Parameter Estimation

The task here is to infer the parameters of the model we are proposing in order to explain the data.

More concretely: imagine we have some data $\vec{x} = \{x_i\}_{i=1}^{N}$, then if this data is i.i.d., the likelihood:

$$p(\vec{x} \mid \vec{\theta}) = \prod_{i=1}^{N} p(x_i \mid \vec{\theta})$$

This can be considered just a function of $\vec{\theta}$ when we have fixed $\vec{x}$ to the observed values.

Assuming a particular distrib. for the likelihood, the task is to obtain the parameters that are more compatible with the observations.

- The estimator of the true parameter θ is denoted as $\hat{\theta}$

- $\hat{\theta}$ is said to be a consistent estimator of θ if

$$\lim_{n \to \infty} P\left(|\hat{\theta} - \theta| > \varepsilon \right) = 0$$

n: # of measurements of the experiment

- $\hat{\theta}$ is said to be unbiased if

$$\text{Bias} = \mathbb{E}[\hat{\theta}] - \theta = 0$$

$\mathbb{E}$ wrt different experiments

Before using the likelihood for estimating the parameters, it is useful to see some basics of parameter estimation:

- Even if the distrib. of $\{x_i\}$ is not known, the quantity

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(the sample mean)}$$

is a consistent estimator of $\mathbb{E}[x]$, and also unbiased:

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{n}\sum_i x_i\right] = \frac{1}{n}\sum_i \mathbb{E}[x_i]$$

$$= \frac{1}{n}(n\mu) = \mu$$

- The sample variance s^2 is defined by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

instead of $1/n$, to be an unbiased estimator of the true variance σ^2

Proof:

$$\mathbb{E}\left[\frac{1}{n-1}\sum_i (x_i-\bar{x})^2\right] = \mathbb{E}\left[\frac{1}{n-1}\left(\sum_i x_i^2 - 2\bar{x}\sum_j x_j + n\bar{x}^2\right)\right]$$

$$= \mathbb{E}\left[\frac{1}{n-1}\left(\sum_i x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right)\right]$$

$$= \mathbb{E}\left[\frac{1}{n-1}\left(\sum_i x_i^2 - n\bar{x}^2\right)\right]$$

$$= \frac{1}{n-1}\left(\sum_i \mathbb{E}[x_i^2] - n\,\mathbb{E}[\bar{x}^2]\right)$$

but $\mathbb{E}[x_i^2] = \mu^2 + \sigma^2 \quad \forall i$

and $\mathbb{E}[\bar{x}^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_i x_i\right)^2\right] = \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i,j=1}^{n} x_i x_j\right]$

$$= \frac{1}{n^2}\left\{\sum_i \mathbb{E}[x_i^2] + \sum_{i\neq j} \mathbb{E}[x_i x_j]\right\}$$

$$= \frac{1}{n^2}\left\{n(\mu^2+\sigma^2) + n(n-1)\mu^2\right\}$$

$$= \frac{1}{n^2}\left\{n^2\mu^2 + n\sigma^2\right\} = \mu^2 + \sigma^2/n$$

$$\Rightarrow \frac{1}{n-1}\left\{n(\mu^2+\sigma^2) - n\mu^2 - \sigma^2\right\} = \sigma^2 \quad \square$$

- Given an estimator, one can compute its variance

$$\mathrm{Var}[\hat{\theta}] = \mathbb{E}[\hat{\theta}^2] - \left(\mathbb{E}[\hat{\theta}]\right)^2$$

e.g. the variance of the sample mean $\bar{x}$:

$$\mathrm{Var}[\bar{x}] = \mathbb{E}\left[\left(\frac{1}{n}\sum_i x_i\right)^2\right] - \mu^2$$

$$= \mu^2 + \sigma^2/n - \mu^2 = \sigma^2/n \quad \square$$

As it is well known, the std of the mean of n measurements of x goes like $1/\sqrt{n}$

* Maximum likelihood estimators

Going back to the likelihood of a given iid data

$$X = \{ \vec{x}_i \}_{i=1}^{N}$$

$$p(X|\vec{\theta}) = \prod_i p(\vec{x}_i|\vec{\theta})$$

the idea is that the probability of these data under the correct hypothesis $\vec{\theta}_{true}$ will be larger than under a different hypothesis

So the procedure is to maximise the likelihood:

$$\vec{\theta}_{MLE} = \underset{\vec{\theta}}{\operatorname{argmax}}\ p(X|\vec{\theta})$$

Let's see an example:

Suppose $\mathcal{L}(\mu, \sigma^2) = \prod_i \mathcal{N}(x_i|\mu, \sigma^2)$

what is μ_{MLE} and σ^2_{MLE} ?

Convenient to work with $\ln \mathcal{L}$ instead:

$$\ln \mathcal{L} = -\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - N \ln(\sqrt{2\pi\sigma^2})$$

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \Rightarrow -\sum_i \frac{(x_i-\mu)}{\sigma^2} = 0$$

$$\frac{1}{\sigma^2}\left(\sum_i x_i - N\mu\right) = 0$$

$$\boxed{\mu_{MLE} = \frac{1}{N}\sum_i x_i}$$ → as expected

$$\frac{\partial \ell}{\partial \sigma^2} = 0 \Rightarrow +\sum_i \frac{(x_i-\mu)^2}{2\sigma^4} - \frac{N}{\sqrt{\sigma^2}}\frac{1}{2\sqrt{\sigma^2}} = 0$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_i (x_i - \mu_{MLE})^2 \simeq \mu_{MLE}$$

So the MLE of σ^2 is not an unbiased estimator!

Note that in general such optimisation will deal with a large # of parameters in general, and it will not be analytical

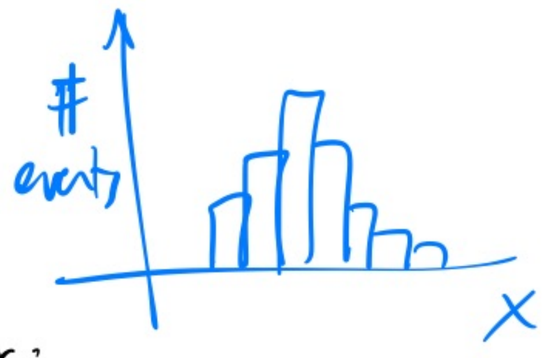
→ There are in the market some public codes for dealing with this (i.e. with automatic differentiation in general)

↳ e.g. Tensorflow and PyTorch

Google Facebook

* "Shape analysis"

Previously, when doing GoF, we saw an example of a counting experiment where one measures a variable x and form a histogram with the # of events.

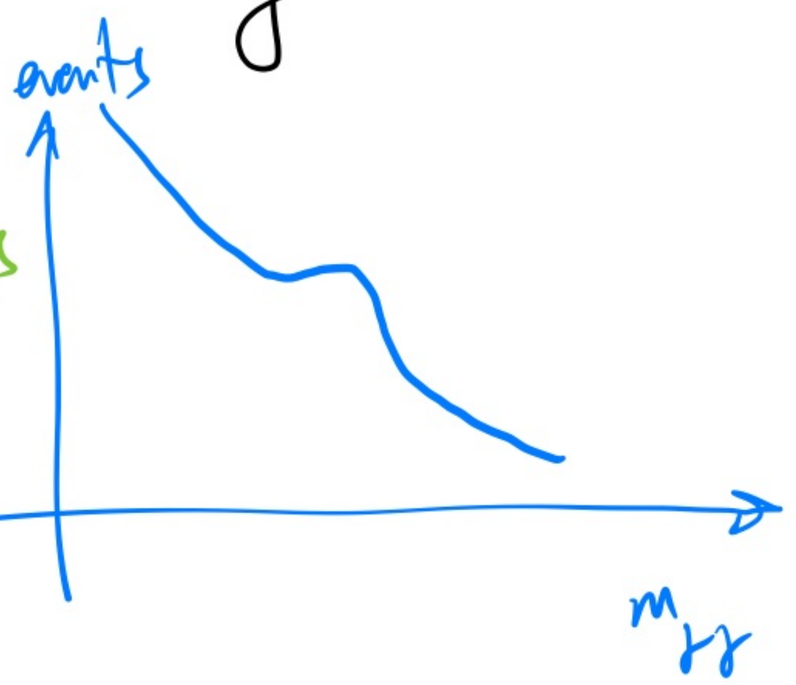


The task is to estimate the parameters $\vec{\theta}$ of the PDF of x:

$p(x|\vec{\theta})$ by MLE

An illustrative example can be given in the context of LHC physics:

Search for Higgs in $H \to \gamma\gamma$ decay channel



Here one can reduce the problem to 2 parameters:

$$\vec{\theta} = \{S, B\}$$

total expected # of events: S and B → Higgs, no Higgs

The likelihood of these data in the form of a histogram like can be expressed as

$$p(\vec{n} \mid S, B) = \prod_{i=1}^{N_{bins}} \text{Pois}(n_i \mid S f_i^S + B f_i^B)$$

here the "shapes" $f_i^{S,B}$ assumed to be known

$S \cdot f_i^S$: # of expected events in bin i from signal

$S \cdot f_i^B$: " from bckg.

The above is an example of a "binned shape analysis"

If instead of taking $\{n_i\}_{i=1}^{N_{bins}}$ as data, we take directly $\{x_i\}_{i=1}^{N}$

N: total # of events

$x = m_{\gamma\gamma}$ in the above example

one should follow a different strategy:

we assume the forms of the PDFs for signal & backg. e.g. in the $H \to \gamma\gamma$ case above one may have

$$N(x \mid m_H, \sigma^2) \quad \} \text{ PDF of the signal}$$

$$\beta e^{-\beta x} \quad \} \text{ exponential PDF, for the backg}$$

s.t. given a value of x:

$$p(x) = \frac{S}{S+B} N(x \mid m_H, \sigma^2) + \frac{B}{S+B} \beta e^{-\beta x}$$

Since the total # of events N is also a random variable in this case, the total likelihood can be written as ("extended")

$$p(\vec{x}) = Pois(N \mid S+B) \prod_{i=1}^{N} \left\{ \frac{S}{S+B} N(x_i \mid m_H, \sigma^2) + \frac{B}{S+B} \beta e^{-\beta x_i} \right\}$$

Note, here S will depend on m_H !

The above is an example of an "unbinned shape analysis"

Note also that, in the above binned analysis, if $N_{bins} \to \infty$, then

$$f_i^S \to N(x_i \mid m_H, \sigma^2)$$

i.e the "shape" parameters tend to the PDFs

$$f_i^B \to \beta e^{-\beta x_i}$$

* Bayesian Parameter Estimation and Inference

The idea here is to use the posterior distrib.

$$p(\vec{\theta}|D) = \frac{p(D|\vec{\theta})\,p(\vec{\theta})}{p(D)}$$

} for a given model

Note that, while in the frequentist MLE approach the outcome is just a best-fit point

$$\vec{\theta}_{MLE} = \text{argmax}\; p(D|\vec{\theta})$$

here one has access to the whole distrib. of $\vec{\theta}$. However a "best fit" could be obtained by

$$\vec{\theta}_{MAP} \equiv \text{argmax}\; p(\vec{\theta}|D)$$

~ "maximum a posteriori"

The problem is that in many practical situations the posterior $p(\vec{\theta}|D)$ is intractable. The reason is that the evidence

$$p(D) = \int d\theta\; p(D|\vec{\theta})\,p(\vec{\theta})$$

is intractable (even if the prior and likelihood are Gaussians, if the dependence of the likelihood with $\vec{\theta}$ is not linear, the integral will not be Gaussian