STAT 516: Multivariate Distributions

Lecture 8: Maximum Likelihood Estimation, Consistency and the Cramer-Rao Lower Bound

Prof. Michael Levine

February 9, 2016

Maximum Likelihood Estimation

- The model: pdf $f(x, \theta)$ with $\theta \in \Omega$
- Information: $\mathbf{X} = (X_1, \ldots, X_n)'$ where each $X_i \sim f(x; \theta)$ and independent
- The likelihood:
$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta)$$
- The log-likelihood:
$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$$
- Important - no loss of information occurs this way!

Regularity conditions

- Identifiability: if $\theta \neq \theta'$, $f(x;\theta) \neq f(x;\theta')$
- Pdfs have common support for all θ
- The true value θ_0 is an interior point in Ω

This is why the MLE makes sense!

- If the identifiability and common support assumptions are true, for all $\theta \neq \theta_0$

$$\lim_{n \to \infty} P_{\theta_0}[L(\theta_0, \mathbf{X}) \geq L(\theta, \mathbf{X})] = 1$$

- Interpretation: in sufficiently large samples, the likelihood achieves its maximum at θ_0
- $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is a **maximum likelihood estimator** (mle) of θ if

$$\hat{\theta} = Argmax L(\theta; \mathbf{X})$$

- The most common **estimating equation** is

$$\frac{\partial l(\theta)}{\partial \theta} = 0$$

Example I

- The birth data: $X_i, i = 1, \ldots, n$, p is the probability of a newborn girl
- Check that $L(p; \mathbf{X}) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$
- The resulting MLE is $\hat{p} = \bar{X}$ which is a rather natural estimator

Example II

- Let now $X_1, \ldots, X_n \sim Unif[0, \theta]$ for unknown $\theta > 0$
- The uniform density is $f_\theta(x) = \frac{1}{\theta}$, $0 \le x \le \theta$; 0 otherwise
- The likelihood is
$$L(\theta; \mathbf{X}) = \frac{1}{\theta^n}$$
for all $\theta \ge \max x_i$ and 0 elsewhere
- The likelihood function is strictly decreasing when $\theta \ge \max x_i$ and so $\hat{\theta} = \max_{1 \le i \le n} x_i$ is the MLE
- Note that you cannot differentiate the likelihood function here

Example III: non-uniqueness of MLE

- Let $X_1, \ldots, X_n \sim Unif[\theta, \theta + 1]$
- We have the pdf $f_\theta(x) = 1$, if $\theta \leq y \leq \theta + 1$ and 0 elsewhere
- Clearly, $L(\theta; \mathbf{X}) = 1$ if $\max x_i - 1 \leq \theta \leq \min x_i$ and 0 elsewhere
- The MLE is then an *entire interval* $(\max_{1 \leq i \leq n} X_i - 1, \min_{1 \leq i \leq n} X_i)$

More complicated examples

- Specific examples:
  - Double exponential (Laplace) distribution: $f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}$
  - Verify that $\hat{\theta} = med(x_1, \ldots, x_n)$
  - Logistic distribution: $f(x;\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}$
  - Can't express in the closed form but can be shown to exist and be unique

Functions of MLE's

- If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $\eta = g(\theta)$
- An example: if the variance of $X \sim b(n,p)$ is $np(1-p)$, the estimated variance is equal to $n\hat{p}(1-\hat{p})$.

Consistency of MLE

- Let *all three* regularity conditions be satisfied, $f(x;\theta)$ is differentiable w.r.t θ in Ω
- There exists an MLE $\hat{\theta}_n \xrightarrow{p} 0$

Unbiased estimation I

- For an estimator $\hat{\theta}$ of a parameter θ, the bias is $\mathbb{E}\hat{\theta} - \theta$
- It is usually impossible to have both low variance and low bias at the same time
- A trivial estimator $\hat{\theta} = \theta_0$ for some constant θ_0 has variance 0 but may have a very large bias if θ_0 is very different from θ

Unbiased estimation II

- An estimator $\hat{\theta}$ is unbiased if $\mathbb{E}\hat{\theta} = 0$ for all possible values of θ
- Classical examples:
  - $\bar{X}$ as an estimator of the mean μ
  - $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ as an estimator of σ^2

Unbiased estimation III

- An unbiased estimator may not exist at all
- Take $\mathbf{X} = (X_1, \ldots, X_n)'$ where $X_i \sim b(1, p)$
- Need to estimate $\theta = \frac{p}{1-p}$ (odds ratio)
- Suppose there exists $\hat{\theta} = \hat{\theta}(\mathbf{X})$ s.t. $\mathbb{E}\hat{\theta} = \frac{p}{1-p}$
- There are 2^n different combinations of 0 and 1; thus, for jth vector $\mathbf{X}_j$,

$$\mathbb{E}\hat{\theta} = \sum_{j=0}^{2^n} \hat{\theta}(\mathbf{X}_j) p^{\sum_{i=1}^n x_{ji}} (1-p)^{n - \sum_{j=1}^n x_{ji}}$$

- One cannot expand a function $\frac{p}{1-p}$ into a finite Taylor series!

Fisher information

- Two additional regularity conditions are needed"
  1. The pdf $f(x;\theta)$ is twice differentiable as a function of θ
  2. The integral $\int f(x;\theta)\,dx$ can be differentiated twice under the integral sign as a function of θ
- Then, two equivalent representations of the **Fisher information** are:

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

or

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

Interpretation

- $\frac{\partial \log f(x;\theta)}{\partial \theta}$ is the **score function**
- Can think of the mle $\hat{\theta}$ as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(x_i;\theta)}{\partial \theta} = 0$$

- The Fisher information is the variance of the score function:

$$I(\theta) = Var\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)$$

Examples

- For $X \sim b(1, \theta)$

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

- The Fisher information is larger for probabilities θ that are close to zero or one
- Another example: consider a random sample

$$X_i = \theta + e_i$$

- If $e_i \sim f(x)$ and independent $X_i \sim f_X(x;\theta) = f(x-\theta)$
- Verify that

$$I(\theta) = \int_{-\infty}^{\infty} \left(\frac{f'(x-\theta)}{f(x-\theta)} \right)^2 f(x-\theta)\, dx = \int_{-\infty}^{\infty} \left(\frac{f'(z)}{f(z)} \right)^2 f(z)\, dz$$

does not depend on θ

Examples

- If $X \sim N(\mu, \sigma^2)$ with known σ^2 - it is a location family
- Check that the Fisher information for μ is

$$I_X(\mu) = \frac{n}{\sigma^2}$$

  and so does not depend on μ
- More information about μ is available if the variance is smaller!

Rao-Cramér lower bound

- For a sample size n, the information is

$$Var\left(\frac{\partial \log L(\theta;\mathbf{X})}{\partial\theta}\right) = nI(\theta)$$

- Let $X_1, \ldots, X_n \sim f(x;\theta)$ and independent
- Let $Y = u(X_1, \ldots, X_n)$ be a statistics with $\mathbb{E}\,Y = k(\theta)$
- Then,

$$Var\,(Y) \geq \frac{[k^{'}(\theta)]^2}{nI(\theta)}$$

- An important special case: if $Y = u(X_1, \ldots, X_n)$ is an unbiased estimator of θ, t.i. $k(\theta) = \theta$,

$$Var\,(Y) \geq \frac{1}{nI(\theta)}$$

Efficiency

- Y is an **efficient estimator** of θ iff the variance of Y attains the Rao-Cramér lower bound
- The ratio of the Rao-Cramér lower bound to the actual variance of any unbiased estimator is called the **efficiency** of that estimator
- Example: for $b(1, \theta)$ the Fisher information is $\frac{1}{nI(\theta)} = \frac{\theta(1-\theta)}{n}$
- The MLE of θ is $\bar{X}$ with the variance $\frac{\theta(1-\theta)}{n}$ - this estimator is efficient!
- $\bar{X}$ as an estimator of the Poisson arrival rate is also efficient - can check directly

Example

- Let $X_1, \ldots, X_n \sim f(x;\theta)$ where $f(x;\theta) = \theta x^{\theta-1}$ for $0 < x < 1$ which is $beta(\theta, 1)$
- Check that $I(\theta) = \theta^{-2}$
- The MLE of θ is
$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log x_i}$$
- How to find the variance of $\hat{\theta}$?

Example

- Verify that $Y_u = \log X_i \sim \Gamma\left(1, \frac{1}{\theta}\right)$ and $W = \sum_{i=1}^{n} Y_i \sim \Gamma\left(n, \frac{1}{\theta}\right)$
- Not hard to find that $\mathbb{E}\, W^k = \frac{(n+k-1)!}{\theta^k (n-1)!}$ and

$$\mathbb{E}\,[\hat{\theta}] = \theta \frac{n}{n-1}$$

- Analogously,

$$Var\,(\hat{\theta}) = \theta^2 \frac{n^2}{(n-1)^2(n-2)}$$

  and the variance of the unbiased estimator $\left[\frac{n-1}{n}\right]\hat{\theta}$ is $\frac{\theta^2}{n-2}$
- For efficiency to be true, it should have been $\frac{\theta^2}{n}$ and so efficiency is

$$\frac{n-2}{n}$$

- The estimator $\left[\frac{n-1}{n}\right]\hat{\theta}$ is **asymptotically efficient**

Asymptotic normality and efficiency

- Two additional regularity conditions:
    1. $f(x;\theta)$ is thrice differentiable as a function of θ. Moreover, for all $\theta \in \Omega$, there is a constant c and a function $M(x)$ s.t.

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x;\theta)\right| \le M(x)$$

       with $\mathbb{E}\,|M(X)| < \infty$ for all $\theta_0 - c < \theta < \theta_0 + c$
- If the Fisher information $0 < I(\theta_0) < \infty$, any consistent sequence of solutions for the mle equations satisfies

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N\left(0, \frac{1}{I(\theta_0)}\right)$$

Asymptotic efficiency and asymptotic relative efficiency

- If $\sqrt{n}(\hat{\theta}_{1n} - \theta_0) \xrightarrow{D} N(0, \sigma^2_{\hat{\theta}_{1n}})$, the **asymptotic efficiency** of $\hat{\theta}_{1n}$ is
$$e(\hat{\theta}_{1n}) = \frac{1/I(\theta_0)}{\sigma^2_{\hat{\theta}_{1n}}}$$
- The estimator $\hat{\theta}_{1n}$ is **asymptotically efficient** if the above ratio is 1
- If $\sqrt{n}(\hat{\theta}_{2n} - \theta_0) \xrightarrow{D} N(0, \sigma^2_{\hat{\theta}_{2n}})$, the **asymptotic relative efficiency** (ARE) of the two estimators is
$$e(\hat{\theta}_{1n}, \hat{\theta}_{2n}) = \frac{\sigma^2_{\hat{\theta}_{2n}}}{\sigma^2_{\hat{\theta}_{1n}}}$$

Sample mean vs. sample median

- For the location model $X_i = \theta + e_i$ where e_i has the Laplace distribution
- Can show that the median Q_2 is asymptotically normal with mean 0 and variance $\frac{1}{n}$
- By CLT, the variance of $\bar{X}$ is $\frac{\sigma^2}{n}$ where $\sigma^2 = Var\, e_i$
- Thus, the asymptotic relative efficiency $\text{ARE}(Q_2, \bar{X}) = \frac{2}{1} = 2$
- Verify that if $e_i \sim N(0,1)$ $\text{ARE}(Q_2, \bar{X}) = \frac{2}{\pi} = 0.636$; thus, asymptotically, $\bar{X}$ is 1.57 times more efficient than Q_2 in the normal case

Large sample confidence intervals based on MLE

- Since $I(\theta)$ is a continuous function of θ, we have

$$I(\hat{\theta}_n) \xrightarrow{p} I(\theta_0)$$

- Thus, for specified $0 < \alpha < 1$, we have an approximate $100(1-\alpha)\%$ confidence interval

$$\hat{\theta}_n \pm z_{\alpha/2} \frac{1}{\sqrt{nI(\hat{\theta}_n)}}$$

- Clearly, for any continuous function $g(x)$ s.t. $g^{'}(\theta_0) \neq 0$

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{D} N\left(0, \frac{g^{'}(\theta_0)^2}{I(\theta_0)}\right)$$

Numerical methods to obtain an MLE

- Typically, Newton's method is used...Let $\hat{\theta}^{(0)}$ is an initial value (guess)
- The next point is the intercept of the tangent line to the curve $l'(\theta)$ at the point $(\hat{\theta}^{(0)}, l'(\hat{\theta}^{(0)})$
- Thus,
$$\hat{\theta}^{(1)} = \hat{\theta}^{(0)} - \frac{l'(\hat{\theta}^{(0)}}{l''(\hat{\theta}^{(0)})}$$
and the process is repeated a number of times