1 Maximum Likelihood Method

Unbinned maximum likelihood is a convenient means of determining parameters of a fit to data. We suppose we are to determine a quantity A by fitting with a dataset $\{t_i\}$. For simplicity, we assume that there is a single distribution $f(A;t)$. It is normalized so that

$$\int dt f(A;t) = 1 \tag{1}$$

independent of A. The log-likelihood function for this dataset is

$$\mathcal{L} = \sum_i \ln f(A,p;t_i) \tag{2}$$

We indicate the true value of the parameter by $\overline{A}$. As the number of data points increases, the likelihood function converges to

$$\mathcal{L} \to N \int dt f(\overline{A},t) \ln f(A,t) \tag{3}$$

since $f(\overline{A},t)$ gives the actual distribution of events in t.

The likelihood function is maximized when

$$\frac{\partial \mathcal{L}}{\partial A} = 0 \tag{4}$$

provided, of course, that

$$\frac{\partial^2 \mathcal{L}}{\partial A^2} < 0 \tag{5}$$

Now as the number of events increases, we will find that

$$\frac{\partial \mathcal{L}}{\partial A} \to N \int dt f(\overline{A},t) \frac{1}{f(A,t)} \frac{\partial f(A,t)}{\partial A} \tag{6}$$

If we evaluate this at $A = \overline{A}$ we find it vanishes as a consequence of

$$\frac{\partial}{\partial A} \int dt f(A,t) = 0 \tag{7}$$

Thus asymtotically the maximum likelihood occurs at the true value of the parameter.

The values of A we find will be distributed in a normal distribution around $\overline{A}$. With $\delta A = A - \overline{A}$, the probability distribution is

$$\frac{dP}{dA} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\delta A^2}{2\sigma^2}} \tag{8}$$

where

$$\frac{1}{\sigma^2} = -\frac{\partial^2 \mathcal{L}}{\partial A^2} \tag{9}$$

Asymptotically

$$\frac{\partial^2 \mathcal{L}}{\partial A^2} \to N \frac{\partial^2}{\partial A^2} \int dt f(\overline{A}, t) \ln f(A, t) = -N \int dt \frac{1}{f} \left(\frac{\partial f}{\partial A} \right)^2 \tag{10}$$

where everything is evaluated at $\overline{A}$. From this we derive the famous asymptotic relation

$$\sigma^{-2} = N \int dt \frac{1}{f} \left(\frac{\partial f}{\partial A} \right)^2 \tag{11}$$

Often we have events that fall into separate categories, each of which has its own distribution. Suppose there are two categories with distributions $f_1(A; t)$ and $f_2(A; t)$. If these are normalized so that

$$\int dt f_1(A; t) + \int dt f_2(A; t) = 1 \tag{12}$$

then everything will work as before. The log-likelihood function is simply the sum of two pieces, where each data point contributes to the appropriate sum:

$$\mathcal{L} = \sum_i \ln f_1(A; t_i) + \sum_j \ln f_2(A; t_j) \tag{13}$$

The corresponding σ is given by

$$\sigma^{-2} = N \left[\int dt \frac{1}{f_1} \left(\frac{\partial f_1}{\partial A} \right)^2 + \int dt \frac{1}{f_2} \left(\frac{\partial f_2}{\partial A} \right)^2 \right] \tag{14}$$

As a pertinent example consider the time distribution in $B \to J/\psi K_S$. If the opposite B is tagged as a B^0 the distribution is

$$f_+(DA; t) = \frac{\Gamma}{4} e^{-\Gamma|t|} (1 + DA \sin \Delta m t) \tag{15}$$

whereas if the tag is a $\overline{B}^0$, the sign in front of DA is negative. We see that these distributions are properly normalized:

$$\int dt f_+(DA;t) + \int dt f_-(DA;t) = 1 \tag{16}$$

We compute directly

$$\begin{aligned} \sigma^{-2} &= N\int_0^\infty dt\Gamma e^{-\Gamma t}\frac{D^2\sin^2\Delta mt}{1-D^2A^2\sin^2\Delta mt} \\ &= ND^2\left[\frac{2x^2}{1+4x^2}+\frac{24x^4}{(1+4x^2)(1+16x^2)}D^2A^2\ldots\right] \end{aligned} \tag{17}$$

If there are several tagging categories i, each with a fraction ϵ_i of the events and with dilution $D_i = 1-2w_i$, there is a contribution to σ^{-2} from each:

$$\sigma^{-2} = N\sum_i \epsilon_i D_i^2\left[\frac{2x^2}{1+4x^2}+\frac{24x^4}{(1+4x^2)(1+16x^2)}D_i^2A^2\ldots\right] \tag{18}$$

With the convenient approximation $x \approx 1/\sqrt{2}$ we obtain the rule of thumb:

$$\sigma(A) = \sqrt{\frac{3}{N\sum_i \epsilon_i D_i^2}} \tag{19}$$

Suppose we have two categories but that we don't know what fraction of the events will fall into each. We can introduce a second variable, ϵ, representing the fraction in the first distribution and maximize the likelihood with respect to both A and ϵ. Here we require unit normalization for the two distributions separately:

$$\int dt f_1(A;t) = 1; \qquad \int dt f_1(A;t) = 1 \tag{20}$$

and take the two functions to be

$$\epsilon f_1; \qquad (1-\epsilon)f_2 \tag{21}$$

Now when we maximize with respect to ϵ we find

$$\frac{\partial}{\partial\epsilon}\left[\sum_{i=1}^{M}\ln\epsilon f_1(A;t_i)+\sum_{j=1}^{N}\ln(1-\epsilon)f_2(A;t_j)\right] = 0 \tag{22}$$

the result is simply

$$\frac{M}{\epsilon} = \frac{N}{1-\epsilon} \tag{23}$$

In otherwords, ϵ is determined simply to reproduce the number of events in the two categories. The equation for A will be just as if we hadn't worried about ϵ. Thus, if f_1 and f_2 have unit normalization, we can forget about ϵ.

When, as in the case above, there is more than one parameter to be determined, say A_i, $i = 1, ..q$, the likelihood equations are simply

$$\frac{\partial \mathcal{L}}{\partial A_i} = 0 \tag{24}$$

For this to be a maximum we need the (negative of the) matrix of second derivatives

$$L_{ij} = -\frac{\partial^2 \mathcal{L}}{\partial A_i \partial A_j} \tag{25}$$

to be positive definite. Asymtoptically the distribution of output A_i for a set of true values $\overline{A}_i$ has a Gaussian shape

$$dP = \sqrt{\det L}(2\pi)^{-q/2} e^{-\frac{1}{2}\delta A_i \delta A_j L_{ij}} dA_1 \dots dA_q \tag{26}$$

From this we can find not just the expected variance in A_i, $\langle \delta A_i^2 \rangle$ but more generally $\langle \delta A_i \delta A_j \rangle$. A direct computation shows that

$$\langle \delta A_i \delta A_j \rangle = [L^{-1}]_{ij} = M_{ij} \tag{27}$$

The inverse of L, the covariance matrix M, is thus the central quantity. In particular

$$\sigma_i = M_{ii}^{-1/2} \tag{28}$$

and the correlation matrix is

$$\rho_{ij} = \frac{M_{ij}}{\sigma_i \sigma_j} \tag{29}$$

The elements of L are given asymptotically by

$$L_{ij} = N \int dt \frac{1}{f} \frac{\partial f}{\partial A_i} \frac{\partial f}{\partial A_j} \tag{30}$$

If, due to fluctuations, the output A_i comes out one sigma high, then we expect A_j to come out ρ_{ij} units of sigma high, as well.

In practice, the CERN workhorse Minuit is generally used to find the minimum of the negative log-likelihood. Since Minuit is looking for χ^2 we need to convert log-likelihood to this quantity. If it is the (positive) loglikelihood that is computed, probabilities vary as $e^{\Delta\mathcal{L}}$ while for χ^2 we have $e^{-\chi^2/2}$ so the coversion is $-2\mathcal{L} \sim \chi^2$. Minuit produces both the covariance matrix and the correlation matrix as part of the regular output.

Aside Probably Not Appropriate for Working Group Report

Suppose our parameters are numbered $0, 1, 2, \ldots q$ and that A_0 is the parameter of greatest interest. The uncertainty in A_0 is given by

$$\sigma_0^2 = [L^{-1}]_{00} = \frac{1}{L_{00} - L_{0i}L_{j0}[\tilde{L}^{-1}]_{ij}} \tag{31}$$

Here $\tilde{L}$ is the $q-1$ by $q-1$ matrix obtained by ignoring the zeroth row and column of L. (It takes some work to prove this relation).

This expression doesn't look particularly familiar. We can see, however, that it is really similar to the usual expression. Suppose we measured A_0 directly with $A_1, \ldots A_q$ as input parameters. We would expect that in addition to whatever direct error there was in measuring A_0, we would append errors in quadrature for the remaining A_i:

$$\sigma^2 \to \sigma^2 + \sum_i \left(\frac{\partial A_0}{\partial A_i}\right)^2 \sigma_i^2 \tag{32}$$

How is this related our full expression?

The log-likelihood $\mathcal{L}$ is a function of the parameters $A_0, A_1. \ldots A_q$. If we impose all the conditions

$$\frac{\partial \mathcal{L}}{\partial A_i} = 0 \tag{33}$$

then every parameter is determined by the maximum likelihood condition. Alternatively, we could leave some of the A_i as free parameters by omitting the maximization condition. Suppose for example we maximize only with respect to A_0 and leave the rest of the As free. Then maximization gives implicitly a function $A_0(A_1, \ldots A_q)$.

Now take the derivative of the relation that defines the maximum:

$$\frac{\partial}{\partial A_i}\frac{\partial \mathcal{L}(A_0(A_1,\ldots A_q),A_1,...A_q)}{\partial A_0}=\frac{\partial^2\mathcal{L}}{\partial A_0 \partial A_i}+\frac{\partial^2\mathcal{L}}{\partial A_0^2}\frac{\partial A_0}{\partial A_i}=0 \tag{34}$$

Now let us consider expanding Eq. (31)

$$\sigma_0^2=[L^{-1}]_{00}=\frac{1}{L_{00}}[1+\frac{L_{0i}L_{j0}}{L_{00}}[\tilde{L}^{-1}]_{ij}] \tag{35}$$

Without the error due to the parameters $A_1,\ldots A_q$, the uncertainty in A_0 would be $\sigma=1/\sqrt{L_{00}}$. We write

$$\sigma_0^2=[L^{-1}]_{00}=\sigma^2[1+\frac{L_{0i}L_{j0}}{L_{00}}[\tilde{L}^{-1}]_{ij}] \tag{36}$$

But we have in our case, from Eq. (34),

$$L_{0i}=L_{i0}=-L_{00}\frac{\partial A_0}{\partial A_i} \tag{37}$$

so in the approximation where the errors due to the extra parameters are small,

$$\begin{aligned}\sigma_0^2 &= [L^{-1}]_{00}=\sigma^2[1+\frac{\partial A_0}{\partial A_i}\frac{\partial A_0}{\partial A_j}[\tilde{L}^{-1}]_{ij}/\sigma^2] \\ &= \sigma^2+\frac{\partial A_0}{\partial A_i}\frac{\partial A_0}{\partial A_j}[\tilde{L}^{-1}]_{ij}\end{aligned} \tag{38}$$

which corresponds to the traditional formula since $\tilde{L}^{-1}$ is related to the correlation matrix for the parameters $A_1,\ldots A_q$

$$\tilde{L}^{-1}_{ij}=\tilde{\rho}_{ij}\sigma_i\sigma_j \tag{39}$$