Math 502AB - Lecture 22

Dr. Jamshidian

November 8, 2017

1 Lecture - Part 1

1.1 Maximum Likelihood Continued

1.1.1 A Justification for Using MLE

Let θ_0 be the true value of θ, and consider the following regularity conditions:

1. For $\theta \neq \theta' \Rightarrow f(x|\theta) \neq f(x|\theta')$. That is, the parameter identifies the distribution. This is not a restrictive assumption since distributions with 2 different sets of parameters are different.

2. The *pdf*'s have common support for all θ. In other words, the support is independent of θ.

3. The point θ_0 is in the interior of the support of f.

1.1.2 Theorem:

Let θ_0 be the true parameter value, then, under the given assumptions above, we have

$$\lim_{n\to\infty} P\left[\mathcal{L}(\theta_0|X) > \mathcal{L}(\theta|X)\right] = 1$$

for all $\theta \neq \theta_0$ where $\mathcal{L}(\theta|X)$ is a likelihood function based on a sample $X = (X_1, ..., X_n)$. This says that, for large n, the likelihood is maximized at the true value θ_0, with probability 1.

Proof:

We start with the likelihood function

$$\begin{aligned}\mathcal{L}(\theta_0|X) > \mathcal{L}(\theta|X) &\iff \log\mathcal{L}(\theta_0|X) > \log\mathcal{L}(\theta|X)\\ &\iff \sum_{i=1}^{n}\log f(x_i|\theta_0) > \sum_{i=1}^{n}\log f(x_i|\theta)\\ &\iff \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right) < 0\end{aligned}$$

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right) \xrightarrow{P} E_{\theta_0}\left[\log\frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right]$$

By Jensen's Inequality,

$$E_{\theta_0}\left[\log\frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right] < \log E_{\theta_0}\left[\frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right]$$

But now,

$$E_{\theta_0}\left[\frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right] = \int_{-\infty}^{\infty}\frac{f(x|\theta)}{f(x|\theta_0)}\cdot f(x|\theta_0)dx = 1$$

We have thus shown that,

$$\lim_{n\to\infty} P\left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i|\theta)}{f(x_i|\theta_0)} < 0\right] = 1$$

1.1.3 Examples

1. Obtain the *MLE* for λ if $X_1, ..., X_n \sim Poisson(\lambda)$

$$\begin{aligned}\ell(\lambda) &= \sum_{i=1}^{n}\log f(x_i|\lambda)\\ &= \sum_{i=1}^{n}\log\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right)\\ &= \sum_{i=1}^{n}[-\lambda + x_1\log\lambda - \log x_i!]\\ &= -n\lambda + \log\lambda\sum_{i=1}^{n}x_i\\ \ell'(\lambda) &= -n + \frac{\sum_{i=1}^{n}x_i}{\lambda} = 0 \Rightarrow \lambda = \overline{x}\end{aligned}$$

2. Obtain the *MLE* of μ and σ^2 if $X_1, ..., X_n \sim N(\mu, \sigma^2)$ (**iid**)

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}\right] \\
&= C + \sum_{i=1}^{n}\left[-\log\sigma - \frac{1}{2\sigma^2}(x_i-\mu)^2\right] \\
&= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\ell}{\partial\hat{\mu}} &= \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0 \\
\frac{\partial\ell}{\partial\hat{\sigma}^2} &= -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2}\left(\frac{1}{\hat{\sigma}^2}\right)^2\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0
\end{aligned}$$

$$\sum(x_i-\hat{\mu}) = 0 \Rightarrow \sum x_i = n\hat{\mu} \Rightarrow \hat{\mu} = \frac{1}{n}\sum x_i = \overline{x}$$

$$-n\hat{\sigma}^2 + \sum_{i=1}^{n}(x_i-\overline{x})^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

An aside (for next quarter): Recall from linear algebra that, to show that these are maximums, we would need to verify the Hessian matrix is *negative definite* at that point.

$$\mathcal{H} = \begin{bmatrix} \frac{\partial^2\ell}{\partial\mu\partial\mu} & \frac{\partial^2\ell}{\partial\mu\partial\sigma^2} \\ \frac{\partial^2\ell}{\partial\mu\partial\sigma} & \frac{\partial^2\ell}{\partial\sigma^2\partial\sigma^2} \end{bmatrix}$$

3. Let $X_1, ..., X_n \sim Unif(1, 2, ..., \theta)$. Obtain the *MLE* of θ. (Estimating population size).

   Recall:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & x = 1, 2, ..., \theta \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned}
\mathcal{L}(\theta) &= \frac{1}{\theta^n}\prod_{i=1}^{n}\mathcal{I}_{[1,\ldots,\theta]}(x_i) \\
&= \frac{1}{\theta^n}\mathcal{I}_{[1,\ldots,\theta]}\left(x_{(n)}\right)
\end{aligned}$$

To *maximize* this function, we have to find the *smallest* θ possible, so that the indicator function holds. But the indicator is true for $1 \leq X_{(n)} \leq \theta$, so this value is $\hat{\theta} = X_{(n)}$, the maximum.

4. Let $X_1, ..., X_n \sim gamma(\alpha, \beta)$. Obtain *MLE* for α and β. Recall that:

$$f(X|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}x^{\alpha-1}e^{-\beta x}$$

So, we have

$$\begin{aligned}\ell(\alpha,\beta) &= \sum_{i=1}^{n}\{\alpha\log\beta + (\alpha-1)\log x_i - \beta x_i - \log\Gamma(\alpha)\} \\ &= n\alpha\log\beta + (\alpha-1)\sum_{i=1}^{n}\log x_i - \beta\sum_{i=1}^{n}x_i - n\log(\Gamma(\alpha))\end{aligned}$$

$$\begin{aligned}&(1) \quad \frac{\partial\ell}{\partial\alpha} = n\log(\beta) + \sum_{i=1}^{n}\log(x_i) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0 \\ &(2) \quad \frac{\partial\ell}{\partial\beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n}x_i = 0 \\ &(2) \Rightarrow \hat{\beta} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n}x_i} = \frac{\hat{\alpha}}{\overline{x}} \\ &(1) \Rightarrow 0 = n\log\left(\frac{\hat{\alpha}}{\overline{x}}\right) + \sum\log x_i - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}\end{aligned}$$

Here, there is no closed form solution for $\hat{\alpha}$. This is an example that you may at times have to use iterative methodds to get your answer.

2 Lecture - Part 2

2.1 Examples Continued

1. Suppose that there are n independent trials, each resulting in one of m outcomes with respect to probabilities $p_1, p_2, ..., p_m$. Let X_i be the number of i^{th} outcomes. Use $X_1, ..., X_m$ to obtain *MLE* for $p_1, ..., p_m$.

   In this case, the joint density is:

$$f(x_1,...,x_m|p_1,...,p_m) = \binom{n}{x_1,...,x_m}p_1^{x_1}\cdots p_m^{x_m}$$

This gives us the log likelihood:

$$\ell(p_1, ..., p_m) = c + x_1 \log p_1 + \cdots + x_m \log p_m$$

$$\ell(p_1, ..., p_m) = c + x_1 \log p_1 + \cdots + x_{m-1} \log p_{m-1} + x_m \log \left(1 - \sum_{i=1}^{m-1} p_i\right)$$

$$\frac{\partial \ell}{\partial p_1} = \frac{x_1}{p_1} - \frac{x_m}{1 - \sum_{i=1}^{m} p_m} = 0 \Rightarrow \frac{x_1}{x_m} = \frac{\hat{p}_1}{\hat{p}_m}$$

$$\frac{\partial \ell}{\partial p_2} = \frac{x_2}{\hat{p}_2} - \frac{x_m}{\hat{p_m}} = 0 \Rightarrow \frac{x_2}{x_m} = \frac{\hat{p}_2}{\hat{p}_m}$$

$$\vdots$$

$$\frac{\partial \ell}{\partial p_{m-1}} = \frac{x_{m-1}}{\hat{p}_{m-1}} - \frac{x_m}{\hat{p_m}} = 0 \Rightarrow \frac{x_{m-1}}{x_m} = \frac{\hat{p}_{m-1}}{\hat{p}_m}$$

$$\Rightarrow \frac{\sum_{i=1}^{m-1} x_i}{x_m} = \frac{\sum_{i=1}^{m-1} \hat{p}_i}{\hat{p}_m} \Rightarrow \frac{n - x_m}{x_m} = \frac{1 - \hat{p}_m}{\hat{p}_m} \Rightarrow \hat{p}_m = \frac{x_m}{n}$$

This implies: $\hat{p}_1 = \frac{x_1}{n}, \cdots, \hat{p}_{m-1} = \frac{x_{m-1}}{n}$

2.2 Invariance Property of MLE

2.2.1 Theorem

Let $X_1, ..., X_n$ be **iid** with *pdf* $f(x|\theta)$, for $\theta \in \Omega$. For a specified function g, let $\eta = g(\theta)$ be a parameter of interest. Suppose that $\hat{\theta}$ is the *MLE* of θ. Then, $g(\hat{\theta})$ is the *MLE* of $\eta = g(\theta)$.

Proof:

First, suppose that g is a one to one function. The likelihood of interest is $\mathcal{L}(g(\theta))$. Since $g(\theta)$ is one to one, we have $\theta = g^{-1}(\eta)$, since the inverse exists. And, the likelihood of $g(\theta)$, written as a function of η, is given by

$$\mathcal{L}^*(\eta) = \prod_{i=1}^{n} f\left(x_i | g^{-1}(\eta)\right) = \mathcal{L}\left(g^{-1}(\eta)\right) = \mathcal{L}(\theta)$$

2.2.2 Examples

1. (**Olkin, et al. (1981)**: Suppose $X_1, ..., X_5 \sim binomial(k, p)$. Suppose we have two sets of data:
   (a) $(16, 18, 22, 25, 27)$ with $\hat{k} = 99$
   (b) $(16, 18, 22, 25, 28)$ with $\hat{k} = 199$

As we can see, a small change in the observed data has a *huge* effect on the *MLE*. This can be mitigated by increasing the sample size. The issue with the *MLE* is that we have a large flat part of the graph causing a massive change in estimate with a small change in data.

2.3 The Bayesian Approach to Estimation

2.3.1 An Example:

Suppose that we have a Poisson distribution with parameter $\theta > 0$. Moreover, suppose that we know that θ is either equal to 2, or equal to 3.

So, what is different here? In *Bayesian* estimation, we treat the parameters as *random variables.*

Suppose that, based on our knowledge of the problem ("prior" knowledge), we know that

$$P(\Theta = 2) = \frac{1}{3} \quad P(\Theta = 3) = \frac{2}{3}$$

We take a sample of size 2, and we get $X_1 = 2$ and $X_2 = 4$. With *MLE*, we would find the sample mean and find $\hat{\theta} = 3$. But with *Bayesian*, now we need to take the prior and, given the data, compute the posteriior:

$$\begin{aligned}
P(\Theta = 2 | x_1 = 2, x_2 = 4) &= \frac{P(\Theta = 2, x_1 = 2, x_2 = 4)}{P(x_1 = 2, x_2 = 4)} \\
&= \frac{P(x_1 = 2, x_2 = 4 | \Theta = 2) P(\Theta = 2)}{P(x_1 = 2, x_2 = 4 | \Theta = 2) P(\Theta = 2) + P(x_1 = 2, x_2 = 4 | \Theta = 3) P(\Theta = 3)} \\
&= \frac{\frac{1}{3}\left[\frac{1}{2!}e^{-2}2^2 \cdot \frac{1}{4!}e^{-2}2^4\right]}{\frac{1}{3}\left[\frac{1}{2!}e^{-2}2^2 \cdot \frac{1}{4!}e^{-2}2^4\right] + \frac{2}{3}\left[\frac{1}{2!}e^{-3}3^2 \cdot \frac{1}{4!}e^{-3}3^4\right]} \\
&= 0.245
\end{aligned}$$

$$\begin{aligned}
\Rightarrow\ & P(\Theta = 2 | x_1 = 2, x_3 = 4) = 0.245 \\
& P(\Theta = 3 | x_1 = 2, x_3 = 4) = 0.755
\end{aligned}$$

The *prior* told us that the distribution would be roughly 0.33 and 0.66, but the *posterior* has higher probability of $\Theta = 3$. Why? Because the *data* caused it to go more in that direction.