansions
- Fisher information

Reading: van der Vaart, Chapter 5.1–5.6

Notation

We'll use empirical process notation, which is very convenient. Given a distribution P on $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}^d$, we write

$$Pf := \int f dP = \int_{\mathcal{X}} f(x) dP(x)$$

Example (Empirical distributions)

If $X_i \stackrel{\text{iid}}{\sim} P$, define $P_n = \frac{1}{n} \sum_{i=1}^n 1_{X_i}$ as the *empirical distribution*, so

$$P_n(A) = \frac{1}{n} \text{card}(\{i \in [n] : X_i \in A\}) \quad \text{and} \quad P_n f = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

"Simple" asymptotic normality arugment

idea: often the log-likelihood of a model is smooth enough that a Taylor expansion and ignoring higher-order terms gives asymptotic normality

setting: model family $\{P_\theta\}_{\theta \in \Theta}$ of distributions on $\mathcal{X}$ with $\theta \in \mathbb{R}^d$, each with density $p_\theta = dP_\theta / d\mu$

the log likelihood is

$$\ell_\theta(x) := \log p_\theta(x)$$

observe: observations $X_i \stackrel{\text{iid}}{\sim} P_{\theta_0}$, but θ_0 unknown, and typically use *maximum likelihood estimator* (MLE)

$$\widehat{\theta}_n := \operatorname*{argmax}_{\theta \in \Theta} P_n \ell_\theta(X)$$

Questions about the MLE

For

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, P_n \ell_\theta(X),$$

would like to know about

(1) consistency
(2) asymptotic distribution
(3) optimality

Consistency

Definition

A model $\{P_\theta\}_{\theta \in \Theta}$ is *identifiable* if $P_\theta \neq P_{\theta'}$ for all $\theta \neq \theta' \in \Theta$. Equivalently, $D_{\rm kl}\left(P_\theta \| P_{\theta'}\right) > 0$.

Theorem (Consistency for finite Θ)

Assume that $\{P_\theta\}$ is identifiable and $\text{card}(\Theta) < \infty$. *Then $\widehat{\theta}_n \stackrel{p}{\to} \theta$ under P_θ*

A few remarks

- Consistency may fail for Θ infinite, but usually doesn't
- Often, consistency the "hardest" part of the argument
- Many sufficient conditions (see exercises); some include
  - Uniform convergence $\sup_{\theta \in \Theta} |P_n \ell_\theta - P \ell_\theta| \stackrel{p}{\to} 0$ for $X_i \stackrel{\text{iid}}{\sim} P$
  - Convexity, i.e. when $\theta \mapsto \ell_\theta(x)$ is convex (or concave when maximizing)

Asymptotic normality: notation and setting

notation: have log-likelihood ℓ_θ, with *score* and Hessian of log likelihood

$$\nabla \ell_\theta(x) = \left[\frac{\partial}{\partial \theta_j} \log p_\theta(x)\right]_{j=1}^d \in \mathbb{R}^d$$

$$\nabla^2 \ell_\theta(x) = \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(x)\right]_{i,j=1}^d \in \mathbb{R}^{d \times d},$$

(sometimes write $\dot{\ell}_\theta = \nabla \ell_\theta$ and $\ddot{\ell}_\theta(x) = \nabla^2 \ell_\theta$)

assumptions: we have a smooth model

$$\left\|\nabla^2 \ell_{\theta_1}(x) - \nabla^2 \ell_{\theta_0}(x)\right\|_{\mathrm{op}} \le M(x) \left\|\theta_0 - \theta_1\right\| \ \text{ where } \mathbb{E}_{\theta_0}[M^2(X)] < \infty$$

and $\mathbb{E}_{\theta_0}[\nabla \ell_{\theta_0}(X) \nabla \ell_{\theta_0}(X)^\top]$ exists

The basic asymptotic normality result

Theorem

Let $X_i \stackrel{\text{iid}}{\sim} P_{\theta_0}$ *and assume* $\widehat{\theta}_n = \text{argmax}_\theta P_n \ell_\theta(X)$ *is consistent.*
Define the covariance

$$\Sigma_\theta := (P_\theta \nabla^2 \ell_\theta(X))^{-1} \text{Cov}_\theta(\nabla \ell_\theta(X))(P_\theta \nabla^2 \ell_\theta(X))^{-1}$$

Under the previous assumptions,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \stackrel{d}{\to} \mathcal{N}(0, \Sigma_{\theta_0})$$

- "typically" $\Sigma_\theta = -(P_\theta \nabla^2 \ell_\theta(X))^{-1} = \text{Cov}_\theta(\dot{\ell}_\theta)$

Proof of Theorem

Additional comments

- proof of result never used log-likelihood, so completely identical result holds for "M-estimation" problems
- loss function (criterion) $\ell(\theta, x)$ and *risk* (population loss)

$$R_P(\theta) := P\ell(\theta, X)$$

- completely parallel derivation for $\widehat{\theta}_n = \text{argmin}_\theta R_{P_n}(\theta)$

Fisher information

Definition (Fisher information)

For a model family $\{P_\theta\}$ on $\mathcal{X}$, the *Fisher information* is

$$I(\theta) := \mathbb{E}_\theta[\nabla \ell_\theta(X) \nabla \ell_\theta(X)^\top]$$

- when $\mathbb{E}$ and ∇ are interchangable, then $I(\theta) = -\mathbb{E}[\nabla^2 \ell_\theta(X)]$

Examples

Example (Normal location family)

For $p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\theta)^2}{2\sigma^2})$, $I(\theta) = \frac{1}{\sigma^2}$

Example (Reparameterization)

If we are interested in $h(\theta)$ instead of θ, then $I(h(\theta)) = \frac{I(\theta)}{h'(\theta)^2}$

Example (Normal location for θ^2)

In this case, $I(\theta^2) = \frac{1}{4\sigma^2\theta^2}$

Properties of Fisher Information

- Additivity: If $X_1 \sim P_\theta$ and $X_2 \sim Q_\theta$ have information $I_1(\theta)$ and $I_2(\theta)$, then information $I(\theta)$ from both is $I_1(\theta) + I_2(\theta)$

- i.i.d. sampling: if $X_i \overset{\text{iid}}{\sim} P_\theta$, then information $I_n(\theta)$ in $\{X_i\}_{i=1}^n$ is $n \cdot I(\theta)$

Information inequalities (or, the biggest con in statistics)

idea: Fisher information should tell us something about how hard problems are

starting point: a covariance lower bound: for any decision procedure $\delta : \mathcal{X} \to \mathbb{R}$ and any function ψ,

$$\mathsf{Var}(\delta) \geq \frac{\mathsf{Cov}(\delta, \psi)^2}{\mathsf{Var}(\psi)}$$

The information inequality

Theorem (The generic information inequality)

Assume that $\delta : \mathcal{X} \to \mathbb{R}$ *is any estimator and* $\ell_\theta = \log p_\theta$ *is "regular enough." Then*

$$\mathsf{Var}(\delta) \geq \frac{(\frac{\partial}{\partial\theta} P_\theta \delta)^2}{I(\theta)}.$$

Cramér Rao bounds

Suppose we wish to estimate $g(\theta)$ and $P_\theta[\delta] = b(\theta) + g(\theta)$, which are $\mathcal{C}^1$. Then we have

Corollary (Cramér Rao Bound)

$$\mathsf{Var}_\theta(\delta) \geq \frac{(b'(\theta) + g'(\theta))^2}{I(\theta)}.$$

Example (Information inequality)

If $g(\theta) = \theta$ and δ is unbiased, then $\mathbb{E}[(\delta - \theta)^2] \geq \frac{1}{I(\theta)}$.

Multi-dimensional Cramér Rao bounds

Lemma

Let $\delta : \mathcal{X} \to \mathbb{R}$ *and* $\psi : \mathcal{X} \to \mathbb{R}^d$*, where* $P_\theta \psi = 0$*. For* $\gamma = \text{Cov}_\theta(\psi, \delta) = P_\theta \psi(\delta - \mathbb{P}_\theta \delta)$ *and* $C = \text{Cov}_\theta(\psi)$*,*

$$\text{Var}(\delta) \geq \gamma^T C^{-1} \gamma$$

A multi-dimensional information bound

Theorem

Let $g(\theta) = P_\theta \delta$ be differentiable in θ and $I(\theta) = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^\top \succ 0$. Then

$$\mathsf{Var}_\theta(\delta) \geq \nabla g(\theta)^\top I(\theta)^{-1} \nabla g(\theta).$$

Corollary (Fisher information bound)

If $\widehat{\theta}$ is unbiased for θ, then $\mathbb{E}_\theta[\|\widehat{\theta} - \theta\|_2^2] \geq \operatorname{tr} I(\theta)^{-1}$ and $\mathbb{E}_\theta[(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^\top] \succeq I(\theta)^{-1}$

Comments on information bounds

- say nothing about biased estimators
- say little about only asymptotically unbiased estimators
- apply to squared error and little else
- extensions via *Van Trees inequality* to arbitrary estimators possible