Statistics for Particle Physicists

Lecture 2: Parameter Estimation

Summer Student Lectures
CERN
9 – 12 July 2024

https://indico.cern.ch/event/1347523/timetable/

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

Lecture 1: Introduction, probability,

→ Lecture 2: Parameter estimation

See exercises on fitting with iminuit here
and on least squares with curve_fit here.

Lecture 3: Hypothesis tests

Lecture 4: Introduction to Machine Learning

Hypothesis, likelihood

Suppose the entire result of an experiment (set of measurements) is a collection of numbers $\boldsymbol{x}$.

A (simple) hypothesis is a rule that assigns a probability to each possible data outcome:

$P(\mathbf{x}|H)$ = the likelihood of H

Often we deal with a family of hypotheses labeled by one or more undetermined parameters (a composite hypothesis):

$P(\mathbf{x}|\boldsymbol{\theta}) = L(\boldsymbol{\theta})$ = the “likelihood function”

Note:

1) For the likelihood we treat the data $\boldsymbol{x}$ as fixed.
2) The likelihood function $L(\boldsymbol{\theta})$ is not a pdf for $\boldsymbol{\theta}$.

The likelihood function for i.i.d.* data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: $x_1, \ldots, x_n$, where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

In this case the likelihood function is

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

Parameter estimation

The parameters of a pdf are any constants that characterize it,

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

r.v. parameter

i.e., θ indexes a set of hypotheses.

Suppose we have a sample of observed values: $\boldsymbol{x} = (x_1, \dots, x_n)$

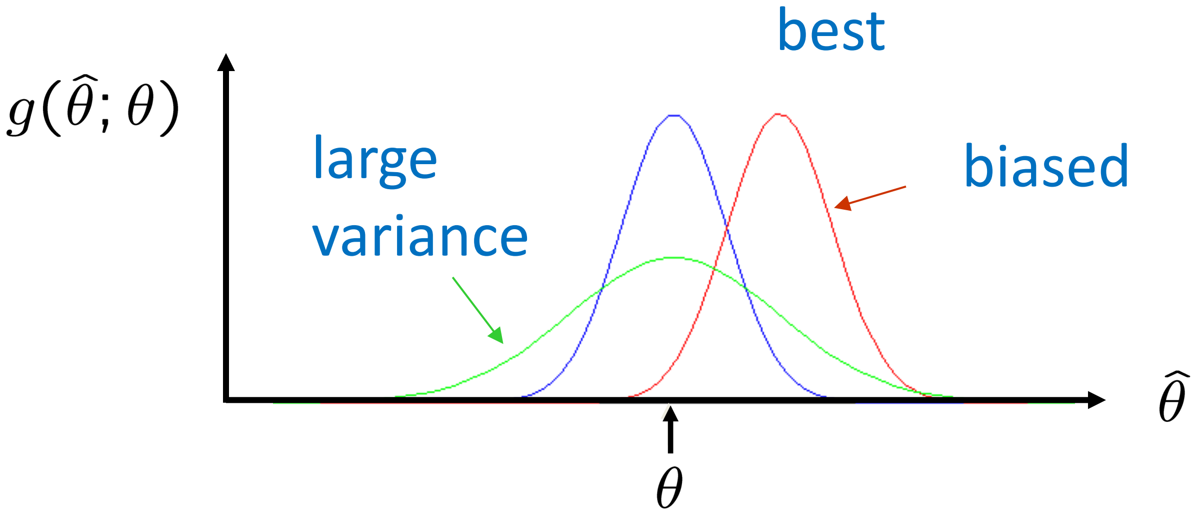
We want to find some function of the data to estimate the parameter(s):

$$\boxed{\widehat{\theta}(\vec{x})} \quad \leftarrow \text{ estimator written with a hat}$$

Sometimes we say 'estimator' for the function of $x_1, \dots, x_n$;
'estimate' for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error): $b = E[\widehat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

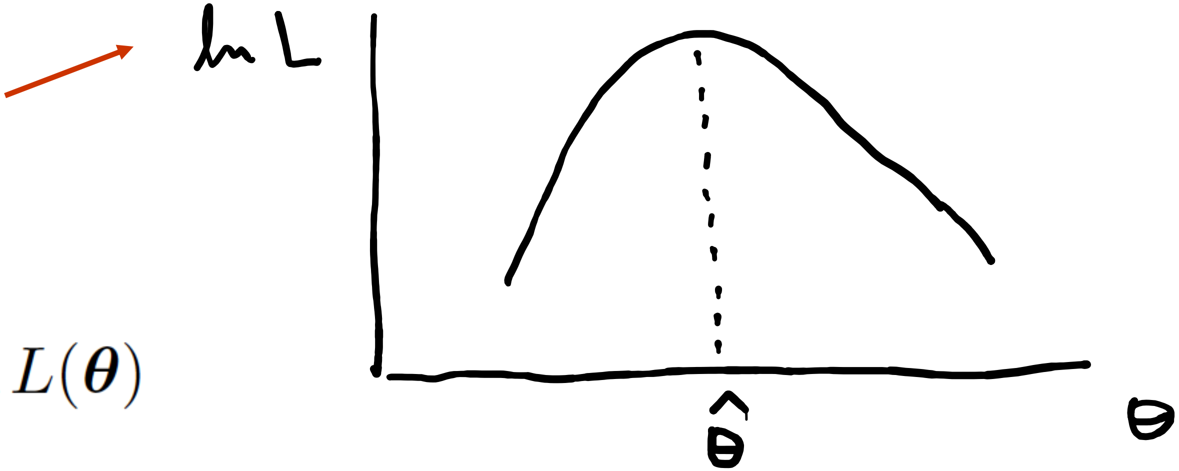
And we want a small variance (statistical error): $V[\widehat{\theta}]$

→ small bias & variance are in general conflicting criteria

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$



$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$$

Could have multiple maxima (take highest).

MLEs not guaranteed to have any ‘optimal’ properties, (but in practice they’re very good).

MLE example: parameter of exponential pdf

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

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

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

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

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

MLE example: parameter of exponential pdf (2)

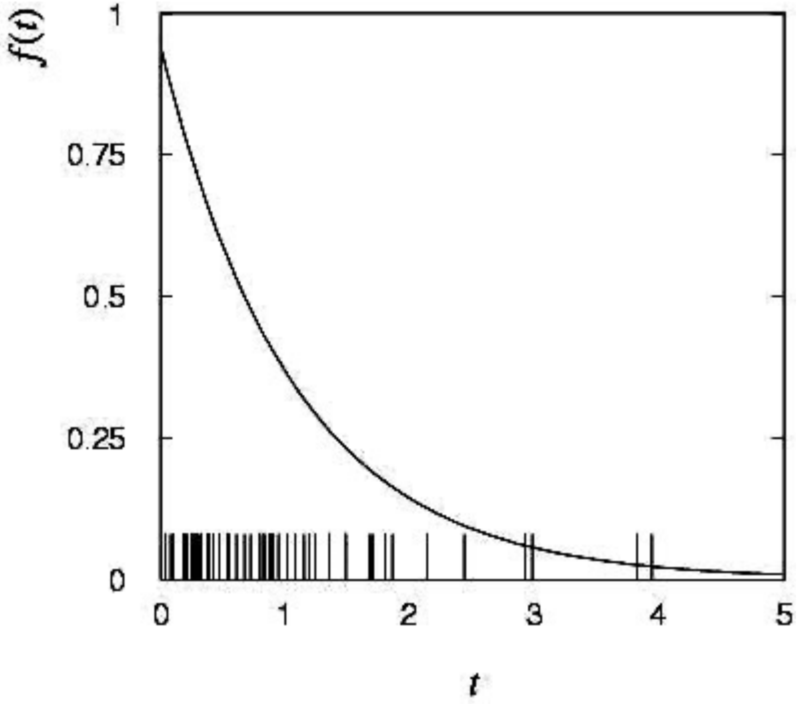
Find its maximum by setting $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

$$\rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Monte Carlo test:

generate 50 values

using $\tau = 1$:



We find the ML estimate:

$$\hat{\tau} = 1.062$$

MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t-\tau)^2 \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the MLE $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n} \sum_{i=1}^{n} t_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n} \sum_{i=1}^{n} t_i \right] = \frac{1}{n^2} \sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

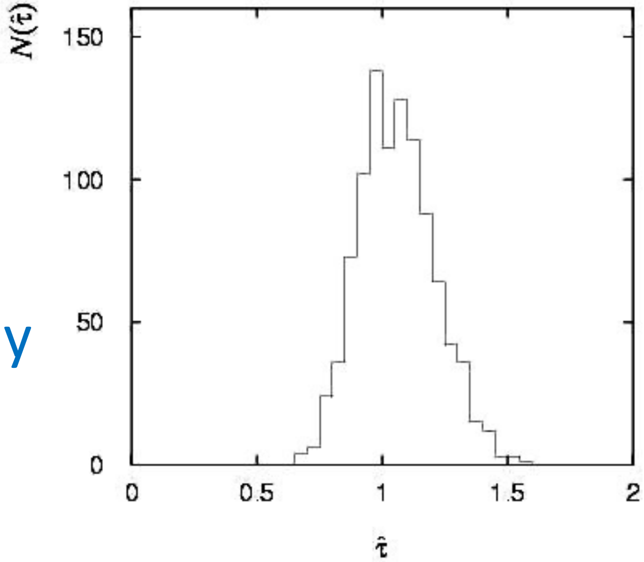
Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\widehat{\sigma}_{\widehat{\tau}} = 0.151$$



Note distribution of estimates is roughly Gaussian − (almost) always true for ML in large sample limit.

Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

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

Minimum Variance Bound (MVB)

$$(b = E[\widehat{\theta}] - \theta)$$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\widehat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

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

MVB for MLE of exponential parameter

Find $$\text{MVB} = -\left(1+\frac{\partial b}{\partial \tau}\right)^2 \Bigg/ E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right]$$

We found for the exponential parameter the MLE $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$

and we showed $b = 0$, hence $\partial b / \partial \tau = 0$.

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

and since $E[t_i] = \tau$ for all i, $$E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right] = -\frac{n}{\tau^2} \,,$$

and therefore $\text{MVB} = \frac{\tau^2}{n} = V[\hat{\tau}]$. (Here MLE is "efficient").

Variance of estimators: graphical method

Expand $\ln L(\theta)$ about its maximum:

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

First term is $\ln L_{\max}$, second term is zero, for third term use information inequality (assume equality):

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

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

→ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by 1/2.

Example of variance by graphical method

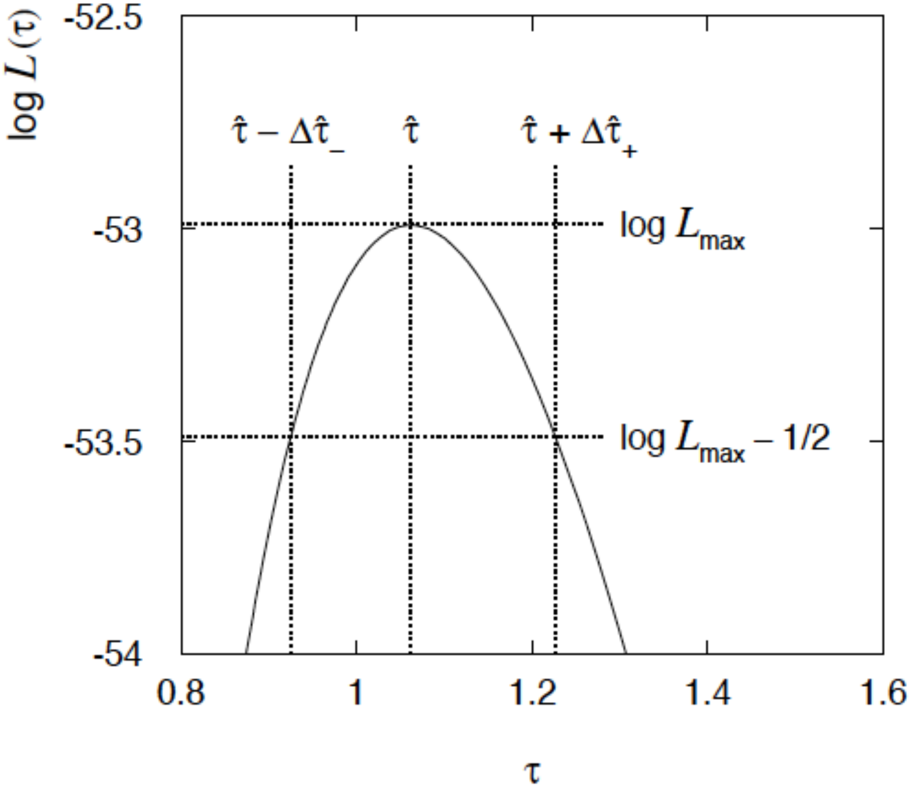
ML example with exponential:

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$

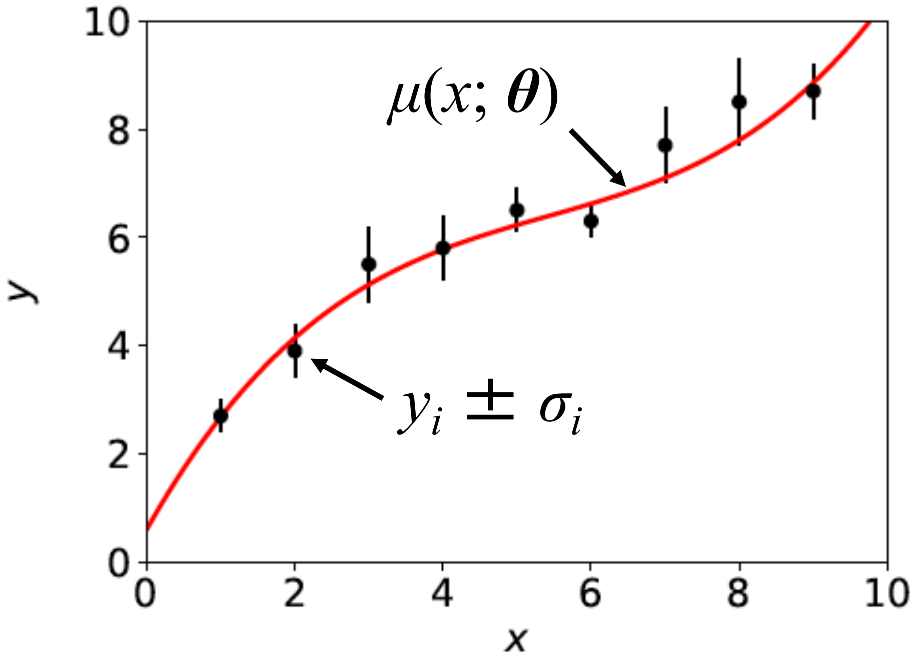


Not quite parabolic $\ln L$ since finite sample size (n = 50).

Curve fitting

Consider N independent measured values y_i, $i = 1,.., N$.

Each y_i has a standard deviation σ_i, and is measured at a value x_i of a control variable x known with negligible uncertainty:



The goal is to find a curve $\mu(x; \boldsymbol{\theta})$ that passes “close to” the data points.

Suppose the functional form of $\mu(x; \boldsymbol{\theta})$ is given; goal is to estimate its parameters $\boldsymbol{\theta}$ (= “curve fitting”).

Gaussian likelihood function → LS estimators

Suppose the measurements $y_1, \ldots, y_N$, are independent Gaussian r.v.s with means $E[y_i] = \mu(x_i; \boldsymbol{\theta})$ and variances $V[y_i] = \sigma_i^2$, so the the likelihood function is

$$L(\boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i;\boldsymbol{\theta}))^2/2\sigma_i^2}$$

The log-likelihood function is therefore

$$\ln L(\boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} + \text{const.}$$

So maximizing the likelihood is equivalent to minimizing

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} = -2 \ln L(\boldsymbol{\theta}) + \text{const.}$$

The minimum of $\chi^2(\boldsymbol{\theta})$ defines the least squares (LS) estimators $\hat{\boldsymbol{\theta}}$.

Information inequality for N parameters

Suppose we have estimated N parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_N)$

The *Fisher information matrix* is

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} P(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}$$

and the covariance matrix of estimators $\hat{\boldsymbol{\theta}}$ is $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

$$M_{ij} = V_{ij} - \sum_{k,l} \left(\delta_{ik} + \frac{\partial b_i}{\partial \theta_k}\right) I^{-1}_{kl} \left(\delta_{lj} + \frac{\partial b_l}{\partial \theta_j}\right)$$

is positive semi-definite:

$z^{\mathrm{T}} M z \geq 0$ for all $z \neq 0$, diagonal elements ≥ 0

Information inequality for N parameters (2)

In practice the inequality is ~always used in the large-sample limit:

bias → 0

inequality → equality, i.e, $M = 0$, and therefore $V^{-1} = I$

That is, $V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$

This can be estimated from data using $\widehat{V}_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\hat{\boldsymbol{\theta}}}$

Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

$$\widehat{V}_{ij} = \widehat{\text{cov}}[\hat{\theta}_i, \hat{\theta}_j]$$

Variance of LS estimators for Gaussian data

If $y_i \sim$ Gauss, then we found $\ln L(\boldsymbol{\theta}) = -\frac{1}{2}\chi^2(\boldsymbol{\theta}) + \text{const.}$

To the extent this (approximately) holds, we can use

$$U_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

and so we estimate the inverse covariance matrix with

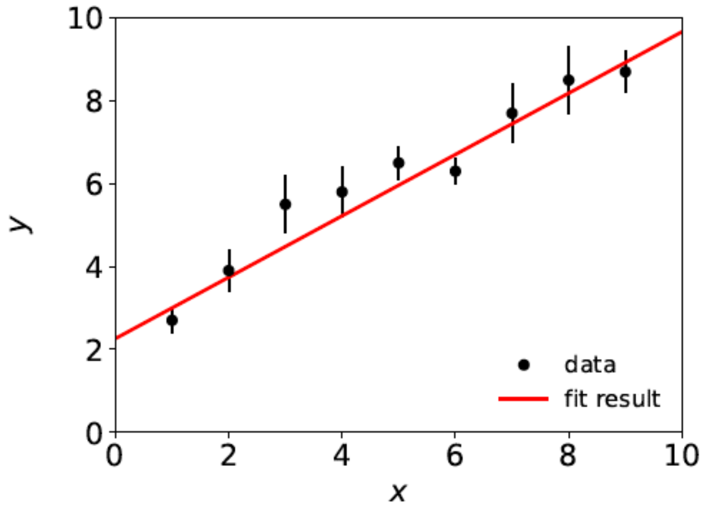
$$\widehat{U}_{ij}^{-1} = - \left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{1}{2} \left.\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

and invert to estimate the covariance matrix U.

For Gaussian data with the linear LS problem, U is the minimum variance bound (the LS estimators are unbiased and efficient).

Covariance from derivatives of $\chi^2(\boldsymbol{\theta})$

This is what programs like **curve_fit** and **MINUIT** do (derivatives computed numerically). Example with straight-line fit gives:



$$\hat{\theta}_0 = 2.258$$

$$\hat{\theta}_1 = 0.741$$

$$\sigma_{\hat{\theta}_0} = 0.29\ ,$$

$$\sigma_{\hat{\theta}_1} = 0.057\ ,$$

$$\mathrm{cov}[\hat{\theta}_0, \hat{\theta}_1] = -0.014\ ,$$

$$\rho = -0.86\ .$$

$$U = \begin{pmatrix} 0.08537 & -0.01438 \\ -0.01438 & 0.003275 \end{pmatrix}$$

The contour $\chi^2(\boldsymbol{\theta}) = \chi^2_{\min} + 1$

If $\mu(x;\, \boldsymbol{\theta})$ is linear in the parameters, then $\chi^2(\boldsymbol{\theta})$ is quadratic:

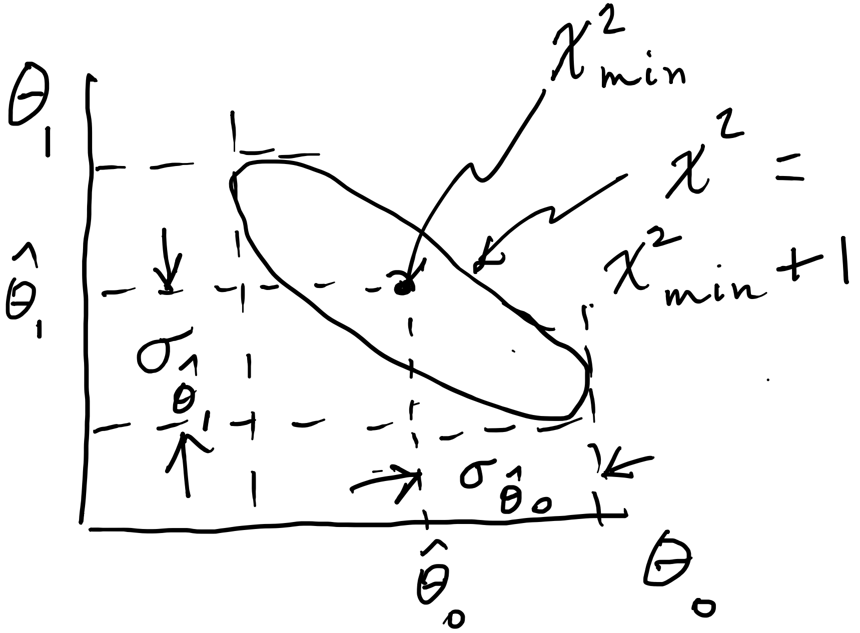
$$\chi^2(\boldsymbol{\theta}) = \chi^2(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \sum_{i,j=1}^{M} \left. \frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

$$= \chi^2_{\min} + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T U^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

Standard deviations from tangents to (hyper-) planes of

$$\chi^2(\boldsymbol{\theta}) = \chi^2_{\min} + 1$$

(corresponds to $\ln L(\boldsymbol{\theta}) = \ln L_{\max} - ½$)



χ^2_{min}
$\chi^2 = \chi^2_{min} + 1$
θ_1
$\hat{\theta}_1$
$\sigma_{\hat{\theta}_1}$
$\sigma_{\hat{\theta}_0}$
$\hat{\theta}_0$
θ_0

Extra slides

LS with correlated measurements

If $\boldsymbol{y}$ ~ multivariate Gaussian with covariance matrix $V_{ij} = \mathrm{cov}[y_i, y_\mathrm{j}]$

$$f(\mathbf{y};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right]$$

where $\boldsymbol{\mu}^\mathrm{T} = (\mu(x_1),\ldots,\mu(x_N))$, then maximizing the likelihood is equivalent to minimizing

$$\chi^2(\boldsymbol{\theta}) = (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

$$= \sum_{i,j=1}^{N} (y_i - \mu(x_i;\boldsymbol{\theta})) V_{ij}^{-1} (y_j - \mu(x_j;\boldsymbol{\theta}))$$

LS with correlated measurements (2)

For the special case of a diagonal covariance matrix, i.e., uncorrelated measurements. Then

$$V = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} \quad \rightarrow \quad V^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & 0 & \dots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_n^2 \end{pmatrix}$$

$V^{-1}{}_{ij} = \delta_{ij}/\sigma_i^2$, carry out one of the sums, $\chi^2(\boldsymbol{\theta})$ same as before:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i,j=1}^{N} (y_i - \mu(x_i;\boldsymbol{\theta}))\frac{\delta_{ij}}{\sigma_i^2}(y_j - \mu(x_j;\boldsymbol{\theta})) = \sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2}$$