What Value are Measurements?

Outline

- Maximum Likelihood Estimators (MLE)
- Maximum A Posteriori (MAP) Estimators
- Cramér-Rao Inequality
- Asymptotic Properties of MLE
- Probabilistic Models of Dynamical Systems
- Fisher Information Matrix
- Entropy Measure of Information
- Regularization.

1. Maximum Likelihood Estimator (MLE)

Principle: To estimate parameters $\underline{\theta}$, form $p(\text{observations} \mid \underline{\theta})$ and choose $\underline{\theta}^*$ to maximize this conditional probability density.

Note: Once observations Z_t are known, $p(Z_t \mid \underline{\theta})$ is just a deterministic function of $\underline{\theta}$. Therefore

$$\underline{\theta}^* = \arg\max_{\underline{\theta}} p(Z_t \mid \underline{\theta}).$$

The book uses the notation $\hat{\theta}_{ML}(y_*^N)$.

Example: Suppose a die (six sided) is weighted so that

$$p(k) = \frac{1}{6} + (k-3.5)\cdot\frac{\theta}{2.5}\cdot\frac{1}{6}$$

where $\theta \in [-1, 1]$, and that die rolls are independent events.

Given N rolls of the die, what is the MLE of θ?

Note that $p(k) = p(k|\theta)$ here, and for N outcomes $\{d_i\}_{i=1}^{N}$,

$$p(\{d_i\}_{i=1}^{N}|\theta) = \prod_{i=1}^{N} p(d_i|\theta) = q(d_i, \theta).$$

The function $q(d_i, \theta)$ is a polynomial in θ, and its maximum value occurs either at $q(d_i, -1)$, $q(d_i, +1)$, or $q(d_i, a)$ where

$$a \in (-1, 1)$$

and

$$\left.\frac{\partial q(d_i, \theta)}{\partial \theta}\right|_{\theta=a} = 0.$$

For the first roll, suppose $d_1 = 5$.

Then

$$p(5|\theta) = \frac{1}{6}\left\{1 + \frac{5-3.5}{2.5}\theta\right\}$$
$$= \frac{1}{6}\{1 + 0.6\theta\}$$

which is linear, so $\frac{\partial p}{\partial \theta}$ is constant $(= 0.1)$.
Therefore $\theta^* = \pm 1$. To determine which, evaluate $p(5|\theta)$:

$$p(5|-1) = \frac{1}{6}\{1 + (0.6)(-1)\} = \frac{0.4}{6} = 0.067$$
$$p(5|+1) = \frac{1}{6}\{1 + (0.6)(+1)\} = \frac{1.6}{6} = 0.267$$

Therefore, $\theta^* = +1$.

Suppose the die is rolled a second time, and $d_2 = 3$. Then

$$p(\{d_i\}_{i=1}^{2}, \theta) = p(5|\theta)\,p(3|\theta)$$
$$= \frac{1}{36}\{1 + 0.6\theta\}\{1 - 0.2\theta\}$$
$$= \frac{1}{36}\{1 + 0.4\theta - 0.12\theta^2\}$$

Taking $\frac{\partial p}{\partial \theta}$:

$$\frac{1}{36}\{0.4 - 0.24\theta\} = 0$$

$$\text{or} \quad 0.24\theta = 0.4 \quad \Rightarrow \theta = \frac{0.4}{0.24} > 1$$

Therefore, $\theta^* = \pm 1$ once again.

$$p(\{d_i\}_{i=1}^{2}, \theta) = \frac{1}{36}\{1 + 0.4\theta - 0.12\theta^2\}$$

So

$$p(\{d_i\}_{i=1}^{2}, 1) = \frac{1}{36}\{1 + 0.4 - 0.12\} = \frac{1}{36}\{1.28\}$$

$$p(\{d_i\}_{i=1}^{2}, -1) = \frac{1}{36}\{1 - 0.4 - 0.12\} = \frac{1}{36}\{0.48\}$$

$$\Rightarrow \theta^* = +1.$$

Try an exercise outside class: Write a program to calculate θ^* for an arbitrary sequence $\{d_i\}_{i=1}^{N}$, and see if you can detect a pattern. (Hint: Count the number of times each number comes up.)

2. Maximum a Posteriori (MAP) Estimation

Suppose we know <u>in advance</u> (a priori) the probability density $p(\theta)$. Then we can calculate $p(\theta|z)$ for observations z using Bayes' Rule:

$$p(\theta|z) = \frac{p(z|\theta)p(\theta)}{p(z)}$$

$$= \frac{p(z|\theta)p(\theta)}{\int_\theta p(z|\theta)p(\theta)\,d\theta}$$

where $p(z|\theta)$ is the parameterized (probabilistic) model that predicts observations z given known θ.

We can then choose to estimate θ by maximizing the conditional density $p(\theta|z)$ (which is only a function of θ):

$$\hat{\theta}_{MAP}(z) = \arg\max_\theta \{p(\theta|z)\}$$

Referring to the example of an unfair die, suppose we give the other player the benefit of the doubt and assume an a priori distribution

$$p(\theta) = \begin{cases} 1 \cdot \delta(\theta) \text{ if } \theta = 0 & \text{(fair die)} \\ 0 \quad \text{otherwise} & (\text{for } \theta \in [-1,1]) \end{cases}$$

As before, for $Z = \{d_i\}_{i=1}^{N}$,

$$p(Z|\theta) = \prod_{i=1}^{N} p(d_i|\theta) = \frac{1}{6}\left\{1 + \left(\frac{d_i - 3.5}{2.5}\right)\theta\right\}$$

for N = 1

Now,

$$p(\tfrac{d_i}{Z}|\theta)\,p(\theta) = \begin{cases} \frac{1}{6}\cdot\delta(\theta) \text{ for } \theta = 0 \\ 0 \quad \text{otherwise} \end{cases}$$

And,

$$\int_{\theta} p(\tfrac{d_i}{Z}|\theta)p(\theta)d\theta = \int_{-1}^{1} p(\tfrac{d_i}{Z}|\theta)p(\theta)d\theta = \frac{1}{6} = p(\tfrac{d_i}{Z})$$

$\Rightarrow$

$$p(\theta|Z) = \delta(\theta)$$

~~Or, if we assume the die is fair, no amount of observations will allow us to conclude it is not!~~

This example illustrates an important point about Bayesian estimation: An a priori zero probability of an event's occurrence precludes an observer from ever forming a Bayesian estimate that it has occurred! Be careful about assuming away possible outcomes.

Exercise: Rework this with $p(\theta) = \frac{1}{2}$ for all $\theta \in [-1, 1]$.

3. Cramér - Rao Inequality

The error covariance of a parameter estimate is a reasonable measure of the estimate's quality. For <u>unbiased</u> estimators and observations whose domain does not depend on the estimate, a lower bound of the error covariance can be found:

$$E\left\{(\underline{\hat{\theta}}(Z) - \underline{\theta}_0)(\underline{\hat{\theta}}(Z) - \underline{\theta}_0)^T\right\} > \underline{M}^{-1}$$

where

$$\underline{M} = E\left\{\left(\frac{d}{d\theta}\log p(Z|\theta)\right)\left(\frac{d}{d\theta}\log p(Z|\theta)\right)^T\right\}\Bigg|_{\underline{\theta}=\underline{\theta}_0}$$

$$= -E\left\{\frac{d^2}{d\theta^2}\log p(Z|\theta)\right\}\Bigg|_{\underline{\theta}=\underline{\theta}_0}$$

$\underline{M}$ is the Fisher Information Matrix.

In these equations, $\hat{\underline{\theta}}(z)$ is an unbiased estimate of $\underline{\theta}$ given observations Z, and $\underline{\theta}_0$ is the exact (true) value of the parameter vector (which may be unknown).

Exercise: Read the proof in Appendix 7A.