ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak

Lecture 14: Maximum Likelihood Estimation

The maximum Likelihood (ML) Estimate is given by

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} p(x|\theta)$$

where $p(x|\theta)$ as a function of x with the parameter θ fixed is the probability density function or mass function. And $p(x|\theta)$ as a function of θ with x fixed is called the "likelihood function".

1 ML Estimation and Density Estimation

ML Estimation is equivalent to density estimation. Assume

$$x_i \overset{\text{iid}}{\sim} q, \quad i = 1, \cdots, n, \quad \text{where } q \text{ is an unknown probability density}$$

The ML Estimation is equivalent to finding the density in $\{p_\theta\}_{\theta\in\Theta}$ that best fits the data. i.e., "The generative model with the highest density/probability value at the point $\{x_i\}$." The true generating density q may not be a member of the parametric family under consideration.

1.1 ML Estimation as Minimization

$$\begin{aligned}\widehat{\theta} &= \arg\min_\theta \frac{1}{p(x|\theta)} \\ &= \arg\min_\theta -\log p(x|\theta)\end{aligned}$$

Thus, we can view the MLE as minimizing the loss

$$\boxed{\ell(q, p_\theta) := -\log p(x|\theta)}$$

where dependence on q is embodied in $x \sim q$.

Example 1.

$$p(x|\theta) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x - H\theta)^T\Sigma^{-1}(x - H\theta)\}\ ,\ x \in \mathbb{R}^n \textit{ and } \theta \in \mathbb{R}^k$$

The value of $\widehat{\theta}$ is given by,

$$\begin{aligned}\widehat{\theta} &= \arg\min_\theta -\log p(x|\theta) \\ &= \arg\min_\theta (x - H\theta)^T\Sigma^{-1}(x - H\theta) \\ &= (H^T\Sigma^{-1}H)^{-1}H^T\Sigma^{-1}x\end{aligned}$$

2 MLE and Risk

The risk associated to the MLE is also known as a "expected loss"

$$\begin{aligned} R_{\text{MLE}}(q, p_\theta) &= \mathbb{E}[\ell(q, p_\theta)] \\ &= \mathbb{E}\left[-\log p(x|\theta)\right] \\ &= \int q(x)\left(-\log p(x|\theta)\right) dx \end{aligned}$$

2.1 Excess Risk ("Regret")

Let θ be any value of the parameter. Then we can compare

$$R_{\text{MLE}}(q, p_\theta) - R_{\text{MLE}}(q, q)$$

which quantifies how much larger the expected loss is when we use θ instead of θ^*. Note that

$$\begin{aligned} R_{\text{MLE}}(q, p_\theta) - R_{\text{MLE}}(q, q) &= \mathbb{E}\left[\log q(x) - \log p(x|\theta)\right] \\ &= \mathbb{E}\left[\log \frac{q(x)}{p(x|\theta)}\right] \\ &= \int q(x) \log \frac{q(x)}{p(x|\theta)} dx \\ &= D\left(q \| p_\theta\right) \\ &= \geq 0 \end{aligned}$$

with equality if $p_\theta = q$. Thus the "optimal" value of θ is

$$\theta^* = \arg\min_\theta D\left(q \| p_\theta\right) .$$

The density p_{θ^*} the member of the parametric class that is closest in KL divergence to the data-generating distribution q.

If we have multiple iid observations then

$$x_i \stackrel{\text{iid}}{\sim} q, \quad i = 1, \cdots, n$$

the loss is given by

$$\begin{aligned} \ell(q, p_\theta) &= -\log\left(\prod_{i=1}^n p(x_i|\theta)\right) \\ &= -\sum_{i=1}^n \log p(x_i|\theta) \end{aligned}$$

MLE:

$$\widehat{\theta} = \arg\min_\theta -\sum_{i=1}^n \log p(x_i|\theta)$$

Excess Risk:

$$R_{\text{MLE}}(q, p_\theta) - R_{\text{MLE}}(q, q) = nD\left(q \| p_\theta\right)$$

for any $\theta \in \Theta$

3 Convergence of log likelihood to KL

Assume $x_i \overset{\text{iid}}{\sim} p(x|\theta^*)$, then by strong law of large numbers (SLLN) for any $\theta \in \Theta$

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\theta)} \xrightarrow{\text{a.s.}} D\left(p_{\theta^*} \| p_\theta\right)$$

We would like to show that the MLE

$$\widehat{\theta}_n = \arg\max_\theta \frac{1}{n}\sum_{i=1}^{n} \log p(x_i|\theta)$$

converges to θ^* in the following sense:

$$D\left(p_{\theta^*} \| p_{\widehat{\theta}_n}\right) \longrightarrow 0$$

Note that since $\widehat{\theta}_n$ maximizes $\sum_{i=1}^{n} \log p(x_i|\theta)$ we have

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\widehat{\theta}_n)} \leq 0$$

Thus we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\widehat{\theta}_n)} - D\left(p_{\theta^*} \| p_{\widehat{\theta}_n}\right) + D\left(p_{\theta^*} \| p_{\widehat{\theta}_n}\right) \leq 0 \\ \Longrightarrow D\left(p_{\theta^*} \| p_{\widehat{\theta}_n}\right) \;\leq\; \left|\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\widehat{\theta}_n)} - D\left(p_{\theta^*} \| p_{\widehat{\theta}_n}\right)\right|\end{aligned}$$

So, $D\left(p_{\theta^*} \| p_{\widehat{\theta}_n}\right) \longrightarrow 0$ if $\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\widehat{\theta}_n)} \longrightarrow D\left(p_{\theta^*} \| p_{\widehat{\theta}_n}\right)$
The subtle issue here is that $\widehat{\theta}_n$ is a random variable, not a fixed $\theta \in \Theta$, so we can not just appeal to the SLLN.

Theorem 1. *Assume*

$$x_i \overset{\text{iid}}{\sim} p(x|\theta^*) \quad i = 1, \cdots, n$$

Define

$$\begin{aligned} L_n(\theta) &:= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\theta)}, \quad \forall \theta \in \Theta \\ L(\theta) &:= \mathbb{E}\left[L_n(\theta)\right] = D\left(p_{\theta^*} \| p_\theta\right)\end{aligned}$$

Suppose the following assumptions hold

A1. $\sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| \xrightarrow{\text{P}} 0$

A2. $\inf_{\theta : \|\theta - \theta^*\| \geq \epsilon} L(\theta) > L(\theta^*), \quad \forall \epsilon > 0$

then

$$\widehat{\theta}_n \xrightarrow{\text{P}} \theta^*$$

A1 says that the LR converges uniformly (wrt θ) to the KL divergence.
A2 says that locally θ^* is strictly better (in KL) other θ.

Proof. Since $\widehat{\theta}_n$ minimizes $L_n(\theta)$ we have

$$L_n(\widehat{\theta}_n) \leq L_n(\theta^*)$$

Hence,

$$\begin{aligned}
L(\widehat{\theta}_n) - L(\theta^*) &= L(\widehat{\theta}_n) - L_n(\theta^*) + L_n(\theta^*) - L(\theta^*) \\
&\leq L(\widehat{\theta}_n) - L_n(\widehat{\theta}_n) + L_n(\theta^*) - L(\theta^*) \\
&\leq \sup_\theta |L(\theta) - L_n(\theta)| + L_n(\theta^*) - L(\theta^*) \\
&\xrightarrow{\text{P}} 0, \quad \text{by A1}
\end{aligned}$$

It follows that for any $\delta > 0$

$$\mathbb{P}\left(L(\widehat{\theta}_n) > L(\theta^*) + \delta\right) \longrightarrow 0, \quad \text{as} \;\; n \longrightarrow \infty$$

Now pick any $\epsilon > 0$. By A2 $\exists \delta > 0$ such that

$$\|\theta - \theta^*\| \geq \epsilon \quad \Rightarrow \quad L(\theta) > L(\theta^*) + \delta$$

Hence

$$\mathbb{P}(\|\widehat{\theta}_n - \theta^*\| \geq \epsilon) \leq \mathbb{P}(L(\widehat{\theta}_n) > L(\theta^*) + \delta) \longrightarrow 0$$

□