Statistical Methods in Particle Physics

5. Maximum Likelihood Estimation

Heidelberg University, WS 2020/21

Klaus Reygers (lectures)
Rainer Stamen, Martin Völkl (tutorials)

Estimator

Suppose we have a measurement of *n* independent values

$$\vec{x} = (x_1, x_2, ..., x_n)$$

which follow the same underlying distribution f(*x*; θ),
e.g., f(*x*; θ) = 1/θ exp(–*x*/θ).

i.i.d. random variables = independent, identically distributed

An estimator is a function of the data which provides a numerical estimate of the parameter θ:

$$\hat{\theta}(\vec{x})$$

θ often is not only one parameter but a vector of parameters.

Properties of estimators

Consistency

An estimator is consistent if it converges to the true value

$$\lim_{n\to\infty} \hat{\vec{\theta}} = \vec{\theta}$$

Bias

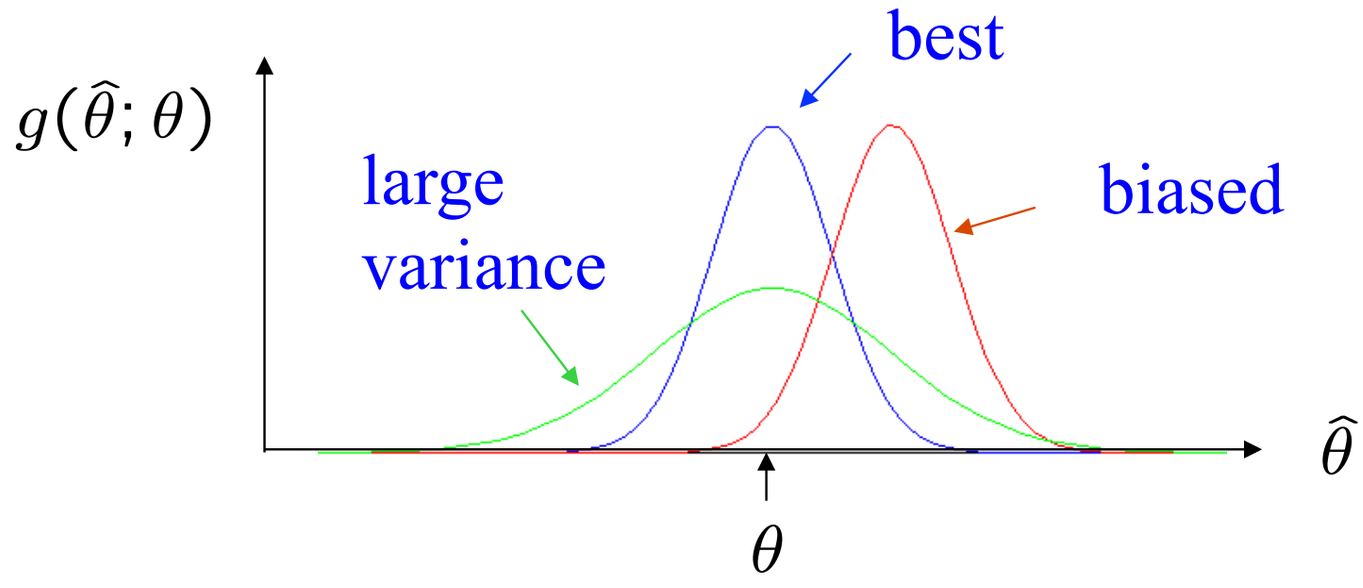
Difference btw. expectation value of estimator and true value

$$\vec{b} := E[\hat{\vec{\theta}}] - \vec{\theta}$$

Efficiency

An estimator is efficient if its variance V[θ] is small

efficient ⇔ Equal-sign in Cramér–Rao inequality holds



http://www.pp.rhul.ac.uk/~cowan/stat_course.html

Example: Estimators for the lifetime of a particle

Estimator	Consistent?	Unbiased?	Efficient?
$\hat{\tau} = \frac{t_1+t_2+...+t_n}{n}$	yes	yes	yes
$\hat{\tau} = \frac{t_1+t_2+...+t_n}{n-1}$	yes	no	no
$\hat{\tau} = t_1$	no	yes	no

http://www.terascale.de/e149980/index_eng.html

Unbiased estimators for mean and variance

Consider n independent and identically distributed measurements x_i drawn from a distribution with mean μ and standard deviation σ:

Estimator for the mean: $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mathrm{E}[\hat{\mu}] = \frac{1}{n}\,\mathrm{E}[\sum_i x_i] = \frac{1}{n}\sum_i \mathrm{E}[x_i] = \mu \quad \rightarrow \text{ estimator is unbiased}$$

$$\mathrm{V}[\hat{\mu}] = \mathrm{V}[\frac{1}{n}\sum_i x_i] = \frac{1}{n^2}\,\mathrm{V}[\sum_i x_i] = \frac{1}{n}\,\mathrm{V}[x] = \frac{\sigma^2}{n}, \text{ i.e., } \sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

Unbiased estimator for the variance:

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

Unbiased estimator of the variance: Derivation (1)

Consider n independent and identically distributed random variables x_i:

$$\mu := E[x_i], \quad \sigma^2 := V[x_i], \quad \bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

We'll use:

$$\sigma^2 = E[x_i^2] - \mu^2 \quad \rightsquigarrow \quad E[x_i^2] = \mu^2 + \sigma^2$$

$$V[\bar{x}] = \frac{1}{n^2} V[\sum_{i=1}^{n} x_i] = \frac{1}{n} V[x_i] = \frac{\sigma^2}{n} \stackrel{!}{=} E[\bar{x}^2] - \mu^2 \quad \rightsquigarrow \quad E[\bar{x}^2] = \frac{\sigma^2}{n} + \mu^2$$

Now we calculate the expectation value of $\sum_{i=1}^{n}(x_i - \bar{x})^2$:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2x_i\bar{x} + \bar{x}^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2$$

$$E[\sum_{i=1}^{n}(x_i - \bar{x})^2] = E[\sum_{i=1}^{n} x_i^2] - E[n\bar{x}^2] = n(\mu^2 + \sigma^2) - \sigma^2 - n\mu^2 = (n-1)\sigma^2$$

Unbiased estimator of the variance: Derivation (2)

This means that

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

is an unbiased estimator of the variance, i.e., $E[s^2] = \sigma^2$.

Multiplying the sample variance by $n/(n–1)$ is known as Bessel's correction.

Note that s is not an unbiased estimator of the standard deviation:

https://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation

Unbiased estimator for the standard deviation for the normal distribution ($E[\hat{\sigma}] = \sigma$):

$$\hat{\sigma} = c_4(n)\sqrt{s^2}, \quad c_4(n) = \sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} = 1 - \frac{1}{4n} - \frac{7}{32n^2} + \ldots,$$

Rule of thumb:

$$\hat{\sigma} \approx \sqrt{\frac{1}{n-1.5}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Likelihood function and maximum likelihood

Suppose we have a measurement of n independent values

$$\vec{x} = (x_1, x_2, ..., x_n)$$

drawn from the distribution

$$f(x; \vec{\theta}), \quad \vec{\theta} = (\theta_1, \theta_2, ..., \theta_m)$$

The joint pdf for the observed values $\vec{x}$ is given by:

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad \text{"likelihood function"}$$

We consider the measured values as fixed and the parameters as variables.

Principle of maximum likelihood

The best estimate of the parameters $\vec{\theta}$ is that value which maximizes the likelihood function

Likelihood function is not a probability density function

The integral of $L(\vec{x}, \vec{\theta})$ with respect to the parameter is not necessarily equal to unity ($L(\vec{x}, \vec{\theta})$ might not be integrable at all).

This is why $L(\vec{x}, \vec{\theta})$ is not a probability density function.

Example: exponential decay, one measurement at t = 1h.

$$L(\tau) = \frac{1}{\tau} e^{-t/\tau} \approx \frac{1}{\tau} \quad \text{as} \quad \tau \to \infty, \qquad \int_0^\infty L(\tau)\, d\tau \quad \text{not defined}$$

Note: With Jeffreys' prior 1/τ the posterior $L(\tau)\ \pi(\tau)$ is normalizable.

Maximum likelihood example 1: Exponential Decay

Consider exponential pdf: $$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

Independent measurements drawn from this distribution: $t_1, t_2, \ldots, t_n$

Likelihood function: $$L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$$

$L(\tau)$ is maximum when $\ln L(\tau)$ is maximum:

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

Find maximum:

$$\frac{\partial \ln L(\tau)}{\partial \tau} = 0 \quad \rightsquigarrow \quad \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \quad \rightsquigarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Maximum likelihood example 2: Gaussian (I)

Consider $x_1, x_2, \ldots, x_n$ drawn from Gaussian(μ, σ^2)

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Log-likelihood function:

$$\ln L(\mu,\sigma^2) = \sum_{i=1}^{n} \ln f(x_i;\mu,\sigma^2) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

Derivatives w.r.t. μ and σ^2:

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2}$$

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)$$

Maximum likelihood example 2: Gaussian (II)

Setting the derivatives w.r.t. μ and σ^2 to zero and solving the equations:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

We find that the ML estimator for σ^2 is biased!

Maximum likelihood uncertainty

Consider maximum likelihood estimate of a parameter θ. Methods to estimate Uncertainty of θ:

1. $\sigma_{\hat{\theta}}$ **from Monte Carlo**
   Generate pseudo-data by sampling the assumed distribution using the ML estimate $\hat{\theta}$ as parameter

2. **Use minimum variance bound**

$$\sigma_{\hat{\theta}} = \frac{1}{\sqrt{-\frac{\partial^2}{\partial^2 \theta} \ln L(\theta)}}$$

3. $\Delta \ln L = -1/2$ **method**:

$$\ln L(\hat{\theta} \pm \sigma) = \ln L(\hat{\theta}) - \frac{1}{2}$$

For a Gaussian likelihood function all methods agree.

Method 3 usually gives asymmetric uncertainties (which are messy).

Likelihood function and minimum variance bound

Let's first consider a likelihood function with only one parameter:

$$L(\vec{x};\theta) = L(x_1, x_2, ..., x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Let $\hat{\theta}(\vec{x})$ be an unbiased estimator of the parameter θ

It can be shown that the variance (of any unbiased estimator) satisfies:

$$V[\hat{\theta}] \geq \frac{1}{E\left[-\frac{\partial^2 \ln L}{\partial^2 \theta}\right]}$$

For a biased estimator this becomes

$$V[\hat{\theta}] \geq \frac{\left(1+\frac{\partial b}{\partial \theta}\right)^2}{E\left[-\frac{\partial^2 \ln L}{\partial^2 \theta}\right]}$$

This bound is called Rao-Cramér-Frechet minimum variance bound (MVB)

MVB example: Exponential decay

Reminder:

$$\frac{\partial \ln L(\tau)}{\partial \tau} = 0 \quad \rightsquigarrow \quad \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \quad \rightsquigarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Minimum variance bound (MVB):

$$\frac{\partial^2 \ln L(\tau)}{\partial^2 \tau} = \sum_{i=1}^{n} \left(\frac{1}{\tau^2} - 2\frac{t_i}{\tau^3} \right) = \frac{n}{\tau^2} - \frac{2}{\tau^3} \sum_{i=1}^{n} t_i = \frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right)$$

$$V[\hat{\tau}] \geq \frac{1}{E\left[-\frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right) \right]} = \frac{1}{-\frac{n}{\tau^2} \left(1 - \frac{2E[\hat{\tau}]}{\tau} \right)} = \frac{\tau^2}{n}$$

Uncertainty of the ML estimator: Approximating the minimum variance bound

In many cases it is impractical to calculate the MVB analytically. Instead, one uses the following approximation which is good for large n:

$$E\left[-\frac{\partial^2 \ln L}{\partial^2 \theta}\right] \approx -\frac{\partial^2 \ln L}{\partial^2 \theta}\bigg|_{\theta=\hat{\theta}}$$

The variance of the ML estimator is given by:

$$V[\hat{\theta}] = -\frac{1}{\frac{\partial^2 \ln L}{\partial^2 \theta}\Big|_{\theta=\hat{\theta}}}$$

Example: Exponential decay

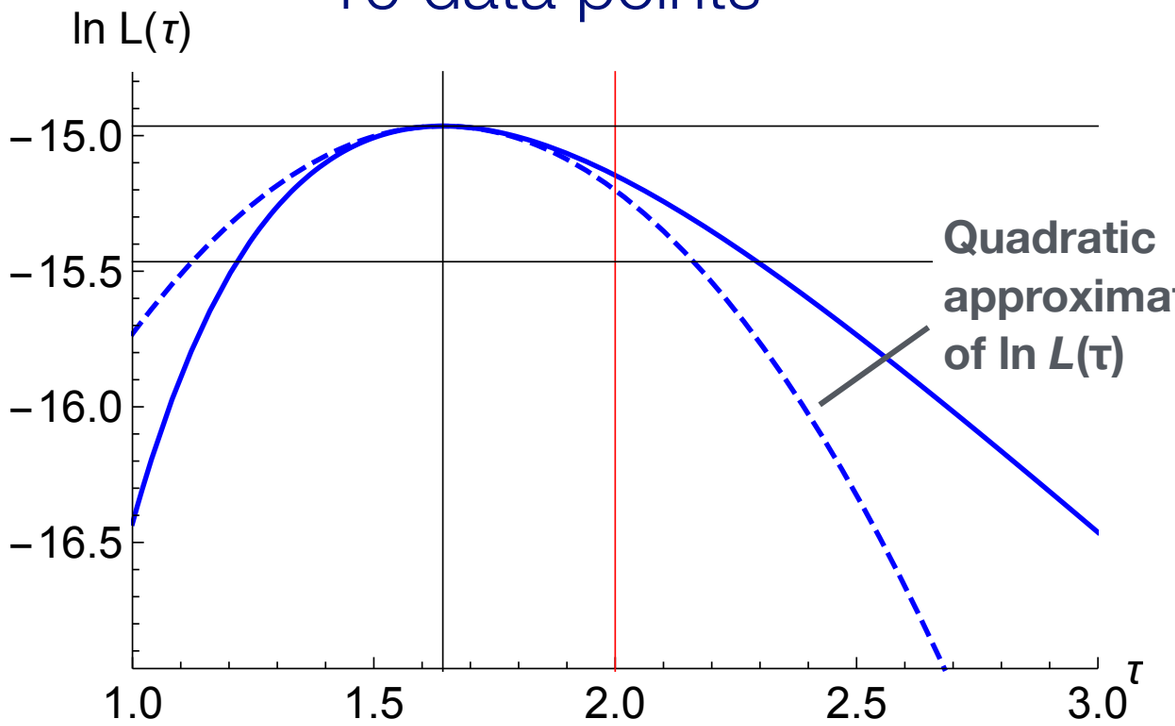
$$\frac{\partial^2 \ln L(\tau)}{\partial^2 \tau} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - 2\frac{t_i}{\tau^3}\right) = \frac{n}{\tau^2} - \frac{2}{\tau^3}\sum_{i=1}^{n} t_i = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

$$V[\hat{\tau}] = -\left(\frac{\partial^2 \ln L}{\partial^2 \theta}\right)^{-1}_{\tau=\hat{\tau}} = \frac{\hat{\tau}^2}{n}$$

Asymptotic normality of the likelihood function

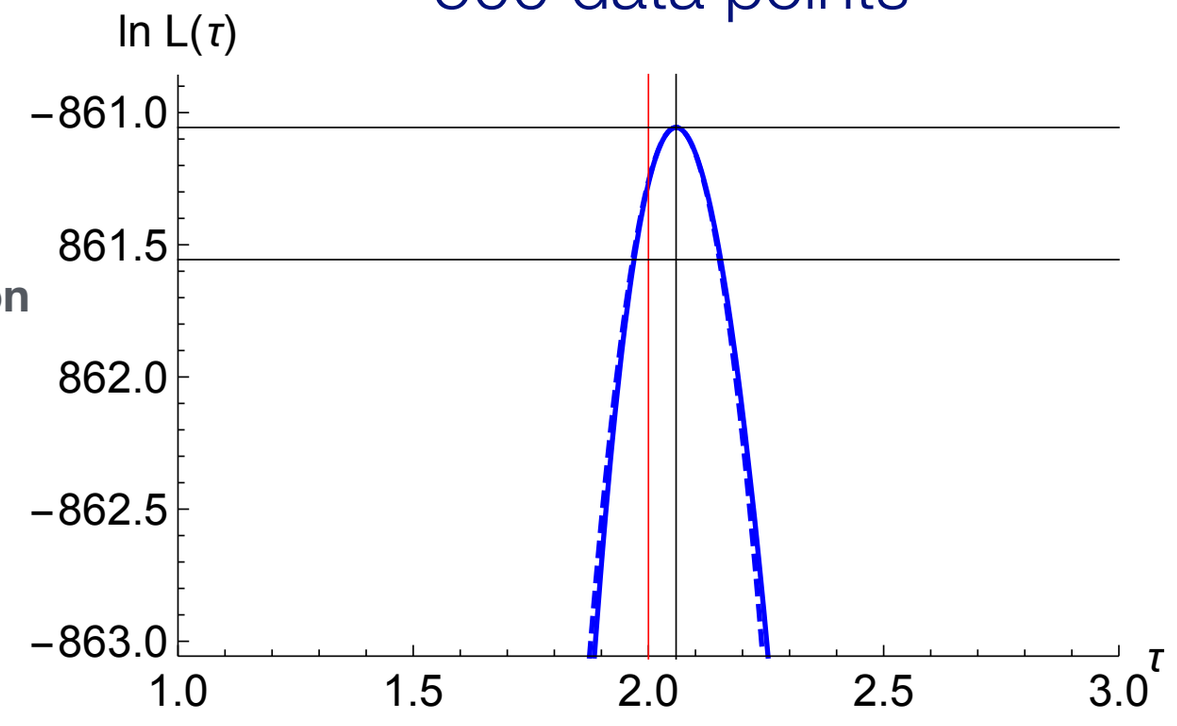
For any probability function $f(x;\theta)$ the likelihood function *L* approaches a Gaussian for large *n*, i.e., for a large number of events, and the variance of the ML estimator reaches the minimum variance bound.

10 data points



quadratic approximation of ln *L*(τ) is not very good

500 data points



quadratic approximation of ln *L*(τ) is excellent

Data points sampled from $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$ with $\tau = 2$

Uncertainty of the ML estimator: $\Delta \ln L = -1/2$ method

Taylor expansion of ln L around the maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \underbrace{\left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})}_{=0} + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial^2 \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

$\left[\frac{\partial^2 \ln L}{\partial^2 \theta}\right]_{\theta=\hat{\theta}} \approx -\frac{1}{\sigma^2}$ [from MVB, or from assuming Gaussian shape]

If $L(\theta)$ is approximately Gaussian ($\ln L(\theta)$ then is a approximately a parabola):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2_{\hat{\theta}}}}$$

good approximation in the large sample limit

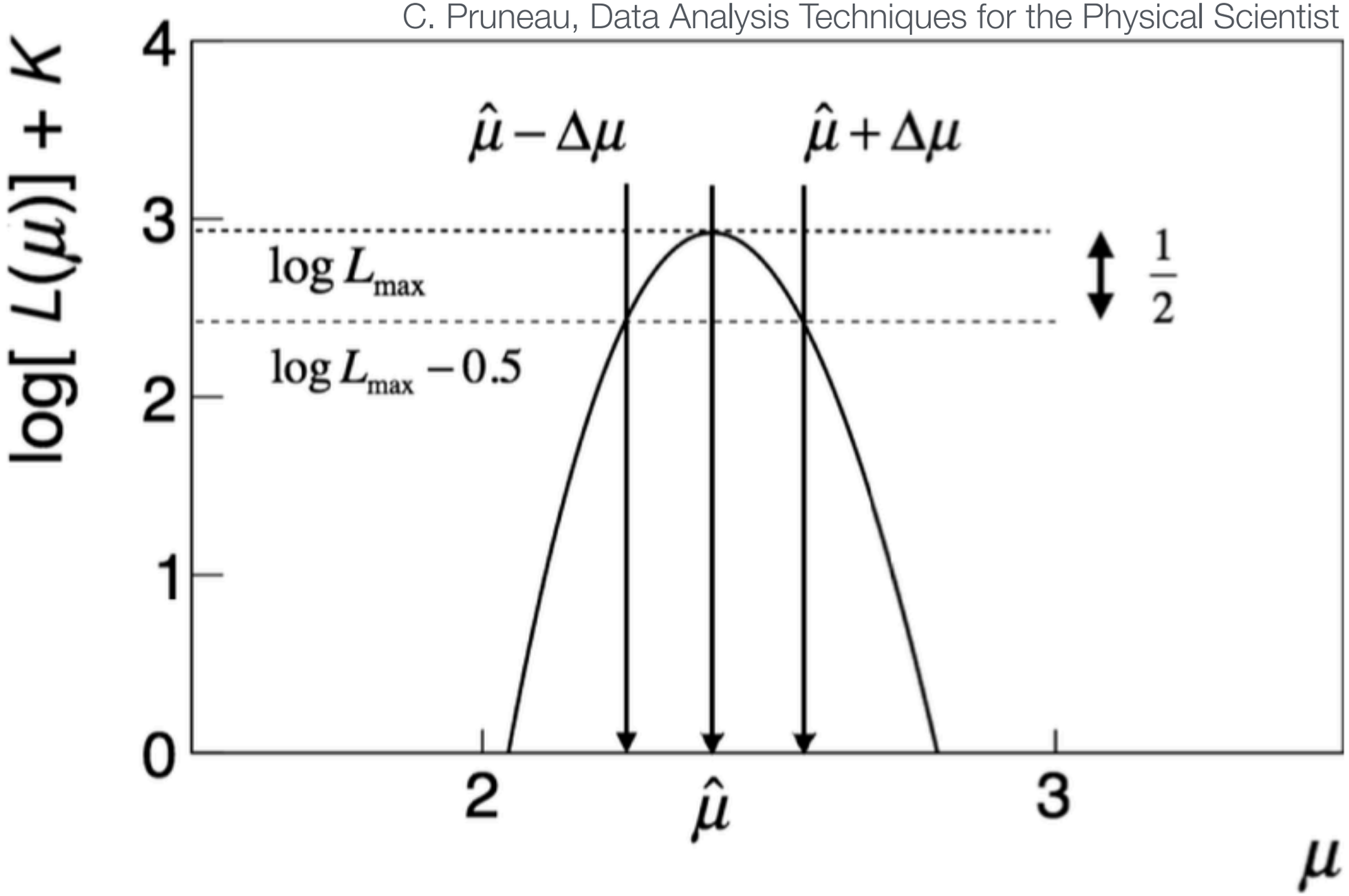
One can then estimate the uncertainties from the points where ln L has dropped by 1/2 from its maximum:

$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

Illustration of the $\Delta \ln L = -1/2$ method

L is Gaussian $\longleftrightarrow$ $\ln L$ is a parabola

C. Pruneau, Data Analysis Techniques for the Physical Scientist



Properties of the ML estimator

The ML estimator is consistent,
i.e., it approaches the true value in the limit of infinite measurements ($n \to \infty$)

ML estimator efficient for large n (you get the smallest possible variance)

For finite n the ML estimator is in general biased

ML efficiency theorem:
the ML estimator will be unbiased and efficient if an unbiased efficient estimator exists

The ML Estimator is invariant under parameter transformation:

$$\psi = g(\theta) \quad \Rightarrow \quad \hat{\psi} = g(\hat{\theta})$$

ML does not provide a goodness-of-fit measure.

Averaging measurements with Gaussian uncertainties

pdf for measurement (same mean, different σ):

$$f(x;\mu,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}} \qquad \ln L(\mu) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma_i - \frac{(x_i-\mu)^2}{2\sigma_i^2} \right)$$

Weighted average = ML estimate

$$\left.\frac{\partial \ln L(\mu)}{\partial \mu}\right|_{\mu=\hat{\mu}} = \sum_{i=1}^{n} \frac{x_i - \hat{\mu}}{\sigma_i^2} \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

Uncertainty? In this case L is Gaussian and we can write it as

$$L(\mu) \propto e^{-\frac{(\mu-\hat{\mu})^2}{2\sigma_{\hat{\mu}}^2}} \quad \text{with} \quad \sigma_{\hat{\mu}}^2 = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}$$

We obtain the formula for the weighted average:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \quad \pm \quad \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}}$$

Minimum variance bound for *m* parameters

$$f(x;\vec{\theta}), \quad \vec{\theta} = (\theta_1, \theta_2, ..., \theta_m)$$

Fisher information matrix $I(\vec{\theta})$ ($m \times m$ matrix):

$$I_{jk}[\vec{\theta}] = -E\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} \ln L(x, \vec{\theta})\right]$$

Covariance matrix of the parameters: $V_{ij} := \mathrm{cov}[\theta_i, \theta_j]$

Cramér-Rao-Frechet bound for an unbiased estimator then states that $V - I^{-1}$ is a positive-semidefinite matrix.

In particular one obtained for the variance:

$$V[\hat{\theta}_j] \geq (I(\vec{\theta})^{-1})_{jj}$$

Variance of the ML estimator for *m* parameters

For any probability function $f(x;\vec{\theta})$ the likelihood function L approaches a multi-variate Gaussian for large n

$$L(\vec{\theta}) \propto e^{-\frac{1}{2}(\vec{\theta}-\widehat{\vec{\theta}})^{\mathsf{T}}\, V^{-1}[\widehat{\vec{\theta}}]\,(\vec{\theta}-\widehat{\vec{\theta}})}$$

The variance of the ML estimator then reaches the MVB:

$$V[\widehat{\vec{\theta}}] \to I(\vec{\theta})^{-1}$$

Covariance matrix of the estimated parameters:

$$V[\widehat{\vec{\theta}}] \approx \left[-\left. \frac{\partial^2 \ln L(\vec{x};\vec{\theta})}{\partial^2 \vec{\theta}} \right|_{\vec{\theta}=\widehat{\vec{\theta}}} \right]^{-1}$$

or equivalently:

$$(V^{-1}[\widehat{\vec{\theta}}])_{ij} = -\left. \frac{\partial^2 \ln L(\vec{x};\vec{\theta})}{\partial\theta_i \partial\theta_j} \right|_{\vec{\theta}=\widehat{\vec{\theta}}}$$

Standard deviation of a single parameters:

$$\hat{\sigma}_{\hat{\theta}_j} = \sqrt{(V[\widehat{\vec{\theta}}])_{jj}}$$

Example: Two-parameter ML fit (from Cowan's book)

Scattering angle distribution, $x = \cos\theta$: $$f(x; a, b) = \frac{1 + ax + bx^2}{2 + 2b/3}$$

Normalization: $$\int_{x_{\text{min}}}^{x_{\text{max}}} f(x; a, b)\, \mathrm{d}x = 1$$

Example: $a = 0.5,\ b = 0.5;\ x_{\text{min}} = -0.95,\ x_{\text{max}} = 0.95,\ 1000\,\text{MC events}$

Numerical minimization with MINUIT:

[link to jupyter notebook]

$$\hat{a} = 0.53 \pm 0.08$$

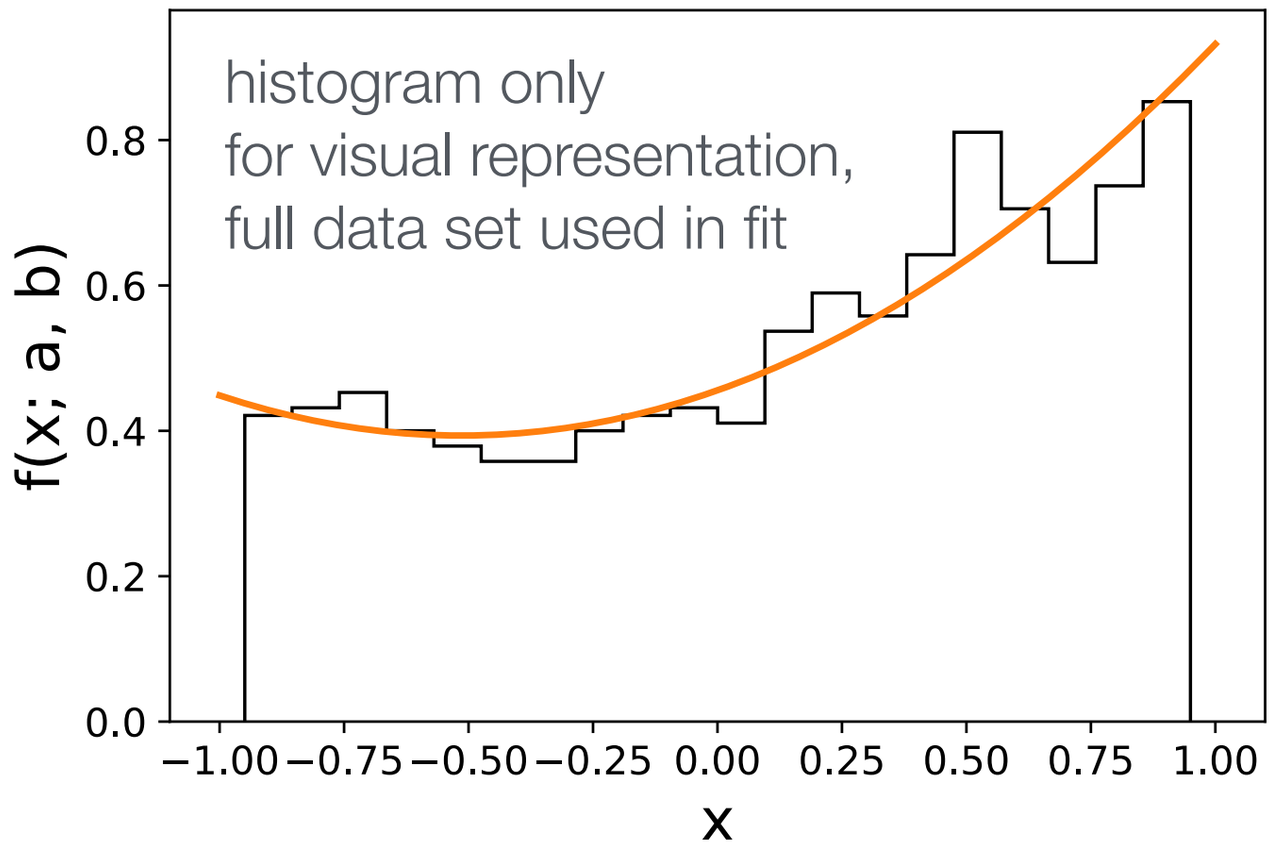
$$\hat{b} = 0.51 \pm 0.16$$

$$\text{cov}[\hat{a}, \hat{b}] = 0.006$$

$$\rho = 0.48$$

Uncertainties and covariance from inverse of Hessian matrix H:

$$\widehat{V} = -H^{-1},\ (H)_{ij} = \left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$$



Example: Two-parameter ML fit (iminuit)

```
import numpy as np
import matplotlib.pyplot as plt
from iminuit import Minuit
```

```
x = np.loadtxt("data.txt")
```

```
def f(x, a, b):
    """normalized fit function"""
    xmin = -0.95
    xmax = 0.95
    return (6 * (1 + a * x + b * x * x)) /
           ((xmax - xmin) * (3 * a * (xmax + xmin) +
           2 * (3 + b * (xmax * xmax + xmax * xmin + xmin * xmin))))
```

```
def negative_log_likelihood(a, b):
    p = np.log(f(x, a, b))
    return -np.sum(p)
```

iminuit uses introspection to detect the parameter names of your function

```
m = Minuit(negative_log_likelihood,
           a=1, b=1, error_a=0.01, error_b=0.01, errordef=Minuit.LIKELIHOOD)
```

```
m.migrad()
```

Example: Two-Parameter ML Fit (iminuit)

```
m.migrad()
```

FCN = 606.5			Ncalls = 10 (146 total)	
EDM = 1.33e-10 (Goal: 0.0001)			up = 0.5	
Valid Min.	Valid Param.	Above EDM	Reached call limit	
True	True	False	False	
Hesse failed	Has cov.	Accurate	Pos. def.	Forced
False	True	True	True	False

	Name	Value	Hesse Error	Minos Error-	Minos Error+	Limit-	Limit+	Fixed
0	a	0.53	0.08					
1	b	0.51	0.16					

https://iminuit.readthedocs.io/en/stable/

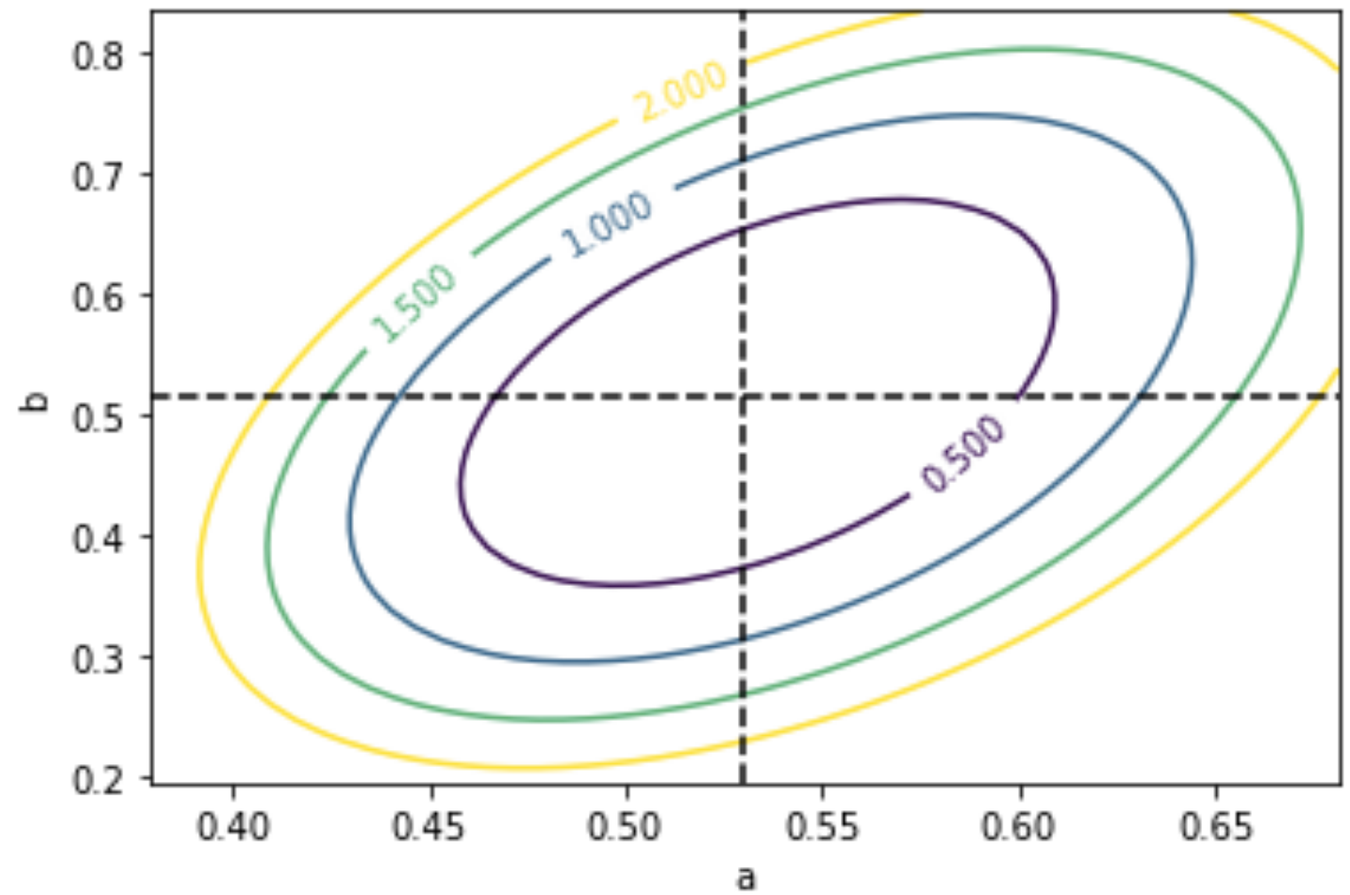
https://nbviewer.jupyter.org/github/scikit-hep/iminuit/blob/master/tutorial/basic_tutorial.ipynb

Example: Two-Parameter ML Fit (iminuit)

```
# covariance matrix
m.matrix()
```

	a	b
a	0.006	0.006
b	0.006	0.026

```
m.draw_contour('a','b');
```



Extended maximum likelihood method (1)

Standard ML fit: information is in the shape of the distribution of the data x_i.

Extended ML fit: normalization becomes a fit parameter

Sometimes the number of observed events contains information about the parameters of interest, e.g., when we measure a rate.

Normal ML method:

$$\int f(x, \vec{\theta})\, \mathrm{d}x = 1$$

Extended ML method:

$$\int q(x, \vec{\theta})\, \mathrm{d}x = \nu(\vec{\theta}) = \text{ predicted number of events}$$

Extended maximum likelihood method (2)

Normalized pdf:

$$\int f(x, \vec{\theta})\, \mathrm{d}x = 1$$

Likelihood function:

$$L(\vec{\theta}) = \frac{\nu^n e^{-\nu}}{n!} \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad \text{where} \;\; \nu \equiv \nu(\vec{\theta})$$

Log-Likelihood function:

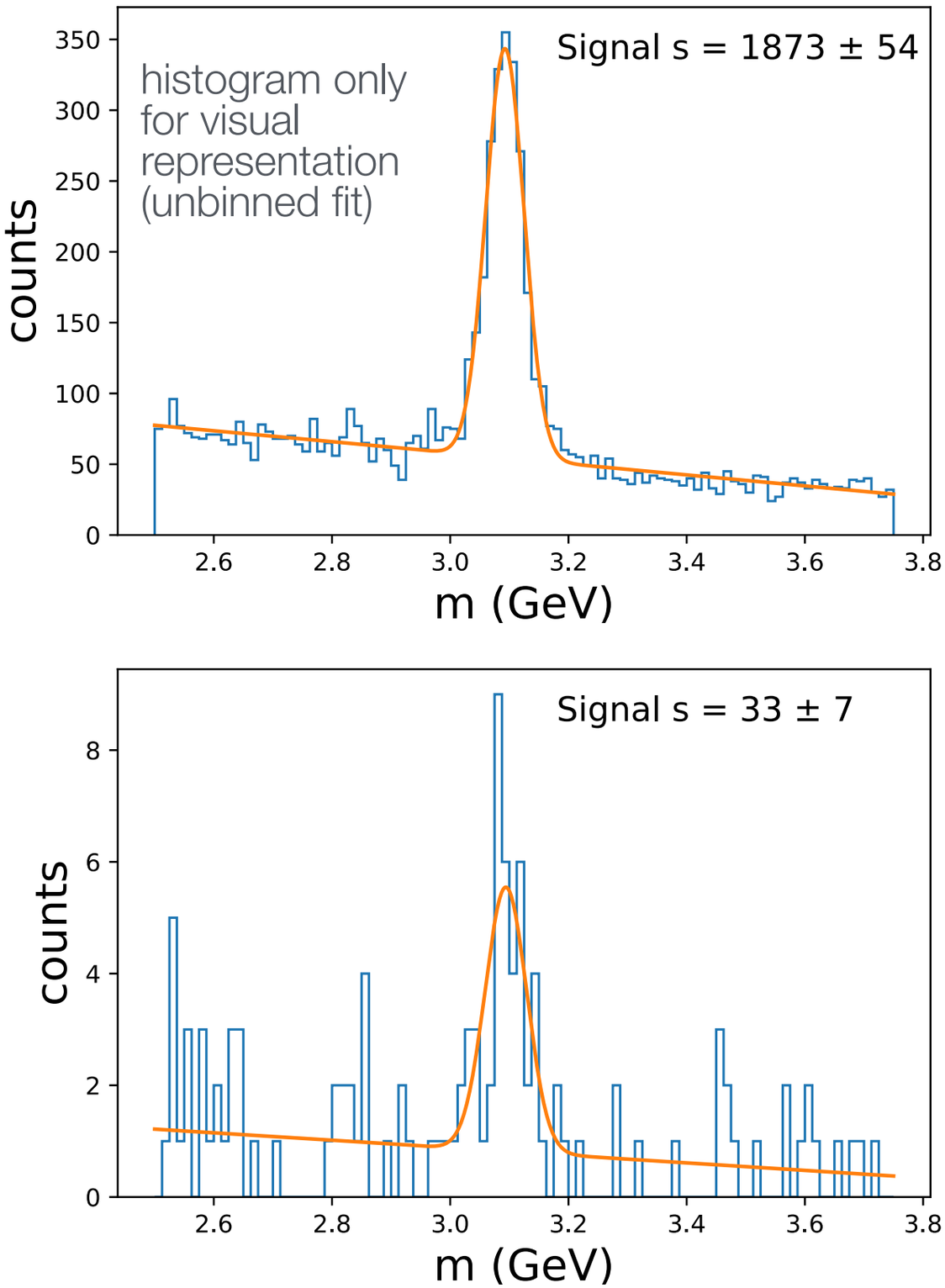
$$\ln L(\vec{\theta}) = -\ln(n!) - \nu(\vec{\theta}) + \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})\nu(\vec{\theta})]$$

ln(n!) does not depend on the parameters. So we need to minimize:

$$-\ln \tilde{L}(\vec{\theta}) = \nu(\vec{\theta}) - \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})\nu(\vec{\theta})]$$

prediction for total number of events

Application of the extended ML method: Linear combination of signal and background PDF (1)



Two-component fit (signal + linear background)

Parameters:
- signal counts *s*
- background counts *b*
- linear background (slope, intercept)
- Gaussian peak: μ, σ

Normalized pdf:

$$f(x; r, \vec{\theta}) = r\, f_s(x, \vec{\theta}) + (1 - r)\, f_b(x, \vec{\theta})$$

negative log-likelihood:

$$-\ln \tilde{L}(\vec{\theta}) = s + b - n \ln(s + b) - \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})]$$

$$\nu(s, b) = s + b, \quad r = \frac{s}{s + b}$$

Unbinned ML fit works fine also in case of low statistics

Application of the extended ML method: Linear combination of signal and background PDF (2)

Discussion:

We could have just fitted the normalized pdf:

$$f(x; r_s, \vec{\theta}) = r\, f_s(x, \vec{\theta}) + (1 - r)\, f_b(x, \vec{\theta})$$

Good estimate of the number of signal events: $n_{\text{signal}} = r\, n$

However, $\sigma_r\, n$ is not a good estimate of the variation of the number of signal events (ignores fluctuations of n)

[C. Blocker, Maximum Likelihood Primer]

(Trivial) example (L. Lyons):
96 protons and 4 heavy nuclei have been measured in a cosmic ray experiment

	protons	heavy nuclei
ML estimate	96 ± 2	4 ± 2
Extended ML estimate	96 ± 10	4 ± 2

Maximum likelihood fits with binned data (1)

Common practice: data put into a histogram: $\vec{n} = (n_1, ..., n_k),\ n_{\text{tot}} = \sum_{i=1}^{k} n_i$

Model prediction for the expected counts in bin i for fixed n_{tot}:

$$\nu_i(\vec{\theta}) = n_{\text{tot}} \int_{\text{bin } i} f(x; \vec{\theta})\, \mathrm{d}x \qquad \nu_i(\vec{\theta}) = (\nu_1, ..., \nu_k)$$

If n_{tot} is fixed the probability to get a certain $\vec{n}$ is given by the multinomial distribution.

Multinomial distribution (generalization of binomial distribution):

→ k different possible outcomes, probability for outcome i is p_i, $\sum_{i=1}^{k} p_i = 1$

$$f(\vec{n}; n_{\text{tot}}, \vec{p}) = \frac{n_{\text{tot}}!}{n_1! \cdot ... \cdot n_k!}\, p_1^{n_1} \cdot ... \cdot p_k^{n_k} \qquad \vec{p} = (p_1, ..., p_k)$$

Maximum likelihood fits with binned Data (2)

With $p_i = \nu_i / n_{\text{tot}}$ we write the likelihood of a certain $n_1, \ldots, n_k$ outcome as:

$$L(\vec{\theta}) = \frac{n_{\text{tot}}!}{n_1! \cdot \ldots \cdot n_k!} \left(\frac{\nu_1}{n_{\text{tot}}}\right)^{n_1} \cdot \ldots \cdot \left(\frac{\nu_k}{n_{\text{tot}}}\right)^{n_k} \qquad \nu_i(\vec{\theta}) = (\nu_1, \ldots, \nu_k)$$

Log-likelihood function:

$$\ln L(\vec{\theta}) = \sum_{i=1}^{k} n_i \ln \nu_i(\vec{\theta}) + C$$

Limit of zero bin width → usual unbinned maximum likelihood method

Maximum likelihood fits with binned Data (3)

Extended log-likelihood fit for binned data:

n_{tot} fluctuates, predicted average: ν_{tot}

$$\nu_{\text{tot}} = \sum_{i=1}^{k} \nu_i, \qquad n_{\text{tot}} = \sum_{i=1}^{k} n_i$$

Likelihood function:

$$L(\vec{\theta}) = \frac{\nu_{\text{tot}}^{n_{\text{tot}}}}{n_{\text{tot}}!} e^{-\nu_{\text{tot}}} \frac{n_{\text{tot}}!}{n_1! \cdot \ldots \cdot n_k!} \left(\frac{\nu_1}{\nu_{\text{tot}}}\right)^{n_1} \cdot \ldots \cdot \left(\frac{\nu_k}{\nu_{\text{tot}}}\right)^{n_k}$$

$$= \prod_{i=1}^{k} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Function that needs to be maximized (dropping terms that do not depend on the parameters):

$$\ln L(\vec{\theta}) = \sum_{i=1}^{k} n_i \ln \nu_i - \nu_i = -\nu_{\text{tot}} + \sum_{i=1}^{k} n_i \ln \nu_i, \qquad \nu_i(\vec{\theta}) = (\nu_1, \ldots, \nu_k)$$

Relation to Bayesian parameter estimation

Bayesian posterior distribution:

$$p(\vec{\theta};\vec{x}) = \frac{L(\vec{x};\vec{\theta})\pi(\vec{\theta})}{\int L(\vec{x};\vec{\theta})\pi(\vec{\theta})\,\mathrm{d}\vec{\theta}}$$

Posterior distribution contains all information about the estimated parameters.

Often the mode (most probable value) of the posterior distribution is reported

→ Coincides with ML estimate for a flat prior distribution

Marginalization in case one is interested in only one parameter of the Bayesian posterior distribution:

$$p(\theta_j;\vec{x}) = \int p(\vec{\theta};\vec{x})\,\mathrm{d}\vec{\theta}_{k\neq j} = \frac{\int L(\vec{x};\vec{\theta})\pi(\vec{\theta})\,\mathrm{d}\vec{\theta}_{k\neq j}}{\int L(\vec{x};\vec{\theta})\pi(\vec{\theta})\,\mathrm{d}\vec{\theta}}$$

Example of a frequentist approach to systematic uncertainties: Profile method

Uncertainty in the probability function for the data described by nuisance parameter ν:

$$L(\theta, \nu) = \prod_i p(x_i|\theta, \nu)$$

If available, can include information on ν from additional measurements y_i:

$$L(\theta, \nu) = \prod_{i,j} p(x_i, y_j|\theta, \nu)$$

Eliminate the nuisance parameter by using the profile likelihood:

$$L_\mathrm{p}(\theta) = L(\theta, \hat{\hat{\nu}}(\theta))$$

$\hat{\hat{\nu}}(\theta)$: value of ν which maximizes $L(\theta, \nu)$ for a given θ