H. The maximum likelihood method

V. Blobel
Universität Hamburg

April 2001

1. The maximum-likelihood principle

A standard data analysis problem:

A measurement is performed in the space of the random variable x. The distribution of the measured values x is assumed to be known to follow the (normalized) *probability density* $p(x;a)$

$$p(x;a) \geq 0 \qquad \text{with} \qquad \int_\Omega p(x;a)\,\mathrm{d}x = 1$$

in the x-space, which depends on a single parameter a.

From a given set of n measured values $x_1, \ldots, x_i, \ldots, x_n$ the optimal value of the parameter a has to be estimated.

The *maximum-likelihood method* starts from the *joint* probability distribution of the n measured values $x_1, \ldots, x_i, \ldots, x_n$, which for *independent* measurements is given by the product of the individual densities $p(x|a)$, which is

$$\mathcal{L}(a) = p(x_1|a) \cdot p(x_2|a) \cdots p(x_n|a) = \prod_{i=1}^{n} p(x_i|a) \; .$$

The function $\mathcal{L}(a)$, for a given set $\{x_i\}$ of measurements considered as a function of the parameter a, is called the *likelihood function* and is a *function* (not a probability density) of the parameter a.

Principle of Maximum Likelihood

The estimate $\widehat{a}$ for the parameters a is the value, which *maximizes* the likelihood function $\mathcal{L}(x|a)$.

For technical and also for theoretical reasons it is easier to work with the logarithm (a monotonically increasing function of its argument) of the likelihood function $\mathcal{L}(\boldsymbol{a})$, or with the *negative* logarithm. In the following the *negative* log-likelihood function[a] is considered,

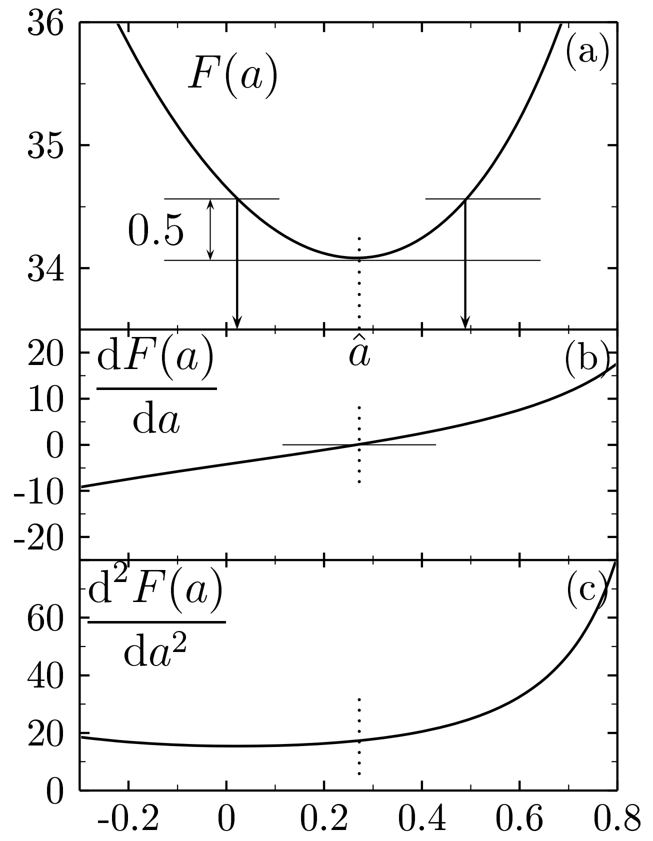
$$\boxed{F(a) = -\sum_{i=1}^{n} \ln p(x_i|a)}$$

and the maximum likelihood estimate $\widehat{a}$ is the value that *minimizes* this function.

Likelihood equation $$\frac{\mathrm{d}F(a)}{\mathrm{d}a} = 0$$

[a] Sometimes a factor of 2 is included in the definition of the negative log-likehood function; this factor makes it similar to the χ^2-expression of the method of least squares in certain applications.

Example



2. Properties of the maximum-likelihood estimates

Maximum-likelihood estimates $\widehat{a}$

Consistency: The estimate $\widehat{a}$ of the MLM is asymptotically ($n \to \infty$) consistent. For finite values of n there may be a bias $B(\widehat{a}) \propto 1/n$.

Normality: The estimate $\widehat{a}$ is, under very general conditions, asymptotical normally distributed with minimal variance $V(\widehat{a})$.

Invariance: The maximum likelihood solution is invariant under change of parameter – the estimate $\widehat{b}$ of a function $b = b(a)$ is given by $\widehat{b} = b(\widehat{a})$. The bias $B(\widehat{a})$ for finite n may be different for different functions of the parameter.

Efficiency: If efficient estimators exist for a given problem the maximum likelihood method will find them.

Normality

Normality: The estimate $\widehat{a}$ is, under very general conditions, asymptotical normally distributed with minimal variance $V(\widehat{a})$, i.e.

$$\lim_{n\to\infty} V\left[\widehat{a}\right] = \frac{1}{n}\left\{ E\left[\frac{\partial \ln p}{\partial a}\right]^2 \right\}^{-1} .$$

Asymptically the likelihood equation becomes a function, which is *linear* in the parameter a (constant second derivative).

Calculation of variance and covariance matrix in practice:

$$V\left[\widehat{a}\right] = \left(\left.\frac{\mathrm{d}^2 F}{\mathrm{d}a^2}\right|_{a=\widehat{a}} \right)^{-1} \qquad \boldsymbol{V}\left[\widehat{\boldsymbol{a}}\right] = \boldsymbol{H} \quad \text{with} \quad H_{jk} = \frac{\partial^2 F}{\partial a_j \partial a_k}$$

Invariance

The Likelihood function is a function of $\boldsymbol{a}$, and is not a probability density of $\boldsymbol{a}$.

Invariance: The maximum likelihood solution is invariant under change of parameter – the estimate $\widehat{b}$ of a function $b = b(a)$ is given by $\widehat{b} = b(\widehat{a})$. The bias $B(\widehat{a})$ for finite n may be different for different functions of the parameter.

Example: Parameter of an exponential distribution

Exponential probability density distribution for the decay of unstable particles

$$p(t;\tau) = \frac{1}{\tau}e^{-t/\tau} \qquad \text{or} \qquad p(t;\lambda) = \lambda e^{-\lambda\, t}$$

dependent on the mean lifetime τ or the decay constant $\lambda = 1/\tau$.

Bayesian parameter estimation

Both the data $\boldsymbol{x}$ and the parameters $\boldsymbol{a}$ are random variables.

- Before the experiment the knowledge about $\boldsymbol{a}$ is summarized by $\pi(\boldsymbol{a})$ – called prior density, for example $\pi(\boldsymbol{a}) =$ const.. $\pi(\boldsymbol{a}) = 0$ outside the physical region.
- Bayes theorem is used to update the prior using the data $\boldsymbol{x}$

$$p(\boldsymbol{a}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{a} = \frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{a})\,\pi(\boldsymbol{a})\,\mathrm{d}\boldsymbol{a}}{\int \mathcal{L}(\boldsymbol{x}|\boldsymbol{a}')\,\pi(\boldsymbol{a}')\,\mathrm{d}\boldsymbol{a}'}$$

  to obtain the posterior density.

3. Application of the maximum-likelihood method

Application of the maximum-likelihood method may be complicated due to imperfections of the measurement:

- limited acceptance,
- finite resolution
- non-negligible background contribution.

In principle there seem to be two possibilities, to take these conditions into account:

- try to *correct* the data, or
- modify the theoretical distribution according to the real properties of the measurement - the correct method.

Limited acceptance

Limited acceptance is described by an acceptance function

$$A(x) \quad = \text{probability of observation} \qquad \text{with} \quad 0 \leq A(x) \leq 1$$

"Acceptance"-correction: assignment of a weight $w_i = 1/A(x_i)$ to the data element x_i.

If for example $A(x_i) = 0.5$, the measurement x_i would get a weight $w_i = 2$. This method may be acceptable for a histogram of the measured distribution; the weighted histogram is then called "acceptance"-corrected.

Weighting method: replace in the likelihood function $f(x_i, \boldsymbol{a})$ by

$$f(x_i, \boldsymbol{a})^{w_i} \qquad \text{with the result} \qquad F(\mathbf{a}) = -\sum_{i=1}^{n} w_i \ln f(x_i|\mathbf{a}).$$

Resulting errors will be wrong. Especially problems with large weights with acceptance $A(x_i) \ll 1$.

Limited acceptance – correct treatment

The correct way is to modify the expectation and to replace $p(x_i|\boldsymbol{a})$ by the properly normalized probability density

$$\mathcal{N}(\boldsymbol{a})^{-1} \cdot A(x_i) \cdot p(x_i, \boldsymbol{a})$$

with the normalization factor $\mathcal{N}$ defined by

$$\mathcal{N} = \int_{\Omega} A(x) p(x, \boldsymbol{a}) \, \mathrm{d}x$$

In general the normalizing factor will depend on the actual parameter value.

Limited acceptance - example

According to the exponential decay law the decay time distribution is proportional to $e^{-t/\tau}$, where τ is the mean lifetime. For a measurement, which is sensitive only in the time region $t_1 \ldots t_2$, the p.d.f is correctly normalized by the condition $\int_{t_1}^{t_2} p(t)\,\mathrm{d}t = 1$, resulting in the expression

$$p(t)\,\mathrm{d}t = \frac{e^{-t/\tau}}{\tau\left(e^{-t_1/\tau} - e^{-t_2/\tau}\right)}\,\mathrm{d}t\ ,$$

valid in the rest system of the particle (the mean decay time τ is defined in the rest system of the particle).

Finite resolution

The measured data are "smeared" by a certain resolution function.

The expected distribution is

$$p(x,a) \quad \text{folded with resolution function} \quad A(x_{meas}, x)$$

Result of folding is new distribution

$$q(x,a) = \int A(x_{meas}, x) p(x,a)\,\mathrm{d}x$$

which has to be used in the Likelihood function.

Background contribution

The measured distribution may contain, in addition to the theoretical distribution $p(x, a)$, an additional contribution due to background. For a M.L. fit the background distribution has to be known, either by a measurement of by a simulation.

$$\begin{array}{rl} p(x,a)\,\mathrm{d}x & \text{signal distribution} \\ q(x)\,\mathrm{d}x & \text{background distribution} \\ (1-\alpha)p(x,a)\,\mathrm{d}x + \alpha q(x)\,\mathrm{d}x & \text{fit distribution} \end{array}$$

The parameter α has either to be known before or has to be fitted.

4. Binned maximum-likelihood

Measurement: sample of real data, each element of which consists of a set of values $\{\boldsymbol{x}\}$.

Binning: dividing the one- or multi-dimensional space of the $\{\boldsymbol{x}\}$ into n bins. This subdivision gives a set of numbers $\{d_1, d_2, \ldots, d_n\}$, where d_i is the number of events in the real data that fall into bin i. The values d_i are integers 0, 1, ...

$$\begin{aligned} d_i &= \text{number of events in the real data that fall into bin } i \\ N_D &= \sum_{i=1}^{n} d_i = \text{total number in the data sample} \end{aligned}$$

The real data arise from a number of sources (or physical processes) and the aim is to determine the proportions P_j of the different sources in the data from the statistical data d_i and from models for the sources.

One has to distinguish the case, where analytic forms are available for the distribution of the sources and where no analytic forms are available.

Model with analytic prediction

Calculate, by integration over the bins, numbers a_{ji} proportional to the expected number of events from source j in bin i; these values have no statistical errors.

$$\begin{aligned}
P_j &= \text{proportion of source } j \text{ in the data (sum to unity)} \\
a_{ji} &= \int_{\text{bin } i} f_j(x)\,\mathrm{d}x = \text{numbers, proportional} \\
& \text{to expected number from source } j \text{ expected in bin } i \\
N_j &= \sum_{i=1}^{n} a_{ji} \\
a_{ji}/N_j &= \text{fraction of events from source } j \text{ expected in bin } i
\end{aligned}$$

The predicted number of events in bin i is, for the proportions P_j, given by the sum

$$f_i = N_D \sum_{j=1}^{m} P_j \, \frac{a_{ji}}{N_j} \qquad \text{or} \qquad f_i = \sum_{j=1}^{m} p_j \, a_{ji}$$

with strength factors $p_j = N_D P_j / N_j$.

Approximate solution by least squares

Measured numbers d_i follow Poisson distribution.

"Standard" method (χ^2 minimization): approximate Poisson distribution by normal distribution with standard deviation $\sigma_i = \sqrt{d_i}$

$$\text{adjust } p_j \text{ to minimize} \qquad \boxed{S(\boldsymbol{p}) = \sum_{i=1}^{n} \frac{(d_i - f_i)^2}{d_i}}$$

This approximation will lead to biased results for small d_i.

Better use correct Poisson distribution.

Poisson maximum likelihood

If the mean value in a bin is f, then the observed values $d = 0, 1, \ldots$ follow the Poisson distribution

$$P_f(d) = e^{-f} \frac{f^d}{d!} \qquad \text{with normalization} \quad \sum_{d=0}^{\infty} P_f(d) = 1 \, .$$

Strength factors are found by maximizing the total likelihood

$$\mathcal{L}(\boldsymbol{p}) = \prod_{i=1}^{n} P_{f_i}(d_i) = \prod_{i=1}^{n} e^{-f_i} \frac{f^{d_i}}{d_i!}$$

with respect to the parameters $\boldsymbol{p}$ or equivalently, by minimizing the negative logarithm of the likelihood $F(\boldsymbol{p}) = -\ln \mathcal{L}(\boldsymbol{p})$:

$$\boxed{F(\boldsymbol{p}) = \sum_{i=1}^{n} (f_i - d_i \cdot \ln f_i)}$$

omitting constant factors in the product like $1/d_i!$.

Poisson maximum likelihood contnd.

This expression correctly accounts for small numbers of data events d_i in a bin or even zero data events in some bins. The method is called *binned maximum likelihood* fit.

There is some advantage in redefining the function F to be minimized by adding some constants as

$$\boxed{F(\boldsymbol{p}) = \sum_{i=1}^{n} g_i \qquad \text{with} \quad g_i = \begin{cases} (f_i - d_i) - d_i \cdot \ln(f_i/d_i) & \text{if } d_i > 0 \\ f_i & \text{if } d_i = 0 \end{cases}} \, .$$

The position of the minimum and the shape of F is not changed by his modification, but now $2F$ is approximately distributed according to the χ^2 distribution (with $(n-m)$ degrees of freedom).