The maximum Likelihood (ML) Estimate is given by

\[ \hat{\theta} = \arg \max_{\theta \in \Theta} p(x|\theta) \]

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

1 ML Estimation and Density Estimation

ML Estimation is equivalent to density estimation. Assume

\[ x_i \overset{\text{iid}}{\sim} q, \quad i = 1, \ldots, n, \]

where \( q \) 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

\[ \hat{\theta} = \arg \min_{\theta} -\log p(x|\theta) \]

Thus, we can view the MLE as minimizing the loss

\[ \ell(q, p_\theta) := -\log p(x|\theta) \]

where dependence on \( q \) is embodied in \( x \sim q \).

Example 1.

\[ p(x|\theta) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x - H\theta)^T\Sigma^{-1}(x - H\theta) \right\}, \quad x \in \mathbb{R}^n \text{ and } \theta \in \mathbb{R}^k \]

The value of \( \hat{\theta} \) is given by,

\[ \hat{\theta} = \arg \min_{\theta} -\log p(x|\theta) \]

\[ = \arg \min_{\theta} (x - H\theta)^T\Sigma^{-1}(x - H\theta) \]

\[ = (H^T\Sigma^{-1}H)^{-1}H^T\Sigma^{-1}x \]
2 MLE and Risk

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

\[
R_{\text{MLE}}(q, \theta) = E[\ell(q, p_{\theta})] = E[-\log p(x|\theta)] = \int q(x) (-\log p(x|\theta)) dx
\]

2.1 Excess Risk (“Regret”)

Let \( \theta \) be any value of the parameter. Then we can compare

\[
R_{\text{MLE}}(q, \theta) - R_{\text{MLE}}(q, q)
\]

which quantifies how much larger the expected loss is when we use \( \theta \) instead of \( \theta^* \). Note that

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

with equality if \( p_{\theta} = q \). Thus the “optimal” value of \( \theta \) is

\[\theta^* = \arg \min_\theta D(q\|p_{\theta}).\]

The density \( p_{\theta^*} \) 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

\[x_i \sim q, \quad i = 1, \ldots, n\]

the loss is given by

\[
\ell(q, p_{\theta}) = -\log \left( \prod_{i=1}^n p(x_i|\theta) \right) = -\sum_{i=1}^n \log p(x_i|\theta)
\]

MLE:

\[\hat{\theta} = \arg \min_\theta -\sum_{i=1}^n \log p(x_i|\theta)\]

Excess Risk:

\[
R_{\text{MLE}}(q, \theta) - R_{\text{MLE}}(q, q) = n D(q\|p_{\theta})
\]

for any \( \theta \in \Theta \)
3 Convergence of log likelihood to KL

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

\[
\frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\theta)} \xrightarrow{a.s.} D(p_{\theta^*} \parallel p_\theta)
\]

We would like to show that the MLE

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

converges to \( \theta^* \) in the following sense:

\[
D(p_{\theta^*} \parallel p_{\hat{\theta}_n}) \rightarrow 0
\]

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

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

Thus we have

\[
\frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\hat{\theta}_n)} - D\left(p_{\theta^*} \parallel p_{\hat{\theta}_n}\right) + D\left(p_{\theta^*} \parallel p_{\hat{\theta}_n}\right) \leq 0
\]

\[
\Rightarrow D\left(p_{\theta^*} \parallel p_{\hat{\theta}_n}\right) \leq \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\hat{\theta}_n)} - D\left(p_{\theta^*} \parallel p_{\hat{\theta}_n}\right) \right|
\]

So, \( D(p_{\theta^*} \parallel p_{\hat{\theta}_n}) \rightarrow 0 \) if \( \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\hat{\theta}_n)} \rightarrow D(p_{\theta^*} \parallel p_{\hat{\theta}_n}) \)

The subtle issue here is that \( \hat{\theta}_n \) is a random variable, not a fixed \( \theta \in \Theta \), so we can not just appeal to the SLLN.

**Theorem 1.** Assume

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

Define

\[
L_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\theta)}, \quad \forall \theta \in \Theta
\]

\[
L(\theta) := \mathbb{E}[L_n(\theta)] = D(p_{\theta^*} \parallel p_\theta)
\]

Suppose the following assumptions hold

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

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

then

\[
\hat{\theta}_n \xrightarrow{P} \theta^*
\]
A1 says that the LR converges uniformly (wrt $\theta$) to the KL divergence. A2 says that locally $\theta^*$ is strictly better (in KL) other $\theta$.

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

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

Hence,

$$L(\hat{\theta}_n) - L(\theta^*) = L(\hat{\theta}_n) - L_n(\theta^*) + L_n(\theta^*) - L(\theta^*) \leq \sup_{\theta} |L(\theta) - L_n(\theta)| + L_n(\theta^*) - L(\theta^*)$$

$$\xrightarrow{P} 0, \text{ by A1}$$

It follows that for any $\delta > 0$

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

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

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

Hence

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