ESE 531 - Midterm 1 → You can bring one sheet of notes (front and back).

03/06/2024 ← Exam date

- Topics: Random samples

  Convergence properties → WLLN, SLLN, CLT

  Point estimators:

  → Method of moments
  → Maximum likelihood

  } - How to compute
  - How to assess bias
  - Their convergence properties

  → Numerical optimization

  → Gradient descent (ex: HW2)

  → EM algorithm

  → Evaluating estimators: (Will post problems on Friday)

  → MVUE

  → CRLB

  → Determining if an estimator is efficient and/or if it is the MVUE

$$X_1, \ldots, X_n \sim p(x; b) = \begin{cases} \frac{x}{b} \exp\left(\frac{-x^2}{2b}\right), & x > 0 \\ 0, & \text{o.w.} \end{cases}$$

$$\mathbb{E}[X] = \sqrt{\frac{\pi b}{2}}$$

a.) Method of moments estimator

$$\mathbb{E}[X] = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \sqrt{\frac{\pi \hat{b}}{2}}$$

$$\hat{b} = \frac{2}{\pi}(\bar{X}_n)^2$$

$$\mathbb{E}[\hat{b}] = \mathbb{E}\left[\frac{2}{\pi}(\bar{X}_n)^2\right]$$

$$= \frac{2}{\pi}\mathbb{E}\left[(\bar{X}_n)^2\right]$$ 0.5

$$\mathbb{V}[\bar{X}_n] + \mathbb{E}[\bar{X}_n]^2$$

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}[X]$$

$$\mathbb{V}[\bar{X}_n] = \frac{\mathbb{V}[X]}{n}$$

(b.) MLE.

$$p(x; b) = \frac{x}{b} \exp\left(\frac{-x^2}{2b}\right), \quad x > 0$$

$$L(b; x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i; b)$$

$$= \frac{\prod_{i=1}^{n} x_i}{b^n} \exp\left(\frac{-1}{2b} \sum_{i=1}^{n} x_i^2\right)$$

$\log(ab) = \log(a) + \log(b)$
$\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$
$\log(a^b) = b \log(a)$

$$\ell(b; x_1, \dots, x_n) = \log L(b; x_1, \dots, x_n)$$

$$= \left(\sum_{i=1}^{n} \log(x_i)\right) - n \log(b) - \frac{1}{2b} \sum_{i=1}^{n} x_i^2$$

$$\frac{\partial \ell}{\partial b} = 0 \quad -\frac{n}{b} + \frac{1}{2b^2} \sum_{i=1}^{n} x_i^2 = 0$$

$$b^2 \cdot \frac{n}{b} = \frac{1}{2b^2} \sum_{i=1}^{n} x_i^2 \cdot b^2$$

$$b = \frac{1}{2n} \sum_{i=1}^{n} x_i^2$$

$$\mathbb{V}[X] = \left(\frac{4-\pi}{2}\right) b \qquad \mathbb{E}[X] = \sqrt{\frac{\pi b}{2}}$$

(c.) Invariance property of MLE

$$\hat{\alpha} = \tau(\hat{\theta}) = \sqrt{\hat{b}}$$

$$= \sqrt{\frac{1}{2N} \sum_{i=1}^{N} x_i^2}$$

$$\hat{\theta} = \arg\max_{\theta} \underbrace{\log L(\theta; x_1, \ldots, x_n)}_{\ell(\theta; x_1, \ldots, x_n)}$$

$\Updownarrow$

Convex optimization

$$\hat{\theta} = \arg\min_{\theta} \underbrace{-\ell(\theta; x_1, \ldots, x_n)}_{\text{convex!}}$$

always negative log-likelihood

$$\nabla_\theta f(\theta)\Big|_{\theta=\theta^*} = 0$$

$\Downarrow$ multivariate

Idea: Start with some θ_0 and perform a first-order Taylor expansion around θ_0

$$g(\theta_0) \triangleq \nabla_\theta f(\theta)\Big|_{\theta=\theta_0}$$

Hessian matrix

$$g(\theta) \approx g(\theta_0) + H(\theta_0)(\theta - \theta_0)$$

$$= 0$$

Newton algorithm update

$$\Rightarrow \boxed{\theta = \theta_0 - H^{-1}(\theta_0)\, g(\theta_0)}$$

Start with θ_0

For $t = 1, \ldots, \infty$:

$$\theta_t = \theta_{t-1} - H^{-1}(\theta_{t-1})\, g(\theta_{t-1})$$

Optimization convergence:

$$f(\theta_t) - f(\theta_*) \longrightarrow 0$$

and at what rate

Gradient descent: $O\left(\frac{1}{t}\right)$

Newton-Raphson: $O\left(\frac{1}{t^2}\right)$ ← Quasi-Newton methods

---

Expectation Maximization

Suppose you have a population distribution for two random variables X, Z

$$(X_i, Z_i) \overset{iid}{\sim} p(x, z; \theta)$$

x: observed, z: latent

What we typically do to find the MLE is we find the likelihood:

$$L(\theta; \underbrace{x_{1:n}}_{x_1,\dots,x_n}, \underbrace{z_{1:n}}_{z_1,\dots,z_n}) = \prod_{i=1}^{n} p(x_i, z_i; \theta)$$

$$\ell(\theta; x_{1:n}, z_{1:n}) = \sum_{i=1}^{n} \log p(x_i, z_i; \theta)$$

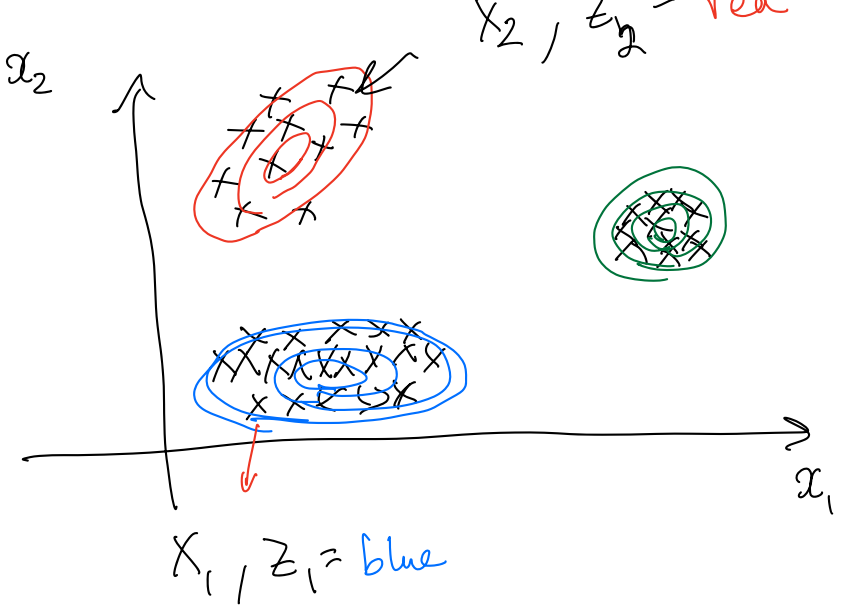
complete data log likelihood

$$\hat{\theta} = \arg\max_{\theta} \underbrace{\ell(\theta; x_{1:n})}$$ } cannot be solved

$$p(x_{1:n}; \theta) = \underbrace{\int p(x_{1:n}, z_{1:n}; \theta)\, dz}$$

cannot be solved!!

Example: Gaussian Mixture Model



$z_i \in \{1, \dots, k\}$

$$p(x_i, z_i; \theta) = \prod_{k=1}^{K} \pi_k^{\mathbb{1}(z_i = k)} \mathcal{N}(x_i; \mu_k, \Sigma_k)^{\mathbb{1}(z_i = k)}$$

π_k: weight of the k^{th} mixture

μ_k: mean

Σ_k: covariance matrix

Population distribution for a GMM with K clusters

$$p(x_i, z_i = 1; \theta) = \pi_1 \mathcal{N}(x_i, \mu_1, \Sigma_1)$$

If $\{z_i\}_{i=1}^{n}$ is known, we can find the MLE of $\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K$

$$p(x_{1:n}, z_{1:n}; \theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} p(x_i, z_i; \theta)$$

$$\ell(\theta; x_{1:n}, z_{1:n}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}(z_i = k) \left(\log \pi_k + \log \mathcal{N}(x_i; \mu_k, \Sigma_k) \right)$$

z_i s.t $z_i = 1$

EM algorithm works as follows:

- Start with a random guess θ_0
- For $t = 1, \ldots, T$ or whenever:

E-step

Find $Q(\theta \mid \theta_{t-1})$

$$= \mathbb{E}_{z \mid x, \theta_{t-1}} \left[\log p(x, z; \theta) \right]$$

M-Step

$$\theta_t = \arg\max_{\theta} Q(\theta \mid \theta^{(t-1)})$$

Known to converge to a stationary point

$$L(\theta_0; x) \leq L(\theta_1; x) \leq \cdots \leq L(\theta_t; x)$$

Let's apply it to GMMs:

$$\ell(\theta; x_{1:n}, z_{1:n}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}(z_i = k) \left(\log \pi_k + \log \mathcal{N}(x_i; \mu_k, \Sigma_k) \right)$$

$$\Rightarrow Q(\theta \mid \theta_{t-1}) = \mathbb{E}_{z_{1:n} \mid x_{1:n}, \theta_{t-1}} \left[\ell(\theta; x_{1:n}, z_{1:n}) \right]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{z_i \mid x_i, \theta_{t-1}} \left[\mathbb{1}(z_i = k) \right] \left(\log \pi_k + \log \mathcal{N}(x_i; \mu_k, \Sigma_k) \right)$$

$\mathbb{E}_{z_i \mid x_i, \theta_{t-1}} \left[\mathbb{1}(z_i = k) \right] = r_{i,k}$ (responsibilities)

$$= \mathbb{P}(z_i = k \mid x_i, \theta_{t-1})$$

Bayes' theorem

$$\mathbb{P}(z_i = k \mid x_i, \theta_{t-1}) = \frac{p(x_i \mid z_i = k, \theta_{t-1}) \, \mathbb{P}(z_i = k \mid \theta_{t-1})}{p(x_i \mid \theta_{t-1})}$$

($p(x_i \mid z_i = k, \theta_{t-1}) = \mathcal{N}(x_i; \mu_k, \Sigma_k)$, $\mathbb{P}(z_i = k \mid \theta_{t-1}) = \pi_k$)

$$= \frac{\pi_k \, \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(x_i; \mu_j, \Sigma_j)} \qquad N \times K$$

Assuming we know $r_{i,k}$, solve

projected gradient descent

$$\hat{\pi}_{1:K} = \arg\max_{\pi_{1:K}:\ \sum_{j=1}^{K}\pi_j = 1,\ \pi_j \geq 0} Q(\theta \mid \theta_{t-1})$$

probability simplex

Lagrange multiplier method:

$$\tilde{Q}(\theta, \lambda \mid \theta_{t-1}) = Q(\theta \mid \theta_{t-1}) + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right)$$

$$\frac{\partial \tilde{Q}}{\partial \theta} = 0 \qquad \frac{\partial \tilde{Q}}{\partial \lambda} = 0 = 1 - \sum_{k=1}^{K}\pi_k$$

$$Q(\theta \mid \theta_{t-1}) = \sum_{i=1}^{n}\sum_{k=1}^{K} r_{i,k} \log \pi_k + \text{terms}$$

$$\frac{\partial Q}{\partial \pi_j} = \sum_{i=1}^{n}\frac{r_{ij}}{\pi_j} \Rightarrow \frac{\partial \tilde{Q}}{\partial \pi_j} = \sum_{i=1}^{n}\frac{r_{ij}}{\pi_j} - \lambda = 0$$

$$\pi_j = \frac{\sum_{i=1}^{n} r_{ij}}{\lambda}$$

$\Rightarrow$ plug this into the $\partial\tilde{Q}/\partial\pi$ part

$$1 - \sum_{k=1}^{K} \pi_k = 0$$

$$\sum_{k=1}^{K} \frac{\sum_{i=1}^{n} r_{ik}}{\lambda} = 1 \Rightarrow \lambda = \sum_{k=1}^{K} \sum_{i=1}^{n} r_{i,k} = n$$

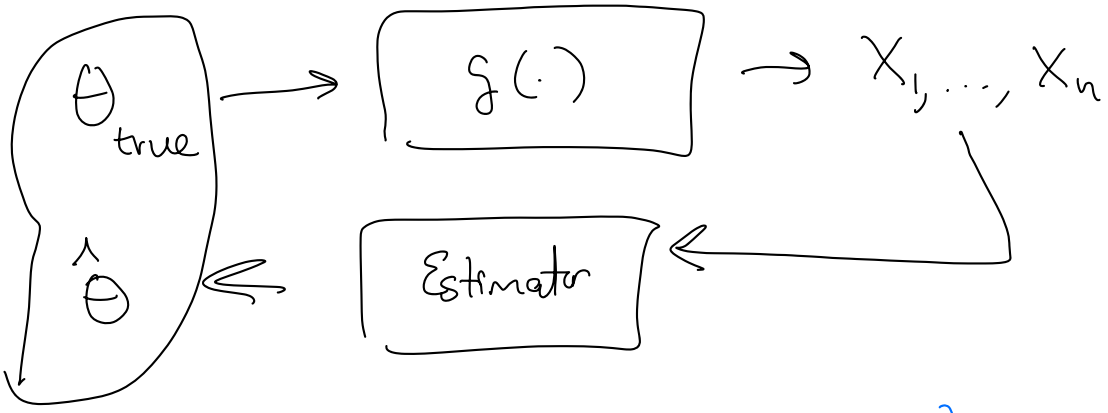
$$= \sum_{i=1}^{n} \sum_{k=1}^{K} r_{i,k}$$

$$= n$$

$$\pi_{j,t} = \frac{\sum_{i=1}^{n} r_{ij}}{n}$$

Point estimation

want to be as close as possible



How do we define this closeness?

1. MSE criterion and the bias-variance tradeoff
2. If the estimator is unbiased, how do I find the "best" estimator
3. Cramer-Rao Lower Bound
4. Biased estimators in the context risk

Mean-square error

$$MSE(\hat{\theta}) = \mathbb{E}\left[(\theta - \hat{\theta})^2\right]$$

$$= \mathbb{E}\left[(\theta + \mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}] - \hat{\theta})^2\right]$$

$$= \mathbb{E}\left[\left((\theta - \mathbb{E}[\hat{\theta}]) - (\hat{\theta} - \mathbb{E}[\hat{\theta}])\right)^2\right]$$

$$= \mathbb{E}\left[(\theta - \mathbb{E}[\hat{\theta}])^2 - 2(\theta - \mathbb{E}[\hat{\theta}])(\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right]$$

$$= \mathbb{E}\left[(\theta - \mathbb{E}[\hat{\theta}])^2\right] - 2\,\mathbb{E}\left[(\theta - \mathbb{E}[\hat{\theta}])(\hat{\theta} - \mathbb{E}[\hat{\theta}])\right] + \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right]$$

$$= \text{bias}^2(\hat{\theta}) + \mathbb{V}[\hat{\theta}]$$

$$\boxed{MSE(\hat{\theta}) = \text{bias}^2(\hat{\theta}) + \mathbb{V}[\hat{\theta}]}$$

- What we want is a $\hat{\theta}$ such that

$$MSE(\hat{\theta}) \leq MSE(\tilde{\theta}), \quad \tilde{\theta} \text{ is another estimator}$$

- What is $MSE(\hat{\theta})$ if $\hat{\theta}$ is an unbiased estimator?

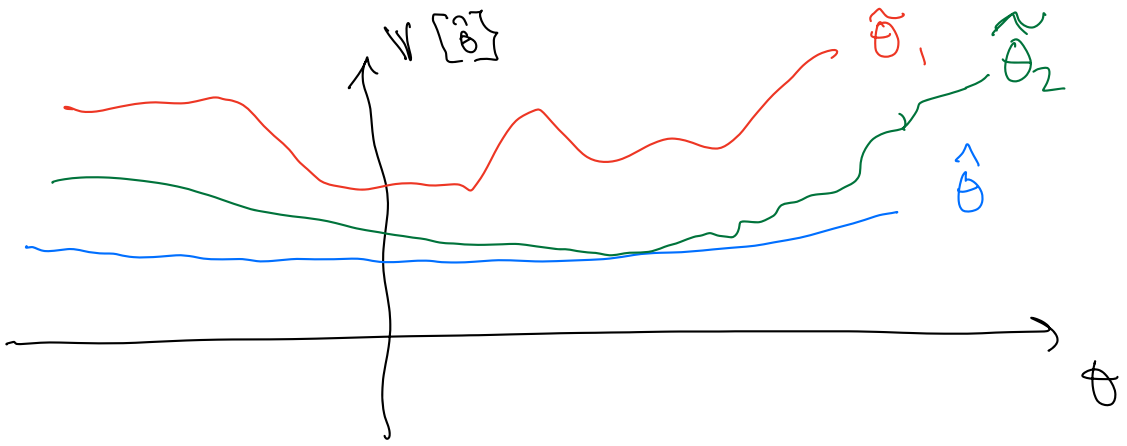
$$MSE(\hat{\theta}) = \mathbb{V}[\hat{\theta}]$$

Definition (Minimum Variance Unbiased Estimator - MVUE)

The MVUE is an estimator $\hat{\theta}$ that satisfies

$$\mathbb{V}[\hat{\theta}] \leq \mathbb{V}[\tilde{\theta}], \quad \forall\, \theta$$

where $\tilde{\theta}$ is any other unbiased estimator



MVUE does not have to exist

→ unbiased estimator may not exist

→ if the population distribution depends on the random sample index

Example: Consider a random sample of size 2 with X_1 and X_2 independently drawn from

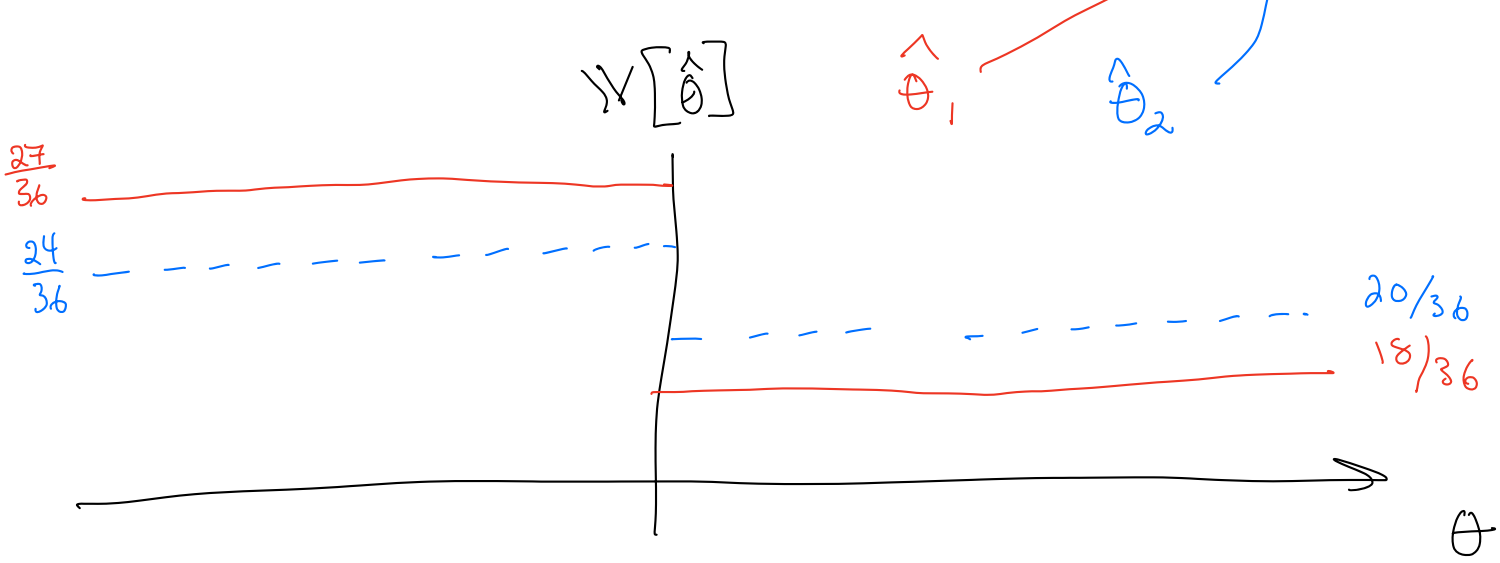
$$X_1 \sim N(\theta, 1)$$

$$X_2 \sim \begin{cases} N(\theta, 2), & \theta \geq 0 \\ N(\theta, 1), & \theta < 0 \end{cases}$$

Two examples of unbiased estimators:

$$\hat{\theta}_1 = \frac{1}{2}X_1 + \frac{1}{2}X_2$$

$$\hat{\theta}_2 = \frac{2}{3}X_1 + \frac{1}{3}X_2$$



- Note: If an unbiased estimator exists and the random sample is i.i.d., then the MVUE always exists

Theorem (Cramer-Rao Lower Bound - CRLB)

Under the assumption that the population pdf $p(x;\theta)$ satisfies the regularity condition:

$$\mathbb{E}\left[\frac{\partial \log p(x;\theta)}{\partial \theta}\right] = 0$$

} change order of integration and differentiation

Then, the variance of any unbiased estimator $\hat{\theta}$ must satisfy

$$\mathbb{V}[\hat{\theta}] \geq -\mathbb{E}\left[\left(\frac{\partial^2 \log p(x;\theta)}{\partial \theta^2}\right)\right]^{-1}$$

$$= I^{-1}(\theta)$$

The $\hat{\theta}$ that attains the CRLB is called an efficient estimator and is the MVUE.

Proof. Cauchy-Schwartz inequality applied to covariance. We know for two random variables Y and Z:

$$\text{Cov}(Y, Z) \leq \sqrt{\mathbb{V}[Y]\,\mathbb{V}[Z]}$$

$$Y = \hat{\theta} \quad \text{and} \quad Z = \frac{\partial}{\partial \theta} \log p(x; \theta)$$

$$\text{Cov}(\hat{\theta}, Z) \leq \sqrt{\mathbb{V}[\hat{\theta}]\,\mathbb{V}[Z]}$$

$$\Rightarrow \mathbb{V}[\hat{\theta}] \geq \frac{\text{cov}^2(\hat{\theta}, Z)}{\mathbb{V}[Z]}$$

$\to 1$ for unbiased estimators

$\to$ Fisher information

$$\text{Cov}(\hat{\theta}, Z) = \mathbb{E}[\hat{\theta} Z] - \mathbb{E}[\hat{\theta}]\,\mathbb{E}[Z]$$

$\mathbb{E}[\hat{\theta}] = \theta$

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log p(x;\theta)\right] = \int \underbrace{\left(\frac{\partial}{\partial\theta}\log p(x;\theta)\right)}\, p(x;\theta)\, d\theta$$

Property: $\frac{\partial}{\partial\theta}\log p(x;\theta) = \frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta}p(x;\theta)$

$$\underbrace{\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta}p(x;\theta)}$$

$$= \int \frac{\partial}{\partial\theta}p(x;\theta)\, d\theta = \frac{\partial}{\partial\theta}\underbrace{\int p(x;\theta)\,d\theta}_{1}$$

$$= 0$$

$$\mathbb{E}[\hat{\theta} z] = \int \hat{\theta}\,\overbrace{\frac{\partial}{\partial\theta}\log p(x;\theta)}^{z}\; p(x;\theta)\, d\theta$$

$$= \int \hat{\theta}\,\frac{\partial}{\partial\theta}p(x;\theta)\, d\theta = \frac{\partial}{\partial\theta}\underbrace{\int \hat{\theta}\,p(x;\theta)\,d\theta}_{\mathbb{E}[\hat{\theta}] \,=\, \theta}$$

$$= \frac{\partial}{\partial\theta}(\theta) = 1$$

$$\mathrm{Cov}(\hat{\theta}, z) = 1 \Rightarrow \mathrm{Cov}^2(\hat{\theta}, z) = 1$$

$$\mathbb{V}[\hat{\theta}] \geq \frac{1}{\mathbb{V}[z]}$$

$$\mathbb{V}[z] = \mathbb{E}[z^2] - \underbrace{\mathbb{E}[z]^2}_{0}$$

$$\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log p(x;\theta)\right)^2\right]$$

Fisher information

$$\mathbb{V}[\hat{\theta}] \geq I^{-1}(\theta)$$

Should know that CRLB is the best possible performance for an unbiased estimator

→ Find the Fisher information
→ Compute CRLB
→ Compare $\mathbb{V}[\hat{\theta}]$ with the CRLB

Example: Gaussian with unknown mean

$$X_1, \dots, X_n \sim N(\mu, 1)$$

$$p(x; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2} (x-\mu)^2 \right)$$

$$L(\mu; X_1, \dots, X_n) = \prod_{i=1}^{n} p(x_i; \mu)$$

$$= \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(\frac{-1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

$$\ell(\mu; X_1, \dots, X_n) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$- \mathbb{E}\left[\frac{\partial^2 \log p(x_1, \dots, x_n; \theta)}{\partial \theta^2} \right] \quad \text{General formula}$$

$$\frac{\partial \ell}{\partial \mu} = 0 - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu) \, 2 \, (-1)$$

$$= \sum_{i=1}^{n} (x_i - \mu)$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = -n$$

$$I(\mu) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \mu^2}\right] = n \qquad \text{CRLB} : \frac{1}{I(\mu)}$$

$$: \boxed{\frac{1}{n}}$$

The MLE asymptotically attains the CRLB as $n \to \infty$ MLE becomes an efficient estimator