ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak

Lecture 15: MLE: Asymptotics and Invariance

uppose we make n independent observations $x_1, \ldots, x_n$ from some distribution in the set $\{p(x|\theta)\}_{\theta \in \Theta}$. The MLE of θ is

$$\begin{aligned} \widehat{\theta}_n &= \arg\max_\theta \prod_{i=1}^n p(x_i|\theta) \\ &= \arg\min_\theta -\sum_{i=1}^n \log p(x_i|\theta) \end{aligned}$$

In the previous lecture we argued that under reasonable assumptions $\widehat{\theta}_n$ converges in probability to the value of θ that generated the observations, which we called θ^*. In this lecture we will study other properties of the MLE, beginning with a characterization of its asymptotic distribution.

Theorem 1. *(Asymptotic Distribution of MLE) Let $x_1, \ldots, x_n$ be iid observations from $p(x|\theta^*)$, where $\theta^* \in \mathbb{R}^d$. Let $\widehat{\theta}_n = \arg\max_\theta \prod_{i=1}^n p(x_i|\theta) = \arg\max_\theta \sum_{i=1}^n \log p(x_i|\theta)$, define $L(\theta) := \sum_{i=1}^n \log p(x_i|\theta)$, and assume $\frac{\partial L(\theta)}{\partial \theta_j}$ and $\frac{\partial^2 L_n(\theta)}{\partial \theta_j \partial \theta_k}$ exist for all j,k. Then*

$$\widehat{\theta}_n \overset{asymp.}{\sim} \mathcal{N}(\theta^*, n^{-1} I^{-1}(\theta^*))$$

where $I(\theta^*)$ is the Fisher-Information Matrix (FIM), whose elements are given by

$$[I(\theta^*)]_{j,k} = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta_j \partial \theta_k}|_{\theta=\theta^*}\right]$$

Remark: The FIM is the expected value of the negative Hessian matrix of the log-likelihood function at the point θ^*. The Hessian is the curvature of the log-likelihood surface. For example, in the case where θ is scalar, the FIM is simply the second derivative of the log-likelihood function. Since we are maximizing the log-likelihood, the curvature should be negative. The more negative the curvature, the more sharply defined is the location of the maximum. Therefore, more negative curvatures lead to less variable estimates, which is precisely what is revealed by the limiting distribution above.

Proof. We will prove the theorem for the special case when θ is scalar. The proof for multidimensional vectors follows the same steps using multivariable calculus. By the mean value theorem,

$$\frac{\partial L(\theta)}{\partial \theta}\Big|_{\theta=\widehat{\theta}_n} = \frac{\partial L(\theta)}{\partial \theta}|_{\theta=\theta^*} + \frac{\partial^2 L(\theta)}{\partial \theta^2}\Big|_{\theta=\tilde{\theta}} (\widehat{\theta}_n - \theta^*) ,$$

where $\tilde{\theta}$ is some value between θ^* and $\widehat{\theta}_n$. By definition, $\frac{\partial L(\theta)}{\partial \theta}|_{\theta=\widehat{\theta}_n} = 0$, so

$$0 = \frac{\partial L(\theta)}{\partial \theta}|_{\theta=\theta^*} + \frac{\partial^2 L(\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}} (\widehat{\theta}_n - \theta^*)$$

From equation above we have

$$\widehat{\theta}_n - \theta^* = -\frac{\frac{\partial L(\theta)}{\partial \theta}|_{\theta=\theta^*}}{\frac{\partial^2 L(\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}}}$$

Next consider $\sqrt{n}(\widehat{\theta}_n - \theta^*)$. The reason scaling the difference by $\sqrt{n}$ is that this is the normalization needed to stabilize the limiting distribution. For example, if $x_1, \ldots, x_n$ were iid observations from the distribution $N(\theta^*, 1)$, then it is easy to see that $\sqrt{n}(\widehat{\theta}_n - \theta^*) \sim N(0, 1)$. So, from above we have

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \quad = \quad -\frac{\frac{1}{\sqrt{n}}\frac{\partial L(\theta)}{\partial \theta}|_{\theta=\theta^*}}{\frac{1}{n}\frac{\partial^2 L(\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}}} \; . \tag{1}$$

First let's study the numerator.

$$\frac{1}{\sqrt{n}}\frac{\partial L(\theta)}{\partial \theta}|_{\theta=\theta^*} \quad = \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log p(x_i|\theta)}{\partial \theta}|_{\theta=\theta^*}$$

Recall the Central Limit Theorem. If $z_1, \ldots, z_n$ are iid random variables with $\mathbb{E}[z_1] = \mu$ and $\mathbb{E}[(z_1-\mu)^2] = \sigma^2$, then $\frac{1}{\sqrt{n}}\sum_i z_i \xrightarrow{D} \mathcal{N}(\mu, \sigma^2)$, meaning the the random variable defined by summation has a distribution that tends to the Gaussian as $n \to \infty$. Therefore, by the CLT we have

$$\frac{1}{\sqrt{n}}\frac{\partial L(\theta)}{\partial \theta}|_{\theta=\theta^*} \quad = \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log p(x_i|\theta)}{\partial \theta}|_{\theta=\theta^*} \quad \xrightarrow{D} \quad \mathcal{N}\left(\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}|_{\theta=\theta^*}\right], \mathrm{var}\left(\frac{\partial \log p(x|\theta)}{\partial \theta}|_{\theta=\theta^*}\right)\right) \; .$$

The mean is

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}|_{\theta=\theta^*}\right] &= \int \frac{\partial \log p(x|\theta)}{\partial \theta}|_{\theta=\theta^*} p(x|\theta^*)dx \\
&= \int \frac{1}{p(x|\theta^*)}\frac{\partial p(x|\theta)}{\partial \theta}|_{\theta=\theta^*}\, p(x|\theta^*)dx \\
&= \int \frac{\partial p(x|\theta)}{\partial \theta}|_{\theta=\theta^*} dx \\
&= \frac{\partial}{\partial \theta}\left[\int p(x|\theta)dx\right]|_{\theta=\theta^*} = 0 \; ,
\end{aligned}$$

since $\int p(x|\theta)dx = 1$ for all θ and the derivative of a constant is 0. Since the mean is zero, the variance is

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

The variance can be related to the curvature of the log-likelihood function as follows. First observe that

$$\begin{aligned}
\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\left(\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial \theta}\right) \\
&= -\frac{1}{p^2(x|\theta)}\left(\frac{\partial p(x|\theta)}{\partial \theta}\right)^2 + \frac{1}{p(x|\theta)}\frac{\partial^2 p(x|\theta)}{\partial \theta^2}
\end{aligned}$$

Now let's take the expectation

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\,|_{\theta=\theta^*}\right] &= -\int\left(\frac{1}{p(x|\theta^*)}\frac{\partial p(x|\theta)}{\partial \theta}\,|_{\theta=\theta^*}\right)^2 p(x|\theta^*)\,dx \; + \; \int \frac{1}{p(x|\theta^*)}\frac{\partial^2 p(x|\theta)}{\partial \theta^2}\,|_{\theta=\theta^*}\; p(x|\theta^*)\,dx \\
&= -\int\left(\frac{\partial \log p(x|\theta)}{\partial \theta}\,|_{\theta=\theta^*}\right)^2 p(x|\theta^*)\,dx \; + \; \int \frac{\partial^2 p(x|\theta)}{\partial \theta^2}\,|_{\theta=\theta^*}\; dx \\
&= -\mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial \theta}|_{\theta=\theta^*}\right)^2\right] + \frac{\partial^2}{\partial \theta^2}\left(\int p(x|\theta)\,dx\right)|_{\theta=\theta^*} \; .
\end{aligned}$$

Since $\int p(x|\theta)\,dx = 1$ the second term is 0. Therefore, the variance is equal to the negative expected curvature:

$$\mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial\theta}|_{\theta=\theta^*}\right)^2\right] \;=\; -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2}\,|_{\theta=\theta^*}\right] \;=\; I(\theta^*)\;.$$

Now consider the denominator of (1). By the Strong Law of Large Numbers (SLLN), this average converges to its mean value, the negative Fisher Information, almost surely:

$$\frac{1}{n}\frac{\partial^2 L(\theta)}{\partial\theta^2}|_{\theta=\tilde{\theta}} \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log p(x_i|\theta)}{\partial\theta^2}\,|_{\theta=\theta^*} \;\overset{a.s.}{\to}\; \mathbb{E}\left[\frac{\partial^2 \log p(x_i|\theta)}{\partial\theta^2}|_{\theta=\theta^*}\right] \;=\; -I(\theta^*)\;.$$

To summarize, the numerator of (1) converges in distribution to a Gaussian

$$\frac{1}{\sqrt{n}}\frac{\partial L(\theta)}{\partial\theta}|_{\theta=\theta^*} \overset{D}{\to} \mathcal{N}(0, I(\theta^*))\;,$$

and the denominator $\frac{1}{n}\frac{\partial^2 L(\theta)}{\partial\theta^2}|_{\theta=\tilde{\theta}} \overset{a.s.}{\to} -I(\theta^*)$. So for large n, the numerator behaves like a Gaussian random variable and the denominator is almost constant. The ratio therefore converges in distribution to a Gaussian rescaled by the limiting constant of the denominator

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \;\overset{D}{\to}\; \frac{1}{I(\theta^*)}\mathcal{N}(0, I(\theta^*)) \;\equiv\; \mathcal{N}(0, I^{-1}(\theta^*))\;.$$

This type of convergence is rigorously proved by *Slutsky's Theorem* (for more information see `http://en.wikipedia.org/wiki/Slutsky's_theorem`). □

1 Invariance of the MLE

Theorem 2. *Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable with distribution $p(x|\theta^*)$, and let $\tau = g(\theta)$, for some function g. The MLE of τ is*

$$\widehat{\tau} \;=\; \arg\max_{\tau}\left(\max_{\theta\in g^{-1}(\tau)}\sum_{i=1}^{n}\log p(x_i|\theta)\right)\;.$$

It follows that $\widehat{\tau} = g(\widehat{\theta})$ is an MLE, where $\widehat{\theta}$ is the MLE of θ.

Proof. Note that $\max_\tau\left(\max_{\theta\in g^{-1}(\tau)}\sum_{i=1}^n \log p(x_i|\theta)\right) \equiv \max_\theta \sum_{i=1}^n \log p(x_i|\theta)$; i.e., the maximizing values must be the same. Proceed by contradiction. Suppose that $\widehat{\theta} \notin g^{-1}(\widehat{\tau})$ and that $\max_{\theta\in g^{-1}(\widehat{\tau})}\sum_{i=1}^n \log p(x_i|\theta) < \max_\theta \sum_{i=1}^n \log p(x_i|\theta)$. This contradicts the fact above, and so $g(\widehat{\theta})$ must maximize the likelihood. □

Example 1. $X_i \overset{i.i.d}{\sim} Poisson(\lambda)$, i = 1,...,n
Find the MLE of probability that $x \sim Poisson(\lambda)$ is greater than λ. Define

$$\begin{aligned}\rho = g(\lambda) &= P(X > \lambda)\\ &= \sum_{k=\lfloor\lambda+1\rfloor}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}\\ &= 1 - \sum_{k=0}^{\lfloor\lambda\rfloor} e^{-\lambda}\frac{\lambda^k}{k!}\end{aligned}$$

The MLE of ρ is

$$\hat{\rho_n} = 1 - \sum_{k=0}^{\lfloor \hat{\lambda_n} \rfloor} e^{-\hat{\lambda_n}} \frac{\hat{\lambda_n}^k}{k!}$$

where

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