

4 Least Squares and Maximum Likelihood

❖ methods to infer (physics) parameters from a set of measurements

- ▣ properties which are assumed to be satisfied

- ▶ individual measurements fluctuate with **known variance**
- ▶ individual measurements are **unbiased**

- ▣ starting point: measurements of the same physical quantity

- ▶ n measurements y_i with $i = 1, 2, \dots, n$
- ▶ all measurements fluctuate around an unknown true value μ
- ▶ the measurements have the standard deviation σ_i
- ▶ each measurement is an estimate for μ with uncertainty σ_i

combine the individual measurements to obtain a better estimate of $\mu \rightarrow$

The weighted average

❖ linear combination of the measurements

$$\hat{\mu} = \sum w_i y_i \quad \text{with} \quad \sum w_i = 1$$

► satisfies $\hat{\mu} = \mu$ if all measurements are equal to the true value μ

■ optimal weights: minimize the variance of $\hat{\mu}$

$$\sigma^2(\hat{\mu}) = \sum w_i^2 \sigma_i^2 \stackrel{!}{=} \min$$

■ constrained minimization problem

$$\sigma^2(\hat{\mu}) = \sum w_i^2 \sigma_i^2 + \lambda \left(1 - \sum w_i\right) \stackrel{!}{=} \min \quad \text{with solution} \quad w_i \propto \frac{1}{\sigma_i^2}$$

Towards Least Squares - a straight line fit

Consider uncorrelated measurements y_i , $i = 1, \dots, n$ with known variances σ_i^2 , recorded for known values x_i of a control parameter x . The expectation value of the measurements is $\langle y_i \rangle = a_0 + a_1 x_i$, where the parameters a_0 and a_1 are not known.

- wanted: estimates $\hat{a}_0$ and $\hat{a}_1$ for a_0 and a_1
- assumptions used
 - ▶ known control parameters x_i
 - ▶ unbiased measurements y_i
 - ▶ known variances σ_i^2
- implications
 - ▶ the shape of the PDFs describing the fluctuations of the y_i is irrelevant
 - ▶ different measurements can fluctuate with different PDFs

❖ starting point: the case of two measurements

- determine the parameters $\hat{a}_0$ and $\hat{a}_1$ which reproduce the measurements

▶ linear problem

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} \quad \text{or} \quad y = M^{-1} \hat{a}$$

▶ solution

$$\hat{a} = M y \quad \text{with covariance matrix} \quad C(\hat{a}) = M C(y) M^T$$

- the parameter estimates are linear combinations of the measurements
- the parameter estimates are random variables that fluctuate with the measurements
- for the expectation values one finds

$$\langle \hat{a} \rangle = M \langle y \rangle = a \quad \text{since} \quad \langle y \rangle = M^{-1} a$$

▶ the estimates $\hat{a}$ are unbiased

generalization →



BLUE (Best Linear Unbiased Estimator)

❖ estimate the unknown parameters by a linear combination of the measurements

$$\hat{a}_0 = \sum_{i=1}^n p_i y_i \quad \text{and} \quad \hat{a}_1 = \sum_{i=1}^n q_i y_i$$

- ❑ this is a convenient ansatz, not derived from any “first principles”
- ❑ it is not the only possible generalization of the case $n = 2$
- ❑ it will in general not give the best possible estimates for a_0 and a_1
- ❑ it is simple, robust and requires only minimal input
- ❑ it turns out to be surprisingly powerful . . .

find optimal coefficients p_i and q_i →

■ requirement: unbiasedness

$$\langle \hat{a}_0 \rangle = \sum p_i \langle y_i \rangle = \sum p_i (a_0 + a_1 x_i) = a_0 \sum p_i + a_1 \sum p_i x_i \stackrel{!}{=} a_0$$

$$\langle \hat{a}_1 \rangle = \sum q_i \langle y_i \rangle = \sum q_i (a_0 + a_1 x_i) = a_0 \sum q_i + a_1 \sum q_i x_i \stackrel{!}{=} a_1$$

► 4 conditions for $2n$ parameters:

$$\sum p_i = 1 \quad \sum p_i x_i = 0 \quad \sum q_i = 0 \quad \sum q_i x_i = 1$$

■ additional criterion to obtain a unique solution: minimal variance

$$\sigma^2(\hat{a}_k) = \sum_{i=1}^n \left(\frac{\partial \hat{a}_k}{\partial y_i} \right)^2 \sigma_i^2 \rightarrow \begin{pmatrix} \sigma^2(\hat{a}_0) \\ \sigma^2(\hat{a}_1) \end{pmatrix} = \sum \begin{pmatrix} p_i^2 \\ q_i^2 \end{pmatrix} \sigma_i^2 \stackrel{!}{=} \min$$

$$\text{with results: } p_i = \frac{1}{\sigma_i^2} (\alpha_0 + \beta_0 x_i) \quad \text{and} \quad q_i = \frac{1}{\sigma_i^2} (\alpha_1 + \beta_1 x_i)$$

the values of α_0 , α_1 , β_0 and β_1 follow from the 4 unbiasedness-constraints



❖ final result, collecting all terms ...

- the best fit parameters $\hat{a}_k$ are linear combinations of the measurements y_i
- measurements enter with weights proportional to $1/\sigma_i^2$
- the covariance matrix follows from error propagation: $C_{kl} = \sum_{i=1}^n \frac{\partial \hat{a}_k}{\partial y_i} \frac{\partial \hat{a}_l}{\partial y_i} \sigma_i^2$

$$\hat{a}_0 = \frac{1}{D} (S_{xx} S_y - S_x S_{xy}) \quad \text{and} \quad \hat{a}_1 = \frac{1}{D} (S_1 S_{xy} - S_x S_y)$$

$$\text{with } S_{\{1,x,xx,y,xy\}} = \sum_{i=1}^n \frac{\{1, x_i, x_i x_i, y_i, x_i y_i\}}{\sigma_i^2} \quad \text{and} \quad D = S_1 S_{xx} - S_x^2$$

$$\text{covariance matrix elements: } C_{00} = \frac{S_1}{D} \quad C_{11} = \frac{S_{xx}}{D} \quad C_{01} = C_{10} = \frac{-S_x}{D}$$

The cost function

❖ re-write the solution derived above

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \frac{1}{D} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \cdot \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} \rightarrow \begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \cdot \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} - \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

- ▶ the best-fit parameters are defined by the zero of a linear 2-dim function
- ▶ the zero of a function is a stationary point of its primitive
- ▶ with a conveniently chosen constant the primitive of this function is given by

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_i^2}$$

- parameter estimation can be interpreted as minimizing a cost function
 - ▶ the cost function defines a distance between data and model
 - ▶ the best fit parameters minimize the distance
 - ▶ with distances measured in units of σ_i , the parameters have minimum variance



General linear Least Squares

❖ linear data model and a general quadratic form for the cost function

$$\chi^2 = \sum_{i,j=1}^n \Delta_i W_{ij} \Delta_j \quad \text{with} \quad \Delta_i = y_i - \sum_k M_{ik} a_k$$

in matrix notation: $\chi^2 = \Delta^T W \Delta$ with $\Delta = y - M a$

- applicable also for non-linear models when using a linear approximation
- the weight matrix W defines a metric for the difference between data and model
 - ▶ any constant matrix W is allowed
 - ▶ $W = 1$: unweighted fit minimizing the euclidean distance between data and model
 - ▶ $W = C^{-1}$: minimum variance of the pararameter estimates
- general solution

$$\hat{a} = Q y \quad \text{with} \quad Q = [M^T W M]^{-1} M^T W \quad \text{and} \quad C(\hat{a}) = Q C Q^T$$

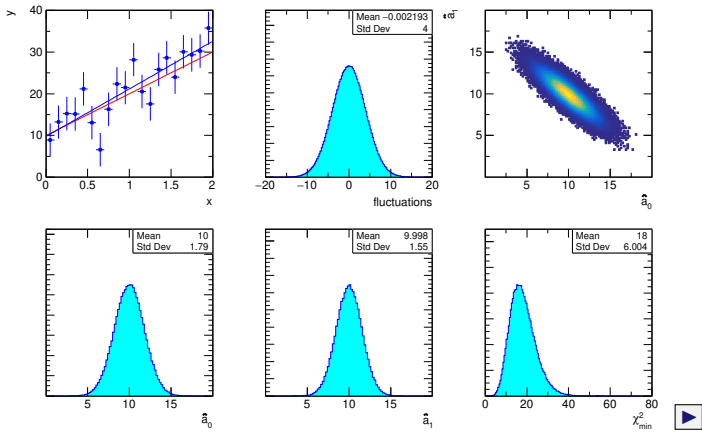


Properties of the linear Least Squares method

- the Least Squares method is distribution-free w.r.t the scatter of the measurements
- the matrix W is constant, i.e. it must be held fixed when minimizing χ^2
- parameter estimates are linear combinations of the measurements
 - ▶ for unbiased data the parameters estimates are unbiased
 - ▶ by the CLT the parameter errors are approximately gaussian distributed
- ▶ additional properties when using $W = C^{-1}$
 - the function χ^2 is invariant under linear transformations of the measurements
 - the parameter estimates have minimum variance
$$C(a) = [M^T C^{-1} M]^{-1} = \left(\frac{1}{2} \frac{\partial \chi^2}{\partial a^2} \right)^{-1}$$
 - a fit of m parameters to n measurements always has $\langle \chi_{\min}^2 \rangle = n - m$
 - for gaussian measurements the distribution of $\chi_{\min}^2$ follows a χ^2 -distribution

Numerical example: straight line fit $y = a_0 + a_1 x$

- expectation values of measurements as a function of x : $\langle y \rangle = 10 + 10 x$
- take 20 equidistant points in the range $0 < x < 2$
- study measurements with x -independent scatter $\text{rms} = 4$ around the expectation value
 - ▶ compare gaussian, exponential and uniform distributions
 - ▶ same covariance matrix and $\langle \chi_{\min}^2 \rangle = 18$ in all cases
 $\sigma(a_0) \approx 1.791$, $\sigma(a_1) \approx 1.551$, negative correlation with $\rho \approx -0.866$
- study poisson distributed measurements with x -dependent scatter $\text{rms} = \sqrt{\langle y \rangle}$
 - ▶ fit with correct weights: $1/\langle y \rangle \rightarrow$ unbiased parameters with minimum variance
 - ▶ fit with estimated weights: $1/y \rightarrow$ biased parameters
 - ▶ fit with constant weights: $1 \rightarrow$ unbiased parameters with non-minimal errors



Least squares with a priori unknown variances

❖ example: different ways to analyze Poisson-distributed measurements

- measurements y_i , $i = 1, \dots, n$ drawn from the same Poisson-PDF
- the variance is equal to the unknown expectation value μ
- inference of μ from a Least Squares fit requires assumptions about the variance

❖ method 1: assume the variance to be known

- ansatz

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2}$$

- minimization

$$\frac{\partial \chi^2}{\partial \mu} = -\frac{2}{\sigma^2} \sum_{i=1}^n (y_i - \mu) \stackrel{!}{=} 0 \rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$$

- ▶ $\hat{\mu}$ is an unbiased estimate for μ

❖ method 2: estimate the variance as part of the fit

■ ansatz

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \mu)^2}{\mu} = \sum_{i=1}^n \frac{y_i^2}{\mu} - 2y_i + \mu$$

■ minimization

$$\frac{\partial \chi^2}{\partial \mu} = \sum_{i=1}^n -\frac{y_i^2}{\mu^2} + 1 \stackrel{!}{=} 0 \quad \rightarrow \quad \hat{\mu} = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2}$$

- ▶ parameter estimate by quadratic average
- ▶ non-linear fit model (non-parabolic cost function)
- ▶ biased estimate $\langle \hat{\mu} \rangle \stackrel{n \rightarrow \infty}{\Rightarrow} \sqrt{\mu(\mu + 1)}$
- ▶ violates that variances are known and thus fixed in the least squares minimisation



❖ method 3: use empirical variance estimates

■ ansatz

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \mu)^2}{y_i} = \sum_{i=1}^n y_i - 2\mu + \frac{\mu^2}{y_i}$$

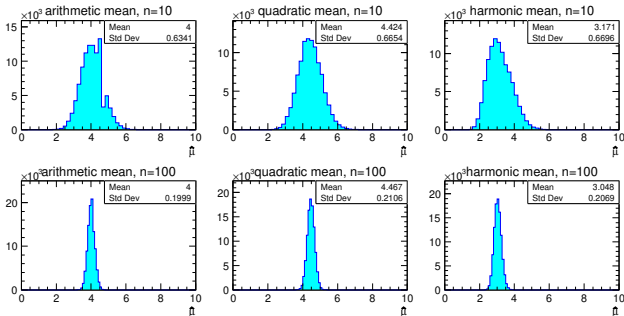
■ minimisation

$$\frac{\partial \chi^2}{\partial \mu} = \sum_{i=1}^n -2 + \frac{2\mu}{y_i} \stackrel{!}{=} 0 \quad \rightarrow \quad \hat{\mu} = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{y_i} \right)^{-1}$$

- ▶ parameter estimate by harmonic average
- ▶ larger weights for downward fluctuations
- ▶ necessity to discard values $y_i = 0$
- ▶ biased estimate $\langle \hat{\mu} \rangle \stackrel{n \rightarrow \infty}{=} 1 / \langle 1/y \rangle$
- ▶ violates that the true variances are independent of the data

problems and exercises

Implement the three schemes discussed above and determine numerically the distributions of $\hat{\mu}$ for a true Poisson average $\mu = 4$ and sample sizes $n = 10$ and $n = 100$.



Least Squares for low statistics

- common wisdom: Least Squares fits need. . .
 - ▶ gaussian fluctuations
 - ▶ sufficiently large event counts for poisson distributed data
- in the derivation of the method none of the above entered
 - ▶ knowledge of the correct variances was assumed
 - ▶ variance estimates should not be correlated to the data
 - ▶ variances must be kept constant in the χ^2 minimisation, which implies an iterative treatment when the true variances are functions of the fit parameters

case study, keeping an eye on those points when doing fits →

Example: determination of the lifetime of an unstable particle

■ lifetime distribution

$$g(t; \tau) = \frac{1}{\tau} e^{-t/\tau}$$

■ MC study of experiments with $\tau = 1$ ns and an expected number of $\langle N \rangle$ decays

- ▶ histogram representation of each experiment with 100 bins for $0 < t < 10$ ns
- ▶ expected bin content $\langle n_i \rangle$

$$\langle n_i \rangle(\tau) = \langle N \rangle \int_{\text{bin } i} dt g(t; \tau)$$

- ▶ Poisson distributed bin contents n_i and total number of events N

■ reference: optimal parameter estimate

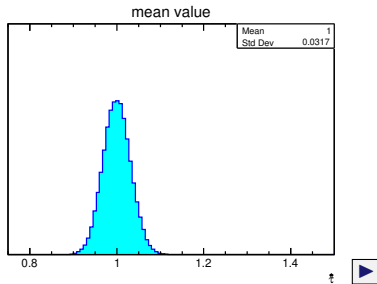
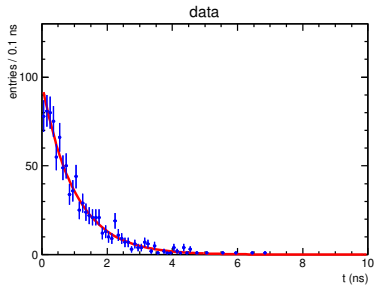
$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N t_i$$

❖ study different weight assignments

- ❑ $w_i = 1$ for all bins
 - ▶ unsophisticated but hopefully robust unweighted fit
- ❑ $w_i = 1$ for all bins with non-zero entries
 - ▶ pretend that empty bins don't contribute information
- ❑ $w_i = 1/n_i$ for all bins with non-zero entries
 - ▶ use empirical variance estimates
- ❑ $w_i = 1/\langle n_i \rangle$ (τ) for all bins
 - ▶ naive way to use the theoretical variances
- ❑ iterative fit with $w_i(0) = 1$ and $w_i(m) = 1/\langle n_i \rangle (\hat{\tau}_{m-1})$ for all bins
 - ▶ proper way to use the theoretical variances
 - ▶ implements that variances must be fixed in minimization
 - ▶ weak correlation between variance estimates and data

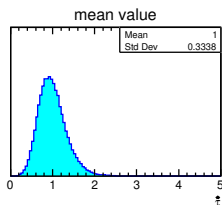
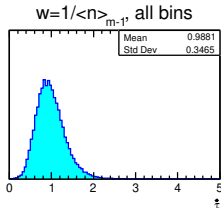
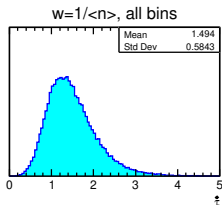
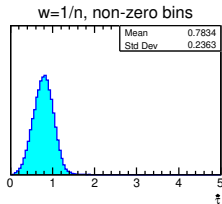
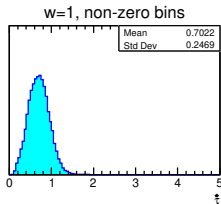
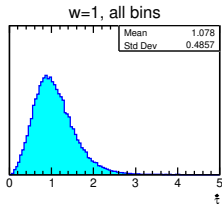


❖ simulation study



- study mean value and standard deviation and of the fitted $\hat{\mu}$ as a function of $\langle N \rangle$ and the different choices of the weight function

❖ performance comparison: parameter estimates for $N = 10$ events



❖ conclusions regarding different weight assignments

- $w_i = 1$ for all bins
 - ▶ unbiased, but not with optimal precision
- $w_i = 1$ for non-zero bins
 - ▶ violates the Least Squares ansatz → biased results, needless loss of information
- $w_i = 1/n_i$ for all bins with non-zero entries
 - ▶ violates the Least Squares ansatz → biased results
- $w_i = 1/\langle n_i \rangle$ for all bins
 - ▶ violates the Least Squares ansatz → biased results
- iterative fit with $w_i(0) = 1$ and $w_i(m) = 1/f_i(\hat{\tau}_{m-1})$ for all bins
 - ▶ close to optimum, even at very low statistics