1. MAXIMUM LIKELIHOOD / ML ESTIMATION

Week 1 : Session 1.

Introduction to Estimation.

⊙ Where do we use Estimation?

Estimation has a lot of applications in signal processing and wireless communication. For instance, consider a simple wireless sensor network, with large number of Wireless sensor Nodes, which are basically transmitting $\underbrace{\text{information / measurements}}$ to a Base station.

observations

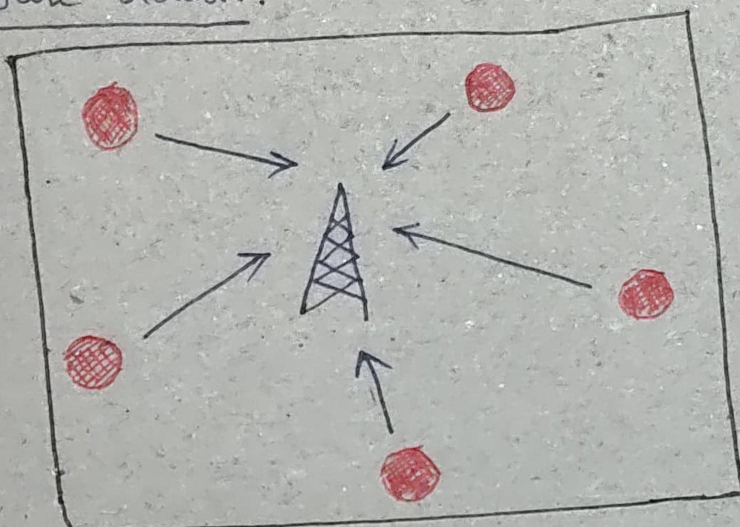


Fig: Wireless Sensor Network.

The Base Station has to estimate a parameter or quantity of interest from these different observations. (i) Each sensor nodes transmits an observation such as Temperature, Pressure, moisture level, etc., and the Base Station has to estimate the quantity. (either single quantity or several quantities together) (we'll study both)

There are various optimal / efficient techniques that deal with the estimation of this unknown quantity. Out of those, Maximum Likelihood Estimation is very popular.

Not just in Wireless Sensor Network, for instance, in any wireless communication system, when we look at the channel, which is an unknown quantity, has to be estimated.

How to estimate the unknown quantity / parameter at the BS?

The parameter estimation problem can be modeled as

Noisy observation

As-of-now, we are considering single parameter. Later on, we'll study more complex model in which we can have multiple parameters (i.e.) Parameter Vector.

$$y = h + v$$

(Measurement/ Observation) (parameter) (Noise)

Typically, we assume the noise (v) to be Gaussian with Mean $(\mu) = 0$ and variance $= \sigma^2$. Therefore, the PDF of v is given as $f_V(v) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{v^2}{2\sigma^2}}$

What is PDF of y?

Since $y = h + v$, (i.e.) when an unknown constant/fixed parameter (h) is added to the gaussian noise v (zero mean), the mean of the gaussian shifts. $\Rightarrow$ y becomes gaussian and the mean is going to be shifted to h and variance is unchanged.

Note:

Initially, we'll model a fixed parameter (h - Fixed unknown constant), and later on subsequently as we go into more advanced topics, we'll model with random parameters.

PDF of y is gaussian (i.e.) $y \sim \mathcal{N}(\underset{\text{Mean}}{h}, \underset{\text{Variance}}{\sigma^2})$ is given as

$$\boxed{f_y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-h)^2}{2\sigma^2}}}$$

Now, we generalize the model. Consider N measurements. Typically, we wont just have single measurement, rather we have multiple measurements, like we saw in WSN, there exists multiple sensor trying to measure the same quantity. Similarly in channel estimation, where we have multiple pilots.

So, this multiple measurements can be modelled as

$$\left.\begin{aligned} y(1) &= h + v(1) \\ y(2) &= h + v(2) \\ &\vdots \\ y(N) &= h + v(N) \end{aligned}\right\} \text{'N' measurements/Observations.}$$

In general, in communication, we use a lot of mathematical tools. Linear Algebra is one such important tool.

This can be denoted in vector form.

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} h \\ h \\ \vdots \\ h \end{bmatrix} + \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(N) \end{bmatrix}$$

Linear Algebra makes our life easy. And this is what we're going to see as we go through estimation as well.

Representing the quantities rather than in scalar, representing them in Vectors and Matrices, makes the notation/presentation/computation very efficient, which is going to be very important as we go forward.

$$\Rightarrow \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = h \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(N) \end{bmatrix}$$

$$\Rightarrow \boxed{\bar{y} = h\,\bar{1} + \bar{v}} \longleftarrow \text{Vector Model}$$

PDF of each $y(k) = h + v(k)$ is given as

$$f_{Y(k)}\left(y(k)\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y(k)-h)^2}{2\sigma^2}}$$

where $v(k) \rightarrow$ Gaussian noise with mean $= 0$, variance $= \sigma^2$

$\rightarrow \mathcal{N}(0, \sigma^2)$.

$h \rightarrow$ constant that is adding to the Gaussian noise.

$y(k) \rightarrow$ Shifted Gaussian with mean $= h$, variance $= \sigma^2$

$\rightarrow \mathcal{N}(h, \sigma^2)$

Joint PDF of observations $y(1), y(2), \ldots y(N)$, considering the noise samples $v(1), v(2), \ldots v(N)$ are Independent Identically Distributed (IID) will be the product of the individual PDFs.

(ii) $$f_{\bar{y}}(\bar{y}) = \left(\text{PDF of } y(1)\right) \times \left(\text{PDF of } y(2)\right) \times \ldots \times \left(\text{PDF of } y(N)\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y(1)-h)^2}{2\sigma^2}} \times \ldots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y(N)-h)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(y(k)-h)^2}$$

The Joint PDF is a function of vector $\bar{y}$. But, in an estimation model, the observation vector $\bar{y}$ is given, and the parameter h is the unknown quantity. So, when the Joint PDF is viewed as a function of the parameter h, we term it as a Likelihood function wrt. h.

(ii) $$p(\bar{y}\,;\, \underset{\text{deterministic unknown}}{h}) = \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(y(k)-h)^2}}_{\text{function of } h}$$

To simply put, for each value of h, what is the probability density of the observation vector $\bar{y}$. This basically indicates the likelihood that the corresponding parameter might have occured. So, what it means is, the likelihood of a parameter is basically indicated by the value of the Joint PDF of the observation vector $\bar{y}$, in the sense that the parameter for which the Joint PDF has a lower value, have a lower likelihood.

We look at this Joint PDF as a function of unknown parameter h, and choose the parameter h for which the likelihood has the maximum value. This is what we called as Maximum Likelihood principle.

The likelihood function $p(\bar{y}; h)$ is a function of 'h', which denotes "How well does 'h' explain observations"

$p(\bar{y}; h) \rightarrow$ Parameter 'h' is deterministic unknown

$p(\bar{y} | h) \rightarrow$ Parameter is Random

$$p(\bar{y}; h) \neq p(\bar{y} | h)$$ CAUTION ⚠

The value of 'h' that best explains the observation vector $\bar{y}$, is the value of the parameter 'h' for which the likelihood is maximum.

The estimate of h is obtained by maximizing the likelihood.

$$\boxed{\hat{h} = \max p(\bar{y}; h)}$$

This is known as <u>Maximum Likelihood Estimate</u> (or) <u>ML Estimate</u> (or) MLE.

The likelihood can be maximized as

constant

$$\max \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^{N} (y(k) - h)^2}$$

constant

due to '$-$' in the exponent.

$$\equiv \underline{\min} \sum_{k=1}^{N} (y(k) - h)^2$$

This can be minimized by differentiating w.r.t h and set to 0.

$$\frac{d}{dh} \sum_{k=1}^{N} (y(k) - h)^2 = 0$$

$$\Rightarrow \sum_{k=1}^{N} 2(y(k) - h)(-1) = 0$$

$$\Rightarrow \sum_{k=1}^{N} (y(k) - h) = 0$$

$$\Rightarrow \sum_{k=1}^{N} y(k) = \sum_{k=1}^{N} h$$

$$\Rightarrow \sum_{k=1}^{N} y(k) = N h$$

$$\Rightarrow \boxed{\hat{h} = \frac{1}{N} \sum_{k=1}^{N} y(k)}$$

This is known as ML Estimate.
(aka) Sample Mean
(aka) Average of observations.

Week 7 : Session 2

Properties of MLE (Noisy Measurements of a Parameter)

We now explore the properties of the ML Estimate.

$$\text{MLE} \Rightarrow \hat{h} = \frac{1}{N}\sum_{k=1}^{N} \underbrace{y(k)}_{\text{Gaussian}} = \frac{1}{N}\left(y(1) + \dots + y(N)\right)$$

$h \rightarrow$ Constant quantity

$\hat{h} \rightarrow$ Random quantity.

- What is the distribution of $\hat{h}$?

  Since $y(k)$ is Gaussian, $\hat{h}$ which is a linear Combination of Gaussian Random variables is also Gaussian. Thus $\hat{h} \sim$ Gaussian.

- What is the Mean of $\hat{h}$?

$$E\{\hat{h}\} = E\left\{\frac{1}{N}\sum_{k=1}^{N} y(k)\right\}$$

$$= \frac{1}{N}\sum_{k=1}^{N} E\{y(k)\}$$

$$= \frac{1}{N}\sum_{k=1}^{N} E\{h + v(k)\}$$

0, since $v(k)$ is Zero mean Gaussian

$$= \frac{1}{N}\sum_{k=1}^{N}\left(h + \cancel{E\{v(k)\}}\right)$$

$$= \frac{1}{N}\sum_{k=1}^{N} h$$

$$= \frac{1}{N} N h$$

$$\boxed{E\{\hat{h}\} = h}$$

Very Interesting Property! Although $\hat{h}$ is Random, the expected value of $\hat{h}$ is equal to the True parameter h, which we do not know.

Note that, Estimate is random. So, estimate will never be equal to the true parameter, unless it is some kind of rare circumstance. But here, On an average, it is behaving in such a fashion that the estimate coincides with the true parameter.

(i.e.) Mean of Estimate = True Parameter.

Such an Estimate is termed as an UNBIASED ESTIMATOR.

⊙ What is the Mean Square Error (MSE) / Variance of $\hat{h}$?

$$E\{(\hat{h}-\underset{\text{Mean}}{h})^2\} = E\left\{\left(\left(\frac{1}{N}\sum_{k=1}^{N} y(k)\right) - h\right)^2\right\}$$

Note that, we are considering all quantities to be REAL. But these can be easily extended to the estimation of complex quantities.

(ie) $E\{|\hat{h}-h|^2\}$

$$= E\left\{\left(\frac{1}{N}\sum_{k=1}^{N}(y(k)-h)\right)^2\right\}$$

$$= E\left\{\frac{1}{N^2}\left(\sum_{k=1}^{N} v(k)\right)^2\right\}$$

$y(k) = h + v(k)$
$\Rightarrow y(k) - h = v(k)$

$$= \frac{1}{N^2} E\left\{\left(\sum_{k=1}^{N} v(k)\right)\left(\sum_{l=1}^{N} v(l)\right)\right\}$$

$$= \frac{1}{N^2} E\left\{\sum_{k=1}^{N}\sum_{l=1}^{N} v(k)\cdot v(l)\right\}$$

$$= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} E\{v(k)\cdot v(l)\}$$

Since $v(k)$ is iid,

$E\{v(k)\cdot v(l)\} = E\{v(k)\}\, E\{v(l)\}$, if $k \neq l$

$= 0 \times 0 = 0$

$E\{v(k)\cdot v(l)\} = E\{v^2(k)\}$, if $k = l$

$= \sigma^2$

In general, $\boxed{E\{v(k)\cdot v(l)\} = \sigma^2\,\delta(k-l)}$

$$\Rightarrow E\{(\hat{h}-h)^2\} = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\sigma^2\,\delta(k-l)$$

$$= \frac{1}{N^2}\sum_{k=1}^{N}\sigma^2$$

$$= \frac{1}{N^2}\, N\sigma^2$$

$$\boxed{E\{(\hat{h}-h)^2\} = \frac{\sigma^2}{N}} \leftarrow \text{MSE / Variance.}$$

observe, MSE decreases as $\frac{1}{N}$.

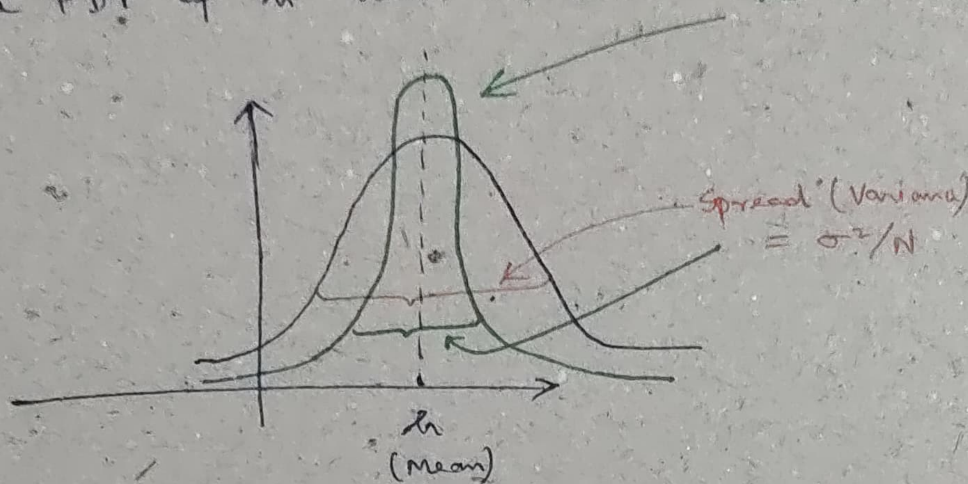
$\text{MSE} = \frac{\sigma^2}{N} \propto \frac{1}{N}$

(ie) $\text{MSE} \to 0$ as $N \to \infty$

Therefore, $\hat{h}$ is Gaussian with Mean $= h$, Variance $= \frac{\sigma^2}{N}$.

$$\boxed{\hat{h} \sim \mathcal{N}\left(h, \frac{\sigma^2}{N}\right)}$$

The PDF of $\hat{h}$ looks as below. As N increases, the variance decreases!



As N increases, the spread (i) variance decreases. This means that the Estimate is getting better and better as N increases.

Example: Consider the observation be $y(1) = 1$, $y(2) = -3$, $y(3) = 2$, $y(4) = -1$.

(i) What is the maximum likelihood estimate $\hat{h}$ of the unknown parameter, h?

$$\hat{h} = \frac{1}{N} \sum_{k=1}^{N} y(k)$$

$$= \frac{1}{4}\left(y(1) + y(2) + y(3) + y(4)\right)$$

$$= \frac{1}{4}\left(1 + (-3) + (2) + (-1)\right) = -\frac{1}{4}$$

(ii) Given the iid Gaussian noise samples of variance $\sigma^2 = 1/4$, what is the variance of the ML estimate?

$$\text{Variance} = \text{MSE} = \frac{\sigma^2}{N}$$

$$= \frac{1/4}{4} = \frac{1}{16}$$