Numerical Methods

- Statistical Methods in Experimental Physics
  - W. T. Eadie, D. Drijard, F. E. James, M. Roos, B. Sadoulet

- Statistics for nuclear and particle Physics
  - Louis Lyon

- Probability and Statistics in Particle Physics
  - A. Frodesen, O. Skeggestad, H. Tofte

- Statistical Data Analysis
  - Glen Cowan

- Monte Carlo theory and practice
  - F. James

- Data Analysis Technique for High Energy Physics
  - M. Regler and R. Fruhwirth

- Statistical and Computational Methods in Data Analysis
  - Sigmund Brandt

Classical Mechanics around 1900

- Thinking in the West:
  - If a problem has a solution, we know how to find it
  - Other systems have no solution
  - Reasonable physicists should move on to other problems

- Thinking in the Soviet Union
  - Most systems have no analytic solution
  - Those systems should have interesting properties of a different kind
  - They are clearly candidates for randomness
  - We need a way to define different degrees of randomness

- Start using ideas of statistics and Monte Carlo techniques in the field of physics

Probability

- In many experiments, even when the conditions are kept the same, repeated measurements can yield different results

The results of individual measurements are unpredictable → possible results of a series of measurements can have a well-defined distribution

- events must be completely uncorrelated (statistical independence)

- The number of trials needs to be large (law of large numbers)

N_i = number of events observed in class i out of a total of N

$\Rightarrow p_i$ = probability of getting an event in class i

= $\lim_{N\to\infty} (N_i/N)$

- for a continuously varying variable

$N\ f(x_i)\ \Delta x_i$ = number of events observed in the interval between x_i and $x_i + \Delta x_i$ for values of x out of a total of N events

element of probability: $dp = f(x)\,dx$

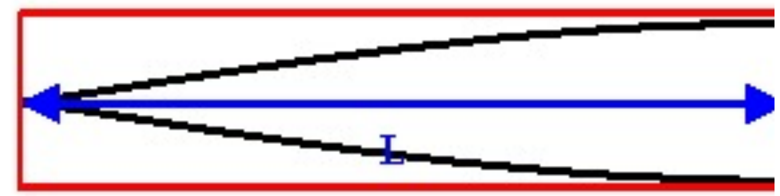
$f(x)$ = probability density function

within the permissible range $\int_{x_{Low}}^{x_{High}} f(x)\,dx = 1$

Examples of Probability

- Throwing a dice:
  the probability that one scores a value n is 1/6
  for a large number N throws n appears $N/6$

- Emission of scintillation light:



  scintillation photons are emitted isotropically
  within a tube having all sides blackened except a small opening of solid angle $d\Omega$ $\rightarrow$ $d\Omega/(4\pi)$ emitted photons will escape

- More common in a physics experiment → determine some parameter from a set of measurements:

$$f(\cos\theta) = \frac{1}{2}(1 + \alpha\cos\theta)$$

  experiment measures $\alpha \rightarrow \alpha^{\star} \pm \Delta\alpha$ Crudely
  - the probability that the true value of α has been in the range $\alpha^{\star} - \Delta\alpha$ to $\alpha^{\star} + \Delta\alpha$ is 68.3% (inverse probability)
  - the probability that the true value of α lies between $\alpha^{\star} - \Delta\alpha$ and $\alpha^{\star} + \Delta\alpha$ is 0 or 1 (direct probability)

Likelihood Ratio

Suppose we are making a measurement of a variable x to justify if the hypothesis A or B is true

- if A is true, the experimental distribution of the variable x must follow $f_A(x)$
- If B is true, the distribution must be $f_B(x)$

The experiment provides N events with measurements $x_1, x_2, \cdots x_N$

If A is true, the joint probability of getting these results:

$$dp_A = \prod_{i=1}^{N} f_A(x_i) dx_i$$

the likelihood ratio:

$$R = \prod_{i=1}^{N} \frac{f_A(x_i)}{f_B(x_i)}$$

is the probability that the particular experimental result with N events turns out the way it did assuming A is true divided by the probability that the experiment turns out the way it did assuming B is true. (Betting odds of A versus B)

Use of Likelihood Ratio

- An experimenter in planning a new experiment must estimate a priori the number of measurements needed to prove a certain hypothesis

use the average logarithm of the likelihood ratio which is better behaved mathematically than the average ratio

$$\begin{aligned}\overline{\log R} &= N \int \log \frac{f_A}{f_B} f_A(x) dx &\quad \text{with } A = .\text{TRUE.}\\ &= N \int \log \frac{f_A}{f_B} f_B(x) dx &\quad \text{with } B = .\text{TRUE.}\end{aligned}$$

e.g. study of K_s^0 spin from its decay to $\pi^+\pi^-$
hypothesis A: if K_s^0 has spin 0, the energy distribution of π^+ in the laboratory frame will be flat $\Rightarrow f_A(x) = 1$
hypothesis B: if K_s^0 has spin 1, the energy distribution will follow $\Rightarrow f_B(x) = 2x$

$$x = \frac{\text{kinetic energy of } \pi^+}{\text{maximum value of kinetic energy}}$$

An experiment is to be designed to establish spin 0 of K_s^0 with odds of 10^4 to 1

$$\Rightarrow \log 10^4 = N \int_0^1 \log \frac{1}{2x} \cdot 1 \cdot dx = -N \int_0^1 \log(2x) dx$$

$$\Rightarrow N = 30$$

Use of Likelihood Ratio

on average 30 events will be required

however, if 1 event is found with $x = 0$, this would make $R \to \infty$

$\Rightarrow$ one such event would be a proof of spin 0 of K_s^0

Fluctuation of $\log R$ for a given N $\Rightarrow$

$$\overline{\left(\log R - \overline{\log R}\right)^2} = N\left[\int \left(\log \frac{f_A}{f_B}\right)^2 f_A(x)dx - \left(\int \left(\log \frac{f_A}{f_B}\right) f_A(x)dx\right)^2\right]$$

Maximum Likelihood Method

- Consider the estimation of the value of a parameter which is a continuous variable

  then the number of hypotheses to be tested is infinite rather than a discrete set

  $\Rightarrow$ use the same basic principle which says that the probability of any two different values of the parameter is the ratio of probabilities of getting a particular experimental result assuming the first parameter value and then the other value

  $f(a,x) =$ truly normalised distribution function

  $a =$ parameter; $x =$ measurement; with $\int f(a,x)\,dx = 1$

  Then the likelihood function $\mathcal{L}(a) = \prod_{i=1}^{N} f(a, x_i)$

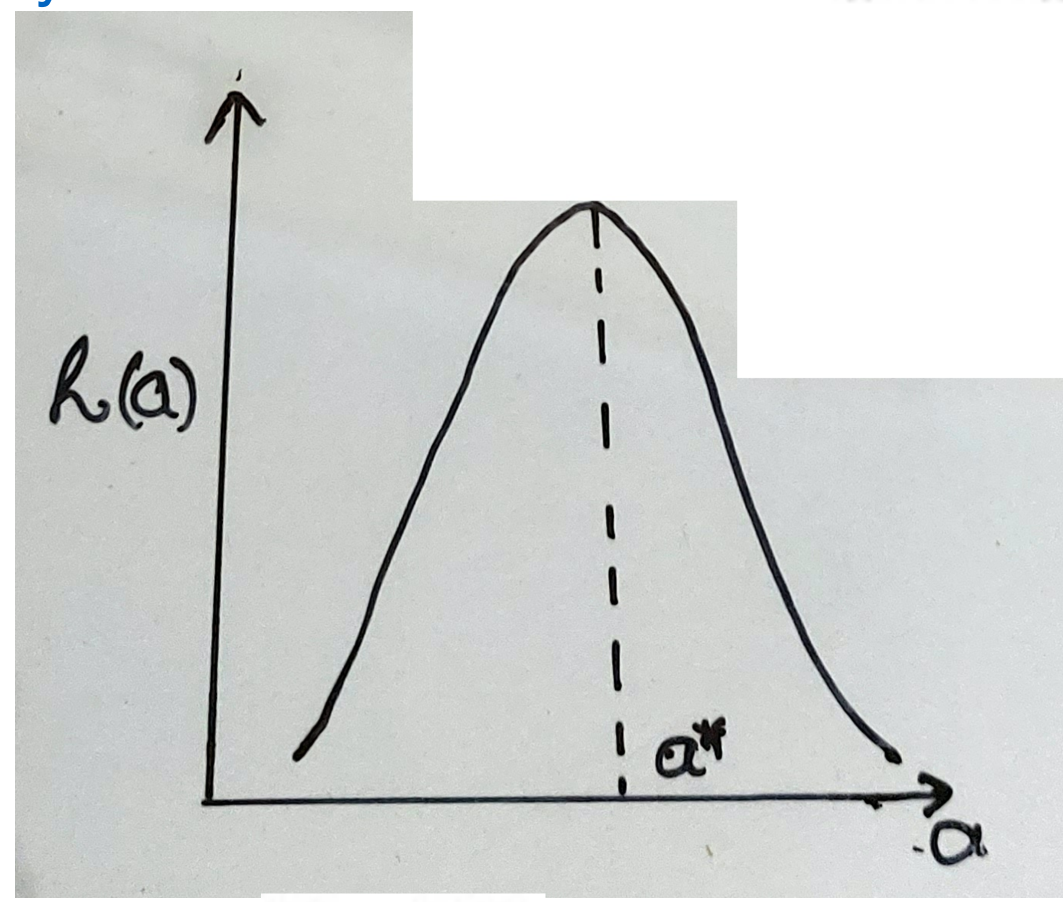
  is the joint probability distribution function of getting a particular set of experimental results $x_1, x_2, \cdots x_N$

  Relative probability of a can be obtained from the distribution of $\mathcal{L}(a)$ vs a

Maximum Likelihood Method

$a^\star$ = the most probable value of (maximum likelihood solution)
Δa = the RMS spread of a about $a^\star$
(conventional measure of accuracy in the determination of $a \rightarrow a^\star$

$$= \left[\frac{\int (a-a^\star)^2 \mathcal{L}(a)da}{\int \mathcal{L}(a)da} \right]^{\frac{1}{2}}$$



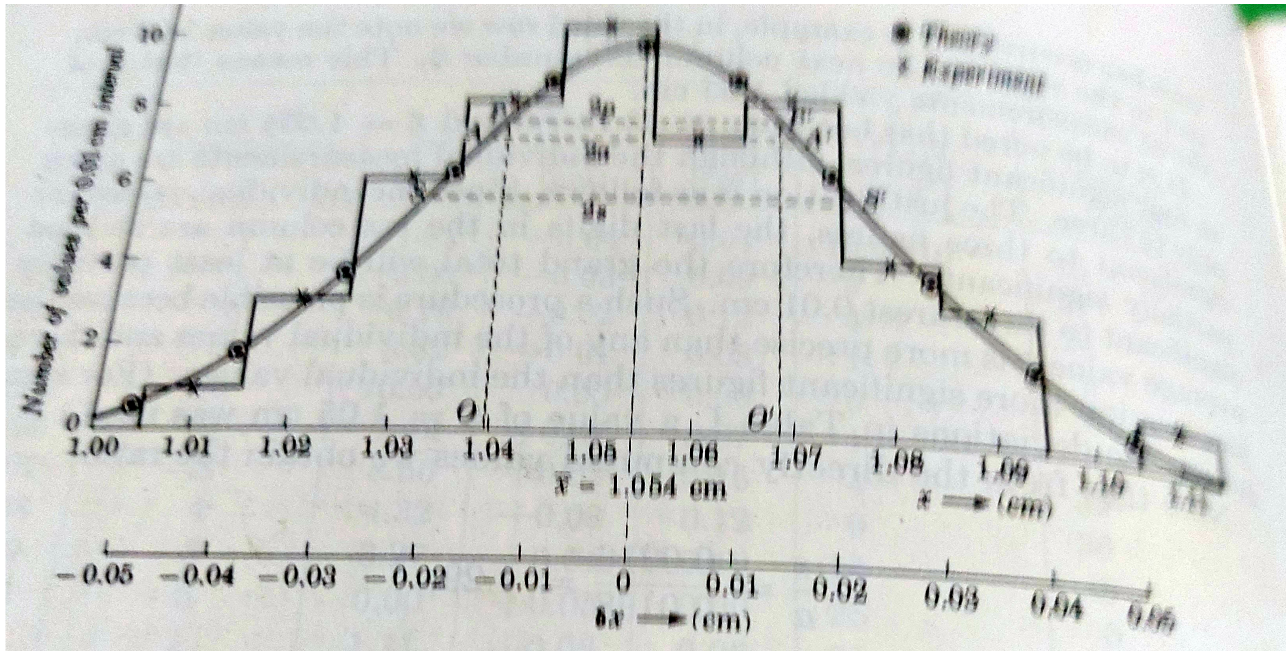
For $N \rightarrow \infty$, $a^\star$ approaches the true value of $a \rightarrow a_0$

To determine m parameters $a_1, a_2, \cdots a_m$ determine $\mathcal{L}(a_1, \cdots a_m)$ and solve m simultaneous equations:

$$\frac{\partial W}{\partial a_i}\Big|_{a_i = a_i^\star} = 0 \qquad [W = \ln \mathcal{L}(a_1, \cdots a_m)]$$

Gaussian Distributions

- If the number of measurements is indefinitely increased, the width of the distribution of the measured parameter steadily decreases and the histogram approaches a smooth curve given by a distribution function known as Gaussian distribution



$$
\begin{aligned}
f(a,x) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-(x-a)^2/(2\sigma^2)\right] \\
x &= \text{measurement} \\
\sigma &= \text{uncertainty in the measurement}
\end{aligned}
$$

for a set of measurements x_i with the corresponding uncertainties σ_i, the likelihood function

$$\mathcal{L}(a) = \prod^{N} \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-(x_i-a)^2/(2{\sigma_i}^2)\right]$$

Then

$$W = -\frac{1}{2}\sum_i^N \frac{(x_i-a)^2}{{\sigma_i}^2} + constant$$

Thus

$$\frac{\partial W}{\partial a} = \sum_i^N \frac{x_i - a}{{\sigma_i}^2}$$

Gaussian Distributions

So the solution of $\frac{\partial W}{\partial a}|_{a=a^\star} = 0$ is given by

$$\sum_i \frac{x_i}{{\sigma_i}^2} - \sum_i \frac{a^\star}{{\sigma_i}^2} = 0$$

$$\Rightarrow a^\star = \left[\sum_i \frac{1}{{\sigma_i}^2} x_i\right] / \left[\sum_i \frac{1}{{\sigma_i}^2}\right]$$ (weighted mean)

when all measurement errors are the same

$$a^\star = \frac{1}{N}\sum_i x_i$$

Maximum Likelihood Error

For large N, $\mathcal{L}(a)$ approaches a Gaussian distribution:

$$\mathcal{L}(a) \quad \sim \quad \exp\left[-\frac{h}{2}(a-a^\star)^2\right]$$

with $1/\sqrt{h}$ as the RMS spread of a around $a^\star$

$$\begin{aligned} W &= -\frac{h}{2}(a-a^\star)^2 + constant \\ \frac{\partial W}{\partial a} &= -h(a-a^\star) \\ \frac{\partial^2 W}{\partial a^2} &= -h \end{aligned}$$

Thus $$\Delta a = \frac{1}{\sqrt{h}} = \left(-\frac{\partial^2 W}{\partial a^2}\right)^{-\frac{1}{2}}$$

If $f(a,x)$ follows a Gaussian distribution

$$\frac{\partial W}{\partial a} = \sum_i \frac{x_i - a}{{\sigma_i}^2}$$

$$\Rightarrow \Delta a = \left[\sum_i \frac{1}{{\sigma_i}^2}\right]^{-\frac{1}{2}}$$

if $\mathcal{L}(a)$ is truly Gaussian, $\frac{\partial^2 W}{\partial a^2}$ is the same for all values of a, otherwise it is better to use the average $\overline{\frac{\partial^2 W}{\partial a^2}}$

Maximum Likelihood Error

$$\overline{\frac{\partial^2 W}{\partial a^2}} = \frac{\int \left(\frac{\partial^2 W}{\partial a^2}\right) \mathcal{L}(a)da}{\int \mathcal{L}(a)da}$$

it is important at the time of designing an experiment to estimate the number of events required to measure a parameter with a given accuracy

$\Rightarrow$ Determine $\overline{\frac{\partial^2 W}{\partial a^2}}$ averaged over many experiments each with N events

For 1 event,

$$\overline{\frac{\partial^2 W}{\partial a^2}} = \int \frac{\partial^2 \ln f}{\partial a^2} f dx$$

For N events

$$\overline{\frac{\partial^2 W}{\partial a^2}} = N \int \frac{\partial^2 \ln f}{\partial a^2} f dx$$

where the parameter is extracted from the distribution $f(a, x)$

Maximum Likelihood Error

$$\frac{\partial^2 \ln f}{\partial a^2} = \frac{\partial}{\partial a}\left(\frac{1}{f}\frac{\partial f}{\partial a}\right) = -\frac{1}{f^2}\left(\frac{\partial f}{\partial a}\right)^2 + \frac{1}{f}\frac{\partial^2 f}{\partial a^2}$$

$$\Rightarrow \int \frac{\partial^2 \ln f}{\partial a^2} f dx = -\int \frac{1}{f}\left(\frac{\partial f}{\partial a}\right)^2 dx + \int \frac{\partial^2 f}{\partial a^2} dx$$

$$= -\int \frac{1}{f}\left(\frac{\partial f}{\partial a}\right)^2 dx + \frac{\partial^2}{\partial a^2}\int f dx$$

$\int f dx = 1 \Rightarrow$ the second term drops out

$$\overline{\frac{\partial^2 W}{\partial a^2}} = -N \int \frac{1}{f}\left(\frac{\partial f}{\partial a}\right)^2 dx$$

$$\Rightarrow \Delta a = \frac{1}{\sqrt{N}}\left[\int \frac{1}{f}\left(\frac{\partial f}{\partial a}\right)^2 dx\right]^{-\frac{1}{2}}$$

Let us consider the decay $\mu \rightarrow e\bar{\nu}_e \nu_\mu$ and study the energy distribution of the decay electron

$$f(a,x) = \tfrac{1}{2}(1 + ax)$$

need to determine a with an accuracy of 1% for $a_0 = -\frac{1}{3}$

Maximum Likelihood Error

$$\frac{\partial f}{\partial a} = \frac{x}{2}$$

Thus $\int_{-1}^{+1} \frac{1}{f}\left(\frac{\partial f}{\partial a}\right)^2 dx = \frac{1}{2}\int_{-1}^{+1} \frac{x^2 dx}{1+ax} = \frac{1}{2a^3}\left[\ln\frac{1+a}{1-a} - 2a\right]$

so $$\Delta a = \frac{1}{\sqrt{N}}\sqrt{\frac{2a^3}{\ln\frac{1+a}{1-a} - 2a}}$$

and $$\frac{\Delta a}{a} = \frac{1}{\sqrt{N}}\sqrt{\frac{2a}{\ln\frac{1+a}{1-a} - 2a}}$$

1% accuracy $\Rightarrow$

$$N = 10^4 \frac{2a}{\ln\frac{1+a}{1-a} - 2a} \quad \text{for} \quad a_0 = -\frac{1}{3}$$

$$\simeq 2.5 \times 10^5$$

Maximum Likelihood Error

An experiment with N events $\Rightarrow$ measures M parameters $a_1, a_2, \cdots a_M$

The earlier formula $\Delta a_i = \left(-\frac{\partial^2 W}{\partial a_i^2}\right)^{-\frac{1}{2}}$ is applicable when the parameters are uncorrelated, i.e., $\overline{(a_i - a_i^\star)(a_j - a_j^\star)} = 0$ for all i, j with $i \neq j$

For the general case, use Taylor's expansion:

$$
\begin{aligned}
W(a) &= W(a^\star) + \sum_i^M \frac{\partial W}{\partial a_i}|_{a_i^\star}\, \beta_i - \frac{1}{2}\sum_i \sum_j H_{ij}\beta_i\beta_j + \cdots \\
\beta_i &= a_i - a_i^\star \\
H_{ij} &= -\frac{\partial^2 W}{\partial a_i\, \partial a_j}|_{a_i{}^\star, a_j{}^\star}
\end{aligned}
$$

For $W(a) = \ln \mathcal{L}(a)$, the most probable value $a^\star$ corresponds to

$$
\begin{aligned}
\frac{\partial W}{\partial a}|_{a^\star} &= 0 \\
\Rightarrow \ln \mathcal{L}(a) &= W(a^\star) - \frac{1}{2}\sum_i \sum_j H_{ij}\beta_i\beta_j + \cdots
\end{aligned}
$$

Neglecting higher order terms.

$$
\mathcal{L}(a) = C \cdot \exp\left[-\frac{1}{2}\sum_i \sum_j H_{ij}\beta_i\beta_j\right]
$$

Maximum Likelihood Error

This gives rise to a M dimensional Gaussian surface.

The formula for the uncertainty depends on the approximation that

$\mathcal{L}(a)$ is Gaussian-like in the region $a_i \approx a_i^\star$

H Is a symmetric matrix

Let U = unitary matrix which diagonalises H

$$U\,H\,U^{-1} = h$$

with h = diagonal matrix

Let $\beta = (\beta_1, \ldots \beta_M)$ and $\gamma = \beta U^{-1}$

Then the element of probability in $\beta -$space

$$d^M P = C \exp\left[-\frac{1}{2}(\gamma U) H (\gamma U)^{-1}\right] d^M \beta$$

Maximum Likelihood Error

The Jacobian relating the volume space $d^M\beta$ and $d^M\gamma$ is $|U| = 1$

$$d^M P = C \exp\left[-\frac{1}{2}\sum h_i \gamma_i^2\right] d^M\gamma$$

M -dimensional Gaussian surface → product of M independent one-dimensional Gaussians

$$\begin{aligned}
\overline{\gamma_a \gamma_b} &= \delta_{ab} h_a^{-1} \\
\overline{\beta_i \beta_j} &= \sum_{a,b} \overline{\gamma_a \gamma_b} U_{a_i} U_{b_j} \\
&= \sum U_{ia}^{-1} h_a^{-1} U_{aj} \\
&= \left(U^{-1} h U\right)_{ij}^{-1}
\end{aligned}$$

Thus $\overline{\left(a_i - a_i^\star\right)\left(a_j = a_j^\star\right)} = H^{-1} ij$ with $H_{ij} = -\frac{\partial^2 W}{\partial a_i \partial a_j}$

Averaging over repeated measurements

$$\overline{H_{ij}} = N \int \frac{1}{f}\left(\frac{\partial f}{\partial a_i}\right)\left(\frac{\partial f}{\partial a_j}\right) dx$$

Error Estimation

Measure the range of mono-energetic particles:

Let it follow Gaussian distribution with mean range a_1 and straggling coefficient a_2

$$\mathcal{L}(a_1, a_2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}a_2} \exp\left[-\frac{(x_i - a_1)^2}{2a_2^2}\right]$$

$$\Rightarrow W = \ln \mathcal{L} = -\frac{1}{2}\sum_i \frac{(x_i - a_1)^2}{a_2^2} - N \ln a_2 - \frac{N}{2}\ln(2\pi)$$

$$\frac{\partial W}{\partial a_1} = \sum_i \frac{(x_i - a_1)}{a_2^2}$$

$$\frac{\partial W}{\partial a_2} = \frac{1}{a_2^3}\sum_i (x_i - a_1)^2 - \frac{N}{a_2}$$

The maximum likelihood solution is obtained using $\left(\frac{\partial W}{\partial a_i}\right)_{a_i^\star} = 0$

$$a_1^\star = \frac{1}{N}\sum_i x_i$$

$$a_2^\star = \sqrt{\frac{\sum_i (x_i - a^\star)^2}{N}}$$

Error Estimation

The matrix H is obtained by evaluating

$$\frac{\partial^2 W}{\partial a_1^2} = -\frac{N}{a_2^2}\,;$$

$$\frac{\partial^2 W}{\partial a_2^2} = -\frac{3}{a_2^4}\sum_i (x_i - a_1)^2 + \frac{N}{a_2^2}\,;$$

$$\frac{\partial^2 W}{\partial a_1 \partial a_2} = -\frac{2}{a_2^2}\sum_i (x_i - a_1)\,.$$

Thus $H = \begin{pmatrix} \frac{N}{{a_2^\star}^2} & 0 \\ 0 & \frac{2N}{{a_2^\star}^2} \end{pmatrix} \Rightarrow H^{-1} = \begin{pmatrix} \frac{{a_2^\star}^2}{N} & 0 \\ 0 & \frac{{a_2^\star}^2}{2N} \end{pmatrix}$

This leads to

$$\Delta a_1 = \frac{a_2^\star}{\sqrt{N}}$$

$$\Delta a_2 = \frac{a_2^\star}{\sqrt{2\,N}}$$

These are errors on error and correlation between $a_1,\ a_2$ is zero

Error Propagation

Let a single physical quantity y be a function of M parameters:

$$y = y(a_1, \cdots a_M)$$

The best estimate of y is

$$y^\star = y(a_i^\star)$$

To the first order in $a_i - a_i^\star$

$$y - y^\star = \sum_i \frac{\partial y}{\partial a_j}(a_j - a_j^\star)$$

$$\overline{(y - y^\star)^2} = \sum_i \sum_j \frac{\partial y}{\partial a_i}\frac{\partial y}{\partial a_j}\overline{(a_i - a_i^\star)(a_j - a_j^\star)}$$

$$\Delta y_{RMS} = \sqrt{\sum_i \sum_j \frac{\partial y}{\partial a_i}\frac{\partial y}{\partial a_j} H_{ij}^{-1}}$$

Error Propagation

In general N physical parameters $y_1, \cdots y_N$ are known functions of M parameters $a_1, \cdots a_M$ whose error matrix is known

$$\overline{(y_l - y_l^\star)(y_m - y_m^\star)} = \sum_i \sum_j \frac{\partial y_l}{\partial a_i} \frac{\partial y_m}{\partial a_j} (H^{-1})_{ij}$$

If $\frac{\partial y_l}{\partial a_i}$ constitutes the derivative matrix D, then

$$(H')^{-1} = D\, H^{-1}\, D^\dagger$$

This is the basic principle of error propagation

Systematic Uncertainties

A variety of processes cause systematic effects in a measurement.

For example, nuclear or high-energy physics experiments study scatterings of particles or nuclei and measure the probability of interaction. In these measurements, there could be effects due to

- background
- selection bias
- detector acceptance
- detector inefficiency
- resolution in the measurement
- dead time
- ………..

In principle one can determine the systematic uncertainty by introducing it in the overall likelihood function through additional parameters $\Rightarrow$ likelihood solution will determine systematic uncertainty

Let us go back to the example of a beam of particles with mean range a_1 and straggling constant a_2 . Let there be an unknown background particle with uniform range distribution be present

Systematic Uncertainties

So the probability density function

$$f(a_1,\ a_2,\ x) \ = \ \frac{1}{\sqrt{2\pi}a_2}\exp\left[-\frac{(x-a_1)^2}{2\ a_2^2}\right]$$

is modified to

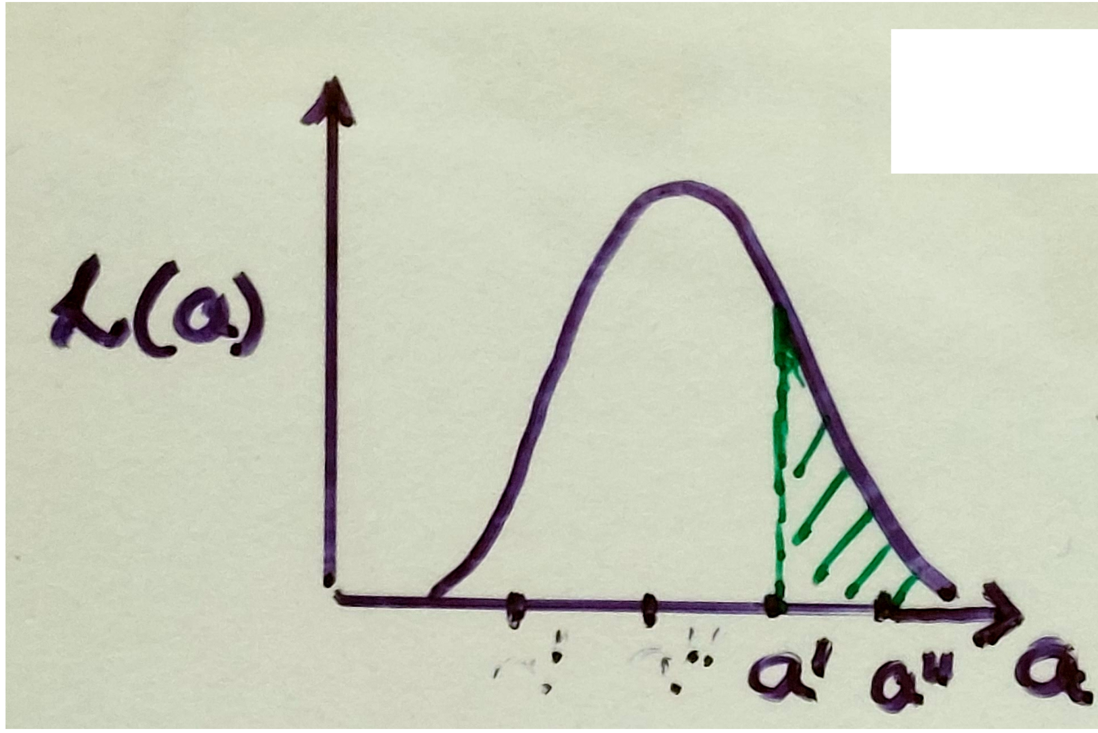
$$f(a_1,\ a_2,\ a_3,\ x) \ = \ \frac{1}{C}\left(a_3 + \frac{1}{\sqrt{2\pi}a_2}\exp\left[-\frac{(x-a_1)^2}{2\ a_2^2}\right]\right)$$

with $$C \ = \ \int_{x_{Min}}^{x_{Max}} f(x)\, dx$$

The solution $a_3^\star$ is related to the percentage of the background.

Confidence Interval

The probability that the parameter value a lies between a' and a'' $\Rightarrow$ Confidence Level



$$P(a' < a < a'') = \frac{\int_{a'}^{a''} \mathcal{L}(a)\, da}{\int_{-\infty}^{\infty} \mathcal{L}(a)\, da}$$

This depends on the arbitrariness of the choice of the parameter. For example, consider the area under the tail

$$P(a > a') = \frac{\int_{a'}^{\infty} \mathcal{L}(a)\, da}{\int_{-\infty}^{\infty} \mathcal{L}(a)\, da}$$

Now choose a physical parameter $\lambda = \lambda(a)$ with $\lambda' = \lambda(a')$

$$\begin{aligned} P(\lambda > \lambda') &= \frac{\int_{\lambda'}^{\infty} \mathcal{L}(\lambda)\, d\lambda}{\int_{-\infty}^{\infty} \mathcal{L}(\lambda)\, d\lambda} \\ &= \frac{\int_{\lambda'}^{\infty} \mathcal{L}(a) \frac{\partial \lambda}{\partial a} da}{\int_{-\infty}^{\infty} \mathcal{L}(\lambda)\, d\lambda} \\ &\neq P(a > a') \end{aligned}$$

Confidence Interval

The numerical value of the confidence interval depends on the choice of the physical parameter

Only the maximum likelihood solution and relative probabilities are unaffected by the choice of the parameter a

Bartlett S Function

This function is defined to avoid the arbitrariness of the confidence interval.

Define a function $S(a)$ with a mean of 0 and a standard deviation of 1
This is independent of the choice of a :

$$S(a) = \frac{1}{C}\frac{\partial W}{\partial a}$$

with

$$C^2 = \int_{a_{Min}}^{a_{Max}} \frac{\partial^2 W}{\partial a^2}\mathcal{L}(a)\, da$$

If $\mathcal{L}(a)$ is Gaussian with a mean $a^\star$ and standard deviation Δa

$$S(a) = \frac{a - a^\star}{\Delta a}$$

68.3% confidence interval in a can be obtained by solving for a in

$S(a') = +1$ and $S(a'') = -1$

95% (2 standard deviations) interval in a is obtained by solving

$S(a) = \pm 2$

Bartlett S Function

The mean:

$$\begin{aligned}\overline{S} &= \frac{1}{C}\int \frac{\partial W}{\partial a}\mathcal{L}(a)da \\ &= \frac{1}{C}\int \frac{\partial \mathcal{L}(a)}{\partial a}da \\ &= \mathcal{L}(a_{Max}) - \mathcal{L}(a_{Min}) = 0\end{aligned}$$

The mean squared: $\overline{S^2}$

$$\begin{aligned}\overline{S^2} &= \frac{1}{C^2}\int \frac{1}{\mathcal{L}^2}\left(\frac{\partial \mathcal{L}}{\partial a}\right)^2 \mathcal{L}(a)da \\ &= \frac{\int \frac{1}{\mathcal{L}}\left(\frac{\partial \mathcal{L}(a)}{\partial a}\right)^2 da}{-\int \frac{\partial}{\partial a}\left(\frac{1}{\mathcal{L}}\frac{\partial \mathcal{L}}{\partial a}\right)\mathcal{L}da} \\ &= \frac{\int \frac{1}{\mathcal{L}}\left(\frac{\partial \mathcal{L}(a)}{\partial a}\right)^2 da}{-\int \frac{\partial^2 \mathcal{L}}{\partial a^2}da + \int \frac{1}{\mathcal{L}}\left(\frac{\partial \mathcal{L}}{\partial a}\right)^2 da}\end{aligned}$$

Now

$$\int_{a_{Min}}^{a_{Max}} \frac{\partial^2 \mathcal{L}}{\partial a^2}da = \frac{\partial \mathcal{L}}{\partial a}|_{a_{Max}} - \frac{\partial \mathcal{L}}{\partial a}|_{a_{Min}} = 0$$

$\Rightarrow$

$$\overline{S^2} = 1$$

Binomial Distribution

Events belong to one of the two possible classes

If P is the probability that an event belongs to class 1, then the probability of observing N_1 out of N events in class 1:

$$P(N_1, N) = \frac{N!}{N_1!\,(N-N_1)!} p^{N_1} (1-p)^{N-N_1}$$

For a given experimental result of the above type, the Likelihood function

$$\begin{aligned}
\mathcal{L}(p) &= \frac{N!}{N_1!\,(N-N_1)!} p^{N_1} (1-p)^{N-N_1} \\
W &= N_1 \ln p + (N-N_1)\ln(1-p) + constant \\
\frac{\partial W}{\partial p} &= \frac{N_1}{p} - \frac{N-N_1}{1-p} \\
&= \frac{N_1 - N\,p}{p\,(1-p)} \\
\frac{\partial^2 W}{\partial p^2} &= -\frac{N_1}{p^2} - \frac{N-N_1}{(1-p)^2}
\end{aligned}$$

Binomial Distribution

$$\Rightarrow \quad \begin{aligned} p^\star &= \frac{N_1}{N} \\ \overline{(p-p^\star)^2} &= \frac{1}{\frac{N_1}{p^{\star 2}} + \frac{N-N_1}{(1-p^\star)^2}} \\ &= \frac{1}{\frac{N}{p^\star} + \frac{N}{1-p^\star}} \\ &= \frac{p^\star\,(1-p^\star)}{N} \\ \Delta p &= \sqrt{\frac{p^\star\,(1-p^\star)}{N}} \end{aligned}$$

There is a measurement of the count of electrons in the decay of muon in the forward hemisphere: (Remember. $f(a,x) = \frac{1}{2}(1+ax)$)

$$p = \int_0^1 \frac{1+ax}{2}dx = \frac{2+a}{4}$$

Observing frequency: $p^\star = \frac{N_1}{N}$

$$\Rightarrow \quad \begin{aligned} a^\star &= 4\,p^\star - 2 = 4\frac{N_1}{N} - 2 \\ \Delta a &= 4\,\Delta p = \sqrt{\frac{4}{N}\,(1-a^{\star 2})} \end{aligned}$$

Poisson Distribution

x/λ = probability of having an event at a distance x

$\Rightarrow$ Probability having 0 event in a length x

$$
\begin{aligned}
dP(0,\, x) &= -P(0,\, x)\frac{dx}{\lambda} \\
P(0,\, x) &= \exp(-x\,\lambda)
\end{aligned}
$$

since $\quad P(0,\, 0) = 1$

$P(N,\, x)$ = probability of having N events in length x

$$
\begin{aligned}
d^N P(N,\, x) &= \prod_{i=1}^{N}\left(\frac{dx_i}{\lambda}\right)\exp(-x\,/\lambda) \\
P(N,\, x) &= \frac{(x/\lambda)^N}{N!}\exp(-x\,/\lambda) \\
\overline{N} &= \sum_{N=1}^{\infty}\frac{N(x/\lambda)^N}{N!}\exp(-x\,/\lambda) = \frac{x}{\lambda}
\end{aligned}
$$

Poisson Distribution

The Likelihood function:

$$\begin{aligned}
\mathcal{L}(a) &= \frac{a^N}{N!}\exp(-a) \\
W &= N\ln a - a + constant \\
\frac{\partial W}{\partial a} &= \frac{N}{a} - 1 \\
\frac{\partial^2 W}{\partial a^2} &= -\frac{N}{a^2}
\end{aligned}$$

which leads to

$$\begin{aligned}
a^\star &= N \\
\Delta a &= \frac{a^\star}{\sqrt{N}}
\end{aligned}$$

while

$$\overline{a} = \frac{\int a\mathcal{L}(a)da}{\int \mathcal{L}(a)da} = \frac{\int a^{N+1}\exp(-a)da}{\int a^N \exp(-a)da} = \frac{(N+1)!}{N!} = N + 1 \neq a^\star$$

Extended Maximum Likelihood

In the standard likelihood formalism, the distribution functions are always normalised to unity

This requirement is strictly not necessary — one needs to use the correct probability of getting experimental distribution

$\Rightarrow$ Estimate the absolute normalisation

$F(x)dx$ = Probability of getting an event within an interval dx

$\Rightarrow$ The average number of events in the experiment when it is repeated many times

$$\overline{N}(a) = \int_{x_{Min}}^{x_{Max}} F(x)\, dx$$

Probability of getting no events in an interval Δx

$$\exp\left(-\int_x^{x+\Delta x} F(x)\, dx\right)$$

Probability of getting no events in the entire interval $x_{Min} < x < x_{Max}$

$$\exp\left(-\int_{x_{Min}}^{x_{Max}} F(x)\, dx\right) = \exp\left(-\overline{N}\right)$$

Extended Maximum Likelihood

The element of probability for a particular experimental result of N events at $x = x_1, \cdots x_N$ is

$$
\begin{aligned}
d^N P &= \exp(-\overline{N}) \prod_{i=1}^{N} F(x_i)\, dx_i \\
\Rightarrow \mathcal{L}(a) &= \exp(-\overline{N}(a)) \prod_{i=1}^{N} F(a, x_i) \\
W(a) &= \sum_{i=1}^{N} \ln F(a, x_i) - \int_{x_{Min}}^{x_{Max}} F(a, x)\, dx
\end{aligned}
$$

The solutions $a_i = a_i^\star$ are still given by M -simultaneous equations

$$\frac{\partial W}{\partial a_i}\Big|_{a_i = a_i^\star} = 0$$

where the error is $\overline{(a_i - a_i^\star)(a_j - a_j^\star)} = (H^{-1})_{ij}$

with $H_{ij} = -\frac{\partial^2 W}{\partial a_i \partial a_j}$

N does not appear explicitly in the formula

$$-\frac{\partial^2 W}{\partial a_i \partial a_j} = \int \frac{1}{F} \frac{\partial F}{\partial a_i} \frac{\partial F}{\partial a_j} dx$$

Maximum Likelihood Method

Maximum Likelihood

Likelihood function : N independent observables $x_1, x_2, ...x_n$ from a thec retical distribution $f(x;\theta)$, where θ is the parameter to be estimated.

Likelihood is defined as $\mathcal{L}(\theta;x) = f(x_1;\theta)f(x_2;\theta)....f(x_1;\theta)$, where θ can b found by solving the equation, $\frac{\partial\mathcal{L}}{\partial\theta} = 0$.
In general use $\frac{\partial ln\mathcal{L}}{\partial\theta} = 0$ and variance, $\sigma^2(\overline{\theta}) = \int(\theta-\overline{\theta})^2\mathcal{L}(\theta;x)dx_1dx_2...dx_n$.

But, very difficult to get analytic solution, thus uses an approximate method i the limit of the large numbers,

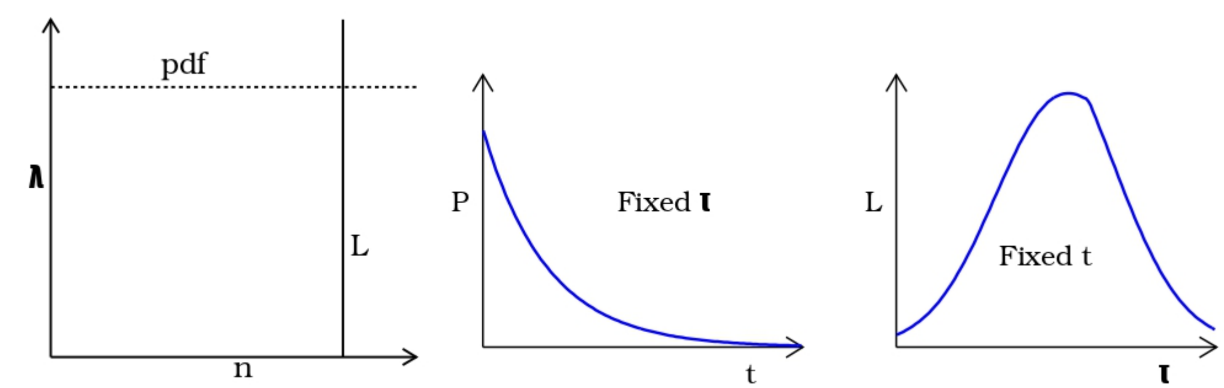
$$\sigma^2(\overline{\theta}) = \left(-\frac{\partial^2 ln\mathcal{L}}{\partial\theta^2}\right)^{-1}, \Longrightarrow U_{ij} = -\frac{\partial^2 ln\mathcal{L}}{\partial\theta_i\partial\theta_j} \Longrightarrow \sigma^2(\overline{\theta_i}) = (U^{-1})_{ii}$$

Maximum Likelihood method

- Angular distribution, $y = \frac{dn}{d\cos\theta} = N(1 + b\cos^2\theta)$
- Normalised y behaves as a probability distribution
- Normalisation factor, $N = \frac{1}{2(1+b/3)}$ is coming from $\int_{-1}^{1} y\, d\cos\theta = 1$
- Normalisation is essential because its dependence on parameter b, factc $1/2$ is not crucial, just a scale, but need for error estimation
- For the ith event, $y_i = N(1+b\cos^2\theta_i)$ is the probability density for observin that event, e.g., θ_i for a given value b.
- Define likelihood, $\mathcal{L}$ as the product of y_i for all events or called as joir probability function, $\mathcal{L}(b;\theta) = \prod y_i$, which is the probability of observin given set of θ_i for that b.
- For true (theory) value of b, $\mathcal{L}(b;\theta)$ is maximum, or inversely maximize as a function of b to find the best value of b.
- Without normalising factor, N, one can make y_i larger simply by increasin b, hence $\mathcal{L}$ would not have any absolute maximum

- Similarly for lifetime fit to $t_1, t_2, ...t_n$, $\tau = \sum_i t_i/n$, if one misses $1/\tau$ in the expression of probability $P(t/\tau) = (1/\tau)exp(-t/\tau)$, probability will larger for any larger value of τ.

Likelihood $\mathcal{L}$ vs probability distribution function (pdf)



- Poisson : pdf for observing n, given λ is $P(n;\lambda) = e^{-\lambda}\lambda^n/n!$
  From this construction $\mathcal{L}$ as $\mathcal{L}(\lambda;n) = \Pi_i\, e^{-\lambda}\lambda^{n_i}/n_i!$
  Use same function for λ and n, but
  for pdf, λ is fixed : $P(n;\lambda)$ exists only at integer $n \geq 0$
  for likelihood, n is fixed : $\mathcal{L}(\lambda;n)$ exists as continuous function of $\lambda \geq 0$
- Lifetime distribution :
  $P(t;\tau) = (1/\tau)e^{-t/\tau}$ is maximum at $t = 0$
  $\mathcal{L}(\tau;t) = \Pi_i\,(1/\tau)e^{-t_i/\tau}$ is maximum at $\tau =< t >$
  Both t and τ are continuous
- Gaussian : Same functional form $(1/(\sqrt{2\pi}\sigma))e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$
  If only consider Gaussian, one can confused between pdf and $\mathcal{L}$
  Integration of pdf = 1, whereas integration of $\mathcal{L}$ is meaningless

Transformation properties of $\mathcal{L}$ and probability densities :

Example of lifetime :
$dn/dt = \lambda e^{-\lambda t}$

Multi-Parameter Fit

change observable from t to y, where y=$\sqrt{t}$
$\therefore dn/dt = dn/dt.dt/dy = \lambda e^{-\lambda y^2} 2y$
$\therefore$

1. pdf changes, but
2. $\int_{t_0}^{\infty}(dn/dt)dt = \int_{y_0}^{\infty}(dn/dy)dy$
3. Maximum probability density, Not very sensitive

In contrast $\mathcal{L}$, which is not a pdf for λ
When parameter change from λ to $\tau = 1/\lambda$,

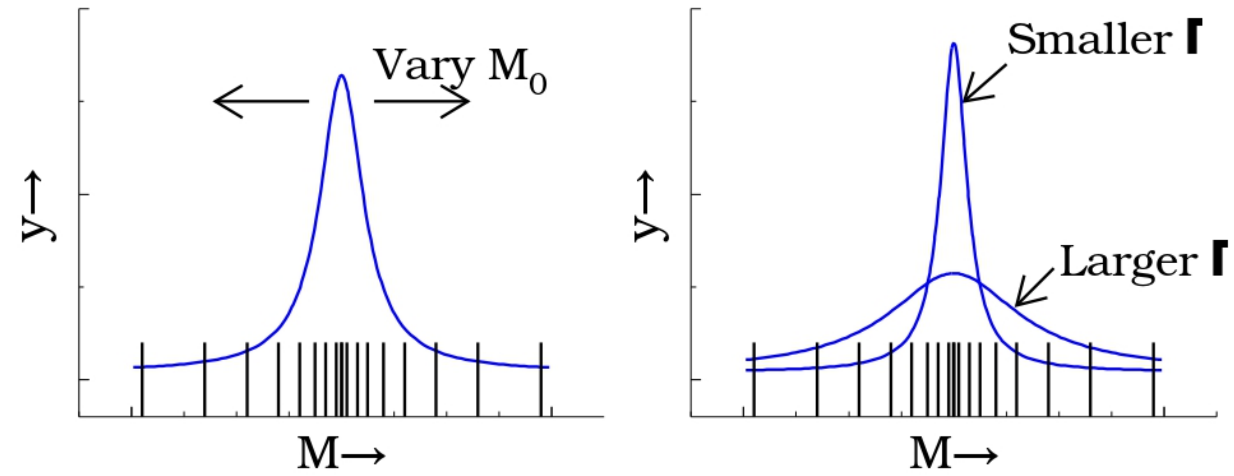
1. $\mathcal{L}$ does not change, pdf changes, $dn/dt = (1/\tau)e^{-t/\tau}$
   $\mathcal{L}(\tau; t) = \mathcal{L}(\lambda = 1/\tau; t)$, because identical number occur in evaluation of two $\mathcal{L}$'s, but
2. $\int_0^{\lambda_0} \mathcal{L}(\lambda; t)d\lambda \neq \int_{\infty}^{1/\tau_0} \mathcal{L}(\tau; t)dt$
3. It is not meaningful to integrate $\mathcal{L}$.

Likelihood function for multi-parameter fit

A case of Breit-Wigner function, where two free parameters in the function (M_0 and Γ) probability density function is

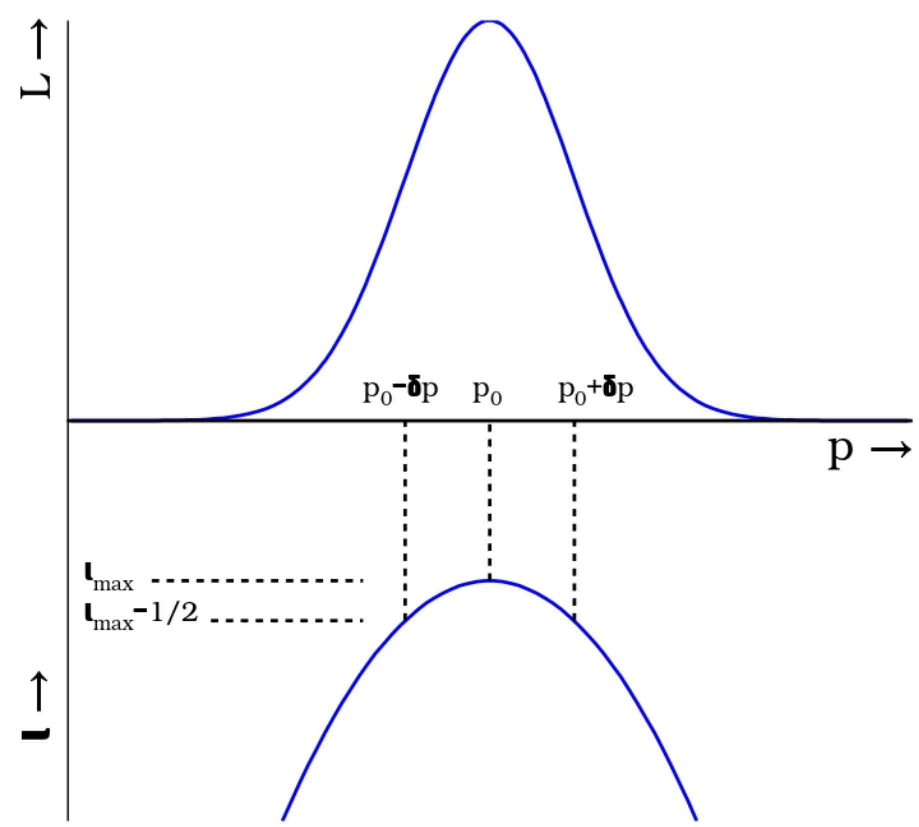
$$\mathcal{L}(M_0, \Gamma; m_i) = \prod_i y_i(M_0, \Gamma), \quad \text{where} \quad y_i(M_0, \Gamma) = \frac{1}{2\pi}\frac{\Gamma}{(m_i^2 - M_0)^2 + (\Gamma/2)^2}$$

The maximum $\mathcal{L}$ give the best value of M_0 and Γ.



$\mathcal{L}(M_0(\Gamma); m_i) = \prod_i y_i(M_0(\Gamma))$ for fixed $\Gamma(M_0)$, one parameter function, when we know second one.

Properties of likelihood function



$\prod_i y_i$ is a very small number for a large number of measurements(N), difficult to calculate.

Error Estimation

Conventional to consider $\ell = ln(\mathcal{L}) = \sum_i y_i$
For large N, likelihood function, $\mathcal{L}$ tends to Gaussian, atleast near the vicinity of maximum,

$$\ell = \ell_{max} + \frac{1}{2!}\frac{\partial^2 \ell}{\partial p^2}(\delta p)^2 + ... = \ell_{max} - \frac{1}{2c}(\delta p)^2 + ...,$$

$$\textbf{where} \quad -\frac{1}{c} = \frac{\partial^2 \ell}{\partial p^2} \Longrightarrow \mathcal{L} = e^{\frac{-(p-p_0)^2}{2c}}$$

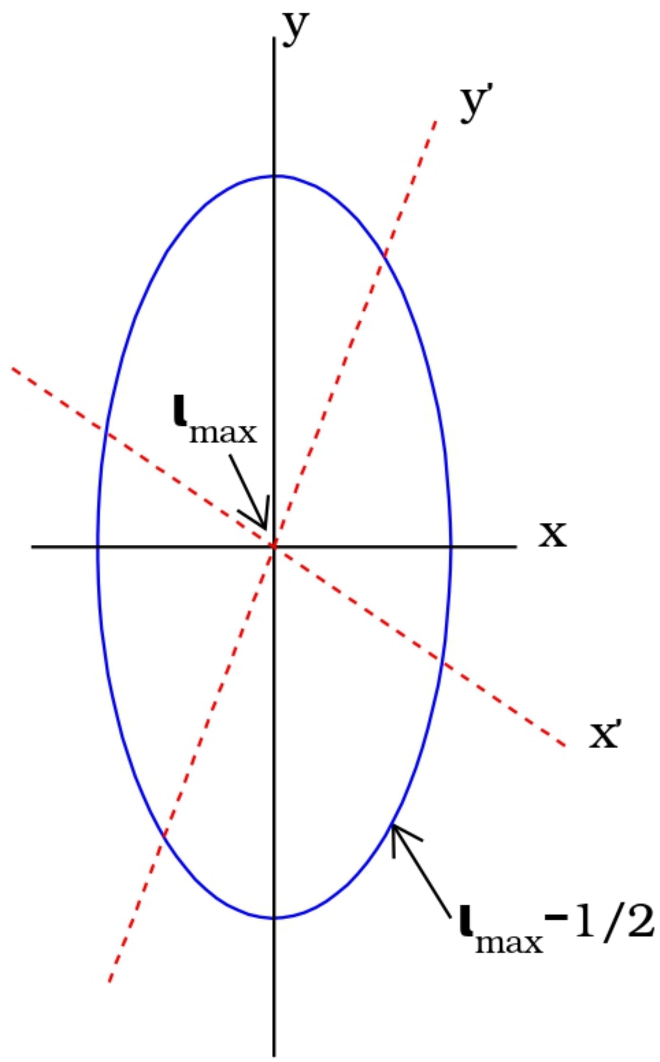
Calculation of p (probable value) is very simple, but what is its error ?

- RMS deviation of the $\mathcal{L}$ distribution about its mean
- $(-\partial^2 \ell/\partial p^2)^{-1/2}$
- The change in p required to reduce ℓ from its maximum value by 1/2, i.e., $\ell(p_0 \pm \partial p) = \ell(p_0) - 1/2$. Error can be different in $+$ve and $-$ve side.
- For non Gaussian $\mathcal{L}$, use second or third option
- Try to avoid non-Gaussian situation, e.g., use decay rate $(1/\tau)$, rather than lifetime, τ or use $1/p$ rather than momentum, p of charge track

Error in maximum likelihood fit

Numerical m

Error Estimation



- For single parameter fit, parameter p is estimated from eqn $\partial\ell/\partial p = 0$ and error, $\sigma = (-\partial^e\ell/\partial p^2)^{-1/2}$
- For multivariate p_i, their best value is obtained from the set of equation $\partial\ell/\partial p_i = 0$
- For error, define $H_{ij} = (-\partial^2\ell/\partial p_i\partial p_j)$ and obtained error matrix as $E_{ij} = (H^{-1})_{ij}$

An example : Two variables x and y. Contours of ℓ are $\ell = -(4x^2 + y^2)$, where ℓ_{max} at $(0,0)$ and $\ell = \ell_{max} - 1/2$ when $8x^2 + 2y^2 = 1$.

Errors on variables, $x = \pm\sqrt{1/8}$ (when y=0)
$y = \pm\sqrt{1/2}$ (when x=0)

Pros and Cons of Likelihood Method

In terms of H_{ij},

$$-\frac{\partial^2 \ell}{\partial x_i \partial x_j} = \begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix} \xrightarrow{invert} \frac{1}{16}\begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}$$

Rotate axes clockwise by 30°

$$-\frac{\partial^2 \ell}{\partial x_i \partial x_j} = \frac{1}{2}\begin{pmatrix} 13 & 3\sqrt{3} \\ 3\sqrt{3} & 7 \end{pmatrix} \xrightarrow{invert} \frac{1}{64}\begin{pmatrix} 7 & -3\sqrt{3} \\ -3\sqrt{3} & 13 \end{pmatrix}$$

Mean and error in Gaussian function through Likelihood function

Measurement of same variable with different error

$$\mathcal{L} = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} exp(-\frac{(x_i-\mu)^2}{2\sigma_i^2}), \quad \ell = ln\mathcal{L} = \sum_i -ln\sigma_i + \sum_i -\frac{(x_i-\mu)^2}{2\sigma_i^2} + const$$

$$\frac{\partial \ell}{\partial \mu} = \sum_i \frac{x_i - \mu}{\sigma_i^2} \Longrightarrow \bar{\mu} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\sum_i \frac{1}{\sigma_i^2} \Longrightarrow \sigma^2(\bar{\mu}) = \frac{1}{\sum_i \frac{1}{\sigma_i^2}} \equiv \frac{\sigma^2}{N}$$

More weight on the measurement with less uncertainty.

What is the average value of count rate 1 ± 1 and 100 ± 10 ?

Calculation 2 ± 1, but simple guess 50.5 ± 5 ?

Assumed true value of σ_i, which are similar in these two measurements and expected same rate.

Don't blindly average the experimental data.

Pros & cons of Likelihood method

- No need of histogramming. Most useful for low statistics.
- Unique answer and error (e.g.,$\lambda_0 \pm \delta\lambda$ or $\tau_0 \pm \delta\tau$, then $\lambda_0 \pm \delta = 1/(\tau_0 \mp \delta\tau)$).
- Able to constrain parameters and ranges.
- But, difficult to tackle background, $\ell = \sum_i ln\,(f_R w_R(m_i) P_R(\theta_i) + f_B w_B(m_i) P_B(\theta_i))$.
- Use weight factor (e.g., efficiency, $\epsilon = 1/w$) for different events, $\ell = \sum_i w_i \ln(y_i)$, but difficult to estimate error, better option is $\ell = \sum ln(N\, y_i\, \epsilon_i)$,
  $y_i\,\epsilon_i$ = distribution with efficiency, N is the normalisation factor with ϵ ??????
- Large computing time, normalisation for each parameter set separately.
- How good the fit is ? No limit on the value of $\mathcal{L}$. Hypothesis testing is not easy, but can compare two hypotheses. $(\ell_a - \ell_b)$ can just give better choices of models.

Relation between likelihood and χ^2 fit

For a measurement of $x_1 \pm \sigma_1, x_2 \pm \sigma_2, ..., x_n \pm \sigma_n$ of an underlying theory of expected value, μ,

$$\mathcal{L} = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} exp(-\frac{(x_i-\mu)^2}{2\sigma_i^2}), \quad \ell = ln\mathcal{L} = \sum_i -ln\sigma_i + \sum_i -\frac{(x_i-\mu)^2}{2\sigma_i^2} + const$$

$$-2\ell = -2\,ln\mathcal{L} = \sum_i \frac{(x_i-\mu)^2}{\sigma_i^2} + const$$

Maximisation of 2ℓ is same as minimisation of

$$\chi^2 = \sum_i \frac{(x_i-\mu)^2}{\sigma_i^2}$$

An Example

Let there be a sample which contains two radioactive species:

$a_1,\ a_2$	lifetime of the two species
$a_3,\ a_4$	the initial decay rates of each of the species
x	denotes the time

The probability density function is then

$$F(a_i, x) = a_3 \exp\left(-\frac{x}{a_1}\right) + a_4 \exp\left(-\frac{x}{a_2}\right)$$

Solution to the Problem

Standard Method:

To determine a_i 's, use the normalised probability distribution function

$$f(a_i, x) = \frac{\exp\left(-\frac{x}{a_1}\right) + a_5\left(-\frac{x}{a_2}\right)}{a_1 + a_5 a_2}$$

with $$a_5 = \frac{a_4}{a_3}$$

$a_3,\ a_4$ would be determined using the auxiliary equation

$$\begin{aligned} \int_0^\infty F dx &= N \\ \Rightarrow \overline{N}(a_i) &= N \end{aligned}$$

Extended maximum likelihood method:

$a_1, \cdots a_4$ are determined directly from F

$\overline{N}(a_i)$ need not be N

The Least Square Method

Let p measurements at points $x_1, \cdots x_p$ lead to the experimental results $(y_1 \pm \sigma_1), \cdots (y_p \pm \sigma_p)$

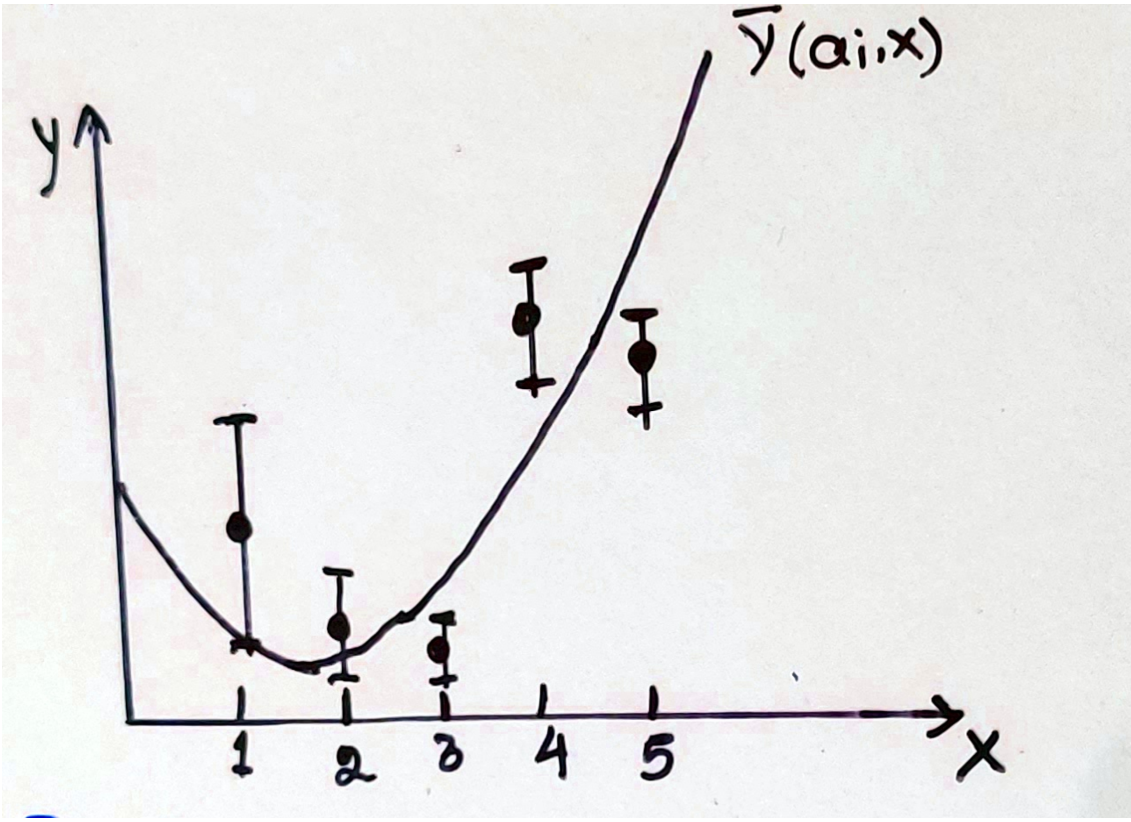
e.g. Each experiment consists of event counting (measurement i with N_i events)

Then $y_i = N_i$ and are Poisson distributed with $\sigma_i = \sqrt{N_i}$

$$\Rightarrow \mathcal{L} = \prod_{i=1}^{p} \frac{[\overline{y}(x_i)]^{N_i}}{N_i!} \exp\left(-\overline{y}(x_i)\right)$$

$$W = \sum_{i=1}^{p} N_i \ln \overline{y}(x_i) - \sum_{i=1}^{p} \overline{y}(x_i) + constant$$

$\overline{y}(a_i, x)$ denotes the curve fitted through the experimental points



The Least Square Method

The best fit corresponds to $a_i = a_i^\star$ coming from M-simultaneous equations:

$$\sum_{j=1}^{p} \frac{\partial \overline{y}}{\partial a_i}(x_j) = \sum_{j=1}^{p} \frac{N_j}{\overline{y}(x_j)} \frac{\partial \overline{y}(x_j)}{\partial a_i}$$

If y_i 's are Gaussian distributed with standard deviations σ_i

$$\mathcal{L} = \prod_{j=1}^{[} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{\left(y_j - \overline{y}(x_j)\right)^2}{2\sigma_j^2}\right]$$

$$W = -\frac{1}{2}M - \sum_{j=1}^{p} \ln \sqrt{2\pi}\sigma_j$$

with $$M = \sum_{j=1}^{p} \frac{\left[y - \overline{y}(x_j)\right]^2}{\sigma_j^2}$$

Solutions $a_i = a_i^\star$ can be obtained by minimising M or maximising W using

$$\frac{\partial M}{\partial a_i} = 0$$

The Least Square Method

Let $M^\star$ = minimum value of M (least square sum)

Values of a_i which minimise M $\Rightarrow$ the least square solution

Here, the least square and maximum likelihood solutions are identical.

Least square errors:

$$\overline{(a_i - a_i^\star)(a_j - a_j^\star)} = \left(H^{-1}\right)_{ij} \quad \text{with} \quad H_{ij} = \frac{1}{2}\frac{\partial^2 M}{\partial a_i \partial a_j}$$

Example of Least Squared Method

$\bar{y}$ = linear in a_i

$$\bar{y}(a_i, x) = \sum_{j=1}^{M} a_j f_j(x)$$

Then $$\frac{\partial M}{\partial a_i} = -2 \sum_{j=1}^{p} \frac{\left[y_j - \sum_{k=1}^{M} a_k f_k(x_j) \right]}{\sigma_j^2} f_i(x_j)$$

and $$H_{ij} = \sum_{k=1}^{p} \frac{f_i(x_k)\, f_j(x_k)}{\sigma_k^2}$$

Let $$U_i = \sum_{k=1}^{p} \frac{y_k f_i(x_k)}{\sigma_k^2}$$

Then $$\frac{\partial M}{\partial a_i} = -2 \left[u_i - \sum_{k=1}^{M} a_k H_{ki} \right]$$

In matrix notation;

$$\begin{aligned} \underline{U} - \underline{a^\star H} &= 0 \\ \underline{a^\star} &= \underline{U}\ \underline{H}^{-1} \\ &= \sum_{j=1}^{M} \sum_{k=1}^{p} \frac{y_k\, f_j(x_k)}{\sigma_k^2} \left(H^{-1}\right)_{ji} \end{aligned}$$

Example of Least Squared Method

Let the curve be a parabola

$$\bar{y} = a_1 + a_2 x + a_3 x^2$$

Then

$$H_{11} = \sum_k \frac{1}{\sigma_k^2}$$

$$H_{12} = \sum_k \frac{x_k}{\sigma_k^2}$$

$$H_{22} = \sum_k \frac{x_k^2}{\sigma_k^2}$$

$$H_{13} = \sum_k \frac{x_k^2}{\sigma_k^2}$$

$$H_{23} = \sum_k \frac{x_k^3}{\sigma_k^2}$$

$$H_{33} = \sum_k \frac{x_k^4}{\sigma_k^2}$$

X	Y
−0.6	5±2
−0.2	3±1
0.2	5±1
0.6	8±2

$\Rightarrow$ Get a_1, a_2, a_3

The Least Square Method

It is customary to denote $\overline{y}$ as an improved set of estimate over the measured values of y

$$\overline{y_j} = y_j + c_j$$

The least squared sum is

$$M \equiv \chi^2 = C^\dagger H C$$

with H^{-1} = the error matrix

In the definition, one assumes independent variables, i.e., H^{-1} has no non-zero off-diagonal terms

We can start with correlated variables y and then transform them to independent variables t using a set of linear transformation

$$t = A\, y \quad \text{such that}$$

$$A\, H_y^{-1}\, A^\dagger = H_t^{-1} \quad \text{is diagonal}$$

The Least Square Method

If H_y^{-1} is non-singular, there must exist A^{-1} such that

$$A^{-1}\,H_t^{-1}\,(A^{-1})^{\dagger} = H_y^{-1}$$

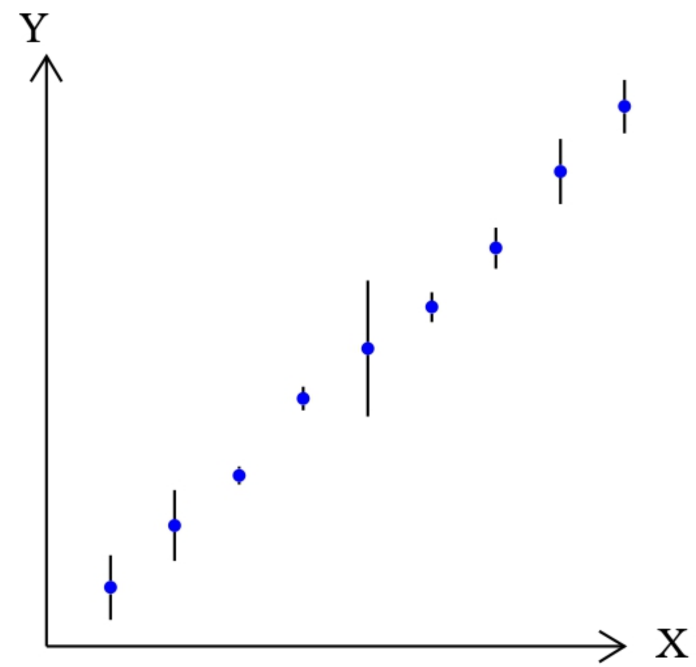
In t -space, one can always get solutions from

$$M \equiv \chi^2 = C_t^{\dagger}\,H_t\,C_t$$

and transform them back to y -space

$$\begin{aligned} C_t^{\dagger}\,H_t\,C_t &= (AC_y)^{\dagger}\,H_t\,(AC_y) \\ &= C_y^{\dagger}A^{\dagger}H_tAC_y \\ &= C_y^{\dagger}H_yC_y \end{aligned}$$

Chi-Square Fit



- Data set $\{x_i, y_i \pm \delta y_i\}$ and theory y=a+bx
- Does it fit in a straight line ? (Hypothesis testing)
- What is gradient and intercept ? (parameter determination)
- $$\chi^2 = \sum_i \frac{(y_i^{obs} - y_i^{th}(\alpha_j))^2}{\sigma_i^2}$$
- σ_i suppose to be error on theory, but in reality it is the error on experimental observation.
  - Simpler : σ_i is free of α_j
  - Different weight on individual point, ..
- If theory and data are consistent with each other, $y^{th} \sim y^{obs}$, χ^2 is small
- Bin size
- Poisson error, ideally +ve and −ve errors are different

Chi-Square Fit

- Not too small, which has a larger error. Also for low statistics Poisson error is different from Gaussian (Interpretation of error interms of Gaussian function).
- Not too large, above detector resolution, peak will not be visible.
- All bins need not have the same size, but preferable.

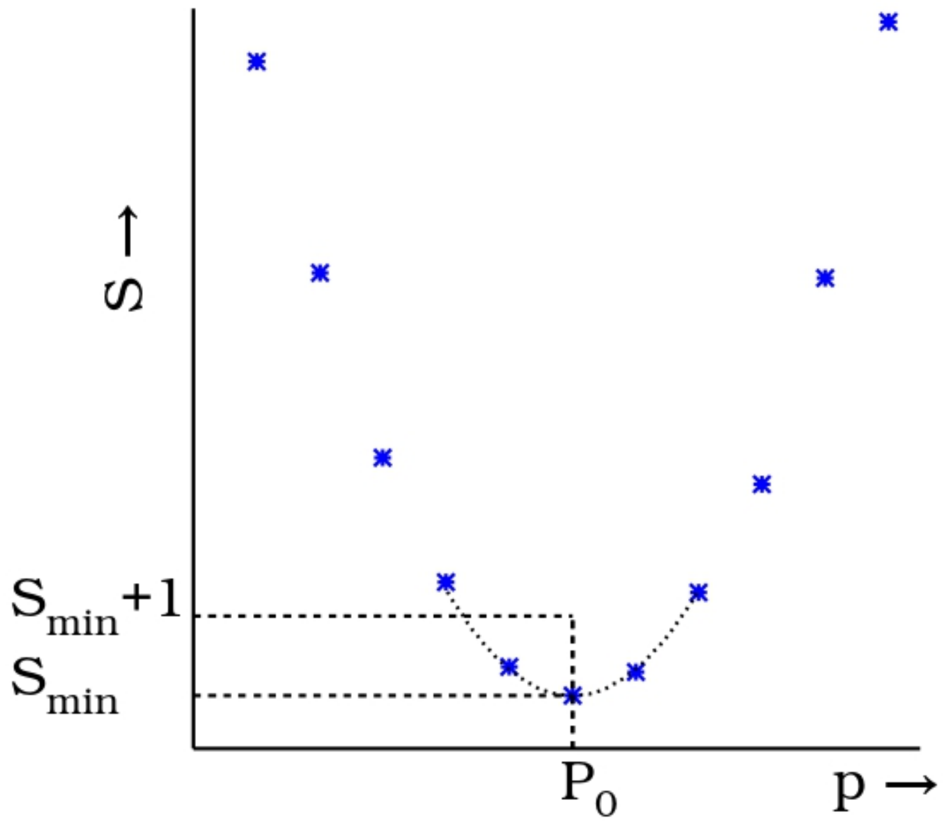
- Minimize χ^2 to obtain the best line (best parameter of the theory)
- The error on parameter $\left(\frac{1}{2}\frac{d\chi^2}{dp^2}\right)^{-1/2}$ or increase χ^2 by one from its minima, χ^2_{min}
- For multi parameters, their best values are obtained from equations, $\partial\chi^2/\partial p_i = 0$ and error matrix is the inverse of $\left(\frac{1}{2}\frac{\partial^2\chi^2}{\partial p_i \partial p_j}\right)$
-
- For single measurement, $y^{obs} \pm \sigma$, $\chi^2 = \frac{(y^{obs}-y^{th})^2}{\sigma^2}$
- Minimises to $\chi^2 = 0$ for $y^{obs} = y^{th}$ and $\chi^2 = 1$ for $y^{obs} = y^{th} \pm \sigma$
- For two measurements, y_1 and y_2 of single quantity with equal error

$$\chi^2 = \frac{(y_1 - y^{th})^2}{\sigma^2} + \frac{(y_2 - y^{th})^2}{\sigma^2}$$

$y^{th} = (y_1 + y_2)/2$ from $d\chi^2/dy^{th} = 0$ and its error $= \sigma/\sqrt{2}$ from $\chi^2 = \chi^2_{min} + 1$, exactly what we expect, error $= \sigma/\sqrt{n}$

Some Examples

Simple example of minimising χ^2



Measurements, $p_1 \pm \sigma_1, p_2 \pm \sigma_2, ...p_n \pm \sigma_n$. The best value $p \pm \sigma$

$$\text{Construct } \chi^2 = \sum_i \frac{(p_i - \bar{p})^2}{\sigma_i^2} \text{ and minimise } \chi^2 \text{ wrt } \bar{p}$$

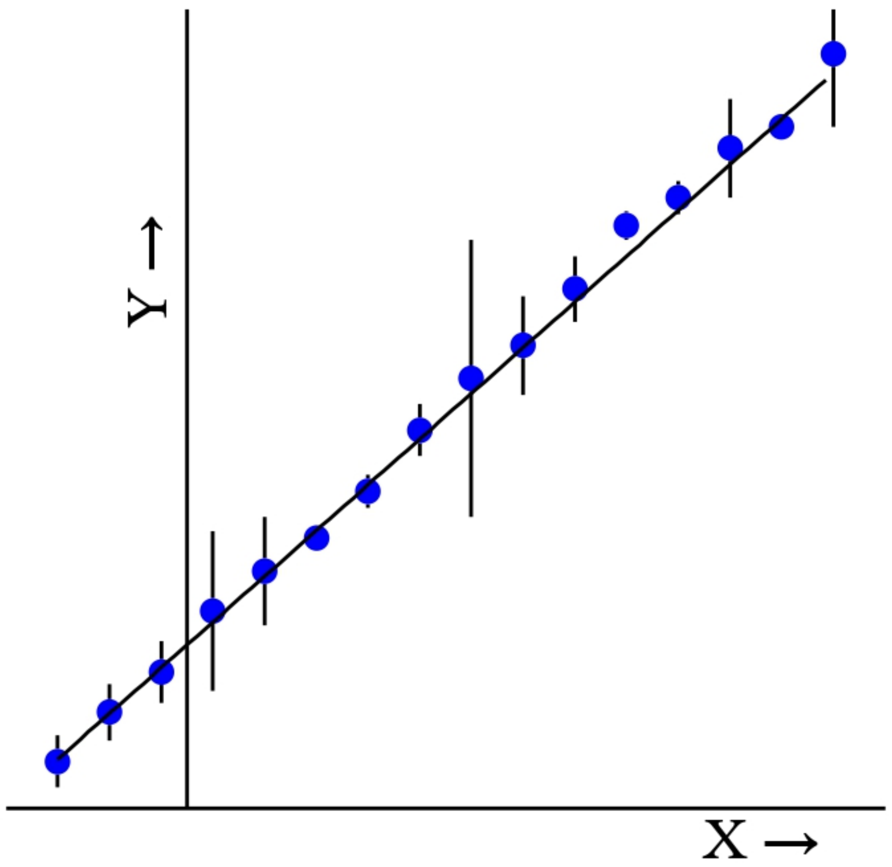
$$\frac{d\chi^2}{dp} = 2\sum_i \frac{p_i - \bar{p}}{\sigma_i^2} = 0 \Longrightarrow \bar{p} = \sum_i \frac{p_i}{\sigma_i^2} / \sum_i \frac{1}{\sigma_i^2}$$

and error on $\bar{p}$ obtained from

$$\sigma = \left(\frac{1}{2}\frac{d^2\chi^2}{dp^2}\right)^{-1/2}; \quad \frac{d^2\chi^2}{dp^2} = 2\sum_i \frac{1}{\sigma_i^2}; \text{ thus } \frac{1}{\sigma^2} = \sum_i \frac{1}{\sigma_i^2}$$

Some Examples

Example of straight line fit



A simple example y=a+bx; simple because y is linear in a and b, not in x.

Data consists of n points $(x_i, y_i \pm \sigma_i)$
Lots of lines for different χ^2. Minimise χ^2 wrt a and b.

$$\chi^2 = \sum_i \frac{(y_i - a - bx_i)^2}{\sigma_i^2};$$

$$\frac{1}{2}\frac{d\chi^2}{da} = -\sum_i \frac{(y_i - a - bx_i)}{\sigma_i^2}; \qquad \frac{1}{2}\frac{d\chi^2}{db} = -\sum_i \frac{(y_i - a - bx_i)x_i}{\sigma_i^2}$$

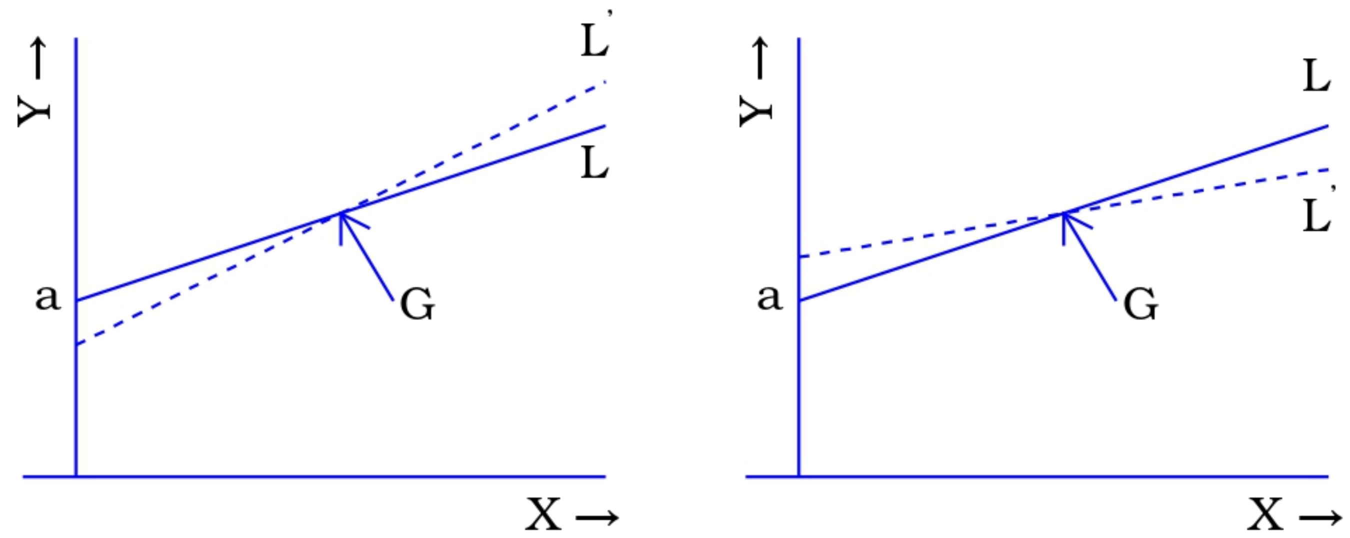
Simultaneous two equations for two unknown a and b. Solution of these equations

$$b = \frac{[1][xy] - [x][y]}{[1][x^2] - [x][x]}, \text{ where } [f] = \frac{1}{n}\sum_i \frac{f_i}{\sigma_i^2}$$

Weighted mean $< f >= [f]/[1]$ and a can be determined from $< y >=< a > +b < x >$

Error Estimation

Error on straight line fit



To evaluate error, one needs an error matrix, a and b are correlated. First evaluate $\left(\frac{1}{2}\frac{\partial^2\chi^2}{\partial p_i \partial p_j}\right)$

$$\frac{1}{2}\frac{\partial^2\chi^2}{\partial a^2} = n[1]; \qquad \frac{1}{2}\frac{\partial^2\chi^2}{\partial b^2} = n[x^2]; \qquad \frac{1}{2}\frac{\partial^2\chi^2}{\partial a \partial b} = n[x]$$

So, inverse error matrix $n\begin{pmatrix} [1] & [x] \\ [x] & [x^2] \end{pmatrix}$

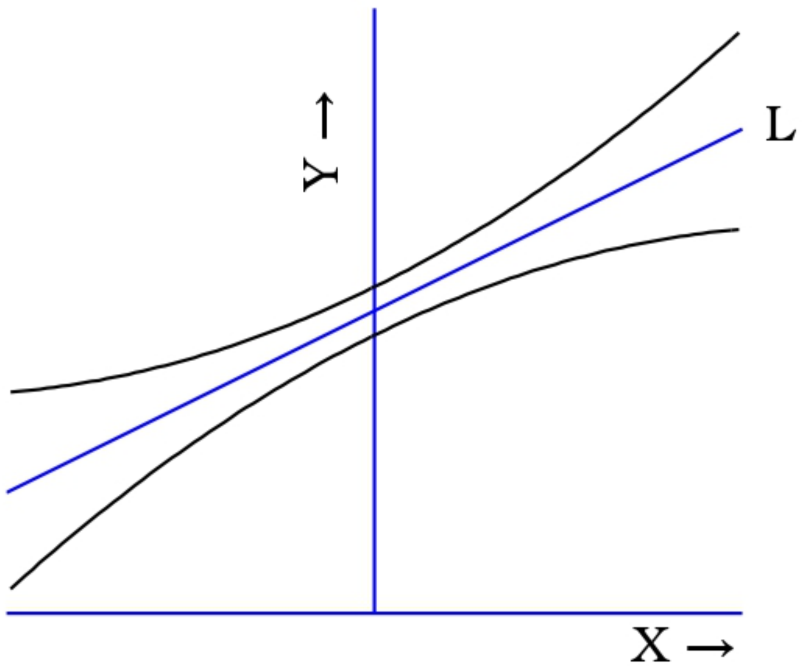
and error matrix $\frac{1}{nD}\begin{pmatrix} [x^2] & -[x] \\ -[x] & [1] \end{pmatrix}$

Where determinant, $D = [x^2][1] - [x][x]$ and for no correlation $\sigma^2(a) = 1/[1]$ and $\sigma^2(b) = 1/[x^2]$

Errors depend only on the measured variable x_i, σ_i, but not on how well the data agrees with the theory.

$\mathrm{cov}(a, b) = - < x >$. Better to use x' $(= x - < x >)$, because errors on a' and b' are uncorrelated.

Error Estimation



- How well is the y-coordinate of the fitted line known for a particular x-value ?

  Variance of y, $\sigma_y^2 = \sigma_a + 2x\ cov(a,b) + x^2\sigma_b^2$.

  For proper shift in x such that $< x >= 0,\ \sigma_y^2 = \sigma_a + x^2\sigma_b^2$

- Hypothesis testing : How well the data points matched with the theoretical expectation ? Probability of getting as large as this in a χ^2 distribution (χ^2/ndf)

$$\chi^2_{min} = \sum_i \frac{y_i^2}{\sigma_i^2} - a\sum_i \frac{y_i}{\sigma_i^2} - b\sum_i \frac{x_i y_i}{\sigma_i^2}$$

- Error on first kind : Reject H when it is true, should happen $x\%$ of time
- Error on second kind : Accept H when something else is true
- Optimise χ^2 criteria for the best result

Minimisation Procedure

Minimisation procedure

Due to complex expression of likelihood/χ^2-function, there is no analytical solution of many experimental data points. There are mainly two different kinds of approaches to look at this problem numerically, e.g.,
1. Grid/Random Search and
2. Gradient search

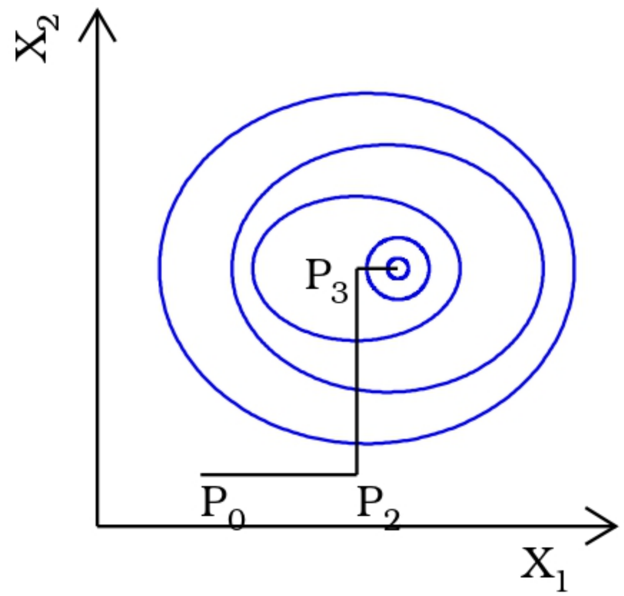
Grid Search :

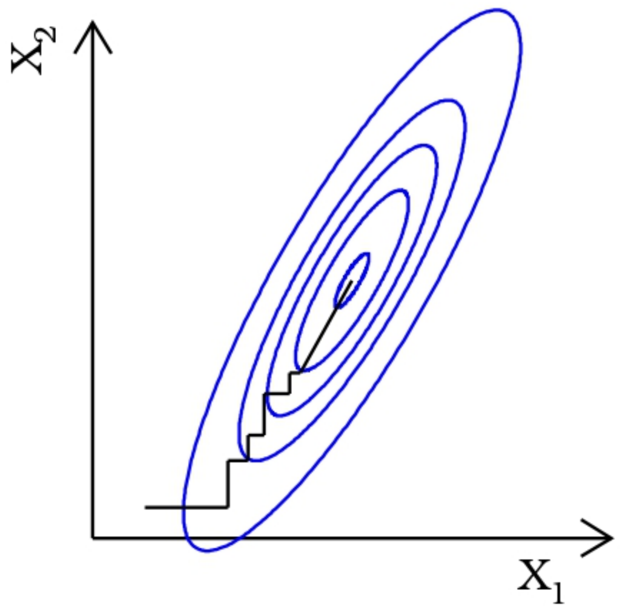
Grid : Evaluate $F(x)$ at $x_0, x_0+\Delta x, x_0+2\Delta x$, ... and look for minimum $F(x)$.
Only suitable for finite range of search and smaller dimension.

Random Search : Instead of equally spaced points, generate points according to some function; better for a large range of parameter space and larger dimensions.
But, may not get any true minima.

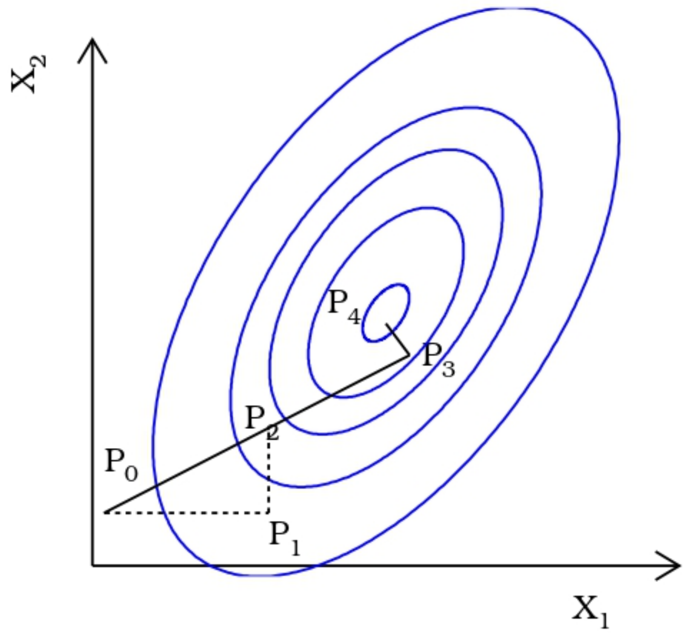
Minimisation Procedure

Coordinate variation method or single parameter or one-by-one variation method. Vary one parameter and get minima, then next one ...
Need large number of steps,
In case of strongly correlated variables, this is unacceptably slow.





Rosenburg method : Single parameter method but get the best direction after the first steps of each dimension.
Efficiency decreases with the number of variables.



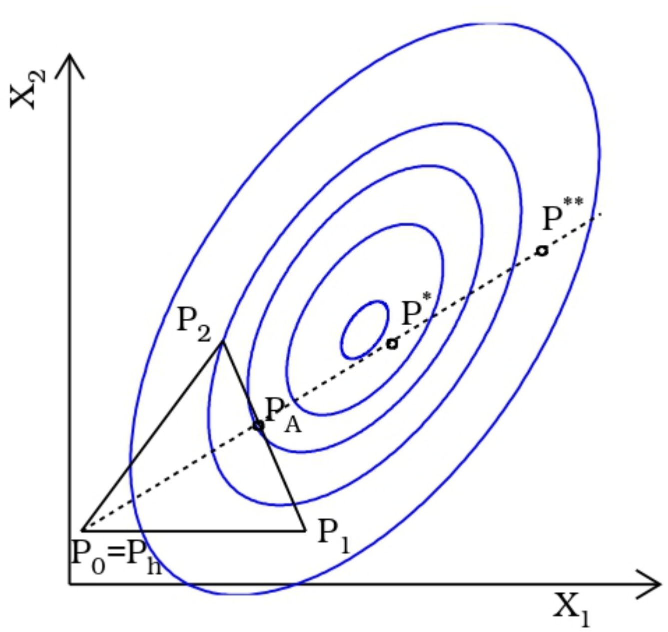
Minimisation Method

Simplex method (default in Minuit): Take n+1 points

$F(P_h) = max(F(p_1), F(P_2), ...)$
$F(P_{lo}) = min(F(p_1), F(P_2), ...)$
New line through P_{lo} and $P_A = \frac{1}{n}\left(\sum_{i=1}^{n+1} p_i - P_h\right)$
Three operations can be used, reflection, contraction and expansion.
Reflecting P_h about P_A, $P^* = (1+\alpha)P_A - \alpha P_h$, where α is a +ve constant

1. if $F(P^*) < F(P_{lo})$, has produces a new minima and see next step
   $P^{**} = \gamma P^* + (1-\gamma)P_A$, where expansion co-efficient, γ >1
   if $F(P^{**}) < F(P_{lo})$, replace P_h by P^{**} and restart
   if $F(P^{**}) \geq F(P_{lo})$, replace P_h by P^* and restart
2. if $F(P_{lo}) \leq F(P^*) \leq F(P_h)$, $P_h = P^*$ and restart
3. if $F(P^*) \geq F(P_h)$, reflection is failed, P^* is unacceptable
   New point P^{**} between P_h and P_A, such that
   $P^{**} = \beta P_h + (1-\beta)P_A, 0 < \beta < 1$
   if $F(P^{**} < F(P_h), P_h = P^{**}$
   if $F(P^{**}) > F(P_h) \& F(P^*)$, failed, all P_i are replaced by $\frac{1}{2}(P_i + P_{lo})$
   and restart whole process.

α, β, γ are free parameters, minimisation depends on those, recommended values are 1, 0.5 and 2 respectively



Minimisation Method

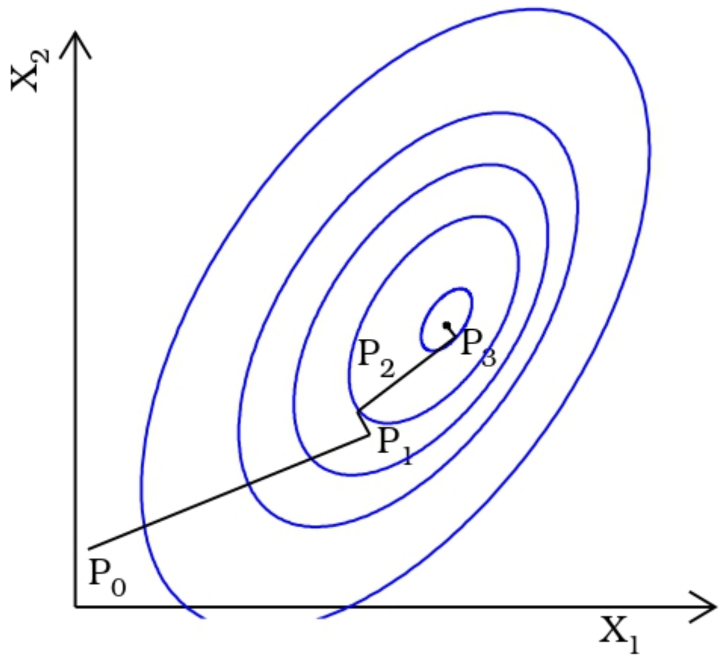
Gradient method : predict new points relatively far from the last point

Steepest descent method : From P_0 seek a minimum of parameter space, where function decreases most rapidly.

$$\xi_i = \frac{\frac{\partial F}{\partial x_i}}{\left[\sum_{j=1}^{n}\left(\frac{\partial F}{\partial x_j}\right)^2\right]^{1/2}}$$

, $i, j = 1, 2, ...n$

In 2D, it is same as 'one-by-one variation method', but with a rotation of co-ordinate axis.



Very slow due to complete interdependency in F and choice of the starting point far from true value.

Minimisation Method

Newton's method : Second degree of taylor expansion

$$
\begin{aligned}
F(x) &= F(x_0) + \frac{\partial F}{\partial x}|_{x_0}(x - x_0) + \frac{1}{2}\frac{\partial^2 F}{\partial^2 x}|_{x_0}(x - x_0)^2 + \ldots \\
F(x) &\approx F(x_0) + g^T(x - x_0) + \frac{1}{2}(x - x_0)^T G(x - x_0) \\
g_i &= \frac{\partial F}{\partial x_i} \quad and \quad G_{ij} = \frac{\partial^2 F}{\partial x_i \partial x_j}
\end{aligned}
$$

Approximate a function around x_0 by a quadratic surface. Calculate the minimum of the n dimensional parabola analytically

$$x_{min} = x_0 - G^{-1}g$$

$$\frac{\partial \chi^2}{\partial \alpha_a}(\gamma_1, ...\gamma_b) = \frac{\partial \chi^2}{\partial \alpha_a}(\beta_1, ...\beta_b) + \sum_c \frac{\partial}{\partial \alpha_c}\frac{\partial \chi^2}{\partial \alpha_a}(\beta_1, ...\beta_b) \times (\gamma_c - \beta_c)$$

Left hand side is zero by definition, where γ_i are true value and corrections are $(\delta = \gamma_c - \beta_c) = -G^{-1}g$

- Is not a true minimum, but forming a new parabolic surface about X_{min} and calculating its minimum.
- It requires G everywhere +ve definite, when it is negative, artificially altered
- Disadvantage : Evaluation of G and inversion : large CPU time

Local versus Global

Local vs Global minima

- Obtained minima may be a local minima, but not a global minima. Change starting value far away from initial one and look for second minima.....
- Errors : From the second derivative of the function in the minimum

$$
\begin{aligned}
F(x) &= F(x_0) + \frac{1}{2}\frac{\partial^2 F}{\partial x^2}(x - x_0)^2 \quad \text{and} \quad \sigma^2 = \left(\frac{1}{2}\frac{\partial^2 F}{\partial x^2}\right)^{-1} \\
\Longrightarrow & \; F(x_0 + \sigma) = F(x_0) + 1
\end{aligned}
$$

For correlated variables, calculate all derivatives and invert that second derivative matrix to obtain error.