

# Statistical Methods in Particle Physics

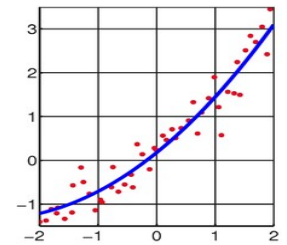
## Lecture 11

*January 9, 2012*

Silvia Masciocchi, GSI Darmstadt  
*s.masciocchi@gsi.de*

*Winter Semester 2011 / 12*

# Method of least squares

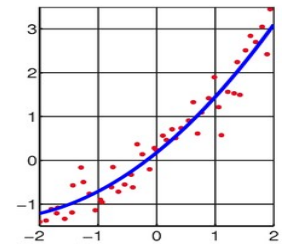


The method of least squares is a standard approach to the approximate solution of **overdetermined systems**, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the errors made in solving every single equation.

The most important application is in **data fitting**. The best fit in the least-squares sense minimizes the **sum of squared residuals**, a residual being the difference between an observed value and the fitted value provided by a model.

Least squares corresponds to the maximum likelihood criterion if the experimental errors have a normal distribution and can also be derived as a method of moments estimator.

# A bit of history

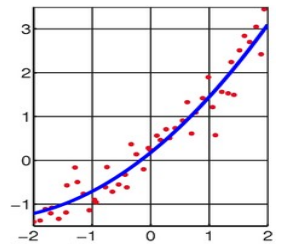


**Carl Friedrich Gauss** is credited with developing the fundamentals of the basis for least-squares analysis in **1795** at the age of eighteen. Legendre was the first to publish the method, however.

An early demonstration of the strength of Gauss's method came when it was used to **predict the future location of the newly discovered asteroid Ceres**. On January 1, 1801, the Italian astronomer Giuseppe Piazzi discovered Ceres and was able to track its path for 40 days before it was lost in the glare of the sun. Based on this data, astronomers desired to determine the location of Ceres after it emerged from behind the sun without solving the complicated Kepler's nonlinear equations of planetary motion. The only predictions that successfully allowed Hungarian astronomer Franz Xaver von Zach to relocate Ceres were those performed by the 24-year-old Gauss using least-squares analysis.



# Data fitting

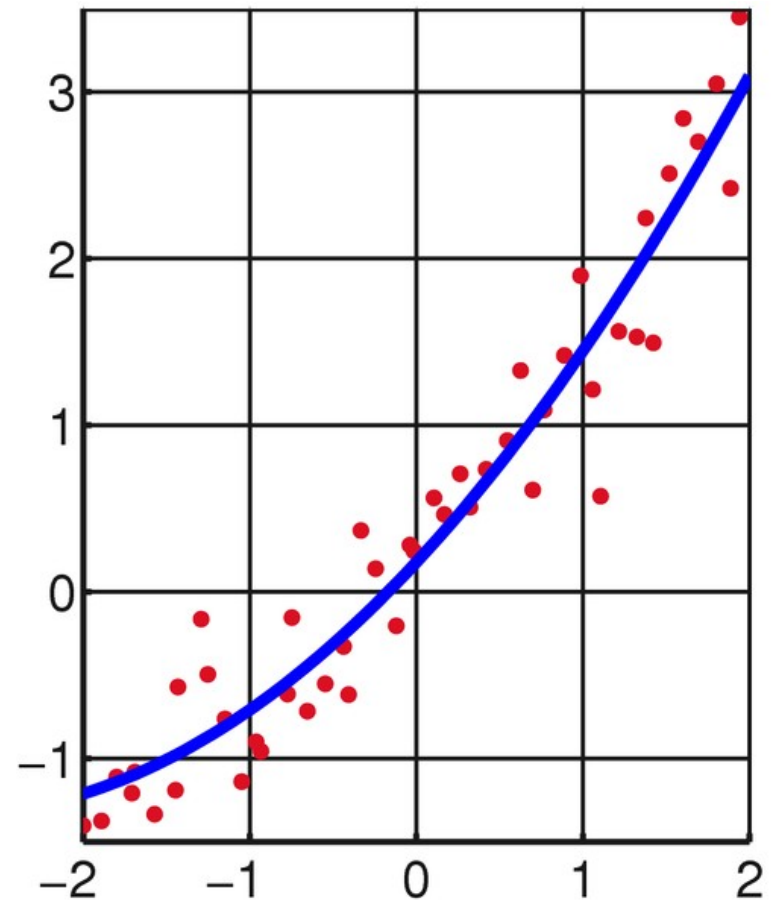


The method of least squares is often used to fit a function through a set of points

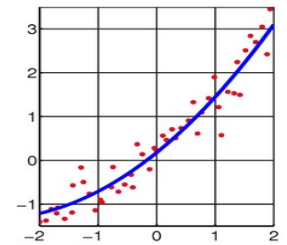
by minimizing the distance between the points and the fitted function

by the least squares

But we need to define well the conditions under which the method can be applied



# Connection with the ML - 1

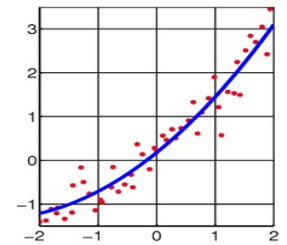


In many situations a measured value  $y$  can be regarded as a Gaussian random variable centered about the quantity's true value  $\lambda$ . This follows from the central theorem as long as the total error (i.e. the deviation from the true value) is the sum of a large number of small contributions.

Extend to  $N$  measurements:

- $N$  measurements  $y_i$  related to another variable  $x_i$ , assumed to be known without error
- The  $y_i$  are  $N$  independent Gaussian random variables
- Each value  $y_i$  has a different unknown mean  $\lambda_i$ , and ...
- .. a different but known variance  $\sigma_i^2$

# Connection with the ML - 2



The  $N$  measurements of  $y_i$  can be equivalently regarded as a single measurement of an  $N$ -dimensional random vector, for which the joint pdf is the product of  $N$  Gaussians:

$$g(y_1, \dots, y_N; \lambda_1, \dots, \lambda_N, \sigma_1^2, \dots, \sigma_N^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - \lambda_i)^2}{2\sigma_i^2}\right)$$

Suppose further that the true value is given as a function of  $x$ :

$$\lambda = \lambda(x; \boldsymbol{\theta})$$

which depends on unknown parameters  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_m)$

The aim of the method of least squares is to estimate the parameters  $\boldsymbol{\theta}$ . In addition, the method allows for a simple evaluation of the goodness-of-fit of the hypothesized function.

# Connection with the ML - 3

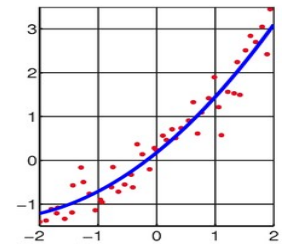
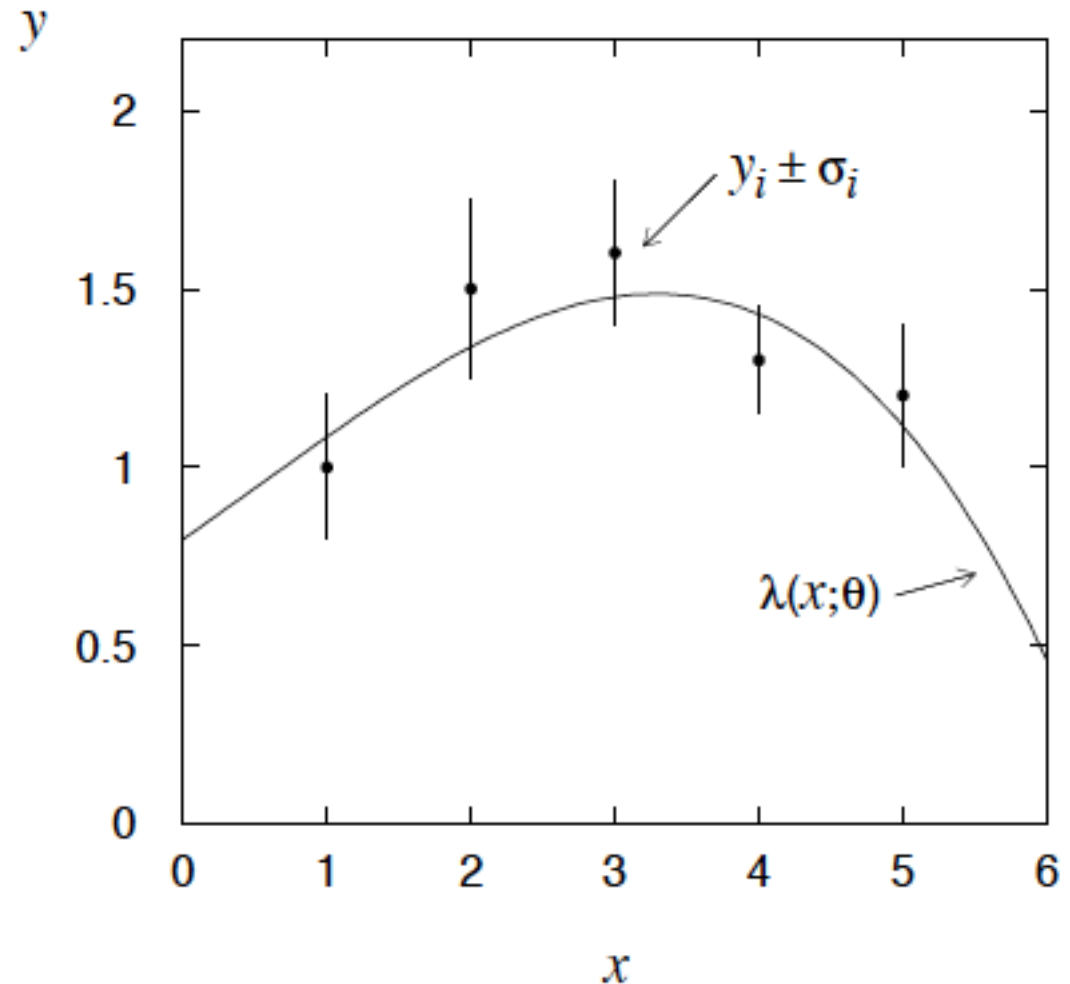


Illustration:

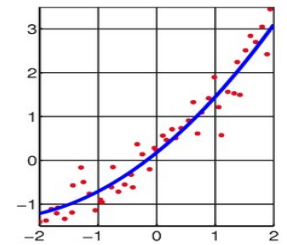
- $x_1, \dots, x_N$  known
- Gaussian random variables  $y_i$
- $E[y_i] = \lambda_i = \lambda(x_i; \theta)$
- $V[y_i] = \sigma_i^2$  known

GOAL:

Estimate the parameters  $\theta$   
→ fit the curve through the points



# Connection to the ML - 4



We take the logarithm of the joint pdf and drop the additive terms that do not depend on the parameters  $\theta$ :

We get the log-likelihood function

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

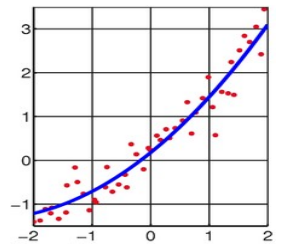
This is maximized by finding the values of the parameters  $\theta$  that minimize the quantity

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

Namely the **quadratic sum of the differences between measured and hypothesized values, weighted by the inverse of the variances**. This is the **basis of the method of least squares**.



# Definition of LS estimators



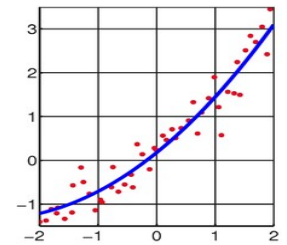
If the  $N$  measurements are not independent but described by an  $N$ -dimensional Gaussian pdf with known covariance matrix  $V$  but unknown mean values, the joint pdf is:

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

The log-likelihood is obtained (again dropping terms not depending on the parameters) as

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

# Definition of LS estimators - 2



The log-likelihood is maximized by minimizing the quantity

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

which reduces to the case on page 8 if the covariance matrix (and hence its inverse) are diagonal (and the variables are independent)

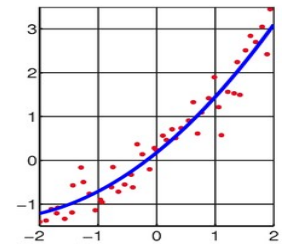
The parameters which minimize the  $\chi^2$  are called the **least square (LS) estimators**

$$\hat{\theta}_1, \dots, \hat{\theta}_m$$

The minimization is most often done numerically.

Example: program MINUIT (in ROOT)

# Example of a LS fit (polynomial)



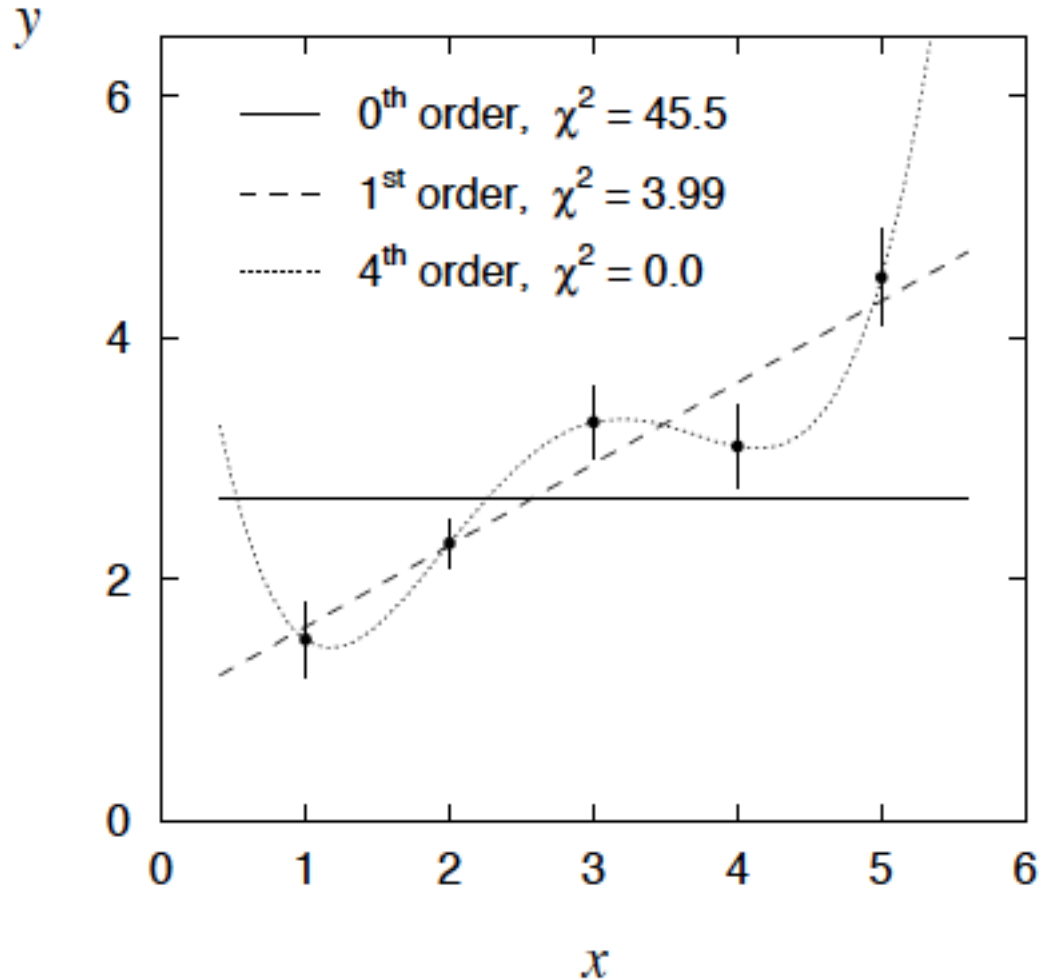
Fit a polynomial of order  $p$ :

$$\lambda(x; \theta_0, \dots, \theta_p) = \sum_{n=0}^p \theta_n x^n$$

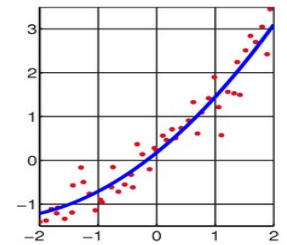
0<sup>th</sup> order: 1 parameter

1<sup>st</sup> order: 2 parameters


4<sup>th</sup> order: 5 parameters



# Linear least squares fit



LS has particularly simple properties of the  $\chi^2$  and of the LS estimators, for the case where  $\lambda(\mathbf{x};\boldsymbol{\theta})$  is a **linear function of the parameters  $\boldsymbol{\theta}$**

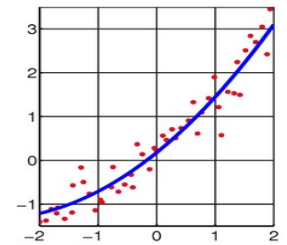
$$\lambda(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^m a_j(\mathbf{x})\theta_j$$


where the  $a_j(\mathbf{x})$  are ANY linearly independent functions of  $\mathbf{x}$  (they are just linearly independent from each other, i.e. one cannot be expressed as a linear combination of the others).

For this case, when we calculate the variance, we obtain that  $\chi^2$  is quadratic in  $\boldsymbol{\theta}$  and it follows that

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

# Variance of least squares estimators



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

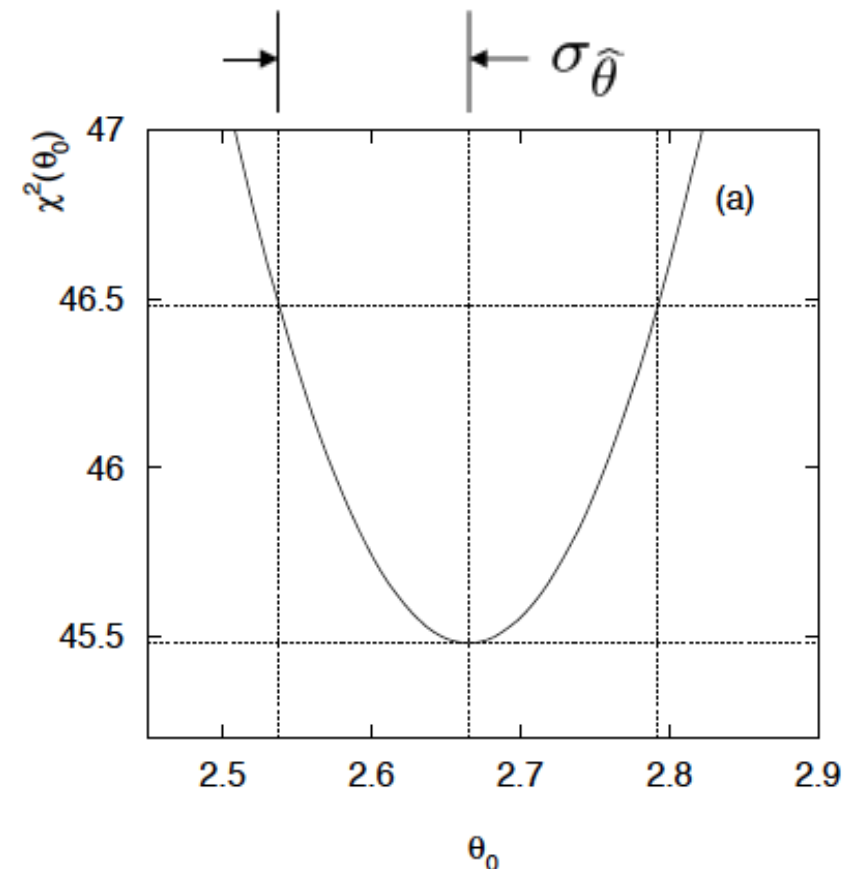
this yields the contours in parameter space whose tangents are at  $\hat{\theta}_i \pm \hat{\sigma}_i$  corresponding to a one standard deviation departure from the LS estimates

Example: polynomial fit, 0<sup>th</sup> order

**1-parameter fit**

Obtain the variance:

$$\sigma_{\hat{\theta}_0} \text{ from } \chi^2(\hat{\theta}_0 \pm \sigma_{\hat{\theta}_0}) = \chi_{\min}^2 + 1$$



# Two parameter LS fit

1<sup>st</sup> order polynomial fit, 2 parameters  
(line with non-zero slope)

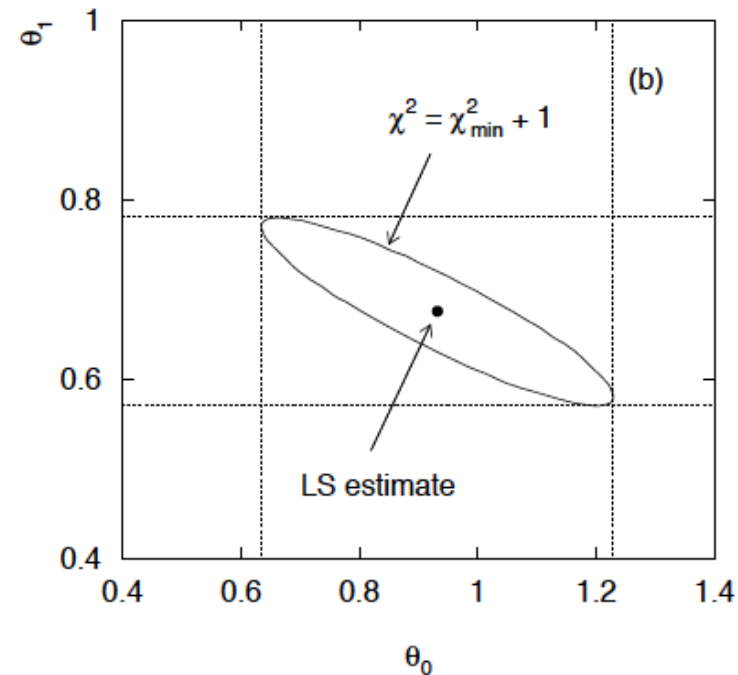
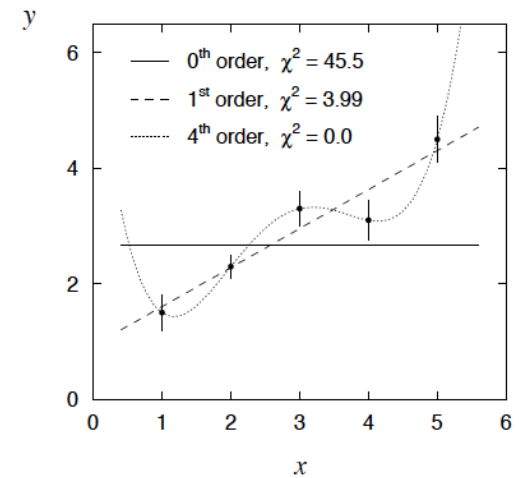
$$\hat{\theta}_0 = 0.93 \pm 0.30,$$

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

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

$$r = -0.90$$

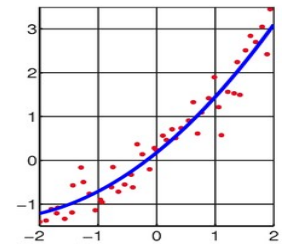
$$\chi^2 = 3.99$$



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

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

# Confidence region



Reminder

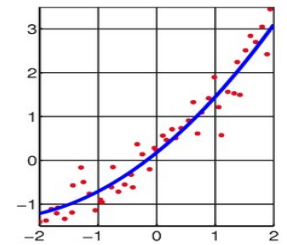
$$\chi^2(\boldsymbol{\theta}) = \chi^2(\hat{\boldsymbol{\theta}}) + 1 = \chi_{\min}^2 + 1$$

this yields the contours in parameter space whose tangents are at  $\hat{\theta}_i \pm \hat{\sigma}_i$  corresponding to a one standard deviation departure from the LS estimates

If the function  $\lambda(x; \boldsymbol{\theta})$  is **NOT linear in the parameters**, then the contour defined above is not in general elliptical, and one can no longer obtain the standard deviations from the tangents. It defines a region in parameter space, however, which can be interpreted as **CONFIDENCE REGION**, the size of which reflects the statistical uncertainty of the fitted parameters

→ Next lecture !!!

# Five parameter LS fit

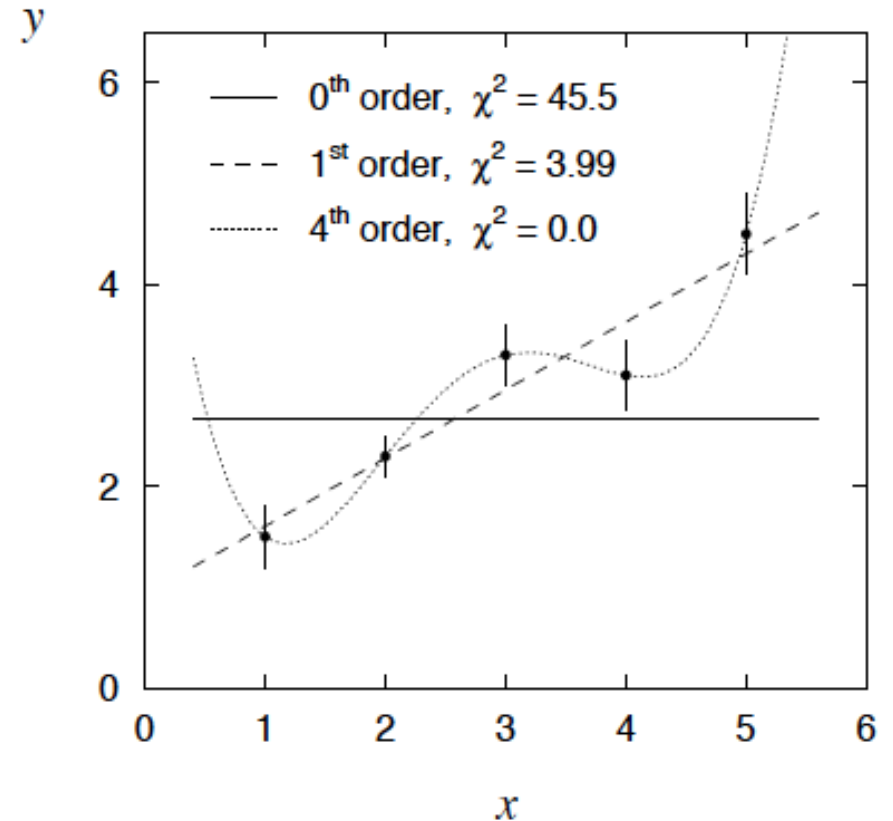


4<sup>th</sup> order polynomial fit, 5 parameters

Curve goes through all points

$$\chi_{\min}^2 = 0$$

Number of parameters = number  
of points

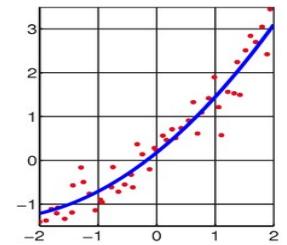


The value of  $\chi_{\min}^2$  reflects the level of agreement between data and hypothesis

→ use as goodness-of-fit test statistics



# Goodness-of-fit with least squares



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

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

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

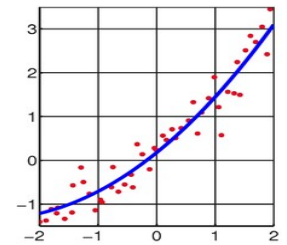
We can show that if the hypothesis is correct, then the statistic  $t = \chi_{\min}^2$  follows the chi-square pdf:

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

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

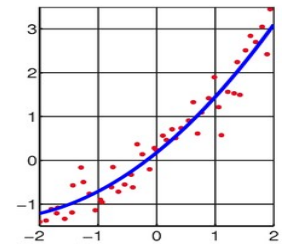


The chi-square pdf has an expectation value equal to the number of degrees of freedom, so if  $\chi_{\min}^2 \approx n_d$  or  $\chi_{\min}^2/n_d \approx 1$  the fit is 'good'.

Or more precisely:

- If  $\chi_{\min}^2/n_d \approx 1$  all is as expected
- If  $\chi_{\min}^2/n_d \ll 1$  then the fit is better than expected given the size of the measurement errors. This is not bad in the sense of providing evidence against the hypothesis, but it is usually grounds to check that the errors  $\sigma_i$  have not been overestimated or are not correlated
- If  $\chi_{\min}^2/n_d \gg 1$  then there is some reason to doubt the hypothesis

# Goodness-of-fit with least squares - 3



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

More generally, find the p-value: 
$$p = \int_{\chi_{\min}^2}^{\infty} f(t; n_d) dt$$

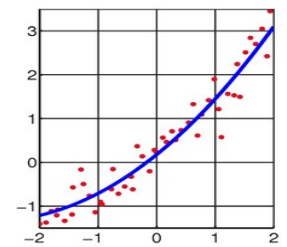
This is the probability of obtaining a  $\chi_{\min}^2$  as high as the one we got, or higher, if the hypothesis is correct.

E.g. for the previous example with 1<sup>st</sup> order polynomial (straight line):

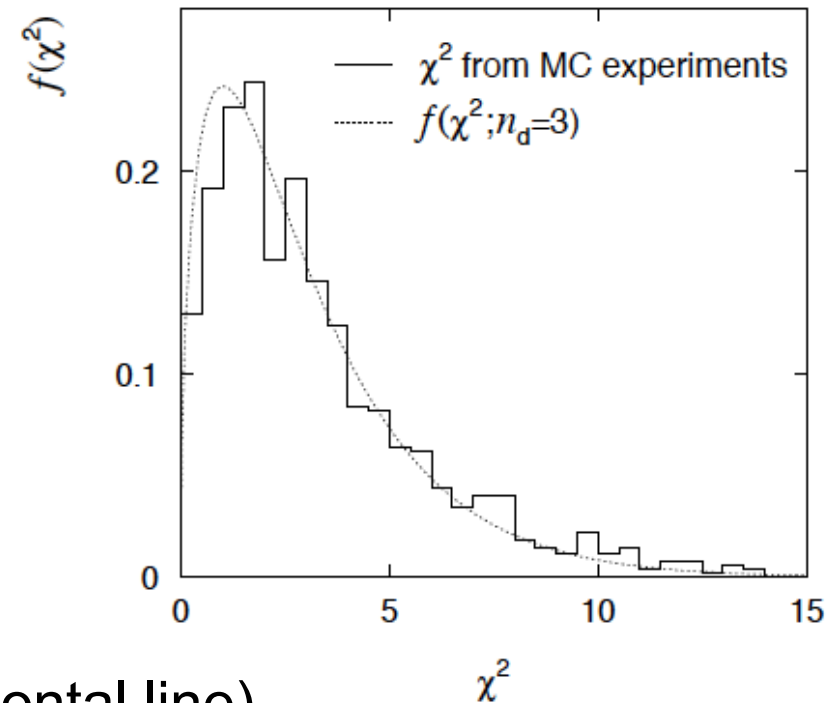
$$\chi_{\min}^2 = 3.99 \quad n_d = 5 - 2 = 3 \quad p = 0.263$$

That is: if the true function  $\lambda(x)$  were a straight line and if the experiment were repeated many times, each time yielding values for the 2 parameters and the  $X^2$ , then one would expect ...

# Goodness-of-fit with least squares - 4



... the  $\chi^2$  values to be worse (i.e. higher) than the one actually obtained in 26.3% of the cases.



Whereas for the 0<sup>th</sup> order polynomial (horizontal line)

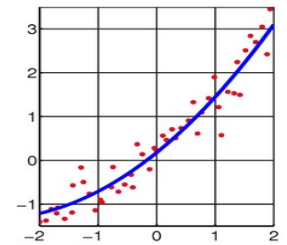
$$\chi_{\min}^2 = 45.5$$

$$n_d = 5 - 1 = 4$$

$$p = 3.1 \times 10^{-9}$$

This hypothesis can be safely ruled out !!!

# Goodness-of-fit vs statistical errors



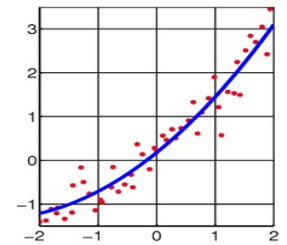
Keep in mind the distinction between having **SMALL STATISTICAL ERRORS** and having a **GOOD (small)  $X^2$** .

The statistical errors are related to the change in  $X^2$  when the parameters are varied away from their fitted values, and not to the absolute value of  $X^2$  itself.

The standard deviation  $\sigma_{\hat{\theta}}$  of an estimator  $\hat{\theta}$  is a measure of how widely estimates would be distributed if the experiment were to be repeated many times.

If the function form of the hypothesis is incorrect, however, then the estimate  $\hat{\theta}$  can still differ significantly from the true value  $\theta$ . That is, if the form of the hypothesis is incorrect, then a small standard deviation is not sufficient to imply a small uncertainty in the estimate of the parameter.

# Goodness-of-fit vs statistical errors



Example: horizontal line fit

First case: page 16

$$\hat{\theta}_0 = 2.66 \pm 0.13$$

$$\chi^2_{\min} = 45.5$$

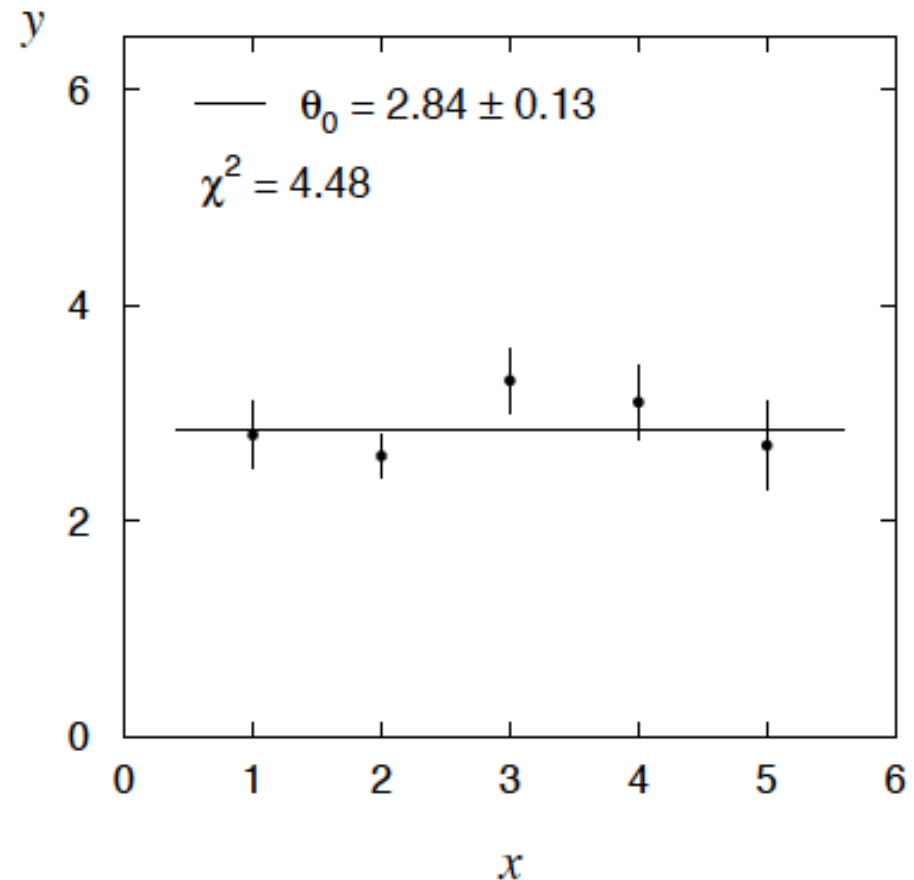
Now keep x and the errors, move the data points wrt previous example

New fit:

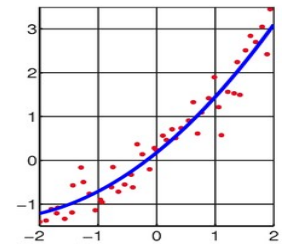
$$\hat{\theta}_0 = 2.84 \pm 0.13$$

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

The variance is the same as before, but now the chi-square is good !!



# Least squares with binned data

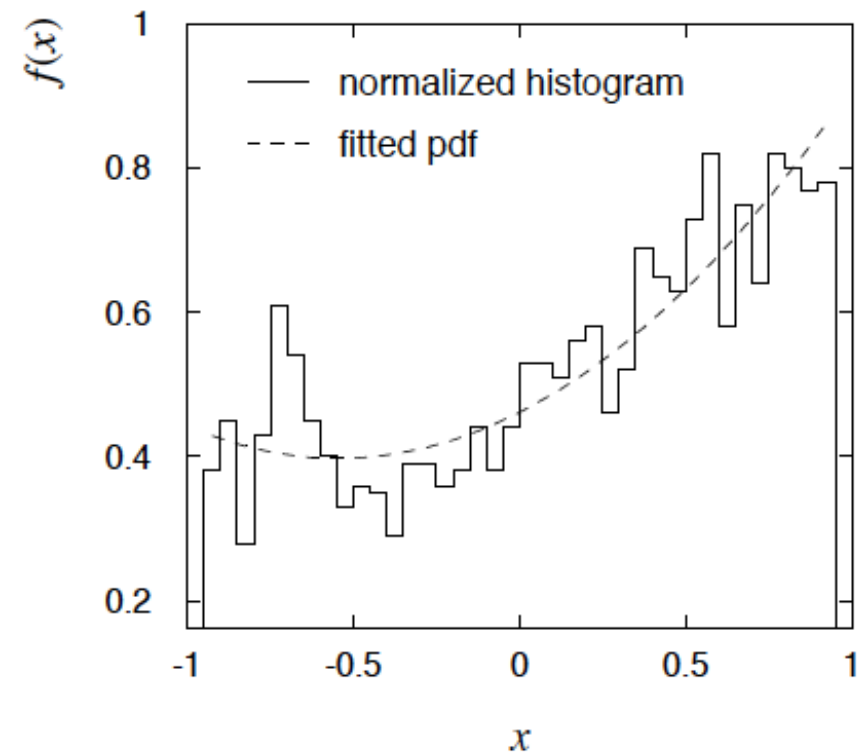


Suppose we have  $n$  observations of a random variable  $x$  from which one makes a histogram with  $N$  bins. Let  $y_i$  be the number of entries in bin  $i$ , and  $f(x; \theta)$  be a hypothesized pdf for which one would like to estimate the parameters  $\theta = (\theta_1, \dots, \theta_m)$ .

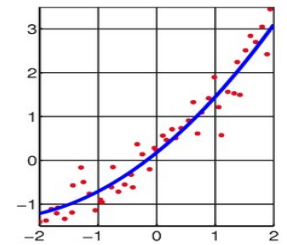
The number of entries predicted in bin  $i$  is:

$$\lambda_i = E[y_i] = n \int_{x_i^{\min}}^{x_i^{\max}} f(x; \theta) dx = np_i(\theta)$$

where  $x_i^{\min}$  and  $x_i^{\max}$  are the bin limits and  $p_i(\theta)$  is the probability to have an entry in bin  $i$ .



# Least squares with binned data



The parameters  $\theta$  are found by minimizing the quantity:

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

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

When the mean number of entries in each bin is small compared to the total number of entries, the contents of each bin are approximately Poisson distributed. The variance is therefore equal to the mean.

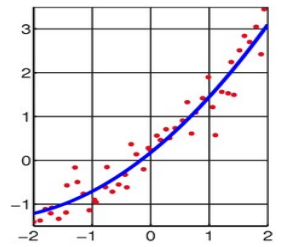
In place of the true variance take either

$$\begin{aligned} \sigma_i^2 &= \lambda_i(\theta) && \text{LS method} \\ \sigma_i^2 &= y_i && \text{Modified LS method} \end{aligned}$$

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



# LS with binned data: normalization



Appreciate the detail:  $f(x; \theta)$  is a pdf and therefore normalized to 1. The function which is fitted to the histogram is  $\lambda_i(\theta)$ .

Often, instead of using the observed total number of entries  $n$  to obtain  $\lambda_i$  (see page 23), an additional adjustable parameter  $\nu$  is introduced as normalization factor.

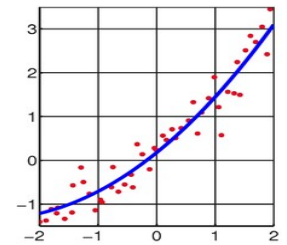
The predicted number of entries in the bin  $i$  then becomes:

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

$\nu$  is fitted along with  $\theta$ .

This might lead to an incorrect estimate of the total number of entries.

# LS with binned data: normalization



$\hat{v}$  is a bad estimator for  $n$ !

Take the formula on the page before, take the derivative of  $X^2$  with respect to  $v$  to 0, and we find:

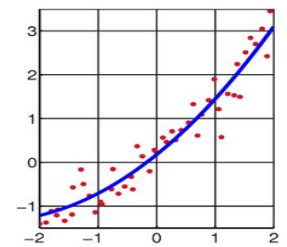
$$\hat{v}_{\text{LS}} = n + \frac{\chi_{\text{min}}^2}{2}$$

$$\hat{v}_{\text{MLS}} = n - \chi_{\text{min}}^2$$

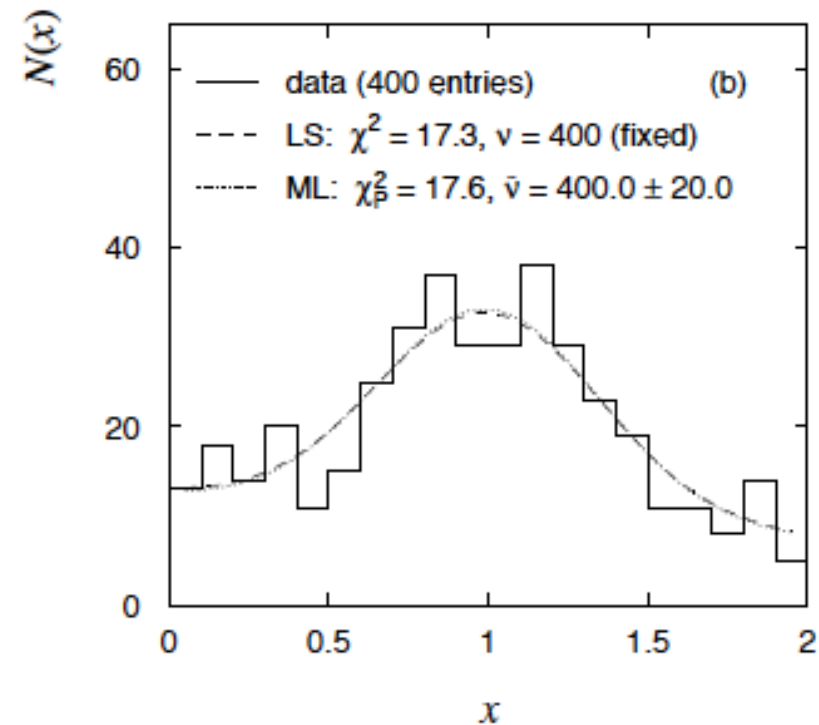
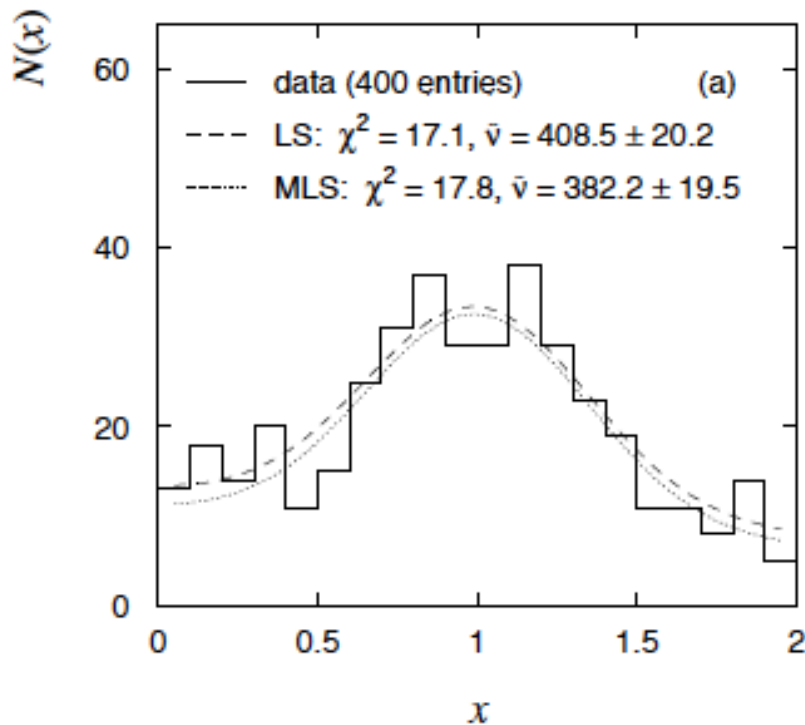
Since one expects a contribution to  $X^2$  on the order of 1 per bin, the relative error in the number of entries is typically  $N/2n$  too high (LS) or  $N/n$  too low (MLS). If one takes as a rule of thumb that each bin should have at least five entries, one could have an (unnecessary) error in the normalization of 10-20%.

→ general preference of maximum likelihood for histogram fitting

# LS with binned data: normalization



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

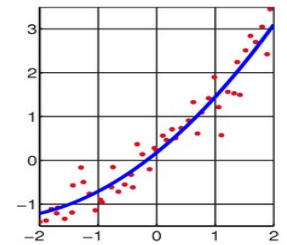


Expect  $\chi^2_{\min}$  around  $N - m$ ,

→ relative error in  $\hat{v}$  large when  $N$  large,  $n$  small

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

# Using LS to combine measurements



## Combine measurements of the same quantity:

Suppose that a quantity of unknown true value  $\lambda$  has been measured  $N$  times (e.g. in  $N$  independent experiments) yielding independent values  $y_i$  and estimated errors (standard deviations)  $\sigma_i$  for  $i = 1, \dots, N$ .

Since the true value is the same for all the measurements, the value  $\lambda$  is a constant, i.e. the function  $\lambda(x)$  is a constant, and thus the variable  $x$  does not actually appear in the problem.

$y_i$  = result of measurement  $i$ ,  $i = 1, \dots, N$

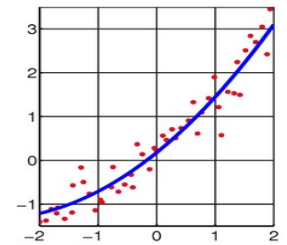
$\sigma_i = V[y_i]$ , assume known

$\lambda$  = true value (plays role of  $\theta$ )

For uncorrelated  $y_i$ , minimize:

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

# Using LS to combine measurements



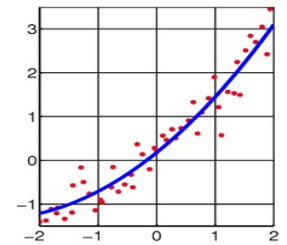
Set  $\frac{\partial \chi^2}{\partial \lambda} = 0$  and solve to get the LS estimator for  $\hat{\lambda}$

$$\hat{\lambda} = \frac{\sum_{i=1}^N y_i / \sigma_i^2}{\sum_{j=1}^N 1 / \sigma_j^2} \quad \text{Weighted average}$$

Variance: 
$$V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^N 1 / \sigma_i^2}$$

The variance of the weighted average is smaller than any of the variances of the individual measurements. Furthermore, if one of the measured  $y_i$  has a much smaller variance than the rest, then this measurement will dominate both in the value and variance of the weighted average.

# Using LS to combine measurements



Generalization to the case where the measurements are not independent.

This would occur, for example, if they are based at least in part on the same data.

Covariance matrix:  $\text{cov}[y_i, y_j] = V_{ij}$ . Minimize:

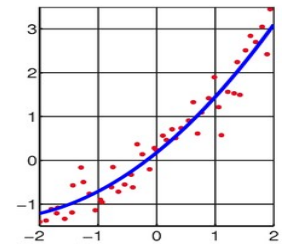
$$\chi^2(\lambda) = \sum_{i,j=1}^N (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda),$$

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

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

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

# Example: averaging two correlated measurements



Suppose we have  $y_1$  and  $y_2$ , two measurements with the covariance matrix:

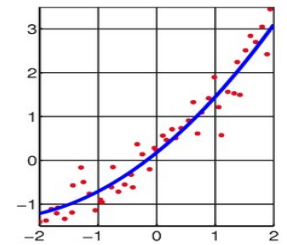
$$V = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}$$

where  $\rho = V_{12}/(\sigma_1 \sigma_2)$  is the correlation coefficient.

The inverse covariance matrix is given by:

$$V^{-1} = \frac{1}{1 - \rho^2} \begin{pmatrix} 1/\sigma_1^2 & (-\rho)/(\sigma_1 \sigma_2) \\ (-\rho)/(\sigma_1 \sigma_2) & 1/\sigma_2^2 \end{pmatrix}$$

# Example: averaging two correlated measurements



Using the formulas on page 30, we determine the weighted average:

$$\hat{\lambda} = w y_1 + (1-w) y_2$$

with

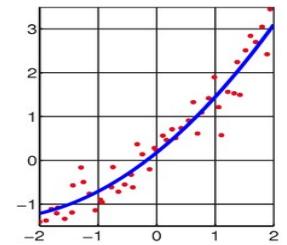
$$w = \frac{\sigma_2^2 - \rho \sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho \sigma_1 \sigma_2}$$

The variance is found to be:

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



# Example: averaging two correlated measurements



The change in the inverse variance due to the second measurement  $y_2$  is

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

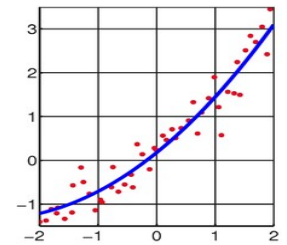
This is always greater than or equal to zero, which is to say that the second measurement always helps to decrease  $\sigma^2$ , or at least it never hurts.

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

$\rightarrow$  the weighted average does not lie between  $y_1$  and  $y_2$  !!!??

Cannot happen is correlation due to common data. But possible for shared random effect. Very unreliable if e.g.  $\rho$ ,  $\sigma_1$ ,  $\sigma_2$  incorrect

# Example



We measure the length of an object with 2 rulers made of different substances, so that the thermal expansion coefficients are different.

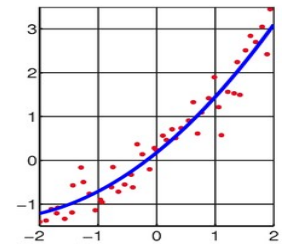
Suppose both rulers have been calibrated to give accurate results at a temperature  $T_0$ , but at any other temperature, a corrected estimate  $y$  of the true (unknown) length  $\lambda$  must be obtained using:

$$y_i = L_i + c_i (T - T_0)$$

$i$  refers to ruler 1 or 2,  $L_i$  is the uncorrected measurement,  $c_i$  is the expansion coefficient, and  $T$  is the temperature, which must be measured.

We will treat the measured temperature as a random variable with standard deviation  $\sigma_T$ , and we assume that  $T$  is the same for the 2 measurements, i.e. they are carried out together. The uncorrected  $L_i$  are treated as random variables, with standard deviation  $\sigma_{L_i}$ . Assume that  $c_i$ ,  $\sigma_{L_i}$  and  $\sigma_T$  are known.

# Example - Formulas



Goal: determine the weighted average of  $y_1$  and  $y_2$ .

We need their covariance matrix:

- Variances of the corrected measurements:  $\sigma_i^2 = V[y_i] = \sigma_{L_i}^2 + c_i^2 \sigma_T^2$

- Diagonal term: if the measurements are unbiased is:

$$E[y_i] = \lambda, \quad E[T] = T_{\text{true}} \quad \text{true (unknown) temperature}$$

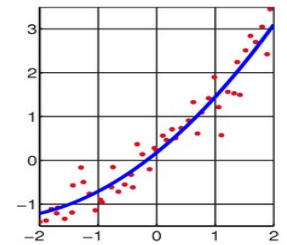
$$\rightarrow E[L_i] = \lambda - c_i (T_{\text{true}} - T_0)$$

$$E[T^2] = \sigma_T^2 + T_{\text{true}}^2$$

$$E[L_i, L_j] = \delta_{ij} \sigma_{L_i}^2$$

$$\begin{aligned} \text{Therefore: } \text{cov}[y_1, y_2] &= V_{12} = E[y_1 y_2] - \lambda^2 \\ &= c_1 c_2 \sigma_T^2 \end{aligned}$$

# Example - Formulas



The correlation coefficient is:

$$\rho = \frac{V_{12}}{\sigma_1 \sigma_2} = \frac{c_1 c_2 \sigma_T^2}{\sqrt{(\sigma_{L_1}^2 + c_1^2 \sigma_T^2) (\sigma_{L_2}^2 + c_2^2 \sigma_T^2)}}$$

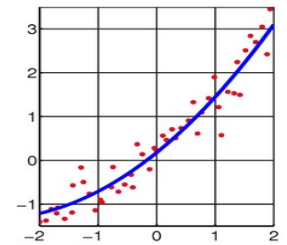
The weighted average is thus:

$$\hat{\lambda} = \frac{(\sigma_{L_2}^2 + (c_2^2 - c_1 c_2) \sigma_T^2) y_1 + (\sigma_{L_1}^2 + (c_1^2 - c_1 c_2) \sigma_T^2) y_2}{\sigma_{L_1}^2 + \sigma_{L_2}^2 + (c_1 - c_2)^2 \sigma_T^2}$$

- If  $\sigma_T$  is very small, then  $\rho$  goes to 0 and  $\hat{\lambda}$  will lie between  $y_1$  and  $y_2$ .
- If  $\sigma_{L_i}$  are very small and  $\sigma_T$  is large, the weight becomes negative!
- In the extreme of  $\rho=1$ , the weighted average becomes:

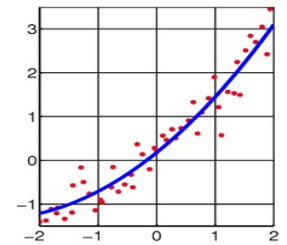
$$\hat{\lambda} = \frac{-c_1}{c_1 - c_2} y_1 + \frac{c_2}{c_1 - c_2} y_2$$

# Example - Illustration



- The 2 diagonal bands represent the possible values of the corrected length  $y_i$ , given measured lengths  $L_i$ , as a function of the temperature
- Suppose  $L_i$  known very accurately, and yet  $y_1$  and  $y_2$  differ by quite a bit
- → THEN the only available explanation is that the true temperature must be **different** from the measured value  $T$  !!!!
- The weighted average  $\hat{\lambda}$  is thus pulled towards the point where the bands of  $y_1(T)$  and  $y_2(T)$  cross.

# Example - Comments



Always question the averaging of correlated quantities!

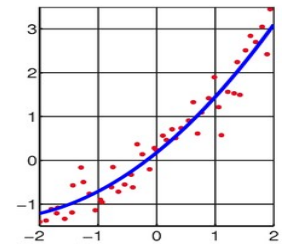
The correlation here stems from the fact that the individual measurements are based in part on a common measurement, here the temperature, which itself is a random quantity. **Averages of highly correlated quantities** should be treated with caution, since a small error in the covariance matrix for the individual measurements can lead to a large error in the average  $\hat{\lambda}$ , as well as an incorrect estimation of its variance.

This is particularly true if the magnitudes of correlated uncertainties are overestimated.

In the example, if the correction to the temperature  $\Delta T = \hat{T} - T$  turns out to be large compared to  $\sigma_T$ , then this means that our assumptions about the measurements are probably INCORRECT!

This is reflected in a high  $\chi^2_{\min}$  value, and a small p-value.

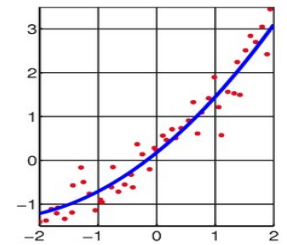
# Example – More comments



Here we would have to revise the model, perhaps reevaluating all standard deviations, or the relation between the measured and the corrected lengths. The correct conclusion is this example, would be that because of the large  $X^2$  the two measurements  $y_1$  and  $y_2$  are incompatible, and the average is therefore not reliable.

→ Comments about under- and over-estimation of errors, **PULLS**

# The next 2 lectures



- Statistical errors, confidence intervals and limits
- Statistical and systematic errors
- Unfolding of distributions

Don't forget to give me **suggestions** if you would like some topic to be discussed !!!!