ESE 402/542 Recitation 5: Demystifying the Estimation Toolkit

Topics

- Estimation review: Method-of-moments and maximum-likelihood
- What is Fisher Information, really?
- Unbiasedness: Cramer-Rao lower bound

Estimation Review

- Method of Moments estimator
- Maximum likelihood estimator

Method of Moments

- In a nutshell: use the fact that $\frac{1}{n}\sum_{i=1}^{n} X_i^k \approx \mathbb{E}\left[X^k\right]$. This means we can use our data to approximate moments.

Method of Moments

- In a nutshell: use the fact that $\frac{1}{n}\sum_{i=1}^{n} X_i^k \approx \mathbb{E}\left[X^k\right]$. This means we can use our data to approximate moments.
- Sometimes, we can write the parameter we want to estimate as function of moments. We can plug in our estimates of the moments to find estimate of the parameter.

Method of Moments

- In a nutshell: use the fact that $\frac{1}{n}\sum_{i=1}^{n} X_i^k \approx \mathbb{E}\left[X^k\right]$. This means we can use our data to approximate moments.
- Sometimes, we can write the parameter we want to estimate as function of moments. We can plug in our estimates of the moments to find estimate of the parameter.
- Example: we know $\text{var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2$. Method of moments estimate will be

$$\text{var}(X) \approx \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

Method of Moments

- Caveats:
  - Purely algebraic technique–basically solving systems of equations where you're given an approximate value of some variables.
  - No guarantees on the quality of the estimate. No guarantee on being minimum-variance (usually not). No guarantee on unbiasedness (usually not). Remember, in general

$$\mathbb{E}\left[f(X)\right] \neq f\left(\mathbb{E}\left[X\right]\right),$$

in particular, $\mathbb{E}\left[\sqrt{X}\right] \neq \sqrt{\mathbb{E}\left[X\right]}$, $\mathbb{E}\left[X^2\right] \neq \left(\mathbb{E}\left[X\right]\right)^2$.

Maximum Likelihood Estimation

- Very different motivation than Method of Moments. MoM is an algebraic technique to write parameter as a function of moments. MLE is direct answer to question: which choice of parameter θ is *most likely* to be correct, given the data $X_1, \ldots, X_n$ that I've observed.

Maximum Likelihood Estimation

- Very different motivation than Method of Moments. MoM is an algebraic technique to write parameter as a function of moments. MLE is direct answer to question: which choice of parameter θ is *most likely* to be correct, given the data $X_1, \ldots, X_n$ that I've observed.
- Mathematical formulation: given data values $X_1, \ldots, X_n$, and the pdf of the distribution $f(\cdot; \theta)$, where the pdf is determined by the parameter (e.g. θ is mean, variance etc). Given a particular value of θ, the probability of a given data value being of being observed is $f(X_i; \theta)$. Therefore, MLE is the θ that maximizes the overall *likelihood* that $X_1, \ldots, X_n$ are observed.

Maximum Likelihood Estimation

- Very different motivation than Method of Moments. MoM is an algebraic technique to write parameter as a function of moments. MLE is direct answer to question: which choice of parameter θ is *most likely* to be correct, given the data $X_1, \ldots, X_n$ that I've observed.
- Mathematical formulation: given data values $X_1, \ldots, X_n$, and the pdf of the distribution $f(\cdot; \theta)$, where the pdf is determined by the parameter (e.g. θ is mean, variance etc). Given a particular value of θ, the probability of a given data value being of being observed is $f(X_i; \theta)$. Therefore, MLE is the θ that maximizes the overall *likelihood* that $X_1, \ldots, X_n$ are observed.

$$\theta_{\text{MLE}} = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} f(X_i; \theta) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log\left(f(X_i; \theta)\right).$$

Maximum Likelihood Estimation

- Caveats
  - How to solve the maximization? For all intents and purposes in this class, the pdfs f we consider in this class are well-behaved enough for us to use calculus to solve for MLE expression (set derivative to 0 etc). In general, this problem is known to be hard and can only be solved (sometimes suboptimally) via numerical algorithms (cf. Expectation-Maximization Algorithm).

Maximum Likelihood Estimation

- Caveats
  - How to solve the maximization? For all intents and purposes in this class, the pdfs f we consider in this class are well-behaved enough for us to use calculus to solve for MLE expression (set derivative to 0 etc). In general, this problem is known to be hard and can only be solved (sometimes suboptimally) via numerical algorithms (cf. Expectation-Maximization Algorithm).
  - Well-behaved f include "log-concave" distributions, where $\log(f(\cdot))$ is a concave function, such that setting derivative to 0 works. You might see this in future classes.

Maximum Likelihood Estimation

- Caveats
  - How to solve the maximization? For all intents and purposes in this class, the pdfs f we consider in this class are well-behaved enough for us to use calculus to solve for MLE expression (set derivative to 0 etc). In general, this problem is known to be hard and can only be solved (sometimes suboptimally) via numerical algorithms (cf. Expectation-Maximization Algorithm).
  - Well-behaved f include "log-concave" distributions, where $\log(f(\cdot))$ is a concave function, such that setting derivative to 0 works. You might see this in future classes.
  - Like MoM estimator, we have no guarantees of unbiasedness (it is usually not). However, MLE usually carries many more attractive asymptotic qualities, which you make use of in the homework (asymptotic variance matches Cramer-Rao, consistency etc)

MLE Example

1. Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p). Find the MLE of p.
2. Hint: X_i has the distribution $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$

MLE Example

1. Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p). Find the MLE of p.
2. Hint: X_i has the distribution $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$

Answer:

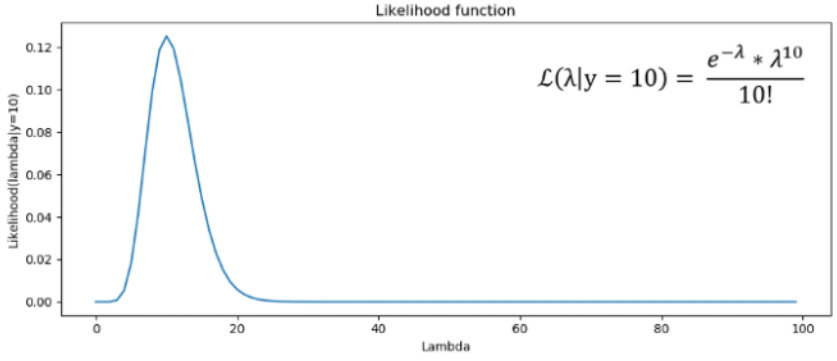
1. $\max_{p \in [0,1]} \sum_{i=1}^{n} \log(f(X_i; p)) =$
   $\max_{p \in [0,1]} y \log p + (n-y) \log(1-p)$ where $y = \sum_{i=1}^{n} x_i$
2. First derivative of log likelihood is $\frac{y}{p} - \frac{n-y}{(1-p)}$
3. Second derivative is $-\frac{y}{p^2} - \frac{ny}{(1-p)^2}$
4. Differentiating and setting to 0 yields $\hat{p}_{ml} = \frac{\sum_{i=1}^{n} x_i}{n}$

Fisher Information

- Recalling the likelihood function, given data $X_1, \ldots, X_n$ drawn from true distribution with parameter θ_0, we consider the value of $L(\theta, \{X_i\}) = \prod_{i=1}^{n} f(X_i; \theta)$. We note that this function should reach its peak at $\theta = \theta_0$. However, the slope of L around θ_0 is also informative.

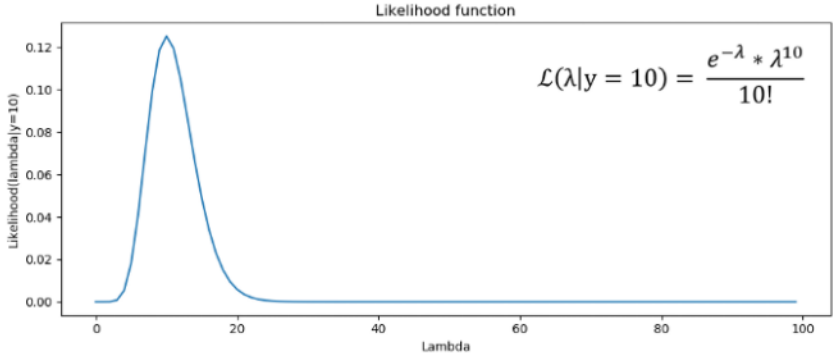
Fisher Information

- Recalling the likelihood function, given data $X_1, \ldots, X_n$ drawn from true distribution with parameter θ_0, we consider the value of $L(\theta, \{X_i\}) = \prod_{i=1}^{n} f(X_i; \theta)$. We note that this function should reach its peak at $\theta = \theta_0$. However, the slope of L around θ_0 is also informative.



Fisher Information

- Recalling the likelihood function, given data $X_1, \ldots, X_n$ drawn from true distribution with parameter θ_0, we consider the value of $L(\theta, \{X_i\}) = \prod_{i=1}^{n} f(X_i; \theta)$. We note that this function should reach its peak at $\theta = \theta_0$. However, the slope of L around θ_0 is also informative.



- If L is very flat with respect to θ, that means that even if we get θ very wrong, the likelihood of observing our data is still pretty good.

Fisher Information

- You can think of this as θ not containing much "information" about our data: in the worst case, if θ is completely irrelevant to our data (Gaussian data X_i,
  $\theta =$ what I should eat for breakfast), then the likelihood L will be completely flat with respect to θ: changing θ will not affect the likelihood of our data at all.

Fisher Information

- You can think of this as θ not containing much "information" about our data: in the worst case, if θ is completely irrelevant to our data (Gaussian data X_i,
  $\theta =$ what I should eat for breakfast), then the likelihood L will be completely flat with respect to θ: changing θ will not affect the likelihood of our data at all.
- Quantify this through the Fisher information: want the average *variance* of the log-slope.

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log(f(X;\theta)\right)^2\right]$$

If θ is informative parameter, the log-slope will be very close to 0 around θ_0, and quickly goes to $\pm$Inf further away – large variance of the slope. We then "average" this informativeness over all X to get the FI.

Cramer-Rao Lower Bound

- Given an unbiased estimator $\hat{\theta}$ of θ_0, how "good" can this estimate be?

Cramer-Rao Lower Bound

- Given an unbiased estimator $\hat{\theta}$ of θ_0, how "good" can this estimate be?
- We will use the mean-squared-error (MSE) $\mathbb{E}[(\hat{\theta} - \theta_0)^2]$ to evaluate our estimate

Cramer-Rao Lower Bound

- Given an unbiased estimator $\hat{\theta}$ of θ_0, how "good" can this estimate be?
- We will use the mean-squared-error (MSE) $\mathbb{E}[(\hat{\theta} - \theta_0)^2]$ to evaluate our estimate
- Recall from HW1:
  $MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2] = \mathrm{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta_0)^2$

Cramer-Rao Lower Bound

- Given an unbiased estimator $\hat{\theta}$ of θ_0, how "good" can this estimate be?
- We will use the mean-squared-error (MSE) $\mathbb{E}[(\hat{\theta} - \theta_0)^2]$ to evaluate our estimate
- Recall from HW1:
  $MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2] = \mathrm{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta_0)^2$
- Since we are considering *unbiased* estimators, $\mathbb{E}[\hat{\theta}] - \theta_0 = 0$ and hence $MSE(\hat{\theta}) = Var(\hat{\theta})$

Cramer-Rao Lower Bound

- Given an unbiased estimator $\hat{\theta}$ of θ_0, how "good" can this estimate be?
- We will use the mean-squared-error (MSE) $\mathbb{E}[(\hat{\theta} - \theta_0)^2]$ to evaluate our estimate
- Recall from HW1:
  $MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2] = \mathrm{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta_0)^2$
- Since we are considering *unbiased* estimators, $\mathbb{E}[\hat{\theta}] - \theta_0 = 0$ and hence $MSE(\hat{\theta}) = Var(\hat{\theta})$
- Cramer-Rao inequality: If $X_1, \ldots, X_n$ i.i.d. with pdf $f(x|\theta_0)$, and let $\hat{\theta}$ be an unbiased estimate of θ_0. Under smoothness assumptions on $f(x|\theta_0)$,

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta_0)}$$

Cramer-Rao Lower Bound

- Given an unbiased estimator $\hat{\theta}$ of θ_0, how "good" can this estimate be?
- We will use the mean-squared-error (MSE) $\mathbb{E}[(\hat{\theta} - \theta_0)^2]$ to evaluate our estimate
- Recall from HW1:
  $MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2] = \mathrm{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta_0)^2$
- Since we are considering *unbiased* estimators, $\mathbb{E}[\hat{\theta}] - \theta_0 = 0$ and hence $MSE(\hat{\theta}) = Var(\hat{\theta})$
- Cramer-Rao inequality: If $X_1, \ldots, X_n$ i.i.d. with pdf $f(x|\theta_0)$, and let $\hat{\theta}$ be an unbiased estimate of θ_0. Under smoothness assumptions on $f(x|\theta_0)$,

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta_0)}$$

- Fact: The MLE is asymptotically unbiased, and asymptotically achieves the CRLB, i.e. $\mathrm{Var}(\hat{\theta}_{ml}) \approx \frac{1}{nI(\theta)}$

Cramer-Rao Example

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

1. Calculate the Cramer-Rao lower bound of any unbiased estimator of σ^2. Hint: $\frac{\partial^2}{\partial(\sigma^2)^2} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})\right) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$
2. Does the unbiased sample variance, i.e. $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, achieve the Cramer-Rao lower bound? Hint: $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$

Cramer-Rao Example

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

1. Calculate the Cramer-Rao lower bound of any unbiased estimator of σ^2. Hint: $\frac{\partial^2}{\partial(\sigma^2)^2} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})\right) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$
2. Does the unbiased sample variance, i.e. $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, achieve the Cramer-Rao lower bound? Hint: $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$

Answer:

1. $I(\sigma^2) = -\mathbb{E}[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}] = \frac{1}{2\sigma^4} \implies \frac{1}{nI(\sigma^2)} = \frac{2\sigma^4}{n}$
2. $\forall n \geq 1, \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$, so it does not achieve the CRLB.