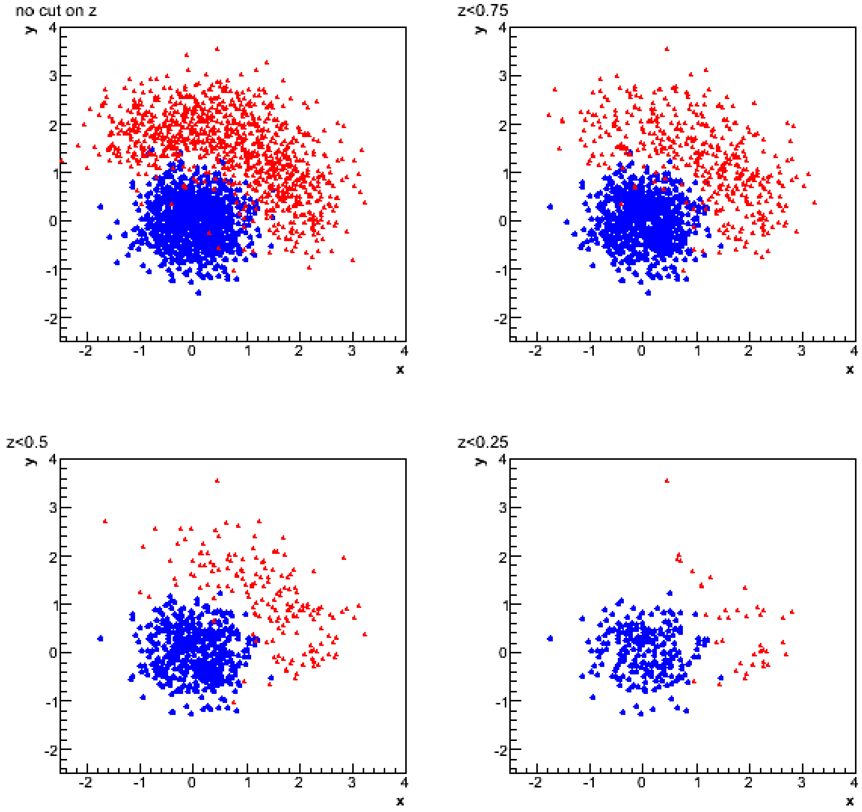
Pre-lecture 11 comments on problem sheet 7

Problem sheet 7 involves modifying some C++ programs to create a Fisher discriminant and neural network to separate two types of events (signal and background):



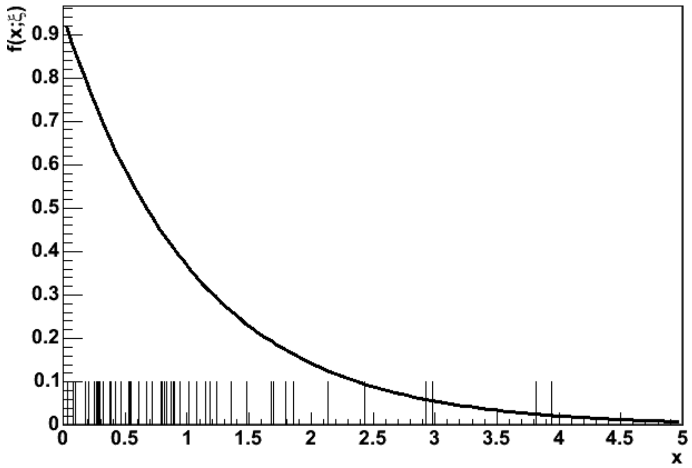
Each event is characterized by 3 numbers: *x*, *y* and *z*.

Each "event" (instance of *x,y,z*) corresponds to a "row" in an *n*-tuple. (here, a 3-tuple).

In ROOT, *n*-tuples are stored in objects of the `TTree` class.

Comments on problem sheet 7

Problem sheet 7 also involves an ML fit using the root class `TMinuit`, which numerically minimizes the (negative) log-likelihood function.



An MC program is used to generate data from exponential, then the parameter is fitted using `TMinuit` (see code).

You then modify the code to do the problem of a mixture of exponentials:

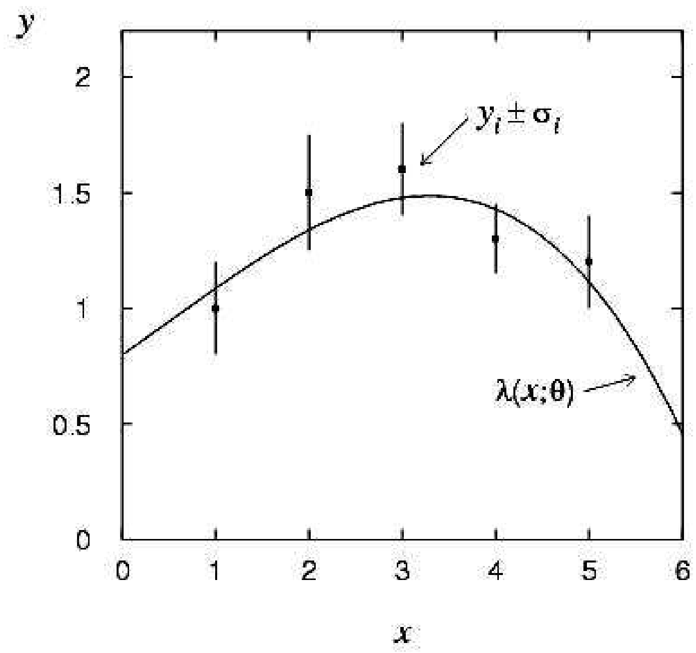
$$f(x;\alpha,\xi_1,\xi_2) = \alpha\frac{1}{\xi_1}e^{-x/\xi_1} + (1-\alpha)\frac{1}{\xi_2}e^{-x/\xi_2}$$

Statistical Data Analysis: Lecture 11

The method of least squares

Suppose we measure N values, $y_1, ..., y_N$, assumed to be independent Gaussian r.v.s with

$$E[y_i] = \lambda(x_i; \theta) .$$



Assume known values of the control variable $x_1, ..., x_N$ and known variances

$$V[y_i] = \sigma_i^2 .$$

We want to estimate θ, i.e., fit the curve to the data points.

The likelihood function is

$$L(\theta) = \prod_{i=1}^{N} f(y_i; \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - \lambda(x_i; \theta))^2}{2\sigma_i^2} \right]$$

The method of least squares (2)

The log-likelihood function is therefore

$$\ln L(\theta) = -\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2} + \text{terms not depending on } \theta$$

So maximizing the likelihood is equivalent to minimizing

$$\chi^2(\theta) = \sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

Minimum defines the least squares (LS) estimator $\hat{\theta}$.

Very often measurement errors are ~Gaussian and so ML and LS are essentially the same.

Often minimize χ^2 numerically (e.g. program `MINUIT`).

LS with correlated measurements

If the y_i follow a multivariate Gaussian, covariance matrix V,

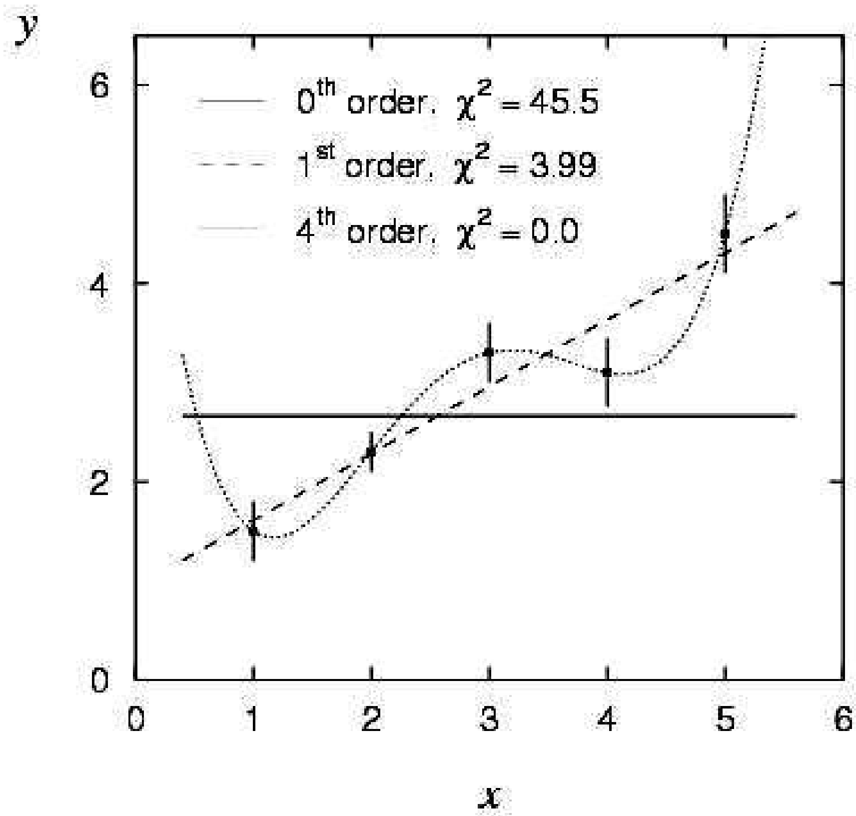
$$g(\vec{y}, \vec{\lambda}, V) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda})\right]$$

Then maximizing the likelihood is equivalent to minimizing

$$\chi^2(\vec{\theta}) = \sum_{i,j=1}^{N} (y_i - \lambda(x_i;\vec{\theta}))(V^{-1})_{ij}(y_j - \lambda(x_j;\vec{\theta}))$$

Example of least squares fit

Fit a polynomial of order p: $\lambda(x;\theta_0,\ldots,\theta_p) = \sum_{n=0}^{p} \theta_n x^n$



Variance of LS estimators

In most cases of interest we obtain the variance in a manner similar to ML. E.g. for data ~ Gaussian we have

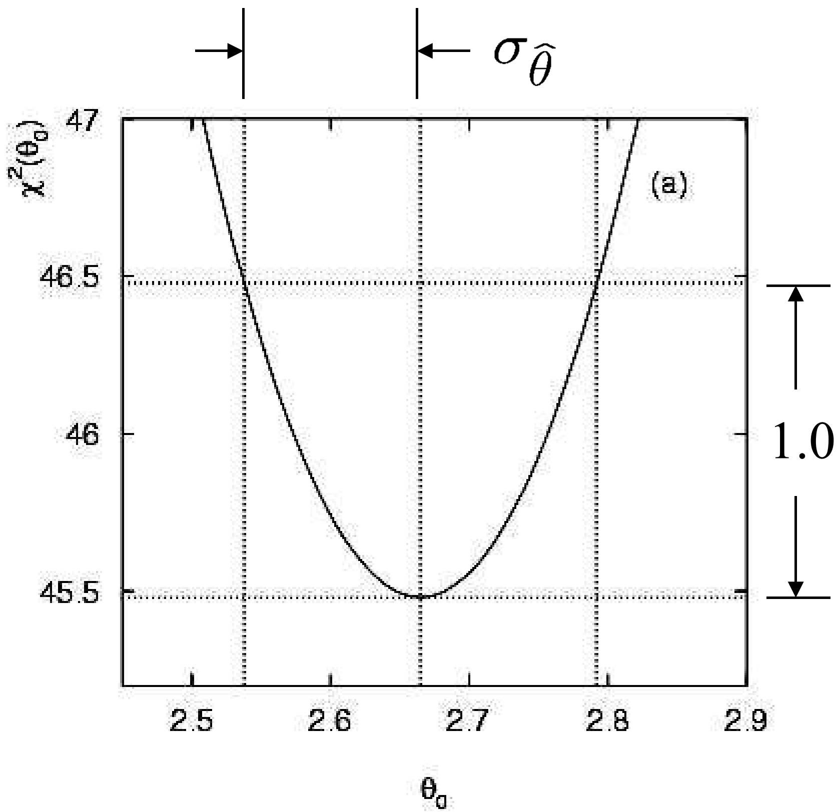
$$\chi^2(\theta) = -2 \ln L(\theta)$$

and so

$$\widehat{\sigma^2}_{\hat{\theta}} \approx 2 \left[\frac{\partial^2 \chi^2}{\partial \theta^2} \right]^{-1}_{\theta = \hat{\theta}}$$

or for the graphical method we take the values of θ where

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



Two-parameter LS fit

2-parameter case (line with nonzero slope):

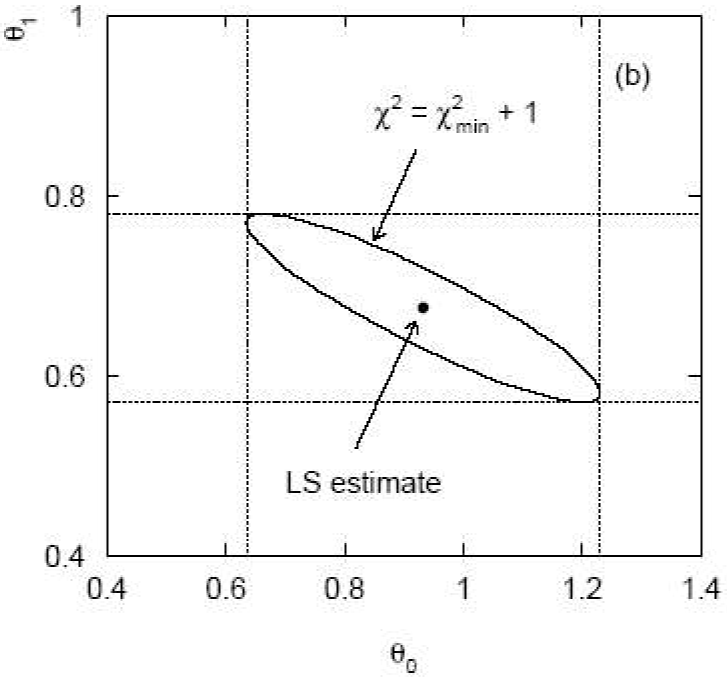
$$\hat{\theta}_0 = 0.93 \pm 0.30,$$

$$\hat{\theta}_1 = 0.68 \pm 0.10$$

$$\widehat{\mathrm{cov}}[\hat{\theta}_0, \hat{\theta}_1] = -0.028$$

$$r = -0.90$$

$$\chi^2 = 3.99$$



Tangent lines $\rightarrow \sigma_{\hat{\theta}_0}, \sigma_{\hat{\theta}_1}$.

Angle of ellipse $\rightarrow$ correlation (same as for ML)

Goodness-of-fit with least squares

The value of the χ^2 at its minimum is a measure of the level of agreement between the data and fitted curve:

$$\chi^2_{\min} = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \hat{\theta}))^2}{\sigma_i^2}$$

It can therefore be employed as a goodness-of-fit statistic to test the hypothesized functional form $\lambda(x; \theta)$.

We can show that if the hypothesis is correct, then the statistic $t = \chi^2_{\min}$ follows the chi-square pdf,

$$f(t; n_{\mathrm{d}}) = \frac{1}{2^{n_{\mathrm{d}}/2}\Gamma(n_{\mathrm{d}}/2)} t^{n_{\mathrm{d}}/2-1} e^{-t/2}$$

where the number of degrees of freedom is

n_{d} = number of data points − number of fitted parameters

Goodness-of-fit with least squares (2)

The chi-square pdf has an expectation value equal to the number of degrees of freedom, so if $\chi^2_{\min} \approx n_\mathrm{d}$ the fit is 'good'.

More generally, find the p-value:

$$p = \int_{\chi^2_{\min}}^{\infty} f(t; n_\mathrm{d})\, dt$$

This is the probability of obtaining a $\chi^2_{\min}$ as high as the one we got, or higher, if the hypothesis is correct.

E.g. for the previous example with 1st order polynomial (line),

$$\chi^2_{\min} = 3.99\,, \qquad n_\mathrm{d} = 5-2 = 3\,, \qquad p = 0.263$$

whereas for the 0th order polynomial (horizontal line),

$$\chi^2_{\min} = 45.5\,, \qquad n_\mathrm{d} = 5-1 = 4\,, \qquad p = 3.1\times 10^{-9}$$

Goodness-of-fit vs. statistical errors

Small statistical error does not mean a good fit (nor vice versa).

Curvature of χ^2 near its minimum → statistical errors ($\sigma_{\hat{\theta}}$)

Value of $\chi^2_{\min}$ → goodness-of-fit

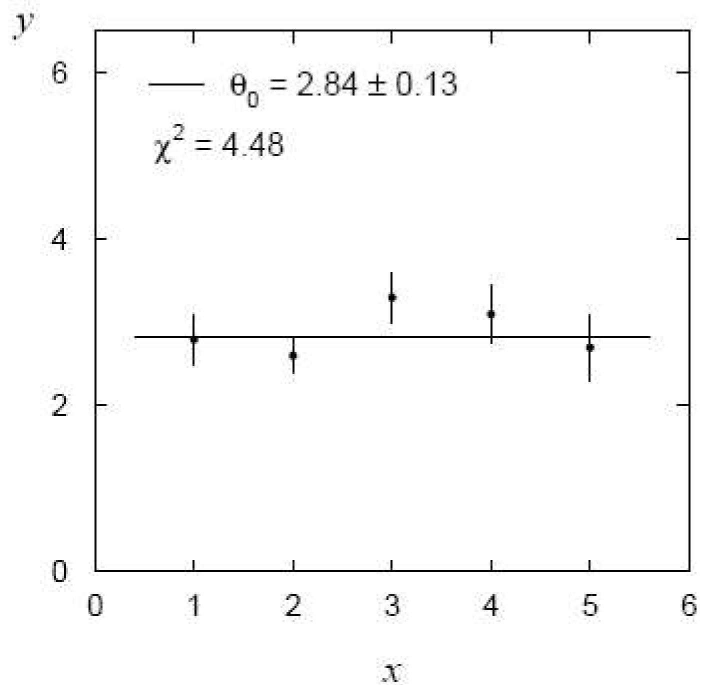
Horizontal line fit, move the data points, keep errors on points same:

$$\hat{\theta}_0 = 2.84 \pm 0.13$$

$$\chi^2_{\min} = 4.48$$

Variance same as before,

now $\chi^2_{\min}$ 'good'.



Goodness-of-fit vs. stat. errors (2)

$\rightarrow \chi^2(\theta_0)$ shifted down, same curvature as before.

Variance of estimator (statistical error) tells us:

if experiment repeated many times, how wide is the distribution of the estimates $\hat{\theta}$. (Doesn't tell us whether hypothesis correct.)

P-value tells us:

if hypothesis is correct and experiment repeated many times, what fraction will give equal or worse agreement between data and hypothesis according to the statistic $\chi^2_{\min}$.

Low P-value $\rightarrow$ hypothesis may be wrong $\rightarrow$ systematic error.

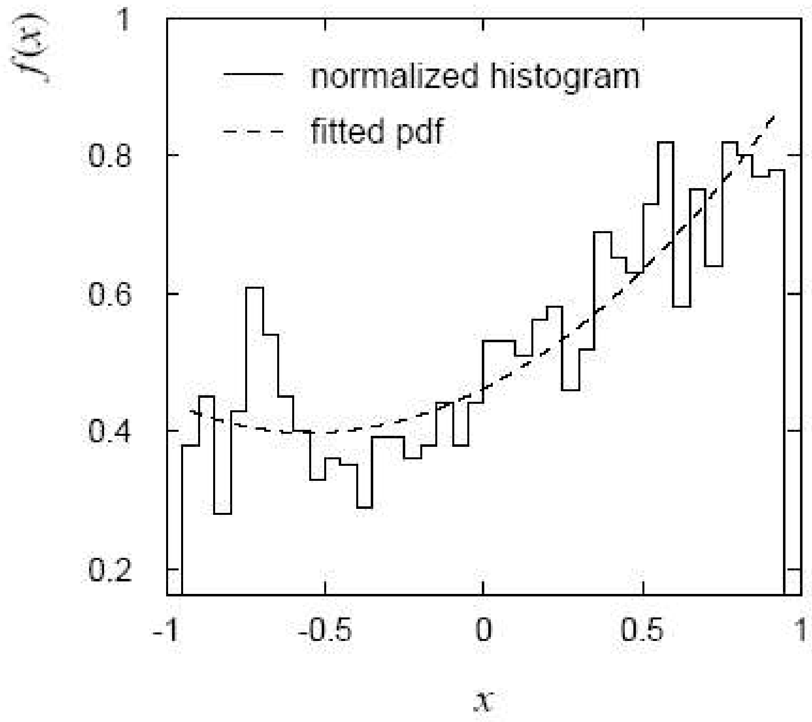
LS with binned data

Histogram:

N bins, n entries.

Hypothesized pdf:

$f(x; \vec{\theta})$



We have

$$y_i = \text{number of entries in bin } i,$$

$$\lambda_i(\vec{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = n p_i(\vec{\theta})$$

LS with binned data (2)

LS fit: minimize

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\vec{\theta}))^2}{\sigma_i^2}$$

where $\sigma_i^2 = V[y_i]$, here not known a priori.

Treat the y_i as Poisson r.v.s, in place of true variance take either

$\sigma_i^2 = \lambda_i(\vec{\theta})$ (LS method)

$\sigma_i^2 = y_i$ (Modified LS method)

MLS sometimes easier computationally, but $\chi^2_{\min}$ no longer follows chi-square pdf (or is undefined) if some bins have few (or no) entries.

LS with binned data — normalization

Do not 'fit the normalization':

$$\lambda_i(\vec{\theta}, \nu) = \nu \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = \nu p_i(\vec{\theta})$$

i.e. introduce adjustable ν, fit along with $\vec{\theta}$.

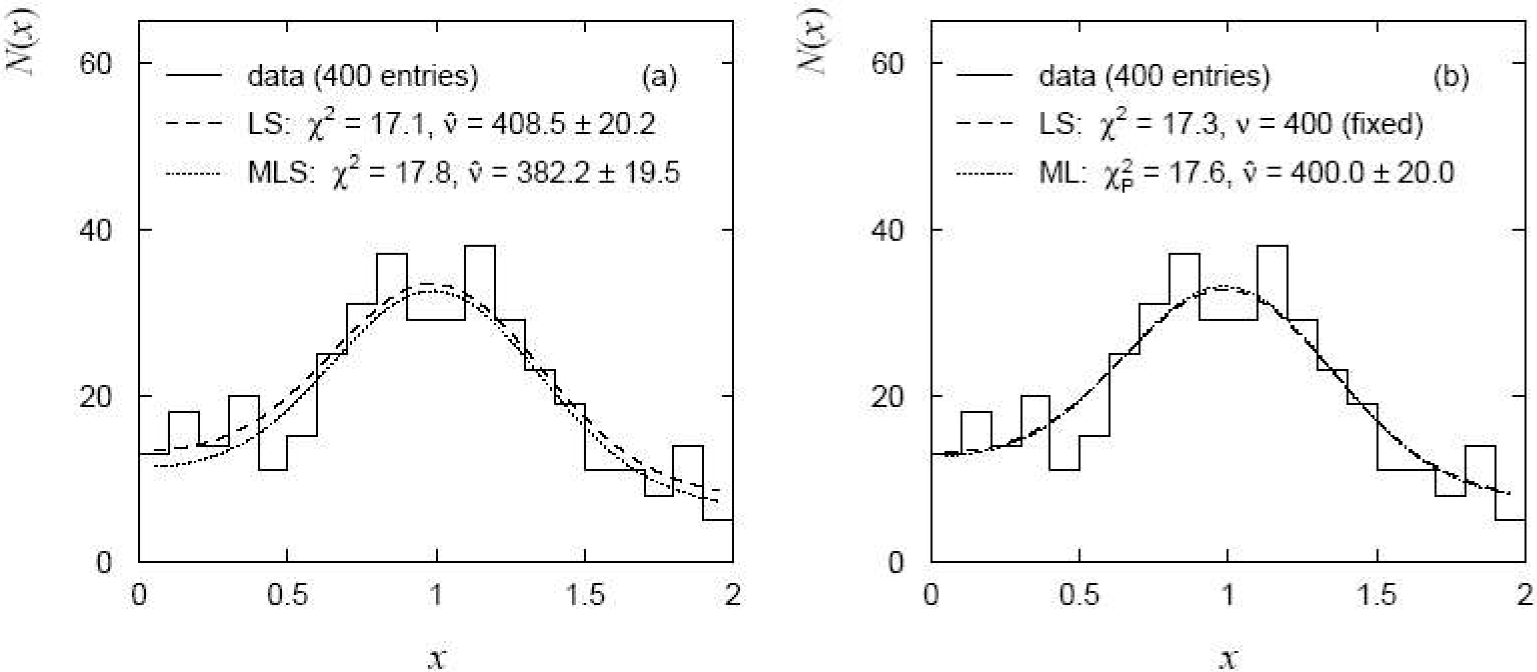
$\hat{\nu}$ is a bad estimator for n (which we know, anyway!)

$$\hat{\nu}_{\mathrm{LS}} = n + \frac{\chi^2_{\min}}{2}$$

$$\hat{\nu}_{\mathrm{MLS}} = n - \chi^2_{\min}$$

LS normalization example

Example with $n = 400$ entries, $N = 20$ bins:



Expect $\chi^2_{\min}$ around $N - m$,

$\rightarrow$ relative error in $\hat{\nu}$ large when N large, n small

Either get n directly from data for LS (or better, use ML).

Using LS to combine measurements

Use LS to obtain weighted average of N measurements of λ:

y_i = result of measurement i, $i = 1, \ldots, N$;

$\sigma_i^2 = V[y_i]$, assume known;

λ = true value (plays role of θ).

For uncorrelated y_i, minimize

$$\chi^2(\lambda) = \sum_{i=1}^{N} \frac{(y_i - \lambda)^2}{\sigma_i^2},$$

Set $\frac{\partial \chi^2}{\partial \lambda} = 0$ and solve,

$$\rightarrow \quad \hat{\lambda} = \frac{\Sigma_{i=1}^{N} y_i / \sigma_i^2}{\Sigma_{j=1}^{N} 1/\sigma_j^2} \qquad V[\hat{\lambda}] = \frac{1}{\Sigma_{i=1}^{N} 1/\sigma_i^2}$$

Combining correlated measurements with LS

If $\text{cov}[y_i, y_j] = V_{ij}$, minimize

$$\chi^2(\lambda) = \sum_{i,j=1}^{N} (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda) \, ,$$

$$\rightarrow \quad \hat{\lambda} = \sum_{i=1}^{N} w_i y_i, \qquad w_i = \frac{\Sigma_{j=1}^{N} (V^{-1})_{ij}}{\Sigma_{k,l=1}^{N} (V^{-1})_{kl}}$$

$$V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j$$

LS $\hat{\lambda}$ has zero bias, minimum variance (Gauss–Markov theorem).

Example: averaging two correlated measurements

Suppose we have y_1, y_2, and $V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

$$\rightarrow \quad \hat{\lambda} = w y_1 + (1-w) y_2, \quad w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$V[\hat{\lambda}] = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \sigma^2$$

The increase in inverse variance due to 2nd measurement is

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} = \frac{1}{1-\rho^2}\left(\frac{\rho}{\sigma_1} - \frac{1}{\sigma_2}\right)^2 > 0$$

$\rightarrow$ 2nd measurement can only help.

Negative weights in LS average

If $\rho > \sigma_1/\sigma_2$, $\rightarrow$ $w < 0$,

$\rightarrow$ weighted average is not between y_1 and y_2 (!?)

Cannot happen if correlation due to common data, but possible for shared random effect; very unreliable if e.g. ρ, σ_1, σ_2 incorrect.

See example in SDA Section 7.6.1 with two measurements at same temperature using two rulers, different thermal expansion coefficients:

average is outside the two measurements; used to improve estimate of temperature.

Wrapping up lecture 11

Considering ML with Gaussian data led to the method of Least Squares.

Several caveats when the data are not (quite) Gaussian, e.g., histogram-based data.

Goodness-of-fit with LS "easy" (but do not confuse good fit with small stat. errors)

LS can be used for averaging measurements.

Next lecture: Interval estimation