Maximum Likelihood (ML)

ML 1

Data $\vec{x} = \{x_i\}$ N of them

Parameters $\vec{\alpha} = \{\alpha_i\}$ M of them

Likelihood

$$\mathcal{L} = \prod_{i=1}^{N} p(x_i|\vec{\alpha}) \longleftarrow \text{for discrete } x_i$$

$$\mathcal{L} = \prod_{i=1}^{N} P(x_i|\vec{\alpha})\,dx_i \longleftarrow \text{for continuous } x_i$$

$p(x_i|\vec{\alpha})\,dx_i$ = PDF for x_i given $\vec{\alpha}$

<u>ML estimation</u> : "estimate" ("fit") $\vec{\alpha}$ by choosing $\vec{\alpha}$ such that $\mathcal{L}$ is maximized.

Makes intuitive sense.
It turns out to have some nice properties in the high N limit eg consistency (unbiased) and optimal

Maximizing $\mathcal{L}$ $\rightarrow$ maximize $\log \mathcal{L}$

It is easier to deal with $\log \mathcal{L}$ than $\mathcal{L}$

$$\log \mathcal{L} = \sum_{i=1}^{N} \log\left[P(x_i|\vec{\alpha})\,dx_i\right]$$

ML2

Numerically more convenient (customary?) to minimize function instead of maximize

Minimize

$$-\log \mathcal{L} = -\sum_{i=1}^{N} \log\left[p(x_i)\,dx_i\right] \qquad p(x_i) = p(x_i|\vec{\alpha})$$

Consider Gaussian PDFS

$$p(x_i|\vec{\alpha})\,dx_i = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{(x_i - \mu_i(\vec{\alpha}))^2}{2\sigma_i^2}\right] dx_i$$

$$\log\left[p(x_i|\vec{\alpha})\,dx_i\right] = -\frac{(x_i - \mu_i(\vec{\alpha}))^2}{2\sigma_i^2} + \log\frac{dx_i}{\sqrt{2\pi}\,\sigma_i}$$

$$= -\frac{\left[x_i - \mu_i(\vec{\alpha})\right]^2}{2\sigma_i^2} + (\text{Const})_i$$

indep of $\vec{\alpha}$

$$-\log \mathcal{L} = +\sum_{i=1}^{N} \frac{(x_i - \mu_i(\vec{\alpha}))^2}{2\sigma_i^2} + \sum(\text{Const})_i$$

another constant

$$\boxed{-\log \mathcal{L} = \frac{1}{2}\chi^2 + \text{Const}}$$

$-\log \mathcal{L} = \text{NLL}$

Negative Log Likelihood

For Gaussian uncertainty

minimize $\chi^2 \longleftrightarrow$ minimize NLL

There is a factor of ½ btw the two

ML3

The power of ML is that you dont have to assume Gaussian uncertainties which is effectively what is done with χ^2 estimates

For example, fit to a histogram of counts where the counts are small enough that we are in the Poisson regime

Then instead of minimizg

$$\cancel{\chi^2 = \sum (x_i - \mu_i(\alpha}\quad \chi^2 = \sum \frac{(x_i - \mu_i(\vec{\alpha}))^2}{x_i}$$

(Remember: counting stats, gaussian regime $\sigma = \sqrt{N}$)

minimize NLL

$$NLL = -\sum_{i=1}^{N} \log\left[p(x_i|\vec{\alpha})\right]$$

But $p(x_i|\vec{\alpha}) = \text{Poisson}(x_i|\vec{\alpha}) = \dfrac{\mu_i(\vec{\alpha})^{x_i} e^{-\mu_i(\vec{\alpha})}}{x_i!}$

$$NLL = \sum_{i=1}^{N} \mu_i(\vec{\alpha}) - x_i \log(\mu_i(\vec{\alpha})) + \underbrace{\log x_i!}_{\text{constant}}$$

(we can drop it)

ML4

~~$\mathcal{L} = \prod p$~~ $\mathcal{L} = \prod p(x_i|\vec{\alpha}) = \mathcal{L}(\vec{\alpha})$

It can be shown that in the large N limit $\mathcal{L}(\vec{\alpha})$ is Gaussian

eg in $M=1$ case, $\vec{\alpha} = \alpha$, fitted value α^0

$$\mathcal{L} \sim \exp\left(-\frac{(\alpha-\alpha^0)}{2\sigma^2}\right)$$

In multidimensional case

$$\vec{\alpha} - \vec{\alpha}^0 = \begin{pmatrix} \alpha_1 - \alpha_1^0 \\ \alpha_2 - \alpha_2^0 \\ \vdots \\ \alpha_M - \alpha_M^0 \end{pmatrix}$$

column vector $= M \times 1$ matrix

$$\mathcal{L} \sim \mathrm{Exp}\left[-\frac{1}{2}\underbrace{(\vec{\alpha})^T}_{1\times M \text{ matrix}} \underbrace{V^{-1}}_{\substack{M\times M \text{ covariance matrix} \\ \text{(inverse of)}}} \underbrace{\vec{\alpha}}_{M\times 1 \text{ matrix}}\right]$$

$$NLL \doteq \frac{1}{2}(\vec{\alpha})^T V^{-1} \alpha + \text{Const}$$

For a single parameter

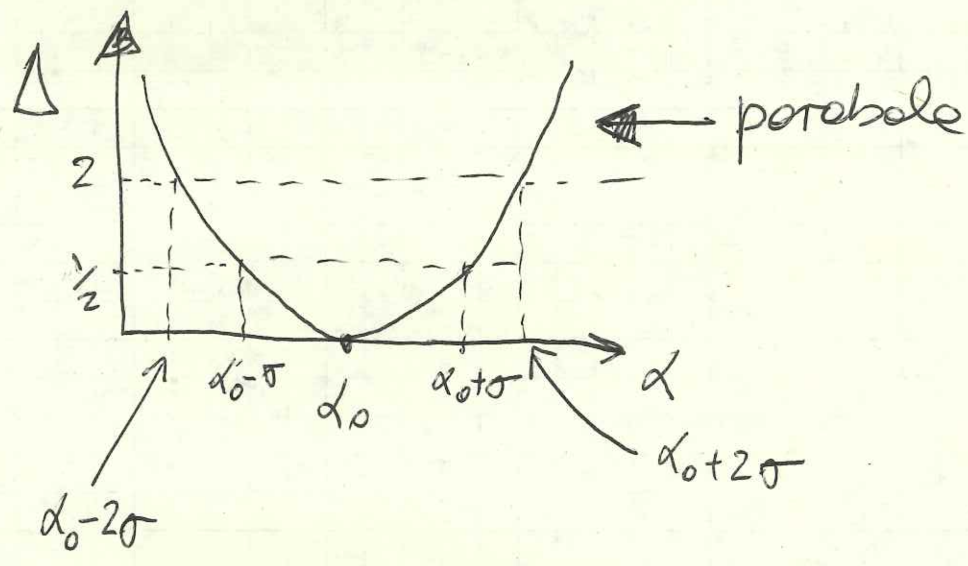
$$NLL = \frac{1}{2}\frac{(\alpha-\alpha^0)^2}{\sigma^2} + \text{const}$$

$$(-\log \mathcal{L})_{MIN} = \text{const}$$

$$= -\log\mathcal{L} - (-\log\mathcal{L})_{MIN}$$

$$\Delta(-\log\mathcal{L}) = \frac{1}{2}\frac{(\alpha-\alpha^0)^2}{\sigma^2} \equiv \Delta$$

~~A change in log $\Delta(-\log\mathcal{L})$ of 1~~



~~One more interesting thing~~

ML6

One more interesting thing

$$\mathcal{L} = \prod p(x_i|\vec{\alpha})$$

$\mathcal{L} = \mathcal{L}(\vec{x}_i, \vec{\alpha})$ is the prob of $\vec{x}$ given $\vec{\alpha}$

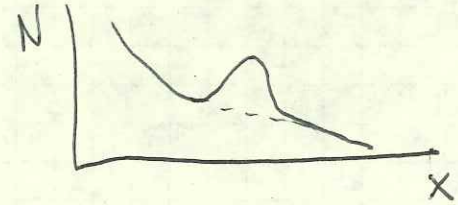
Using Bayes theorem, we can turn it around and (up to multiplicative constant) we can interpret it as the prob of $\vec{\alpha}$ given $\vec{x}$ assuming a flat prior for $\vec{\alpha}$

$$\boxed{p(\vec{\alpha}|\vec{x}) = \mathcal{L}(\vec{x}_i|\vec{\alpha}) \times \text{Const}}$$

This then justifies taking $\Delta = 1$, $\Delta = 2$, etc as the 1σ, 2σ, etc uncertainties

Extended max likelihood formalism

Fit a distribution as sum of signal + background



S = number of signal events (to be fitted)

B = number of background events (to be fitted)

$P_s(x)$ = pdf for signal, normalized to $\underline{1}$

$P_B(x)$ = pdf for background, normalized to $\underline{1}$

$\vec{x} = \{x_1 \ldots x_N\}$ set of N values (events)

N = number of values (events)

$$p(N) = e^{-(S+B)} \frac{(S+B)^N}{N!} = \text{prob of having } N \text{ events}$$

$$p_i = \text{prob of event } i = \frac{S\,P_s(x_i) + B\,P_B(x_i)}{S+B}$$

Likelihood (extended!)

$$\mathcal{L} = p(N) \prod_{i=1}^{N} p_i =$$

$$\mathcal{L} = e^{-(S+B)} \frac{(S+B)^N}{N!} \prod_{i=1}^{N} \left(\frac{S\,P_s(x_i) + B\,P_B(x_i)}{S+B} \right)$$

$$\mathcal{L} = e^{-(S+B)} \frac{(S+B)^N}{N!} \frac{\prod_{i=1}^{N} \left(S\,P_s(x_i) + B\,P_B(x_i)\right)}{(S+B)^N}$$

$$\mathcal{L} = e^{-(S+B)} \frac{1}{N!} \prod_{i=1}^{N} \left[S P_S(x_i) + B P_B(x_i) \right]$$

$$-\log \mathcal{L} = S + B - \sum_{i=1}^{N} \log \left[S P_S(x_i) + B P_B(x_i) \right] + \cancel{\log(N!)}$$

constant
ignore

$$\boxed{-\log \mathcal{L} = S + B - \sum_{i=1}^{N} \log \left[S P_S(x_i) + B P_B(x_i) \right]}$$