ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak , **scribe:** J.W.Yim

Lecture 15: MLE: Theory and Practice

Suppose $X_1, ..., X_n$ independent and identically distributed random variables with $p(x|\theta)$, for some $\theta \in \Theta$. The Maximum Likelihood Estimator (MLE) is the random variable

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

Theorem 1 *(Asymptotic Distribution of MLE) Let $X_1, X_2, ..., X_n$ be iid random variables with $p(x|\theta^*)$, where $\theta^* \in \Theta$ is a vector parameter. And let $\hat{\theta_n} = arg\max_\theta \prod_{i=1}^n p(x_i|\theta) = arg\max_\theta \sum_{i=1}^n log\ p(x_i|\theta)$. Define $log\ p(x|\theta) := \sum_{i=1}^n log\ p(x_i|\theta)$, and assume $\frac{\partial logp(x|\theta)}{\partial \theta_j}$ and $\frac{\partial^2 logp(x|\theta)}{\partial \theta_j \partial \theta_k}$ exist for all j,k. Then*

$$\hat{\theta_n} \overset{asymp.}{\sim} N(\theta^*, \frac{1}{n} I^{-1}(\theta^*))$$

where $I(\theta^*)$ is the Fisher- Information Matrix.

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

PROOF (scalar θ)

By the mean value theorem,

$$\frac{\partial logp(x|\theta)}{\partial \theta}|_{\theta=\hat{\theta}} = \frac{\partial logp(x|\theta)}{\partial \theta}|_{\theta=\theta^*} + \frac{\partial^2 logp(x|\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}}\ (\hat{\theta} - \theta^*)$$

, where $\tilde{\theta}$ is some value between θ^* and $\hat{\theta}$.

By definition, $\frac{\partial logp(x|\theta)}{\partial \theta}|_{\theta=\hat{\theta}} = 0$, so

$$0 = \frac{\partial logp(x|\theta)}{\partial \theta}|_{\theta=\theta^*} + \frac{\partial^2 logp(x|\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}}\ (\hat{\theta} - \theta^*)$$

Now consider $\sqrt{n}(\hat{\theta} - \theta^*)$. The reason for multiplying by $\sqrt{n}$ is that in the case where $X_1, X_2, ..., X_n$ be iid random variables with $N(\theta^*, 1)$, $\sqrt{n}(\hat{\theta} - \theta^*) \sim N(0, 1)$.

From equation above we have

$$\sqrt{n}(\hat{\theta} - \theta^*) = \frac{\frac{1}{\sqrt{n}} \frac{\partial logp(x|\theta)}{\partial \theta}|_{\theta=\theta^*}}{\frac{-1}{n} \frac{\partial^2 logp(x|\theta)}{\partial \theta^2}|_{\theta=\tilde{\theta}}}$$

Consider the numerator.

$$\frac{1}{\sqrt{n}} \frac{\partial logp(x|\theta)}{\partial \theta}|_{\theta=\theta^*} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial logp(x_i|\theta)}{\partial \theta}|_{\theta=\theta^*}$$

By the Central Limit Theorem, we have

$$\frac{1}{\sqrt{n}}\frac{\partial logp(x|\theta)}{\partial\theta}|_{\theta=\theta^*} \overset{distribution}{\rightarrow} Normal$$

with mean

$$E[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial logp(x_i|\theta)}{\partial\theta}|_{\theta=\theta^*}] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}E[\frac{\partial logp(x_i|\theta)}{\partial\theta}|_{\theta=\theta^*}]$$

and

$$\begin{aligned}
E[\frac{\partial logp(x_i|\theta)}{\partial\theta}|_{\theta=\theta^*}] &= \int\frac{\partial logp(x|\theta)}{\partial\theta}|_{\theta=\theta^*}p(x|\theta^*)dx \\
&= \int\frac{1}{p(x|\theta^*)}\frac{\partial p(x|\theta)}{\partial\theta}p(x|\theta^*)dx \\
&= \int\frac{\partial p(x|\theta)}{\partial\theta}dx \\
&= \frac{\partial}{\partial\theta}\int p(x|\theta)d\theta = 0
\end{aligned}$$

and variance

$$E[(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial logp(x_i|\theta)}{\partial\theta}|_{\theta=\theta^*})^2] = \frac{1}{n}\sum_{i=1}^{n}E[(\frac{\partial logp(x_i|\theta)}{\partial\theta})^2|_{\theta=\theta^*}]$$

Note that

$$\begin{aligned}
E[\frac{\partial^2 logp(x_i|\theta)}{\partial\theta^2}] &= \int(\frac{1}{p(x|\theta)}\frac{\partial^2 p(x_i|\theta)}{\partial\theta^2} - (\frac{1}{p(x|\theta)}\frac{\partial p(x_i|\theta)}{\partial\theta})^2)p(x|\theta)d\theta \\
&= -E[(\frac{\partial logp(x_i|\theta)}{\partial\theta})^2]
\end{aligned}$$

So the variance is

$$-\frac{1}{n}\sum_{i=1}^{n}E[\frac{\partial^2 logp(x_i|\theta)}{\partial\theta^2}|_{\theta=\theta^*}] = I(\theta^*)$$

Now consider the denominator.

$$\begin{aligned}
\frac{1}{n}\frac{\partial^2 logp(x|\theta)}{\partial\theta^2}|_{\theta=\tilde{\theta}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 logp(x_i|\theta)}{\partial\theta^2}|_{\theta=\theta^*} \\
&\overset{SLLN}{\rightarrow} E[\frac{\partial^2 logp(x_i|\theta)}{\partial\theta^2}|_{\theta=\theta^*}] \\
&= -I(\theta^*)
\end{aligned}$$

So from the equation below,

$$\sqrt{n}(\hat{\theta}-\theta^*) = \frac{\frac{1}{\sqrt{n}}\frac{\partial logp(x|\theta)}{\partial\theta}|_{\theta=\theta^*}}{\frac{-1}{n}\frac{\partial^2 logp(x|\theta)}{\partial\theta^2}|_{\theta=\tilde{\theta}}}$$

the numerator

$$\frac{1}{\sqrt{n}}\frac{\partial logp(x|\theta)}{\partial\theta}|_{\theta=\theta^*} \overset{distribution}{\rightarrow} N(0, I(\theta^*))$$

and the denominator

$$-\frac{1}{n}\frac{\partial^2 logp(x|\theta)}{\partial\theta^2}|_{\theta=\tilde{\theta}} \overset{SLLN}{\rightarrow} I(\theta^*)$$

Hence, the whole term converges as follows

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta^*) &\overset{distribution}{\rightarrow} \frac{1}{I(\theta^*)}N(0, I(\theta^*)) \\ &\equiv N(0, I^{-1}(\theta^*))\end{aligned}$$

1 Invariance of the MLE

Theorem 2 *Let $\tau = g(\theta^*)$ be a function of θ^*, and let $\hat{\theta_n}$ be the MLE of θ^*. Then $\hat{\tau_n} = g(\hat{\theta_n})$ is the MLE of τ.*

PROOF:
Let $h = g^{-1}$ denote the inverse map of g. Define the induced log-likelihood function

$$L(x|\tau) = \max_{\theta \in h(\tau)} logp(x|\theta)$$

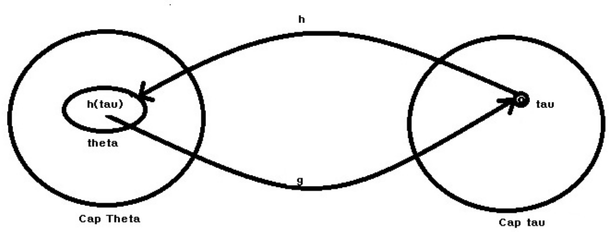

Figure 1: h is the inverse map of g

The MLE of τ is

$$\begin{aligned}\hat{\tau_n} &= arg\max_{\tau} L(x|\tau) \\ &= arg\max_{\tau}\max_{\theta\in h(\tau)} logp(x|\theta) \\ &= g(\hat{\tau_n})\end{aligned}$$

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

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

2 Numerical Methods for Obtaining the MLE

If $X \sim p(x|\theta), \theta \in \Theta$, then the MLE is the solution to the equations $\frac{\partial logp(x|\theta)}{\partial\theta} = 0$. Sometimes these equations have a simple closed form solution, and other times they do not and we must use computational methods to find $\hat{\theta}$.

Example 2 $X_i \overset{i.i.d}{\sim} Poisson(\lambda), \hat{\lambda_n} = \frac{1}{n}\sum X_i$

Example 3 $X \sim N(H\theta, I)$, *where H is $n\times$ k and known and θ is $k\times$ 1 and unknown.*
$\hat{\theta} = (H^T H)^{-1} H^T X$

Example 4 $X_i \overset{i.i.d}{\sim} pN(\mu_0, \sigma_0^2) + (1-p)N(\mu_1, \sigma_1^2), i = 1, ..., n, \theta = [p\ \mu_0\ \sigma_0^2\ \mu_1\ \sigma_1^2]^T$

$$p(x_i|\theta) = \frac{p}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\mu_0)^2}{2\sigma_0^2}} + \frac{1-p}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}}$$

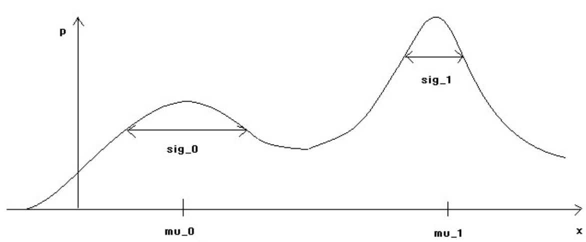

Figure 2: Mixed Gaussian Density

$$p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta), a\ product\ of\ sums\ of\ exponentials$$

$$log\ p(x|\theta) = \sum\ of\ logs\ (sums\ of\ exponentials)\ \leftarrow Messy!$$

Sufficient Statistic: $(X_1, X_2, ..., X_n)$

How to proceed?
1. Gradient/Newton methods

$$\theta^{(t+1)} = \theta^{(t)} + \Delta \frac{\partial}{\partial \theta} log\ p(x|\theta)|_{\theta=\theta^{(t)}},\ where\ \Delta\ is\ a\ step\ size.$$

2. Expectation-Maximization Algorithm (EM algorithm)

3 The EM Algorithm

Suppose the log-likelihood function looks like this:

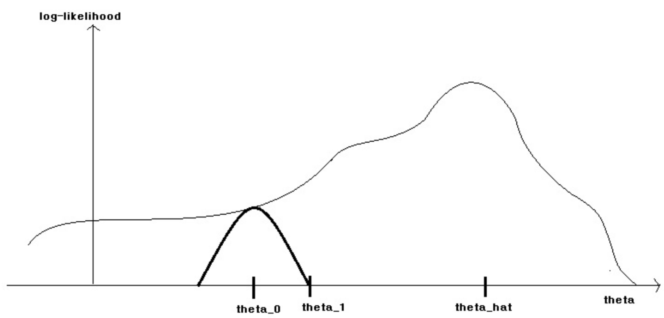

Figure 3: EM algorithm

We would like to find the maximum at the point $\hat{\theta}$. One way is to follow the gradient (uphill) from a random initial guess.
The EM algorithm operates a bit differently. It is an iterative method that constructs a surrogate function at an intial starting point θ_0 as shown above as the C2shed function. This function is designed to "touch" the log-likelihood at θ_0 and t_0 be easy to maximize. The maximizer of the surrogate gives us a new point θ_1 which is guaranteed to have a likelihood value at least as large as θ_0.

We then repeat this process at θ_1 and generate a sequence of values(with increasing likelihood: $\theta_0, \theta_1, ...$ in this fashion). Unfortunately, unless the log-likelihood is concave (negative log-likelihood convex) and hence unimodal, there is no guarantee that any method, besides a global brute-force search, will converge to $\hat{\theta}$.