Nov 8, 2023.

Reference: Vandervaart

Maximization

Extremum estimators are called M-estimators

- $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \theta|$: extremum estimator is not differentiable; pop^n mean would be

(Least Absolute Deviation Estimator)

$X \sim \text{Bernoulli}(p)$

$P(X=x) = p^x (1-p)^{1-x}$ → dist^n.

cannot technically take log.

log likelihood $f^n = \log P(X=x) = x \log p + (1-x)\log(1-p)$

$$\hat{l}(p) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i \log(p) + (1-x_i)\log(1-p)\right)$$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\mathcal{L}}(p) = \exp(n\,\hat{l}(p)) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= \prod_{i=1}^{n} P(X=x_i)$$

→ Joint Dist^n of the data.

maximises wrt to what?

We are maximizing this joint probability $\Rightarrow$ maximizes the probability of our data.

~ MEAN VALUE EXPANSION ~

(*) Empirical Risk

$$\nabla \hat{R}(\hat{\theta}) = \vec{0}$$

Assumed convergence conditions apply.

$$\Rightarrow \hat{\theta} \xrightarrow{P} \theta_0$$

Assumed twice continuity & differentiability

$$\hat{H}(\theta) = \left(\frac{\partial^2 \hat{R}(\theta)}{\partial \theta_i \partial \theta_j} \right)_{ij}$$

$$\nabla \hat{R}(\hat{\theta}) = \nabla \hat{R}(\theta_0) + H(\theta^*)$$

Linearlization

$$\|\theta^* - \theta_0\| < \|\hat{\theta} - \theta_0\|$$

$$\nabla \hat{R}(\hat{\theta}) = \nabla \hat{R}(\theta_0) + H(\theta^*)(\hat{\theta} - \theta_0) = 0$$

$$\hat{\theta} - \theta_0 = -H(\theta^*)^{-1} \nabla \hat{R}(\theta_0)$$

iteration of Newton - Raphson algorithm

$$\theta_{k+1} = \theta_k - H^{-1}(\theta_k)\nabla\hat{R}(\theta_k)$$

$\Big[$ Mean Value Theorem:

$[a,b]$, $c \in [a,b]$

$$f(b) - f(a) = f'(c)(b-a) \Big]$$

$$\hat{\theta} - \theta_0 = -H(\theta^*)^{-1}\nabla\hat{R}(\theta_0)$$

we are trying to find $\hat{\theta}$

$$\nabla\hat{R}(\theta_0) = \nabla\left(\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \theta)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(Y_i, \theta_0)$$

We know risk is minimized at θ_0 by construction i.e.

$$\nabla R(\theta_0) = 0 = \nabla E[\ell(Y, \theta_0)]$$

$$= E[\nabla_\theta \ell(Y, \theta_0)]$$

(we can't always do this).

Fubbini's Theorem: Allows you to change the order of integrals & differentials

By LLN

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(Y_i, \theta_0) \xrightarrow{P} 0$$

If $\nabla \hat{R}(\theta_0) \xrightarrow{P} 0$

$$\Rightarrow \hat{\theta} - \theta_0 = -\hat{H}(\theta^*)^{-1} \nabla \hat{R}(\theta_0) \xrightarrow{P} 0$$

(*) Fubbini's Theorem: Allows you to change the order of integrals & differentials

$$\frac{\partial}{\partial t} \int_a^b \underbrace{f(x,t)\,dx}_{\text{Leibneitz}} = \int_a^b \frac{\partial f(x,t)}{\partial t}\,dx$$

GENERALIZATION OF CLT:-

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \begin{pmatrix} 1/n \sum x_i^1 \\ 1/n \sum x_i^2 \\ \vdots \\ 1/n \sum x_i^d \end{pmatrix}$$

dimension of vector X

Covariance matrix

$$\Sigma = \left(E\left[(x^k - E[x^k])(x^p - E[x^p]) \right] \right)_{kp}$$

kp element of the matrix

$$\sqrt{n}\,(\bar{X}_n - \mu) \longrightarrow N(0, \Sigma)$$

$\sqrt{n}\, \nabla \hat{R}(\theta_0) \xrightarrow{d} N(0, \Sigma) \rightarrow$ Assumption. This is acceptable, 'cause

$$\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta\, l(Y_i, \theta) \xrightarrow{P} 0$$

- $\sup_{\theta} |\hat{H}(\theta) - H(\theta)| \xrightarrow{p} 0$ (Assumption)

  $\theta^* \xrightarrow{p} \theta_0$ (By construction)

$\Rightarrow$ $\hat{H}(\theta^*) \xrightarrow{p} H(\theta_0)$

- Random vector $X \sim N(\mu, \Sigma)$ → Covariance matrix

  $A \rightarrow$ fixed matrix

  $AX \sim$? What would be its distⁿ.

$\Rightarrow$ $AX \sim N(A\mu, A\Sigma A^T)$

$$\Sigma = E[(X - E[X])(X - E[X])^T] \rightarrow \text{Outer product}$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\hat{H}^{-1}(\theta^*)\sqrt{n}\,\nabla\hat{R}(\theta_0) \xrightarrow{d} N(0, H^{-1}(\theta_0)\Sigma H(\theta_0)^{-1})$$