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

$$X \overset{iid}{\sim} p, \quad i = 1, \ldots, n, \quad p \in \{p_\theta\}_{\theta \in \Theta}$$

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 $x$.”

1.1 ML Estimation as Minimization

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{p(x|\theta)}$$

$$\quad = \arg \min_{\theta} -\log p(x|\theta)$$

Thus, we can view the MLE as minimizing the loss

$$\ell(\theta^*, \hat{\theta}) := -\log p(x|\theta)$$

where dependence on $\theta^*$ is embodied in $x \sim p(x|\theta^*)$

Example 1.

$$p(x|\theta) = \frac{1}{(2\pi)^n/2|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x - H\theta)^T \Sigma^{-1} (x - H\theta)\}$$

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

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

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

$$\quad = (H^T \Sigma^{-1} H)^{-1} H^T \Sigma^{-1} x$$
MLE and Risk

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

\[ R_{\text{MLE}}(\theta^*, \hat{\theta}) = \mathbb{E}[\ell(\theta^*, \theta)] \]
\[ = \mathbb{E}[- \log p(x|\theta)] \]
\[ = \int p(x|\theta^*) (- \log p(x|\theta)) \, dx \]

2.1 Excess Risk (“Regret”)

Let \( \theta \) be any value of the parameter and \( \theta^* \) be the true value that generates \( x \). Then we can compare

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

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

Note that

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

with equality if \( \theta = \theta^* \)

Example 2.

\[ X \sim \mathcal{N}(H\theta, \Sigma), \quad \theta \in \mathbb{R}^k, \quad \Sigma, H \text{known} \]

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

3 Likelihood as a Loss function

In general

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

the loss is given by,

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

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

Excess Risk:

$$R_{\text{MLE}}(\theta^*, \theta) - R_{\text{MLE}}(\hat{\theta}, \theta^*) = nD(p(x|\theta^*)||p(x|\theta))$$

for any $\theta \in \Theta$

4 Convergence of log likelihood to KL

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

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

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:

$$D\left(p(x|\theta^*)||p(x|\hat{\theta}_n)\right) \rightarrow 0$$

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

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\hat{\theta}_n)} - D\left(p(x|\theta^*)||p(x|\hat{\theta}_n)\right) \leq 0$$

$$\implies D\left(p(x|\theta^*)||p(x|\hat{\theta}_n)\right) \leq \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta^*)}{p(x_i|\hat{\theta}_n)} - D\left(p(x|\theta^*)||p(x|\hat{\theta}_n)\right) \right|$$

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

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

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

Define

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

\[ L(\theta) := \mathbb{E}[L_n(\theta)] = D(p(x|\theta^*)\|p(x|\theta)) \]

Suppose the following assumptions hold

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

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

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) than \( \theta \).

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

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

Hence,

\[ L(\hat{\theta}_n) - L(\theta^*) = L(\hat{\theta}_n) - L_n(\theta^*) + L_n(\theta^*) - L(\theta^*) \leq L(\hat{\theta}_n) - L_n(\hat{\theta}_n) + L_n(\theta^*) - L(\theta^*) \leq \sup_{\theta} |L(\theta) - L_n(\theta)| + L_n(\theta^*) - L(\theta^*) \xrightarrow{P} 0, \quad \text{by A1} \]

It follows that for any \( \delta > 0 \)

\[ \mathbb{P}\left(L(\hat{\theta}_n) > L(\theta^*) + \delta\right) \xrightarrow{n \to \infty} 0 \]

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

\[ \|\theta - \theta^*\| \geq \epsilon \Rightarrow L(\theta) > L(\theta^*) + \delta \]

Hence

\[ \mathbb{P}(\|\hat{\theta}_n - \theta^*\| \geq \epsilon) \leq \mathbb{P}(L(\hat{\theta}_n) > L(\theta^*) + \delta) \xrightarrow{n \to \infty} 0 \]