

The Maximum Likelihood (ML) method

❖ derivation as a limiting case of the least squares method

■ consider a histogram of N IID measurements of x , assuming ...

- ▶ counting statistics
- ▶ infinitesimal bin widths Δx , i.e. bin contents $y_i \in 0, 1$
- ▶ model described by a PDF $f(x, a)$ with a single parameter a

■ iterative minimization for a with $\hat{a}$ the estimate from the previous step

$$\begin{aligned}\chi^2 &= \sum_{\text{bins}, i} \frac{(y_i - N p_i)^2}{N \hat{p}_i} \quad \text{with} \quad p_i = f(x_i, a) \Delta x \quad \text{and} \quad \hat{p}_i = f(x_i, \hat{a}) \Delta x \\ &= \sum_{\text{bins}, i} \frac{y_i^2}{N \hat{p}_i} - 2 \sum_{\text{bins}, i} \frac{y_i p_i}{\hat{p}_i} + N \sum_{\text{bins}, i} \frac{p_i^2}{\hat{p}_i}\end{aligned}$$

- ▶ the 1st term is arbitrary ($\propto 1/\Delta x$) depends on $\hat{a}$, but is independent of a
- ▶ only the 2nd and 3rd terms contain p_i and are functions of a

for infinitesimal bin widths with $y_i \in \{0, 1\}$:

$$-2 \sum_{\text{bins}, i} \frac{y_i p_i}{\hat{p}_i} = -2 \sum_{\text{events}, i} \frac{p_i}{\hat{p}_i} = -2 \sum_{\text{events}, i} \frac{f(x_i, a) \Delta x}{f(x_i, \hat{a}) \Delta x} = -2 \sum_{\text{events}, i} \frac{f(x_i, a)}{f(x_i, \hat{a})}$$

and

$$N \sum_{\text{bins}, i} \frac{p_i^2}{\hat{p}_i} = N \sum_{\text{bins}, i} \frac{f^2(x_i, a) (\Delta x)^2}{f(x_i, \hat{a}) \Delta x} \rightarrow N \int dx \frac{f^2(x, a)}{f(x, \hat{a})}$$

minimisation of χ^2 with convergence $\hat{a} \rightarrow a$ then leads to:

$$\frac{\partial}{\partial a} \chi^2 = -2 \sum_{\text{events}, i} \frac{f'(x_i, a)}{f(x_i, \hat{a})} + N \int dx \frac{2 f(x, a) f'(x, a)}{f(x, \hat{a})}$$

$$\stackrel{\hat{a} \rightarrow a}{=} -2 \sum_{\text{events}, i} \frac{f'(x_i, a)}{f(x_i, a)} + 2N \int dx f'(x, a) = -2 \sum_{\text{events}, i} \frac{f'(x_i, a)}{f(x_i, a)}$$

the integral vanishes because of $\int dx f'(x, a) = \frac{\partial}{\partial a} \int dx f(x, a) = \frac{\partial}{\partial a} 1 = 0$

► result:

$$\begin{aligned}\frac{\partial}{\partial a} \left(\frac{1}{2} \chi^2 \right) &= - \sum_{\text{events}, i} \frac{f'(x_i, a)}{f(x_i, a)} = \frac{\partial}{\partial a} \left(- \sum_{\text{events}, i} \ln f(x_i, a) \right) \\ &= \frac{\partial}{\partial a} \left(- \ln \prod_{\text{events}, i} f(x_i, a) \right) = - \frac{\partial}{\partial a} \ln L(a)\end{aligned}$$

- the best fit parameter is obtained by maximizing the likelihood of the data
- for uncorrelated measurements it is the estimate with the smallest variance
- going to infinitesimal bin sizes, the χ^2 -minimum becomes arbitrary, i.e.
the maximum of the likelihood contains no information about the quality of the fit
- maximum likelihood and least squares fits have analogous properties

$$\Delta(-\ln L) = \frac{1}{2} \Delta \chi^2$$

❖ estimate also the Poisson average N for n events that were seen:

$$\begin{aligned}\chi^2 &= \sum_{\text{bins}, i} \frac{(y_i - N p_i)^2}{\hat{N} \hat{p}_i} \quad \text{with} \quad p_i = f(x_i, a) \Delta x \quad \text{and} \quad \hat{p}_i = f(x_i, \hat{a}) \Delta x \\ &= \sum_{\text{bins}, i} \frac{y_i^2}{\hat{N} \hat{p}_i} - 2 \frac{N}{\hat{N}} \sum_{\text{bins}, i} \frac{y_i p_i}{\hat{p}_i} + \frac{N^2}{\hat{N}} \sum_{\text{bins}, i} \frac{p_i^2}{\hat{p}_i}\end{aligned}$$

- ▶ only the second and the third term depend on N and a
- ▶ χ^2 function in the limit of infinitesimal bin widths with $y_i \in \{0, 1\}$:

$$\chi^2 = \text{const} - 2 \frac{N}{\hat{N}} \sum_{\text{events}, i}^n \frac{f(x_i, a)}{f(x_i, \hat{a})} + \frac{N^2}{\hat{N}} \int dx \frac{f^2(x, a)}{f(x, \hat{a})}$$

→ derivatives in the limit $\hat{N} \rightarrow N$ and $\hat{a} \rightarrow a$:

$$\frac{\partial \chi^2}{\partial N} = \frac{\partial}{\partial N} \left(-2 \frac{N}{\hat{N}} \sum_{\text{events}, i}^n \frac{f(x_i, a)}{f(x_i, \hat{a})} + \frac{N^2}{\hat{N}} \int dx \frac{f^2(x, a)}{f(x, \hat{a})} \right) \rightarrow -2 \frac{n}{N} + 2$$

$$\frac{\partial \chi^2}{\partial a} = \frac{\partial}{\partial a} \left(-2 \frac{N}{\hat{N}} \sum_{\text{events}, i}^n \frac{f(x_i, a)}{f(x_i, \hat{a})} + \frac{N^2}{\hat{N}} \int dx \frac{f^2(x, a)}{f(x, \hat{a})} \right) \rightarrow -2 \sum_{\text{events}, i}^n \frac{f'(x_i, a)}{f(x_i, a)}$$

which implies (up to a global constant)

$$\frac{1}{2} \chi^2 = N - n \ln N - \sum_{\text{events}, i}^n \ln f(x_i, a) = N - \sum_{\text{events}, i}^n \ln(N f(x_i, a))$$

→ ML & EML method can be derived from the LSQ method – NIMA 940 (2019) 135

Maximum Likelihood versus Least Squares

❖ compare two solutions to the same problem:

■ estimate the mean value of a PDF from a n IID measurements x_i , $i = 1, \dots, n$

➤ case 1: nothing is known → unweighted LSQ fit

■ minimize

$$\chi^2 = \sum_{i=1}^n (x_i - \mu)^2$$

■ estimate from $\partial\chi^2/\partial\mu = 0$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \pm \frac{\sigma}{\sqrt{n}}$$

→ the uncertainty decreases $\propto 1/\sqrt{n}$

► case 2: the PDF is known to be rectangular → ML fit

■ PDF and log-likelihood function:

$$f(x, a, b) = \frac{1}{b-a} \Theta(x-a) \Theta(b-x) \quad \text{and}$$

$$\ln L(a, b) = \sum_{i=1}^n \ln(\Theta(x_i - a)) + \ln(\Theta(b - x_i)) - \ln(b - a)$$

■ maximize $\ln L$:

- the true mean is $\mu = (a + b)/2$
- a should be as large as possible – but must be smaller than $x_{\min}$
- b should be as small as possible – but must be larger than $x_{\max}$

■ best estimate

$$\hat{\mu} = \frac{1}{2}(x_{\min} + x_{\max}) \pm \frac{\sigma \sqrt{6}}{\sqrt{(n+1)(n+2)}}$$

→ asymptotically the uncertainty decreases $\propto 1/n$

Asymptotic behaviour of ML parameter estimates

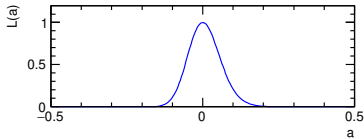
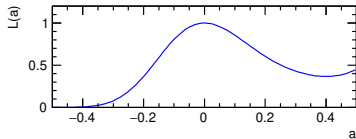
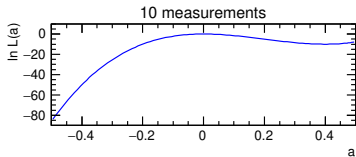
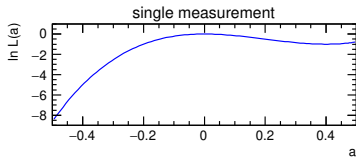
- the ML method is consistent: parameter estimates are asymptotically unbiased
 - ▶ for $N \rightarrow \infty$ the log-likelihood function has a stationary point at the true parameters
 - ▶ proof by the Law of Large Numbers

$$\begin{aligned}\lim_{N \rightarrow \infty} \frac{\partial}{\partial a} \sum_{i=1}^N \ln f(x_i, a) &= N \left\langle \frac{\partial \ln f(x, a)}{\partial a} \right\rangle \\ &= N \int dx f(x, a) \frac{1}{f(x, a)} \frac{\partial f(x, a)}{\partial a} = N \frac{\partial}{\partial a} \int dx f(x, a) = 0\end{aligned}$$

- the ML method is efficient: for $N \rightarrow \infty$ parameter estimates have minimal variance
 - ▶ the covariance matrix is the inverse of the Hessian of the -log-likelihood function
 - ▶ it saturates the Cramer-Rao bound (shown later)

■ **asymptotic normality:** for $N \rightarrow \infty$ parameter estimates have gaussian uncertainties

- ▶ log-likelihood function has a maximum
- ▶ in the limit $N \rightarrow \infty$ only the parabolic terms are relevant
- ▶ sketch:



If a population with a variate x is distributed according to the probability distribution $f(x, \theta_1, \theta_2, \dots, \theta_h)$, such that optimum estimates $\hat{\theta}_i$ of θ_i exist which are distributed in large samples according to (1), then when the hypothesis H is true that $\theta_i = \theta_{0i}, i = m+1, m+2, \dots, h$, the distribution of $-2 \ln \lambda$, where λ is given by (2) is, except for terms of order $1/\sqrt{n}$, distributed like χ^2 with $h - m$ degrees of freedom.

- ▶ (1) a PDF deviating from a d-dim Gaussian only by terms of order $1/\sqrt{n}$
- ▶ (2) the ratio of the best fit likelihoods fitting all or only m parameters fixing the others to the true values

$$\lambda = \frac{L(\hat{\theta}_1, \dots, \hat{\theta}_m, \theta_{0m+1}, \dots, \theta_{0h})}{L(\hat{\theta}_1, \dots, \hat{\theta}_m, \hat{\theta}_{m+1}, \dots, \hat{\theta}_h)}$$

→ likelihoods are meaningless, only likelihood ratios carry information

❖ necessary conditions for Wilks' theorem to apply

- **asymptotic statistics**: sufficient data is observed
 - ▶ ensure that errors of order $1/\sqrt{n}$ are small
- **interior**: all parameters sufficiently far from the boundary of the parameter space
 - ▶ best fit parameters should not be able to fluctuate outside their allowed range
- **indentifiability**: changes in parameters modify the model
 - ▶ e.g. the location of a signal with amplitude zero is not identifiable
- **nestedness**: the model H with fixed parameters is a limiting case of the general model
 - ▶ both models must be defined on the same parameter space
- **correctness**: either the restricted or the general model must correspond to the truth
 - ▶ ensure that χ^2 variables have a well defined behaviour

applications →

Testing for the existence of a signal component in data

- when the location of a signal is known, determine the likelihood ratio λ
 - ▶ from a fit with a model containing a **signal-component** s
 - ▶ from a fit with the same model and with **fixed** $s = 0$

$$\lambda = \frac{L_{\text{best}}(s = 0)}{L_{\text{best}}(s)} \quad \text{and} \quad z_{\text{obs}} = -2 \ln \lambda = 2 \Delta \ln L$$

- by Wilks' theorem convert $\lambda \rightarrow$ **significance** n in units of gaussian standard deviations
 - ▶ calculate the upper tail probability p for $z > z_{\text{obs}}$ for the hypothesis $s = 0$

$$p = \int_{z=z_{\text{obs}}}^{\infty} dz \chi^2(z, \text{ndf} = 1) = \frac{1}{\sqrt{2\pi}} \int_{z=z_{\text{obs}}}^{\infty} dz \frac{e^{-z/2}}{\sqrt{z}}$$

- ▶ determine the significance

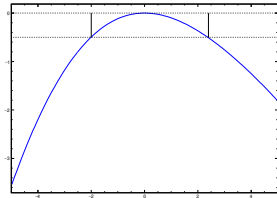
$$\text{solve } p = \frac{1}{\sqrt{2\pi}} \int_{-n}^n dz e^{-z^2/2} \quad \text{for } n$$

$$\text{e.g.: } p = 2.7 \cdot 10^{-3} \rightarrow n = 3 \sigma \quad \text{or} \quad p = 5.7 \cdot 10^{-7} \rightarrow n = 5 \sigma$$

Profiling - constrained maximization of the likelihood function

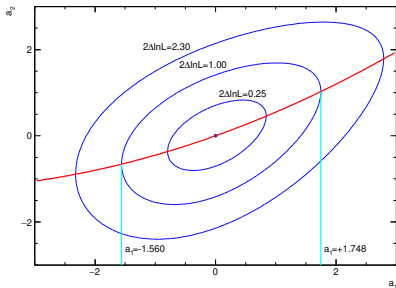
❖ extract information from the likelihood function: the 1 dim case

- best fit parameter $\hat{\theta}$: maximum of the likelihood function
- uncertainty: use Wilks' theorem to construct asymmetric errors
 - ▶ scan tentative true values θ around $\hat{\theta}$ and determine $2\Delta \ln L$
 - ▶ Wilks' theorem: $2\Delta \ln L$ is χ^2 distributed with $\text{ndf} = 1$
 - ▶ asymptotically $\hat{\theta}$ is with 68% probability in a region with $\chi^2 < 1$
 - ▶ $\Delta \ln L < 1/2$ defines a range around $\hat{\theta}$ which contains θ with 68% confidence level



- additional remarks
 - ▶ asymmetric errors and variance of $\hat{\theta}$ are alternative/different concepts
 - ▶ asymptotically both have correct coverage
 - ▶ neither can substitute the Neyman construction

❖ the 2-dim case



- analogous approach as in 1-dim case
- probability content of $2\Delta \ln L$ given by χ^2 distribution with $\text{ndf}=2$
 - ▶ 68% CL requires $2\Delta \ln L < 2.3$
 - ▶ $2\Delta \ln L < 1$ has 39% CL in 2-dim

- implementation in MINUIT: two types of error estimates
 - ▶ **HESSE**: estimate of the covariance matrix of the best fit parameter
 - ▶ **MINOS**: asymmetric errors from likelihood scan (1- and 2-dimensions)

Combination of asymmetric errors

→ R.Barlow et al., arXiv:2411.15499, NIMA 1082(2026)170857

- symmetric errors are interpreted as standard deviations and added in quadrature

$$42 \pm 3 \pm 4 = 42 \pm 5 \quad \text{OK}$$

- ▶ correct if the errors are indeed standard deviations

- asymmetric errors are $\Delta \ln L = 1/2$ points on likelihood functions

$$42_{-8-15}^{+5+12} \stackrel{?}{=} 42_{-17}^{+13}$$

- ▶ questionable to add objects in quadrature, which are not standard deviations
- ▶ correct limit if symmetric errors always correspond to gaussian standard deviations
- ▶ proper combination of errors needs the full likelihood functions as input

even when a rigorous solution is not possible the above method is defensible →

- correct solution when the full likelihood functions are known
 - ▶ each uncertainty corresponds to parameter deviation from a best fit value
 - ▶ combination of two deviations x_1 and x_2 into a u :
maximize the combined likelihood subject to the constraint $u = x_1 + x_2$

$$\ln L = \ln L_1(x_1) + \ln L_2(x_2) - \lambda(u - x_1 - x_2) \stackrel{!}{=} \max$$

- ▶ can be done if $L_1(x_1)$ and $L_2(x_2)$ are known
- ad-hoc assumptions are needed when only the asymmetric errors are given
 - ▶ simplest case, with asymmetric errors $+\sigma, -\eta$

$$\ln L(x) = -\frac{x^2}{2} \begin{cases} 1/\sigma^2 & \text{for } x \geq 0 \\ 1/\eta^2 & \text{for } x < 0 \end{cases}$$

- solution

$$\ln L(u) = -\frac{u^2}{2} \begin{cases} 1/(\sigma_1^2 + \sigma_2^2) & \text{for } u \geq 0 \\ 1/(\eta_1^2 + \eta_2^2) & \text{for } u < 0 \end{cases}$$

- ▶ upper and lower uncertainties are independently added in quadrature