The maximum-likelihood method

1. The maximum likelihood principle
2. Properties of maximum-likelihood estimates
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 $p(x; a)$

$$p(x; a) \geq 0 \quad \text{with} \quad \int_{\Omega} p(x; a) \, dx = 1$$

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 Likelihood function

The maximum-likelihood method starts from the joint probability distribution of the $n$ measured values $x_1, \ldots, x_i, \ldots, x_n$.

For independent measurements this 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.

The likelihood function is a function, it is not a probability density of the parameter $a$ (→ Bayes interpretation).
Principle of Maximum Likelihood

The estimate \( \hat{a} \) for the parameters \( a \) is the value, which \textit{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 \( \mathcal{L}(a) \), or with the \textit{negative} logarithm. In the following the \textit{negative} log-likelihood function is considered,

\[
F(a) = -\ln \mathcal{L}(a) = - \sum_{i=1}^{n} \ln p(x_i|a)
\]

and the maximum likelihood estimate \( \hat{a} \) is the value that \textit{minimizes} this function.

Likelihood equation, defining estimate \( \hat{a} \):

\[
\frac{dF(a)}{da} = 0
\]

Sometimes a factor of 2 is included in the definition of the negative log-likelihood function; this factor makes it similar to the \( \chi^2 \)-expression of the method of least squares in certain applications: \( F(a) = -2 \ln \mathcal{L}(a) \).
Example of angular distribution

The value $x \equiv \cos \vartheta$ is measured in $n$ decays of an elementary particle. According to theory the distribution is

$$p(\cos \vartheta) = \frac{1}{2} \left(1 + a \cos \vartheta\right)$$

This probability density is normalized for all physical values of the parameter $a$, if the whole range of $\cos \vartheta$ can be measured.

The aim is to get an estimate of the parameter $a$.

$$\minimize \mathcal{L}(a) = \prod_{i=1}^{n} \left[\frac{1}{2} (1 + a \cos \vartheta_i)\right]$$

$$\maximize F(a) = -\sum_{i=1}^{n} \ln (1 + a \cos \vartheta_i) + \text{const.}$$

Note: The normalization is parameter dependent, if the measured range of $\cos \vartheta$ is limited.
- shape of $F(a)$ approximately parabolic
- first derivative approximately linear
- second derivative approximately constant
Example: exponential distribution

Measured are $n$ times $t_i$, which should be distributed according to the density

$$p(t; \tau) = \frac{1}{\tau} \exp \left[-\frac{t}{\tau}\right].$$

Log. Likelihood function for parameter $\tau$, to be estimated from the data:

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

By minimization of $F(\tau)$ the resulting estimate is

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i \quad \text{with} \quad E[\hat{\tau}(t_1, t_2, \ldots)] = \tau$$

i.e. the estimator is unbiased.

Note: in general mean values are unbiased.
Instead of parameter $\tau$ the parameter $\lambda$ in the density 

$$p(t; \lambda) = \lambda \exp[-\lambda t] .$$

has to be estimated. Can the previous result be used?

$$\text{yes, because of} \quad \left( \frac{\partial L}{\partial \tau} \right) = \left( \frac{\partial L}{\partial \lambda} \right) \cdot \frac{\partial \lambda}{\partial \tau} = 0$$

the Maximum Likelihood estimate for $\lambda$ is

$$\hat{\lambda} = \frac{1}{\hat{\tau}}$$

(note: $\mathcal{L}(a)$ is a function of $a$, not a density).

But:

$$E \left[ \hat{\lambda}(t_1, t_2, \ldots) \right] = \frac{n}{n-1} \lambda = \frac{n}{n-1} \frac{1}{\tau} \quad \text{biased!}$$

i.e. there is invariance of the Maximum Likelihood estimates w.r.t. transformations, but only one parametrization can be unbiased.
Properties of the maximum-likelihood estimates

Maximum-likelihood estimates \( \hat{a} \)

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

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

**Invariance:** The maximum likelihood solution is invariant under change of parameter – the estimate \( \hat{b} \) of a function \( b = b(a) \) is given by \( \hat{b} = b(\hat{a}) \). The bias \( B(\hat{a}) \) 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.
Information inequality

\[
I(a) = E \left[ \left( \frac{\partial \ln \mathcal{L}}{\partial a} \right)^2 \right] = \int_{\Omega} \left( \frac{\partial \ln \mathcal{L}}{\partial a} \right)^2 \mathcal{L} \, dx_1 \, dx_2 \ldots \, dx_n
\]

This is the definition of information, where \( \mathcal{L} \) is the joint density of the \( n \) observed values of the random variable \( x \).

**Information inequality**  \[ V[\hat{a}] \geq \frac{1}{I} \]

The inverse of the information \( I_n(a) \), or short \( I \), is the lower limit of the variance of the parameter estimate \( \hat{a} \) – minimum variance bound \( \text{MVB} \).

The inequality is also called Rao-Cramér-Frechet inequality, and is valid in this form for any unbiased estimate \( \hat{a} = \hat{a}(x) \).
Alternative expression of information $I$

From the proof of the information inequality in previous chapter:

$$\int_{\Omega} \left( \frac{\partial \ln L}{\partial a} \frac{\partial L}{\partial a} + \frac{\partial^2 \ln L}{\partial a^2} L \right) \, dx_1 \, dx_2 \ldots \, dx_n = 0 ,$$

Rewritten in terms of expectation values:

$$I(a) = E \left[ \left( \frac{\partial \ln L}{\partial a} \right)^2 \right] = -E \left[ \frac{\partial^2 \ln L}{\partial a^2} \right]$$

i.e. either square of first derivative or negative second derivative.

The second derivative is *almost* constant: expectation value is close to value at the minimum

$$I(a) = -E \left[ \frac{\partial^2 \ln L}{\partial a^2} \right] \approx \frac{\partial^2 F(a)}{\partial a^2} \bigg|_{a=\hat{a}}$$
Case of several variables

Case of $m$ variables $a_1, \ldots, a_j, \ldots, a_m$: information $I$ becomes a $m$-by-$m$ symmetric matrix $I$ with elements

$$I_{jk} = E \left[ \frac{\partial \ln L}{\partial a_j} \frac{\partial \ln L}{\partial a_k} \right] = -E \left[ \frac{\partial^2 \ln L}{\partial a_j \partial a_k} \right]$$

The minimal variance $V [\hat{a}]$ of an estimate $\hat{a}$ is given by the inverse of the information matrix $I$:

$$\text{minimal variance} \quad V [\hat{a}] = I^{-1}$$
Normality: The estimate \( \hat{a} \) is, under very general conditions, asymptotical normally distributed with minimal variance \( V(\hat{a}) \), i.e.

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

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:

\[
V[\hat{a}] = \left( \frac{d^2 F}{da^2} \bigg|_{a=\hat{a}} \right)^{-1} \quad \text{V} [ \hat{a} ] = H \quad \text{with} \quad H_{jk} = \frac{\partial^2 F}{\partial a_j \partial a_k}
\]
The maximum-likelihood method

The maximum-likelihood principle
   The Likelihood function .......................... 2
   Principle of Maximum Likelihood ................. 3
   Example of angular distribution .................. 4
   Example: exponential distribution ............... 5

Properties of the maximum-likelihood estimates  9
   Information inequality ........................... 10
   Alternative expression of information $I$ ........ 11
   Case of several variables ......................... 12
   Normality ............................................ 13