Parameter estimation (Fitting)

Terascale Statistics Tools School

$1/2^{th}$ April 2014, DESY Hamburg

Olaf Behnke, DESY

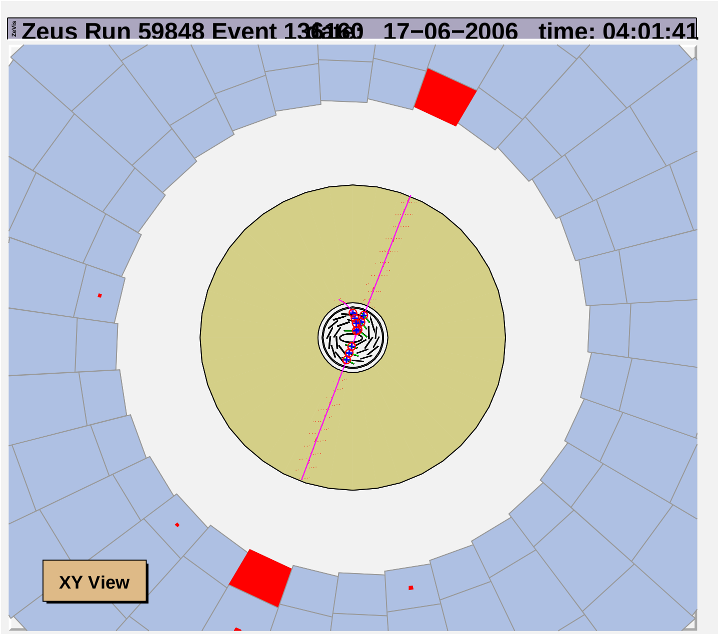
- Today: Least Squares fits
- Tomorrow: Maximum Likelihood fits and Masspeakfit-Tutorial (MINUIT)

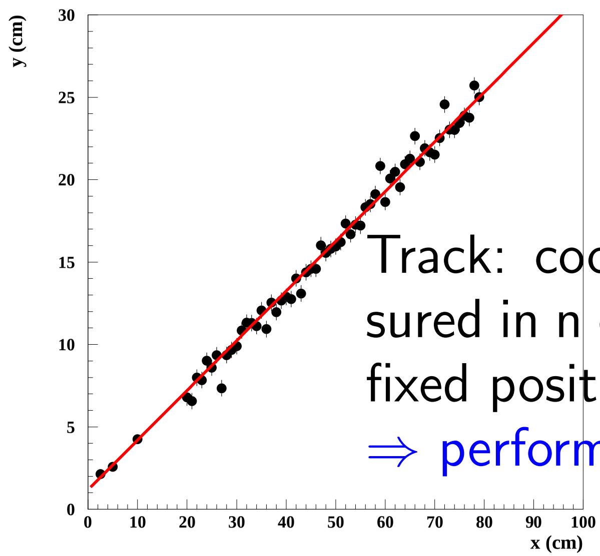
Literature:

- Roger Barlow: “Statistics, A Guide To The Use Of Statistical Methods In The Physical Sciences” Wiley & Sons, 1994
- Jay Orear: “Notes on Statistics for Physicists, Revised”, 1958, *http://www.astro.washington.edu/users/ivezic/Astr507/orear.pdf*
- Olaf Behnke, Kevin Kröninger, Gregory Schott and Thomas Schörner Sadenius: “Data Analysis in-High-Energy-Physics” Wiley & Sons, 2013

Intro: Fitting is essential for our measurements!

Example: for possible discovery
$Z' \to \mu^+\mu^-$ need precise muon track fits

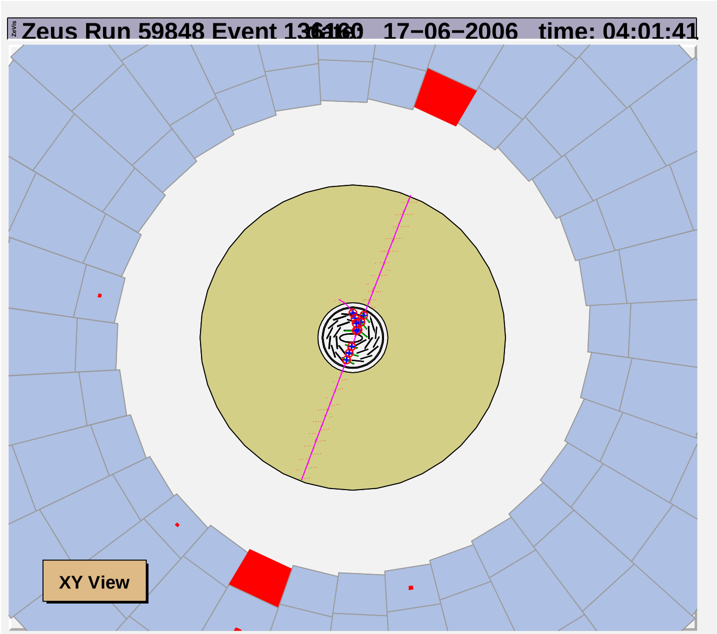


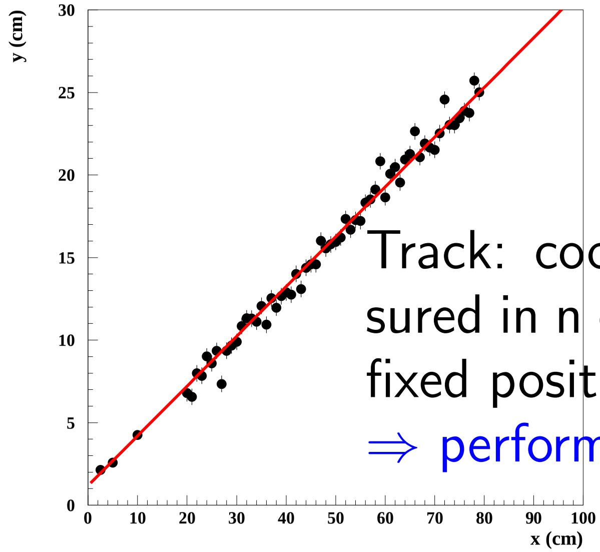


Track: coordinates y_i measured in n detector layers at fixed positions x_i
$\Rightarrow$ perform track fit

Intro: Fitting is essential for our measurements!

Example: for possible discovery
$Z' \to \mu^+\mu^-$ need precise muon track fits

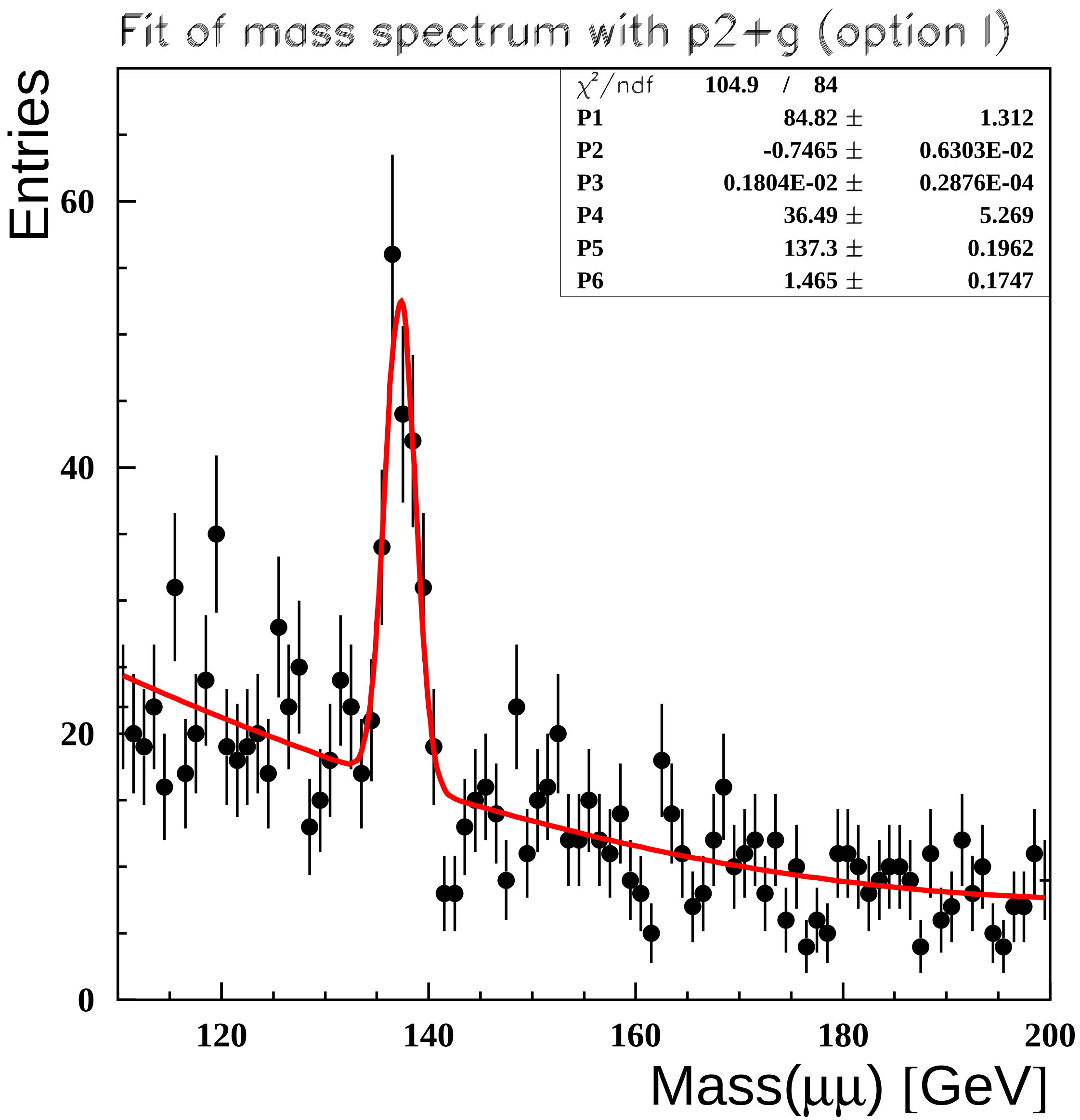




Track: coordinates y_i measured in n detector layers at fixed positions x_i
$\Rightarrow$ perform track fit

- Typical Assumptions:
  - Measurements with gaussian uncertainties
  - Linear(ized) model: $y = a_0 + a_1 x + a_2 x^2$ *(but could also use exact track helix model)*
- Construct χ^2:
  - $\chi^2 = \sum_i \frac{\left[y_i - (a_0 + a_1 x + a_2 x^2)\right]^2}{\sigma_i^2}$
  - Determine a_0, a_1, a_2 by finding χ^2 minimum (normal equations)
- Check consistency:
  - use χ^2 and χ^2-fit probability, reject outliers
- Analyse results:
  - parameters, errors and correlations (error ellipses), track trajectory error band
  - calculate momentum (error propagation)
- Analyse $\mu^+\mu^-$ mass spectrum obtained from many events $\Rightarrow$ see next page

Fitting is essential: Mass peak fit



Relevant questions:

- What you really want to know: Number of signal events, (cross section) mass and width of resonance
- Fit:
  - Use Poisson likelihood or χ^2 fit?
  - Parametrisation for signal and background
  - Choice of binning
  - etc.

All being discussed tomorrow!

Least Squares Fit lecture overview

1) Least square χ^2-fit method

   – most simple example: averaging measurements

2) Check consistency of a fit using χ^2_{min}

   – expected χ^2_{min} distributions, reject outliers

3) Linear least square fits

   – examples, normal equations, straight line fit

4) Non Linear least square fits

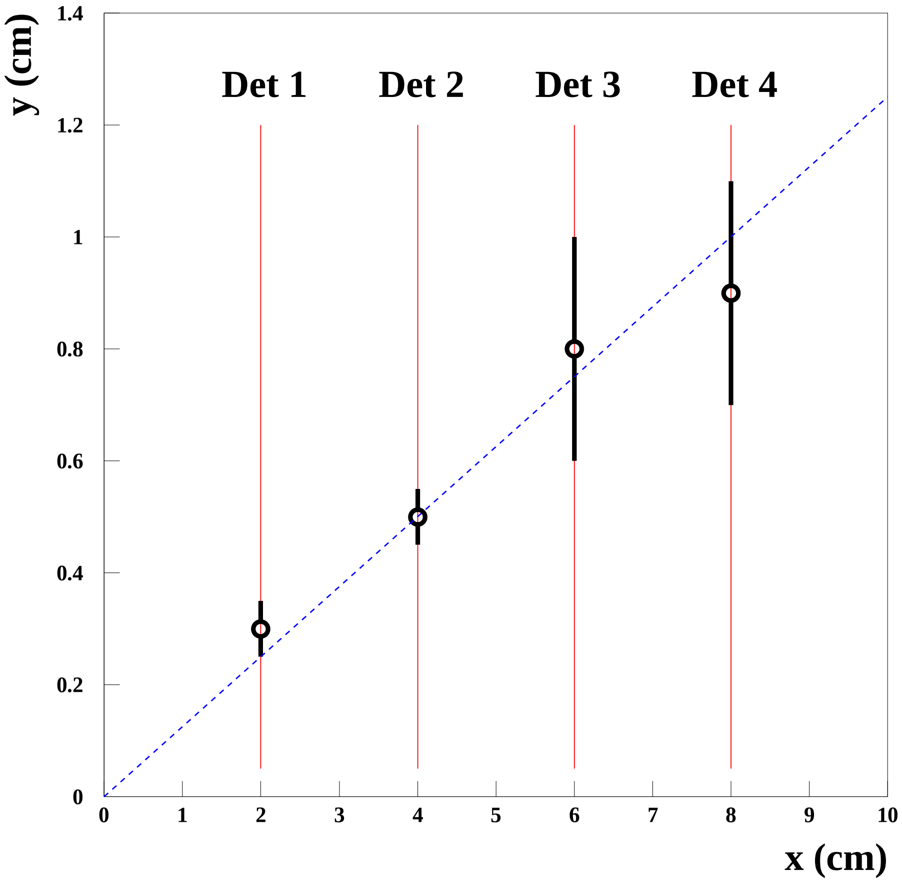
   – fitting a gaussian with unknown peak position

1. Least square χ^2-fit method

Please do not call it “schei..-squared-fit”

Method of least squares fit - Intro

Example: Particle trajectory measurement



n-measurements $y_i \pm \sigma_i$ at fixed x_i

Model: $y = f(x, a)$

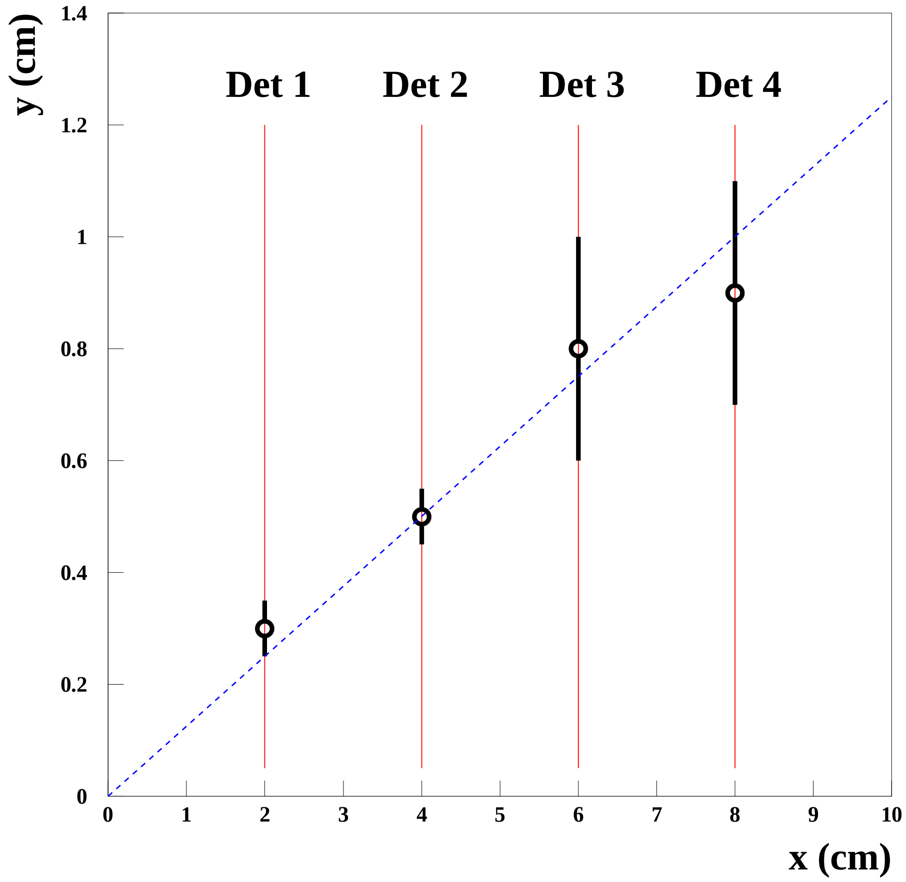
here: $y = ax$

$\Rightarrow$ how to determine a?

$\Rightarrow$

Method of least squares fit - Intro

Example: Particle trajectory measurement



n-measurements $y_i \pm \sigma_i$ at fixed x_i

Model: $y = f(x, a)$

here: $y = ax$

$\Rightarrow$ how to determine a?

$\Rightarrow$ Idea: for correct a one expects: $|y_i - f(x_i, a)| \lesssim \sigma_i$

Method of least squares fit - Intro

$\rightarrow$ $\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i, a))^2}{\sigma_i^2}}$ $\leftrightarrow$ Minimum w.r.t a

$\Rightarrow$ determine estimator $\hat{a}$ from $\frac{d\chi^2}{da} = 0$

$\Rightarrow$

Method of least squares fit - Intro

$\rightarrow$ $\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i,a))^2}{\sigma_i^2}}$ $\leftrightarrow$ Minimum w.r.t a

$\Rightarrow$ determine estimator $\hat{a}$ from $\frac{d\chi^2}{da} = 0$

$$\Rightarrow \boxed{\frac{d\chi^2}{da_{|a=\hat{a}}} = 2 \cdot \sum_{i=1}^{n} \frac{(y_i - f(x_i,a))}{\sigma_i^2} \cdot \frac{df(x_i,a)}{da} = 0}$$

In general not analytically solvable.

$\Rightarrow$ use iterative (numerical) methods (MINUIT, Mathematica)

Method of least squares fit

Most general case

- y_i, y_j correlated measurem. with cov. V_{ij}
- m fitparameters $\vec{a}$

$$\rightarrow \boxed{\begin{aligned} \chi^2 &= \sum_{i,j=1}^{n} (y_i - f(x_i, \vec{a})) V_{ij}^{-1} (y_j - f(x_j, \vec{a})) \\ &= \end{aligned}}$$

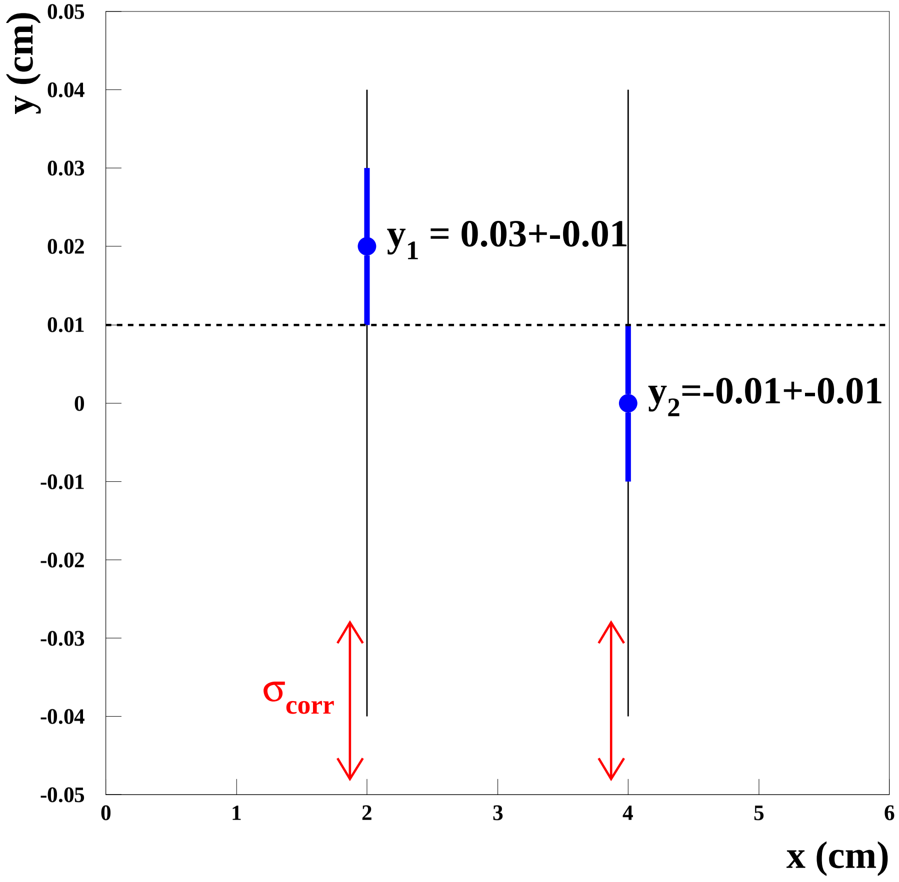
Method of least squares fit

Most general case

- y_i, y_j correlated measurem. with cov. V_{ij}
- m fitparameters $\vec{a}$

$$\rightarrow \boxed{\begin{aligned} \chi^2 &= \sum_{i,j=1}^{n} (y_i - f(x_i, \vec{a})) V_{ij}^{-1} (y_j - f(x_j, \vec{a})) \\ &= (\vec{y} - \vec{f}(\vec{a}))^t V^{-1} (\vec{y} - \vec{f}(\vec{a})) \end{aligned}}$$

Example for two correlated measurements

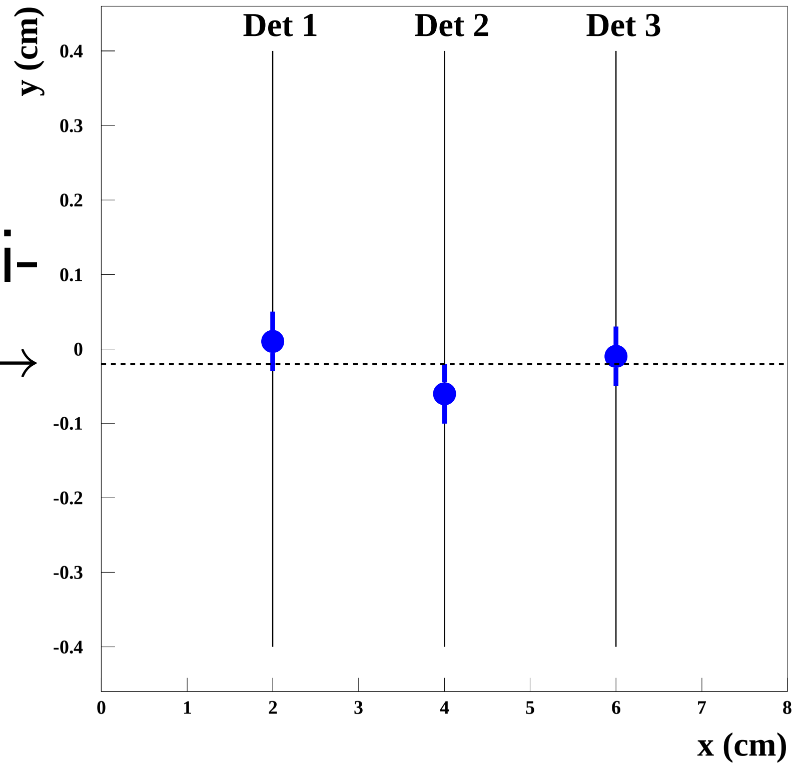


Measure track in two detector layers with global position uncertainty

$$V = \begin{pmatrix} 0.01^2 + \sigma_{corr}^2 & \sigma_{corr}^2 \\ \sigma_{corr}^2 & 0.01^2 + \sigma_{corr}^2 \end{pmatrix}$$

Fit of a constant

Measure position of horizontally flying particle $\longrightarrow$

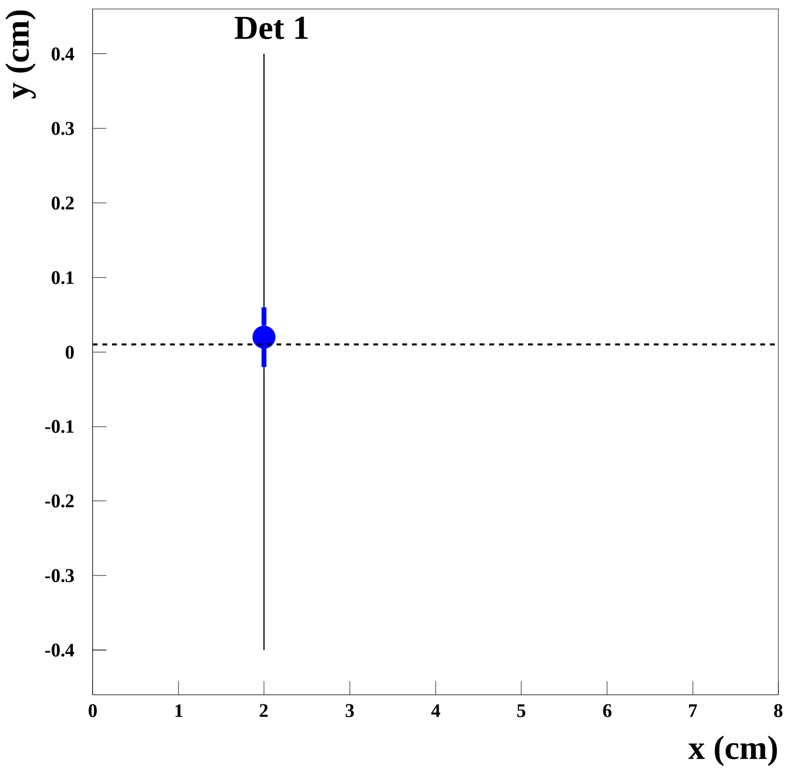


$\rightarrow$ Averaging of n measurements $y_i \pm \sigma_i$

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

Fit of a constant (one measurement)

"Idiot example" of one measurement $y_1 \pm \sigma_1$:



$$\chi^2 = \frac{(y_1 - a)^2}{\sigma_1^2}$$

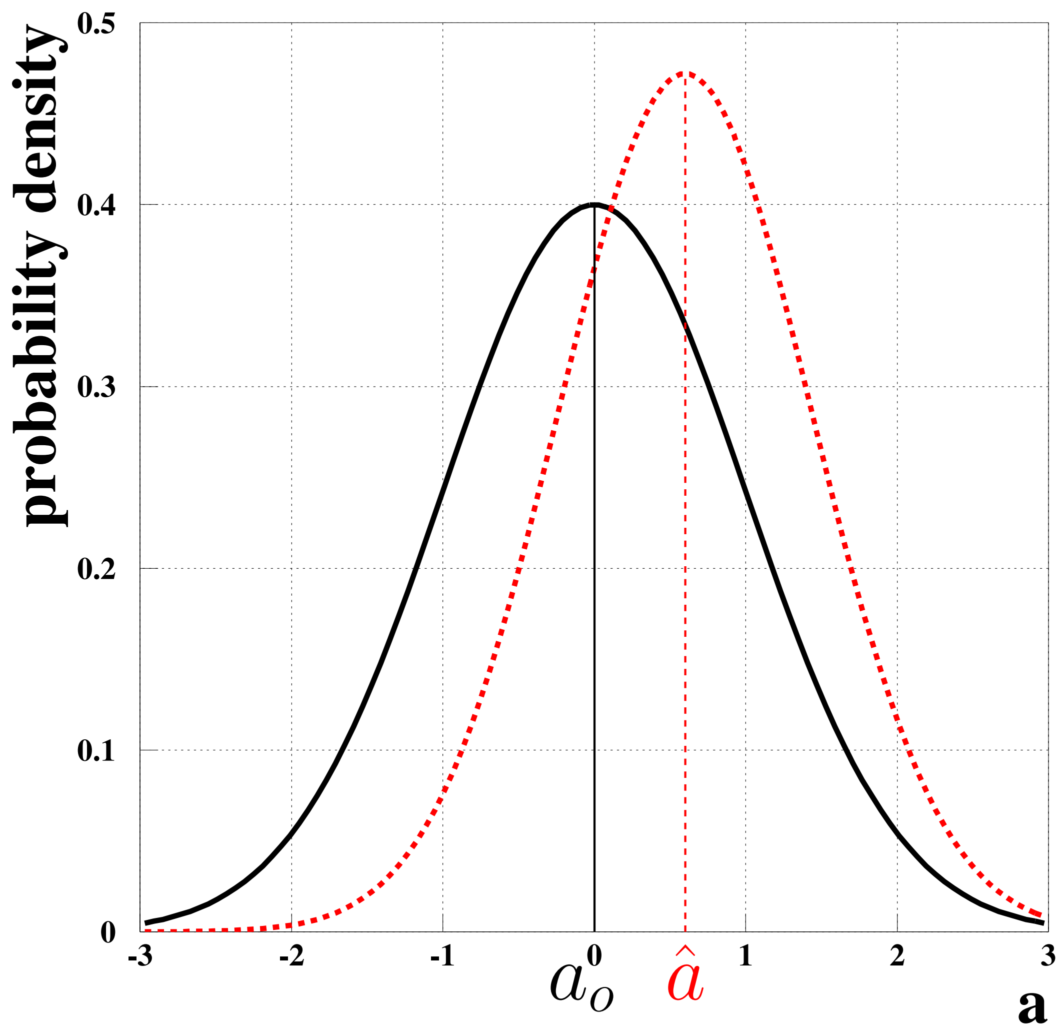
$$Min.\chi^2 : \quad \frac{d\chi^2}{da} = 0$$

$\longrightarrow$ Estimated value: $\hat{a} = y_1$

$\longrightarrow$ Error propagation: $\sigma_{\hat{a}} = \sigma_1$

True and inverse probability densities for one measurement

with gaussian uncertainty: $\hat{a} = y_1$, $\sigma_{\hat{a}} = \sigma_1$



True probability density to observe $\hat{a}$ for given true value a_o:

$$p = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(\hat{a}-a_O)^2}{2\sigma^2}}$$

But what if we don't know a_0?

Estimate "inverse probability density" for a_o from measurement $\hat{a} \pm \sigma_{\hat{a}}$:

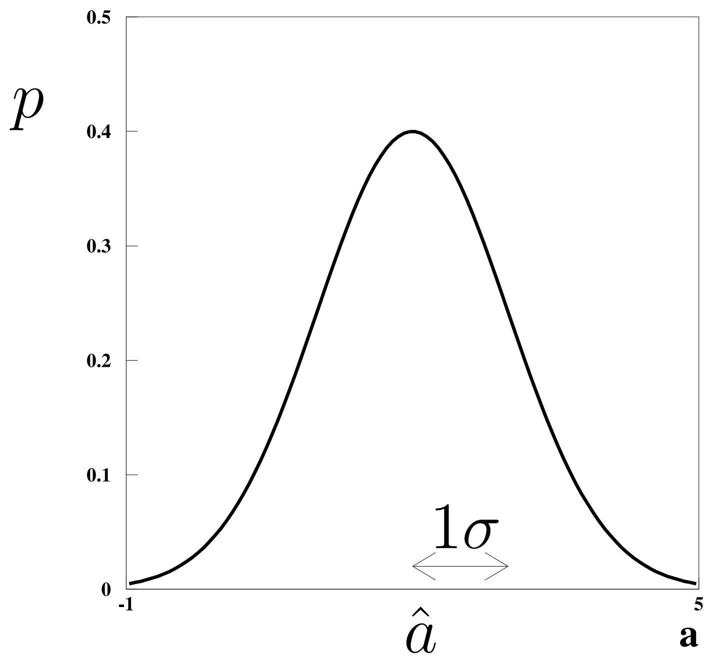
$$p = \frac{1}{\sqrt{2\pi}\sigma_{\hat{a}}} \cdot e^{-\frac{(\hat{a}-a_O)^2}{2\sigma_{\hat{a}}^2}}$$

Note: this is the Bayesian posterior probability density and not a real prob. density!

from now on we will use a as synonym for a_0!

Fit of a constant (one measurement)

Inverse probability density for true a:



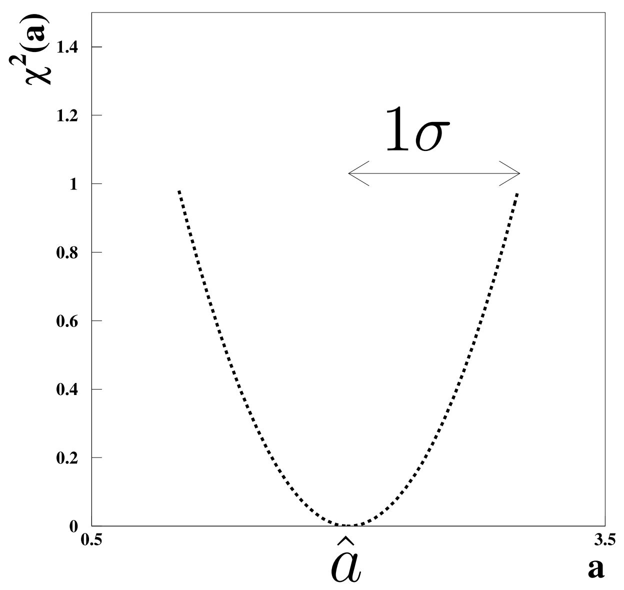
$$p \sim e^{-\frac{(a-\hat{a})^2}{2\sigma_{\hat{a}}^2}}$$

with $\chi^2 = \frac{(a-\hat{a})^2}{\sigma_{\hat{a}}^2} \Rightarrow p \sim e^{-\chi^2/2}$

Two methods to determine $\sigma_{\hat{a}}^2$:

$$\frac{1}{\sigma_{\hat{a}}^2} = \frac{1}{2}\frac{d^2\chi^2}{da^2}_{|a=\hat{a}}$$ or from

$$\chi^2(\hat{a} \pm \sigma_{\hat{a}}) - \chi^2(\hat{a}) = 1$$



Note: These are the two standard error determination methods for χ^2 fits! In general the second one is more reliable (See Max. Likelihood lecture)

Fit of a constant - many measurements

Probability density for true value a to observe measurements y_i, with $i = 1, n$:

$$p(y_1, y_2, ..., y_n|a) \propto \prod_{i=1}^{n} e^{-\frac{(y_i-a)^2}{2\sigma_i^2}}$$

$$= e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-a)^2}{\sigma_i^2}} = e^{-\frac{\chi^2}{2}}$$

but we don't know true a,

so *let's turn the whole thing around* to estimate probability density for true a from the measurements

Fit of a constant - many measurements

$$p(y_1, y_2, ..., y_n|a) = e^{-\chi^2/2}$$

Expand χ^2 around its minimum at $\hat{a}$:

$$\chi^2 = \chi^2(\hat{a}) + \underbrace{\frac{d\chi^2}{da}_{|a=\hat{a}}}_{=0} \cdot (a - \hat{a}) + \frac{1}{2}\frac{d^2\chi^2}{da^2}_{|a=\hat{a}} \cdot (a - \hat{a})^2$$

Fit of a constant - many measurements

$$p(y_1, y_2, \ldots, y_n | a) = e^{-\chi^2/2}$$

Expand χ^2 around its minimum at $\hat{a}$:

$$\chi^2 = \chi^2(\hat{a}) + \underbrace{\frac{d\chi^2}{da}_{|a=\hat{a}}}_{=0} \cdot (a - \hat{a}) + \frac{1}{2}\frac{d^2\chi^2}{da^2}_{|a=\hat{a}} \cdot (a - \hat{a})^2$$

$$= \chi^2(\hat{a}) + H \cdot (a - \hat{a})^2 \quad \text{with } H = \frac{1}{2}\frac{d^2\chi^2}{da^2}_{|a=\hat{a}}$$

'Hesse matrix'
(for one par. a number)

$$\Rightarrow \quad p(y_1, y_2, \ldots, y_n | a) \propto \underbrace{e^{-\frac{\chi^2(\hat{a})}{2}}}_{\text{Fit consistency}} \cdot \underbrace{e^{-\frac{1}{2} H \cdot (\hat{a} - a)^2}}_{\text{gaussian density}}$$

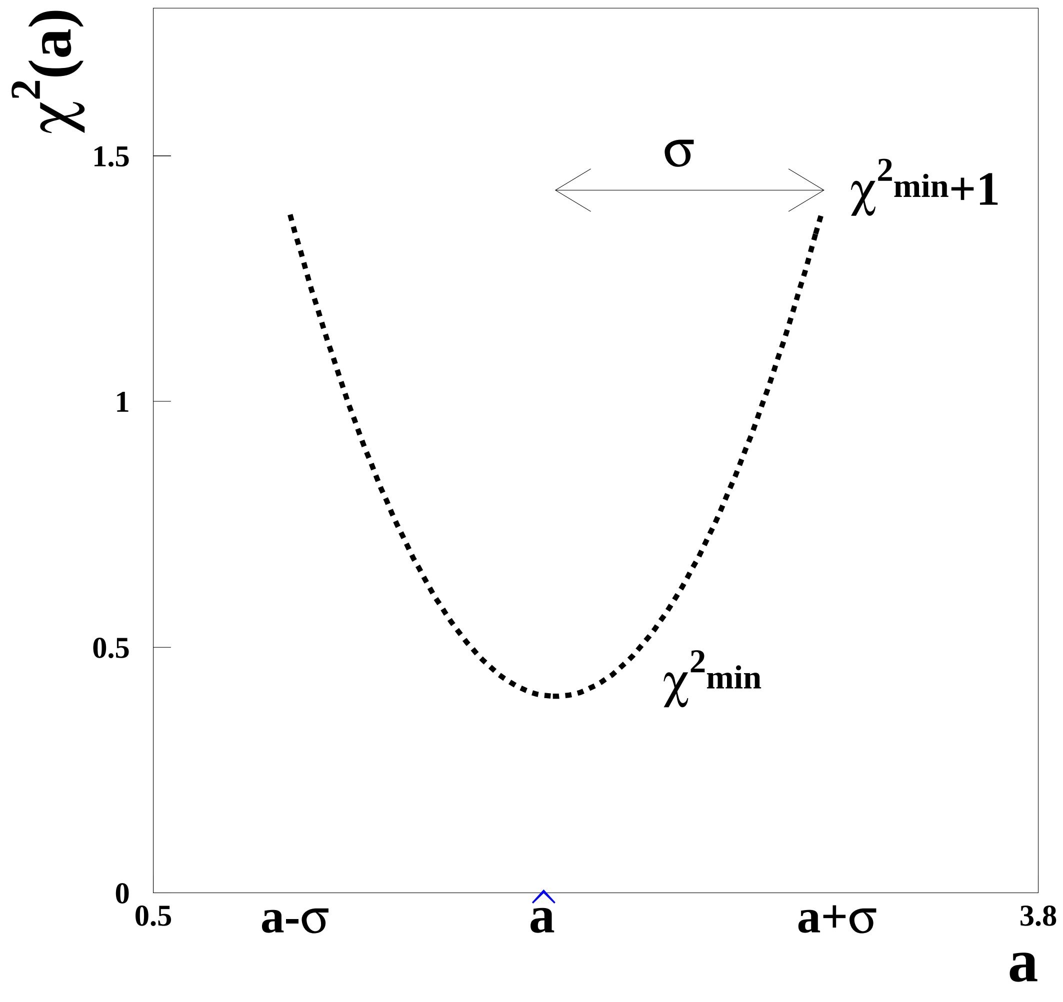
$\Rightarrow$ *interpreted as inverse probability density for true* a*:*

Gaussian distribution around $\hat{a}$ with width $\sigma = H^{-1/2}$

Generalisation to any one-parameter (linear) fit

$$\chi^2(a) = \chi^2(\hat{a}) + \frac{(a-\hat{a})^2}{\sigma_{\hat{a}}^2}$$

$$\rightarrow \quad \chi^2(\hat{a} \pm 1\sigma_{\hat{a}}) = \chi^2(\hat{a}) + 1 = \chi^2_{min} + 1$$



→ Read error directly from χ^2 curve

Mini-exercise

Averaging of two meas. via χ^2 parabolas

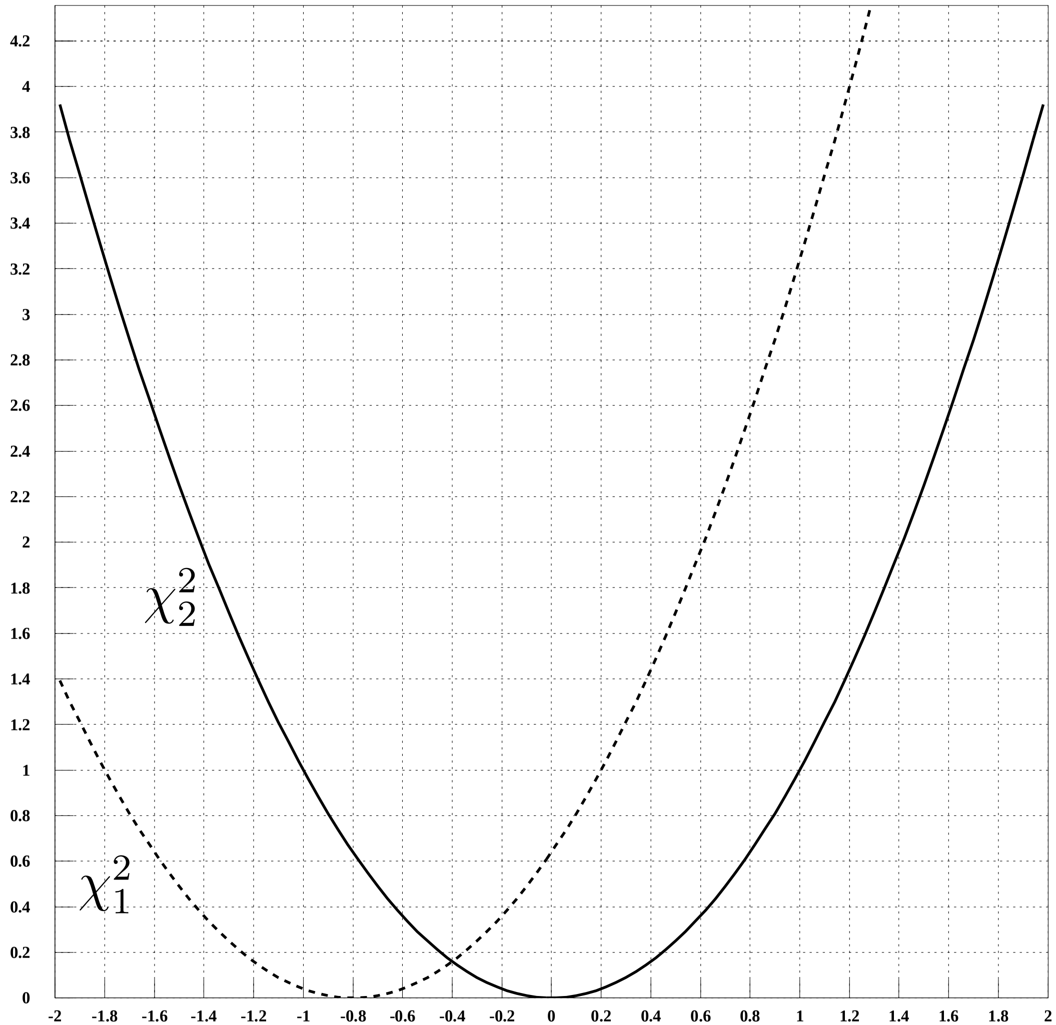
Paper exercise

Mini-exercise Averaging of two meas. via χ^2 parabolas

Two measurements y_1 and y_2 of the observable a are represented in the figure by χ^2 parabolas:
$\chi_i^2 = (y_i - a)^2/\sigma_i^2; \quad i = 1, 2$

- Determine (yes by eye!) from the two χ^2 curves the values y_1, σ_1 and y_2, σ_2
- Draw the total χ^2, i.e. the sum of the two parabolas (yes, do it by hand :-)) and determine $\hat{a}$ and $\sigma_{\hat{a}}$

  (use χ^2_{min} and $\chi^2 = \chi^2_{min} + 1$)
- How much is the error $\sigma_{\hat{a}}$ reduced compared to σ_1 and σ_2?
- Relax your eyes and hands ;-)



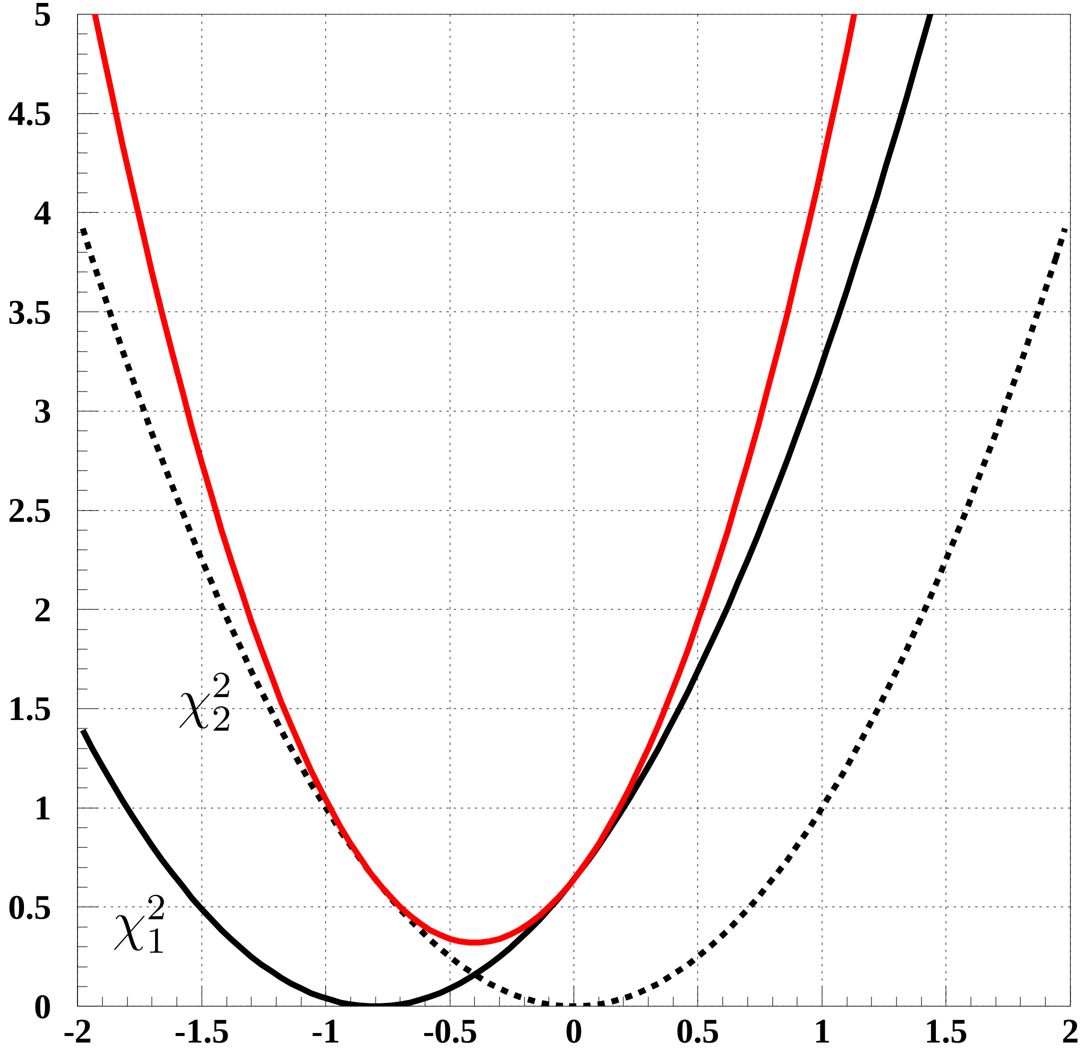
Mini-exercise Averaging of two meas. via χ^2 parabolas

Two measurements y_1 and y_2 of the observable a are represented in the figure by χ^2 parabolas:

$$\chi_i^2 = (y_i - a)^2/\sigma_i^2; \quad i = 1, 2$$

- Determine (yes by eye!) from the two χ^2 curves the values y_1, σ_1 and y_2, σ_2
- Draw the total χ^2, i.e. the sum of the two parabolas (yes, do it by hand :-)) and determine $\hat{a}$ and $\sigma_{\hat{a}}$

  (use χ^2_{min} and $\chi^2 = \chi^2_{min} + 1$)
- How much is the error $\sigma_{\hat{a}}$ reduced compared to σ_1 and σ_2?
- Relax your eyes and hands ;-)



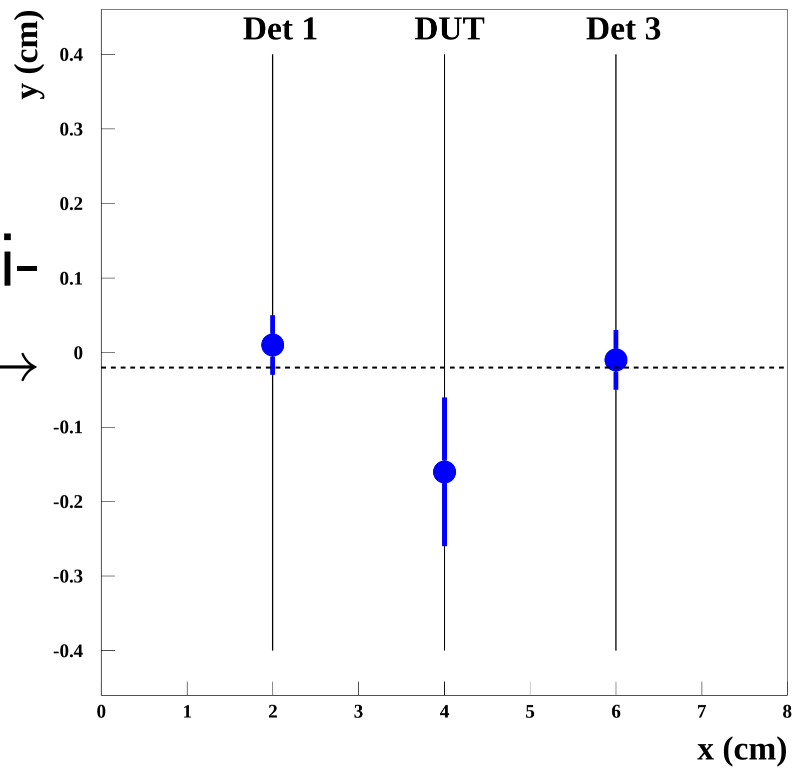
$$y_1 = -0.8 \pm 1$$
$$y_2 = 0 \pm 1$$
$$\hat{a} = -0.4 \pm 0.7$$

Averaging several measurements

n measurements $y_i \pm \sigma_i$ (Note: $\sigma_1 \neq \sigma_2$, etc.)

(Quiz question: Why is $\frac{1}{n}\Sigma y_i$ not the best average?

Measure position of horizontally flying particle $\longrightarrow$



Averaging several measurements

n measurements $y_i \pm \sigma_i$:

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

$$\frac{d\chi^2}{da} = 0 = \sum_{i=1}^{n} \frac{-2(y_i - a)}{\sigma_i^2} = -2 \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} + 2a \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$

Averaging several measurements

n measurements $y_i \pm \sigma_i$:

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

$$\frac{d\chi^2}{da} = 0 = \sum_{i=1}^{n} \frac{-2(y_i - a)}{\sigma_i^2} = -2 \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} + 2a \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$

$$\rightarrow \quad \boxed{\hat{a} \; = \; \sum_{i=1}^{n} \left[\frac{y_i}{\sigma_i^2}\right] \Big/ \sum_{i=1}^{n} \left[\frac{1}{\sigma_i^2}\right]}$$

$$\frac{1}{\sigma_{\hat{a}}^2} = \frac{1}{2}\frac{d^2\chi^2}{da^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$

Averaging - just reformulated

$\rightarrow$ Single measurements contribute with weight $G_i = \frac{1}{\sigma_i^2}$;

Define $G_s := \sum_{i=1}^{n} G_i$;   Hesse matrix $H = \frac{1}{2}\frac{d^2\chi^2}{da^2} = G_s$

$$\rightarrow \quad \hat{a} = \frac{1}{\sum_{i=1}^{n} G_i} \cdot \sum_{i=1}^{n} G_i y_i = \frac{1}{G_s} \cdot \sum_{i=1}^{n} G_i y_i$$

$\sigma_{\hat{a}}$ from simple error propagation:

$$\sigma_{\hat{a}}^2 = \sum_{i=1}^{n} \left(\frac{d\hat{a}}{dy_i}\right)^2 \cdot \sigma_i^2 = \sum_{i=1}^{n} \left(\frac{G_i}{G_s}\right)^2 \cdot \sigma_i^2$$

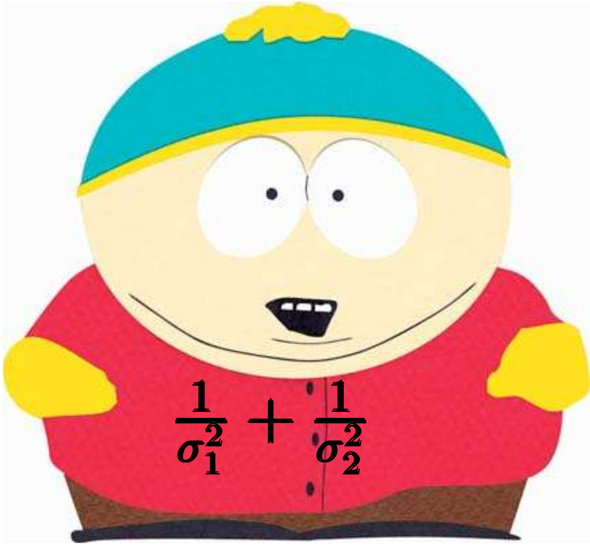
$$= \frac{1}{G_s^2} \cdot \sum_{i=1}^{n} G_i = \frac{1}{G_s} = \frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2}$$

$\Rightarrow$ Corollar: least square fitting is nothing else than a clever mapping of measurements to the fitparameters and obtaining fitparameter uncertainties using error propagation

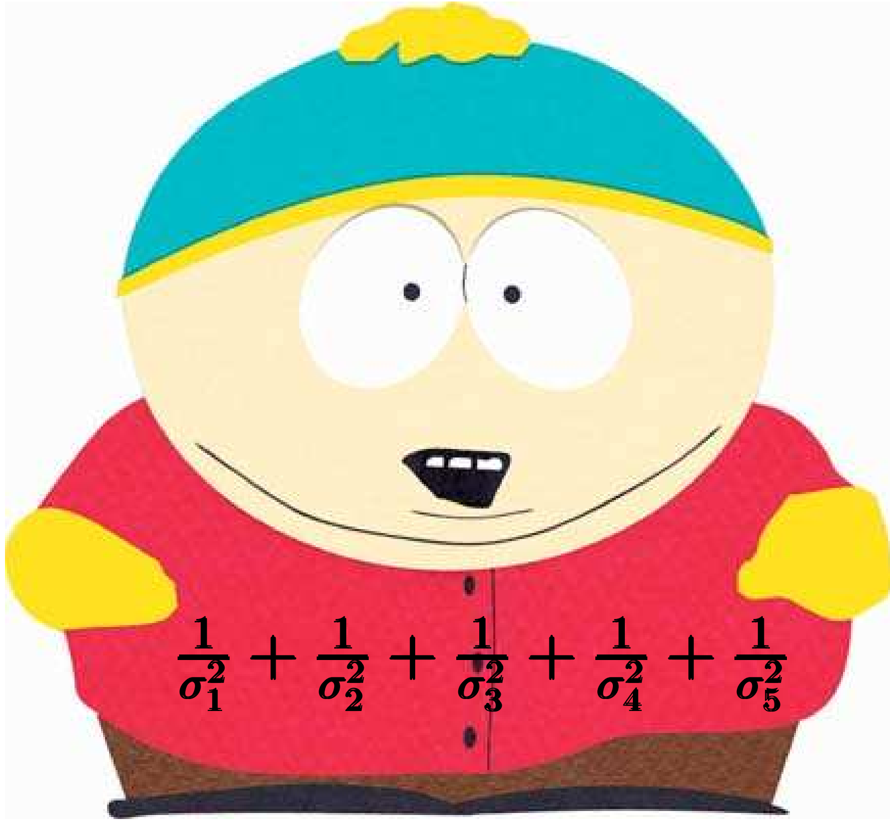
The role of the Hesse matrix

illustrated for weighted average (just a number)

$$H = \frac{1}{2}\frac{d^2\chi^2}{da^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$



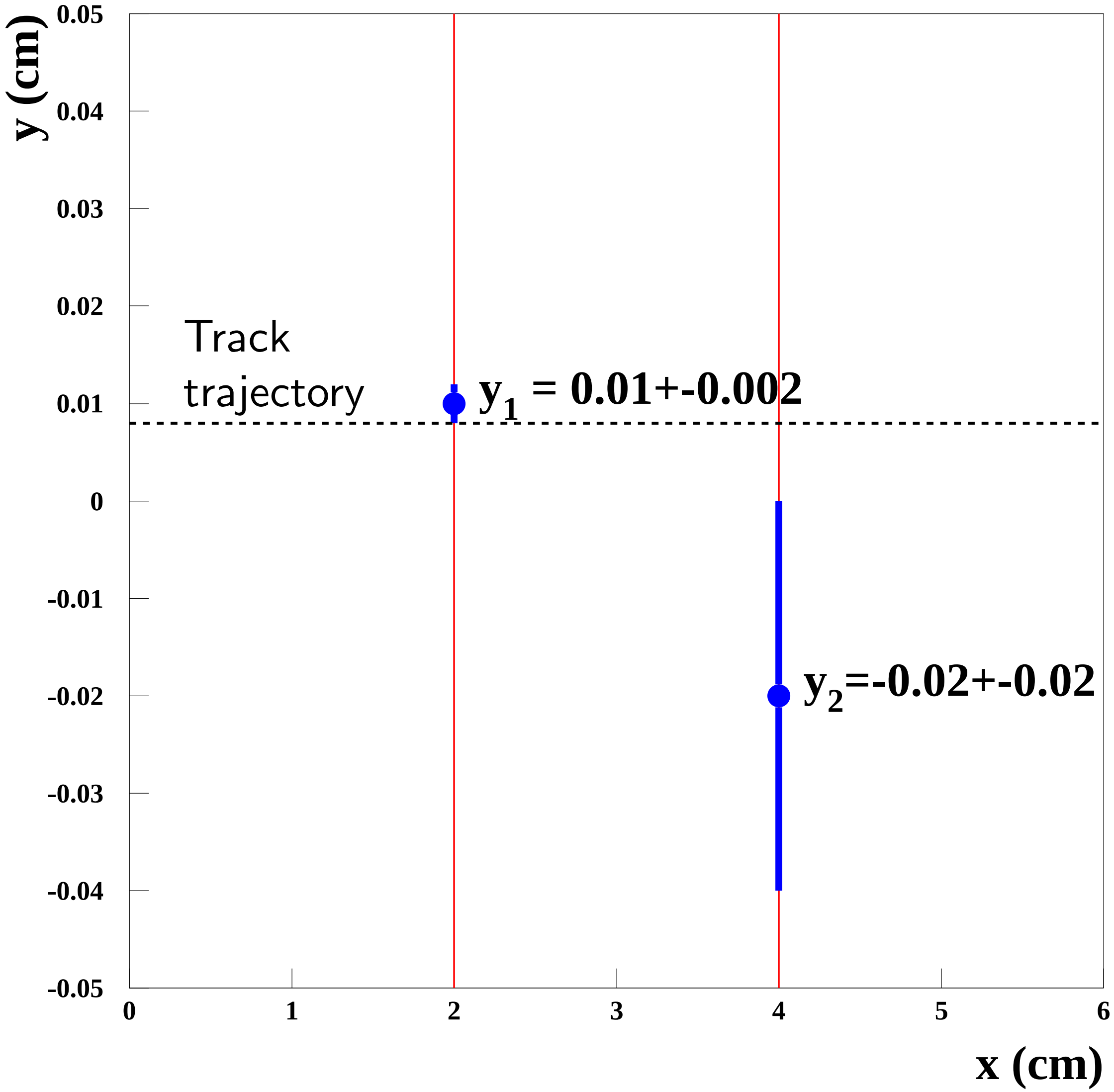
H **"grows" with each measurement** $\longrightarrow$



H is "counting the information" from the measurements

Finally $V = H^{-1}$ **Note: all this holds also for fits with many parameters**

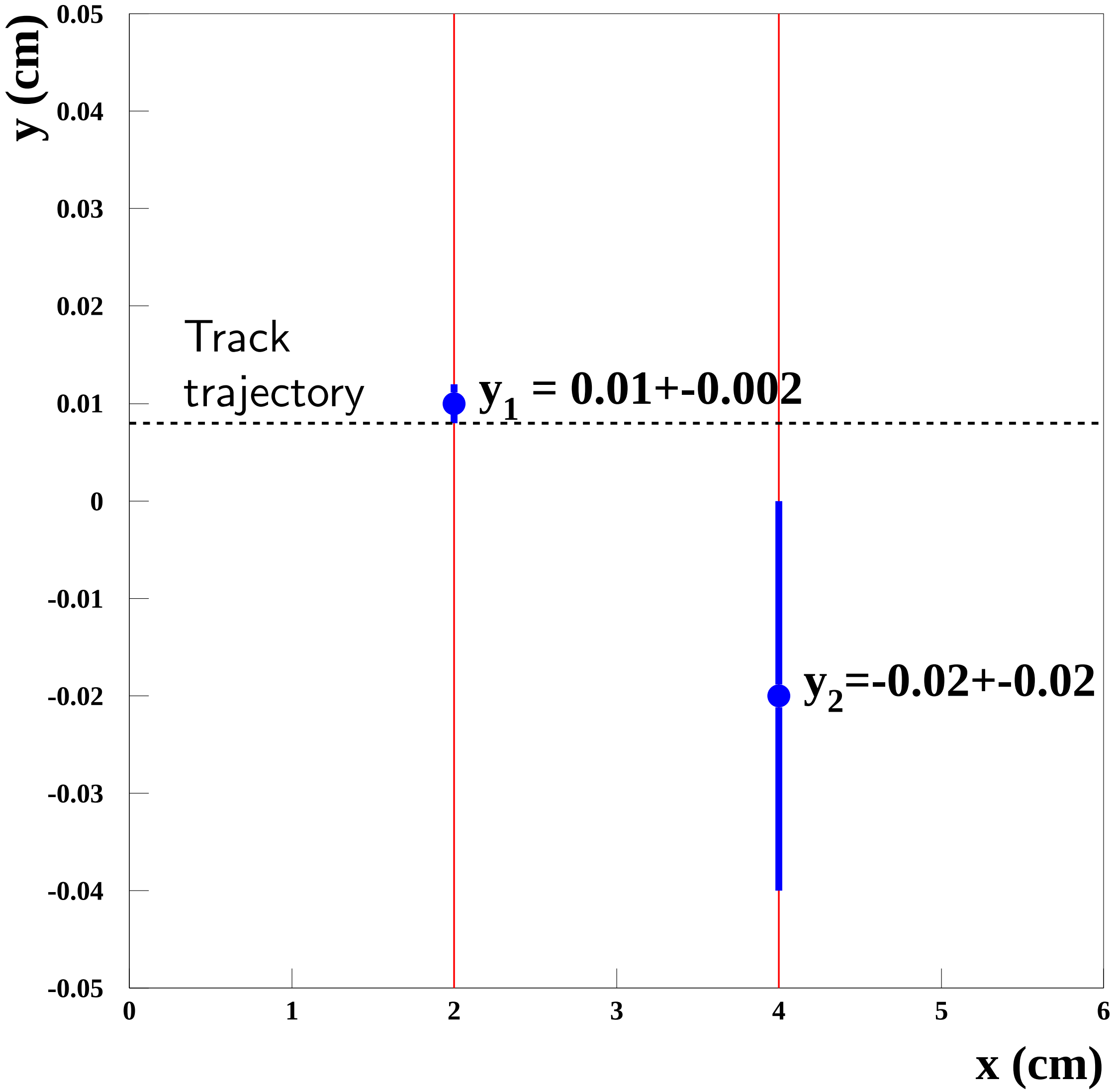
Mini-exercise Weighted average



Determine weighted average of two measurements:

Mini-exercise

Weighted average



Determine weighted average of two measurements:

$$\hat{y} = \frac{1}{\frac{1}{0.002^2} + \frac{1}{0.02^2}} \left(\frac{0.01}{0.002^2} + \frac{-0.02}{0.02^2} \right) = 0.0097$$

$$\sigma_{\hat{y}} = \sqrt{\frac{1}{\frac{1}{0.002^2} + \frac{1}{0.02^2}}} = 0.00199$$

Mini summary of what we have learnt

One parameter fits:

- **Least square expression for independent measurements:**
  $\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i,a))^2}{\sigma_i^2}$
  $\Rightarrow$ **get estimator** $\hat{a}$ **from minimum** $\chi^2 \Leftrightarrow d\chi^2/da_{|a=\hat{a}} = 0$
- **True physics parameters have a definite value, so true probability densities exist only for the measurements, fitting means in the Bayesian interpretation estimating posterior probability densities for the true parameters**
- $\frac{1}{\sigma_{\hat{a}}^2} = \frac{1}{2}\frac{d^2\chi^2}{da^2}_{|a=\hat{a}}$ **(general relation)**
- $\chi^2(\hat{a} \pm \sigma_{\hat{a}}) = 1$ **(general relation, more powerful)**
- **Averaging several measurements can be easily done graphically by adding individual** χ^2 **parabolas**
- **Least square fitting is nothing else than clever mapping of measurements to fitparameters; errors of fitparameters can be obtained from simple errorpropagation**
- **The Hesse matrix** $H = \frac{1}{2}\frac{d^2\chi^2}{da^2}_{|a=\hat{a}}$ **"counts the information" from the measurements**

2. Check consistency of a fit using χ^2_{min}

... (when) does it fit like a glove? ...

Consistency of measurements

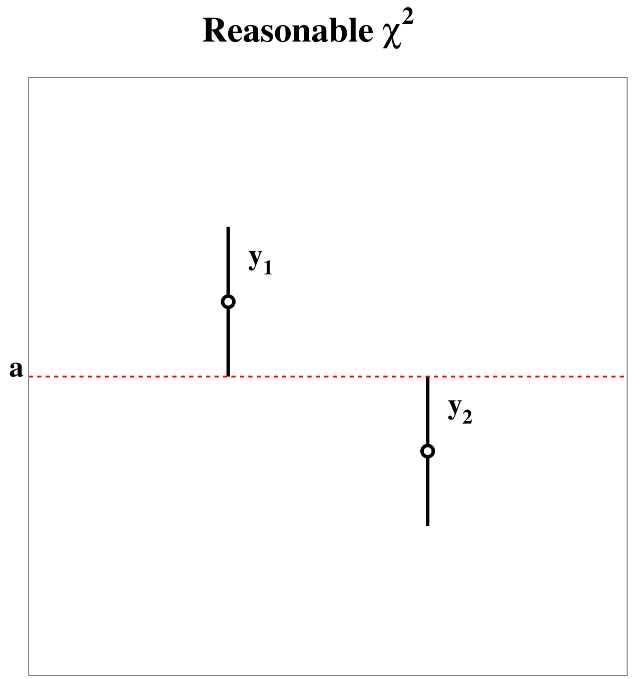
Recall "inverse probability density" for averaging n measurements:

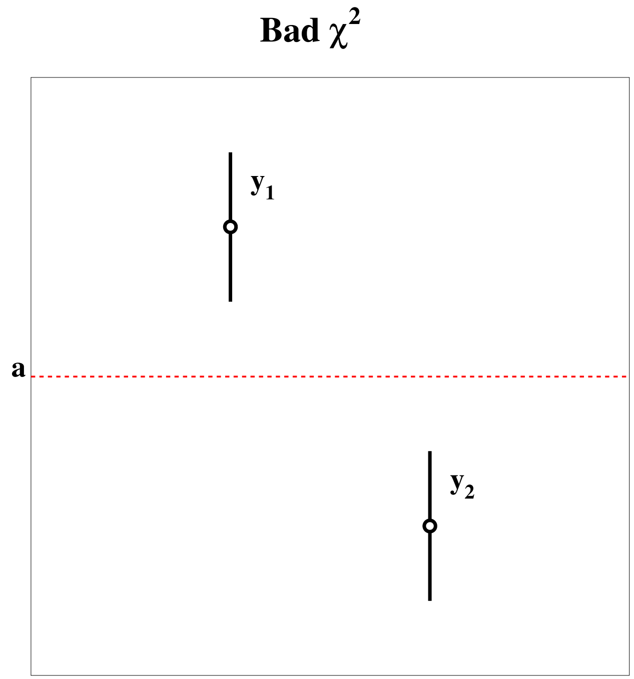
$$\Rightarrow p(y_1, y_2, \ldots, y_n | a) \propto \underbrace{e^{-\frac{\chi^2(\hat{a})}{2}}}_{\text{Fit consistency}} \cdot \underbrace{e^{-\frac{1}{2} H \cdot (\hat{a}-a)^2}}_{\text{gaussian density}}$$

Now lets have a closer look at the first term

Consistency of measurements

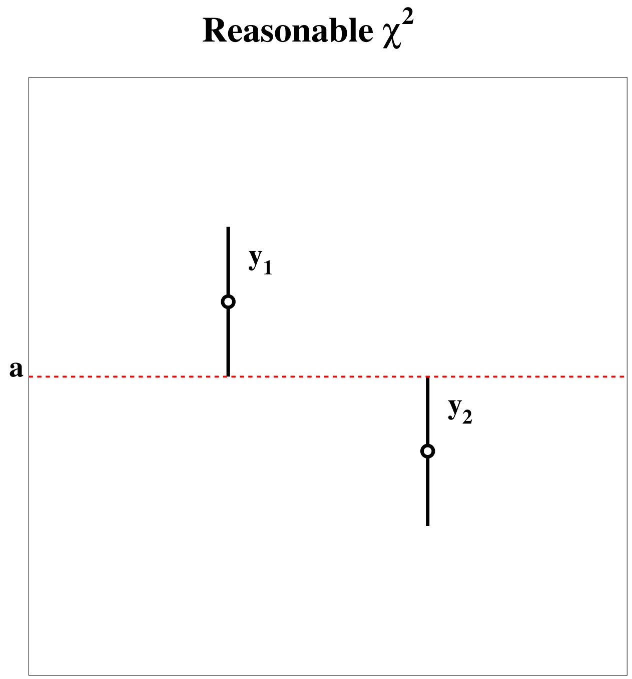
Example: Two measurements $y_1 \pm \sigma_1$ and $y_2 \pm \sigma_2$; the true value a be known, are the measurements consistent with a?:

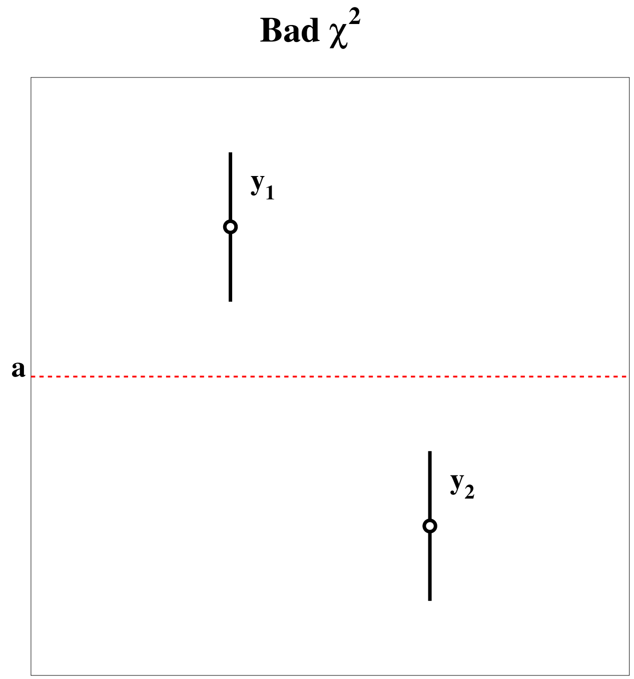




Consistency of measurements

Example: Two measurements $y_1 \pm \sigma_1$ and $y_2 \pm \sigma_2$; the true value a be known, are the measurements consistent with a?:





$\chi^2 = 2$ $\chi^2 = 8$

$\rightarrow \chi^2$ is a measure of consistency

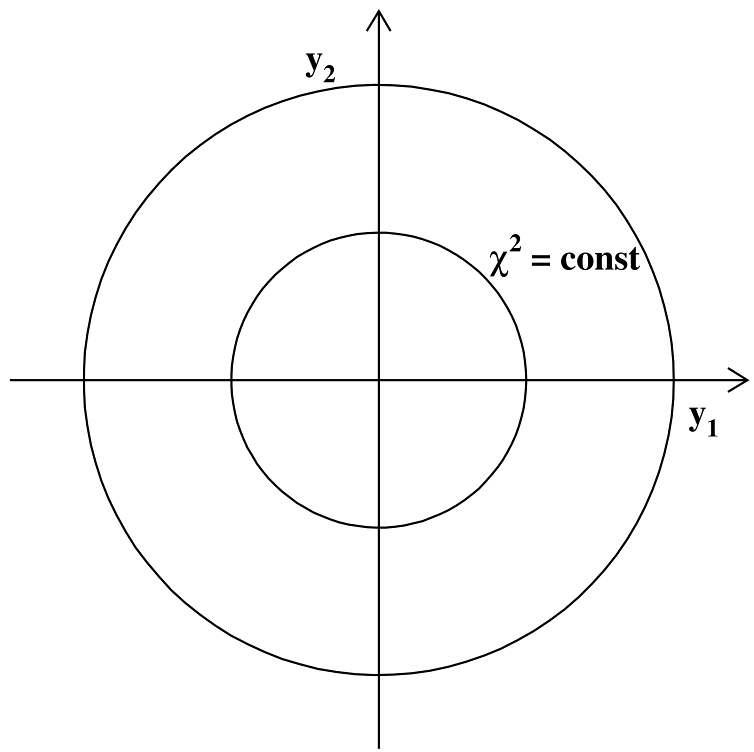
But how should χ^2 be distributed?

χ^2 for two measurements and known true value

Expected density for (y_1, y_2) (simple case $a = 0; \sigma_1 = \sigma_2 = 1$):

$$f(y_1, y_2) = \frac{1}{2\pi} e^{-y_1^2/2} e^{-y_2^2/2} = \frac{1}{2\pi} e^{-r^2/2}$$

with $r = \sqrt{y_1^2 + y_2^2} = \sqrt{\chi^2}$

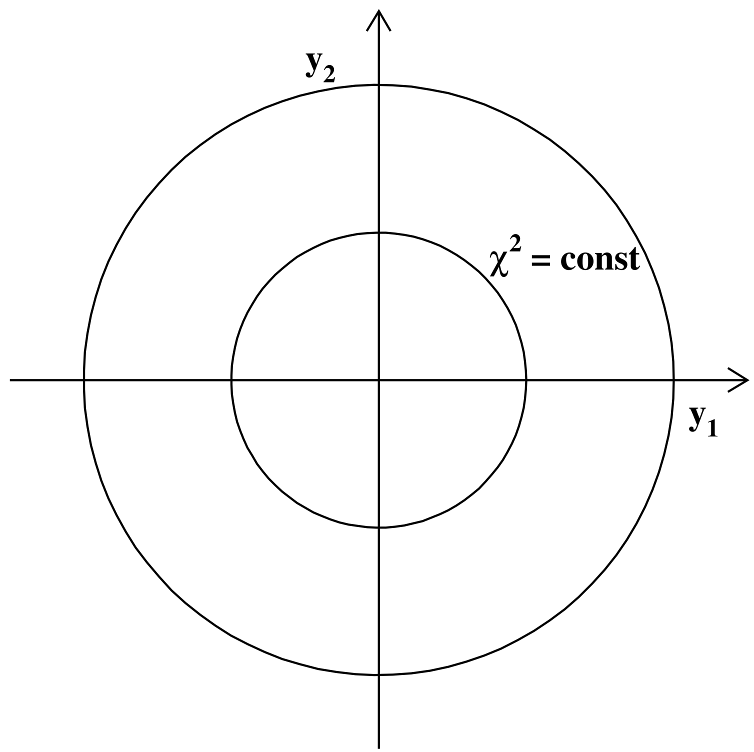


χ^2 for two measurements and known true value

Expected density for (y_1, y_2) (simple case $a = 0; \sigma_1 = \sigma_2 = 1$):

$$f(y_1, y_2) = \frac{1}{2\pi} e^{-y_1^2/2} e^{-y_2^2/2} = \frac{1}{2\pi} e^{-r^2/2}$$

with $r = \sqrt{y_1^2 + y_2^2} = \sqrt{\chi^2}$



Probability to find value between r and $r + dr$:

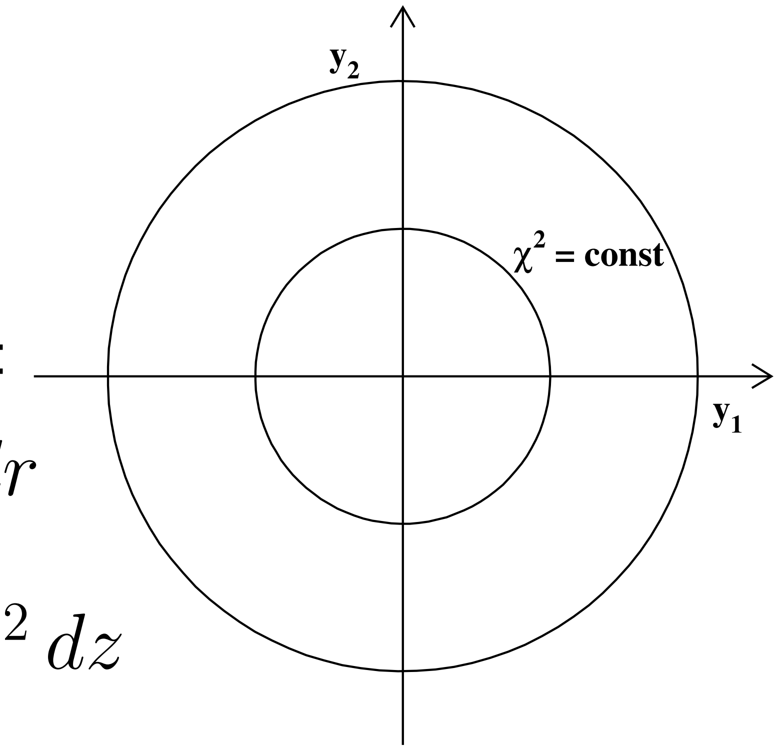
$$f(r)\, dr = 2\pi r dr \, \frac{1}{2\pi} e^{-r^2/2} = r e^{-r^2/2} \, dr$$

χ^2 for two measurements and known true value

Expected density for (y_1, y_2) (simple case $a = 0; \sigma_1 = \sigma_2 = 1$):

$$f(y_1, y_2) = \frac{1}{2\pi} e^{-y_1^2/2} e^{-y_2^2/2} = \frac{1}{2\pi} e^{-r^2/2}$$

with $r = \sqrt{y_1^2 + y_2^2} = \sqrt{\chi^2}$



Probability to find value between r and $r + dr$:

$$f(r)\,dr = 2\pi r dr\,\frac{1}{2\pi} e^{-r^2/2}\,dr = r e^{-r^2/2}\,dr$$

$$z = r^2 : \longrightarrow f(z)\,dz = f(r)\frac{dr}{dz}dz = \frac{1}{2}e^{-z/2}\,dz$$

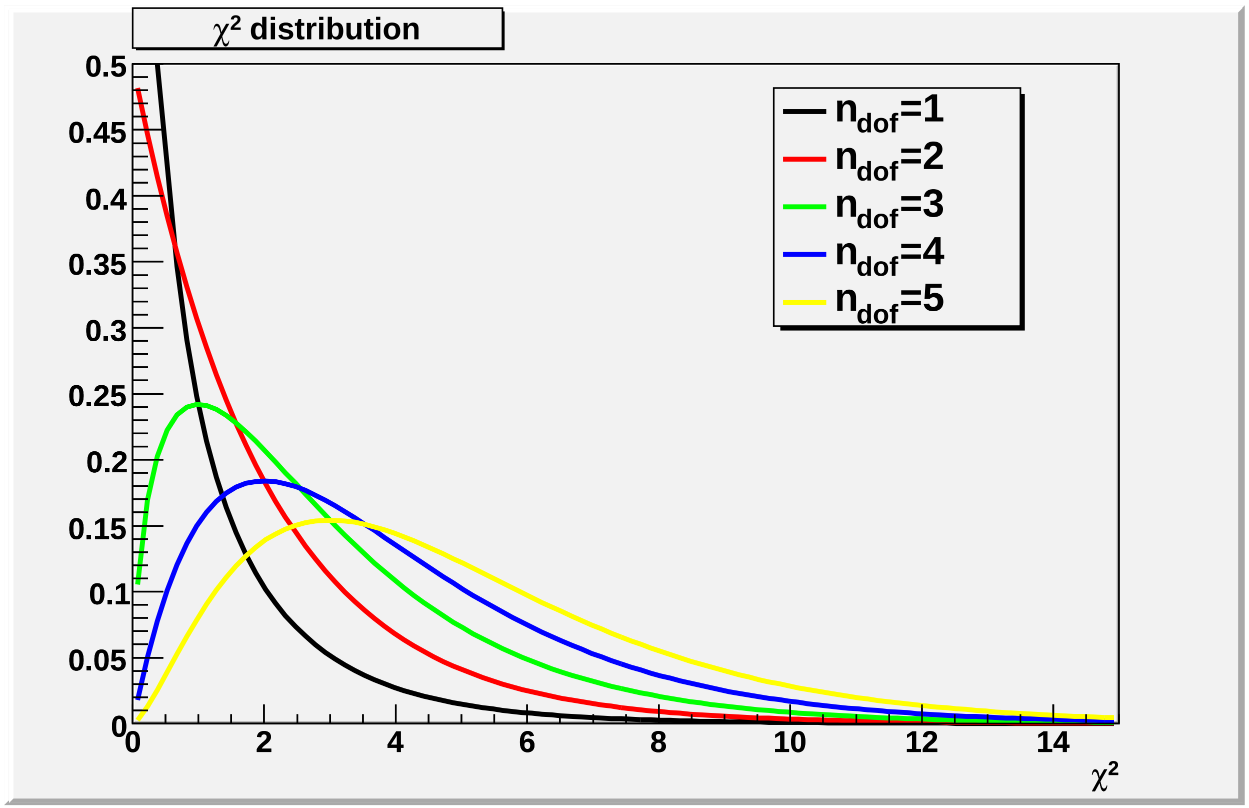
$\longrightarrow$ introduces χ^2-distribution for $z = \chi^2$ and two dimensions (ndf=2): $f(z, 2) = \frac{1}{2}e^{-z/2}$

χ^2-function for n degrees of freedom

$\rightarrow$ maps the χ^2 in n dimensions into probability density for χ^2

$$f(\chi^2, n) = \frac{1}{\Gamma(n/2)2^{n/2}} \cdot (\chi^2)^{n/2-1} \cdot e^{-\chi^2/2}$$

$$\text{with} \quad \Gamma(n/2) = \int_0^\infty dt e^{-t} t^{n/2-1}$$



Properties:

$\int_0^\infty f(\chi^2, n) d\chi^2 = 1$

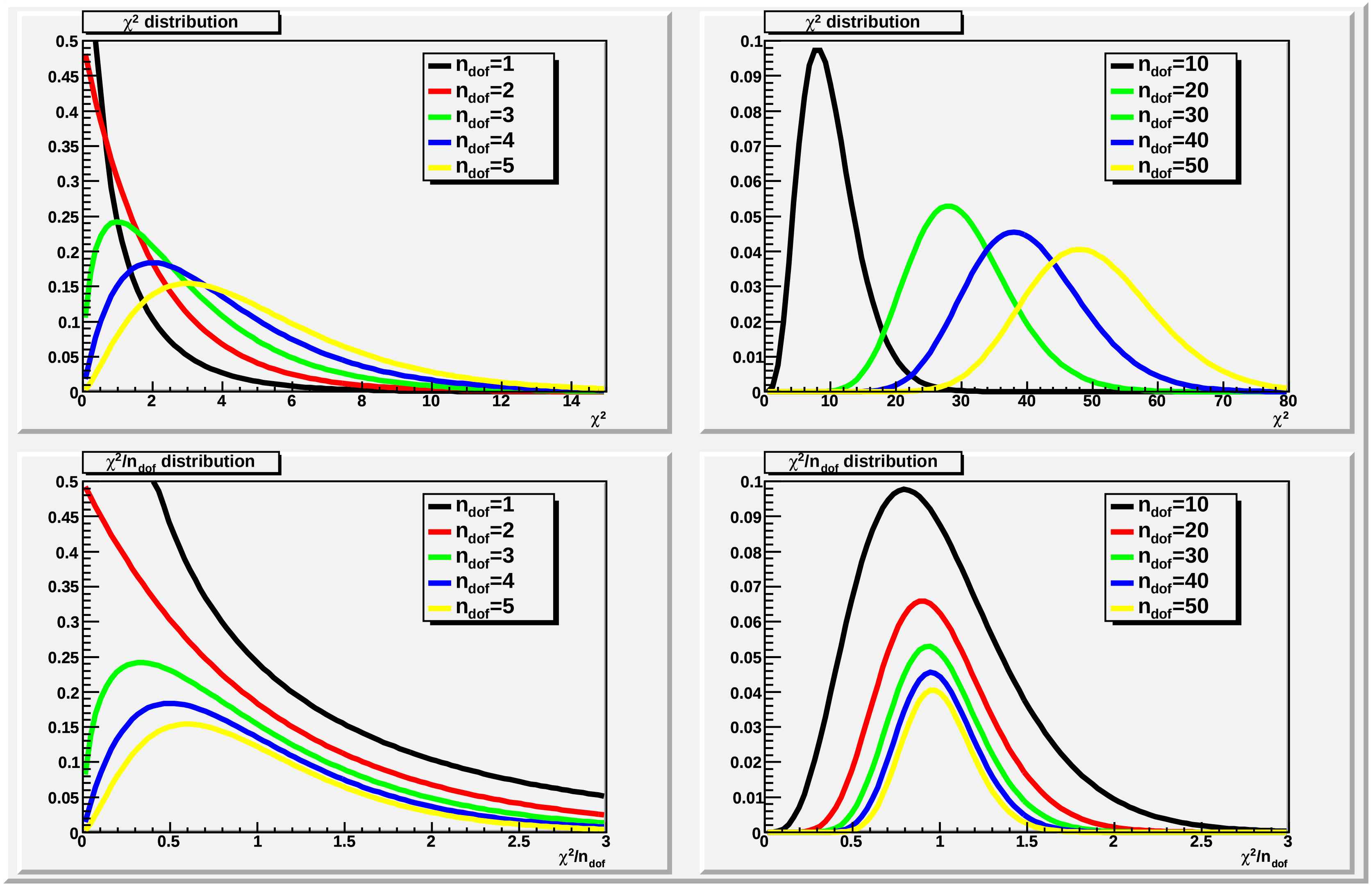
$\langle \chi^2 \rangle = n$

$V(\chi^2) = 2n;\ \sigma(\chi^2) = \sqrt{2n}$

$\langle \chi^2/n \rangle = 1$

$V(\chi^2/n) = 2;\ \sigma(\chi^2/n) = \sqrt{2/n}$

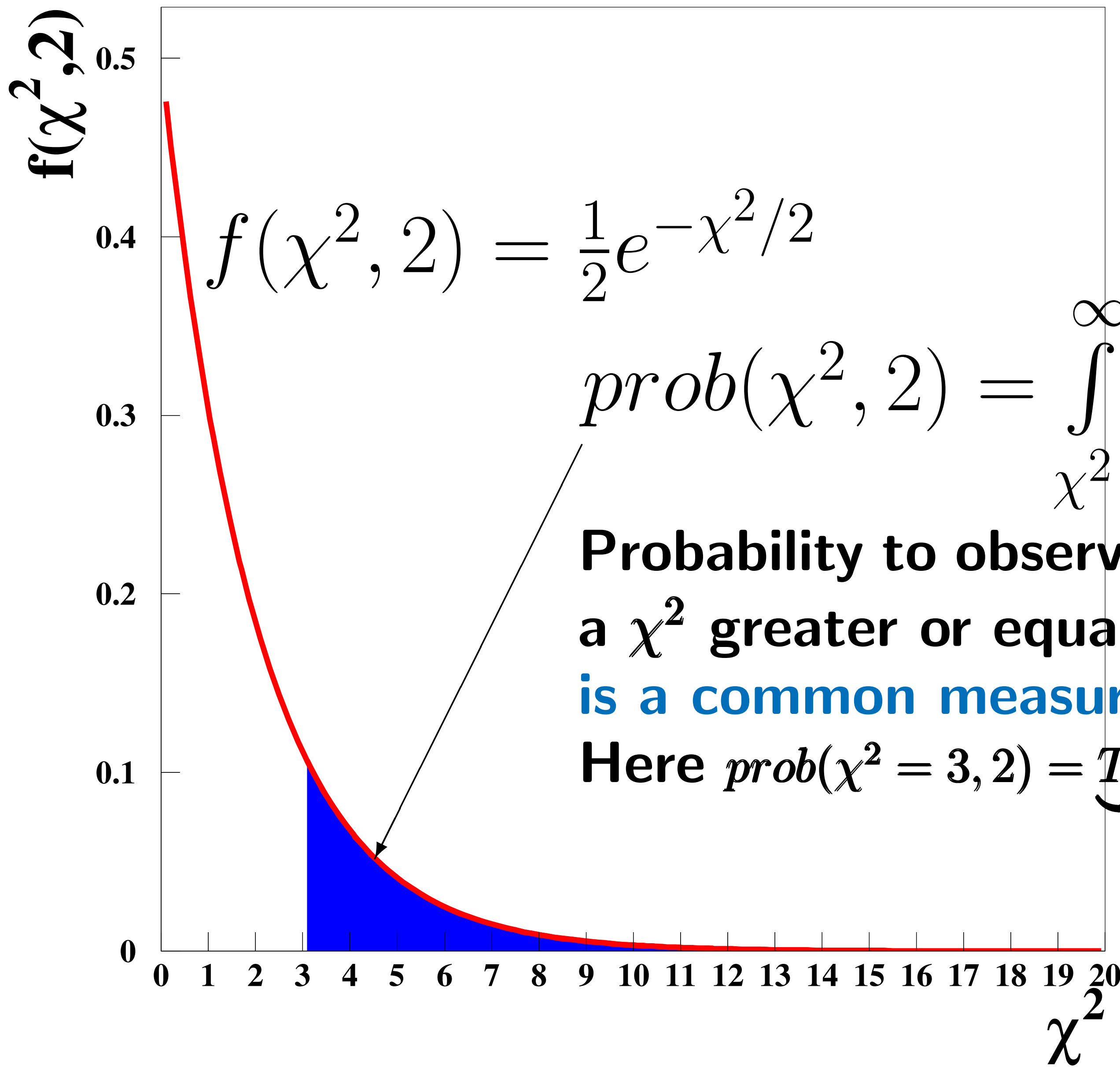
χ^2 distributions for various n



χ^2 distr.

χ^2/n distr.

$f(\chi^2, 2)$ function and $prob(\chi^2, 2)$



$$f(\chi^2, 2) = \frac{1}{2} e^{-\chi^2/2}$$

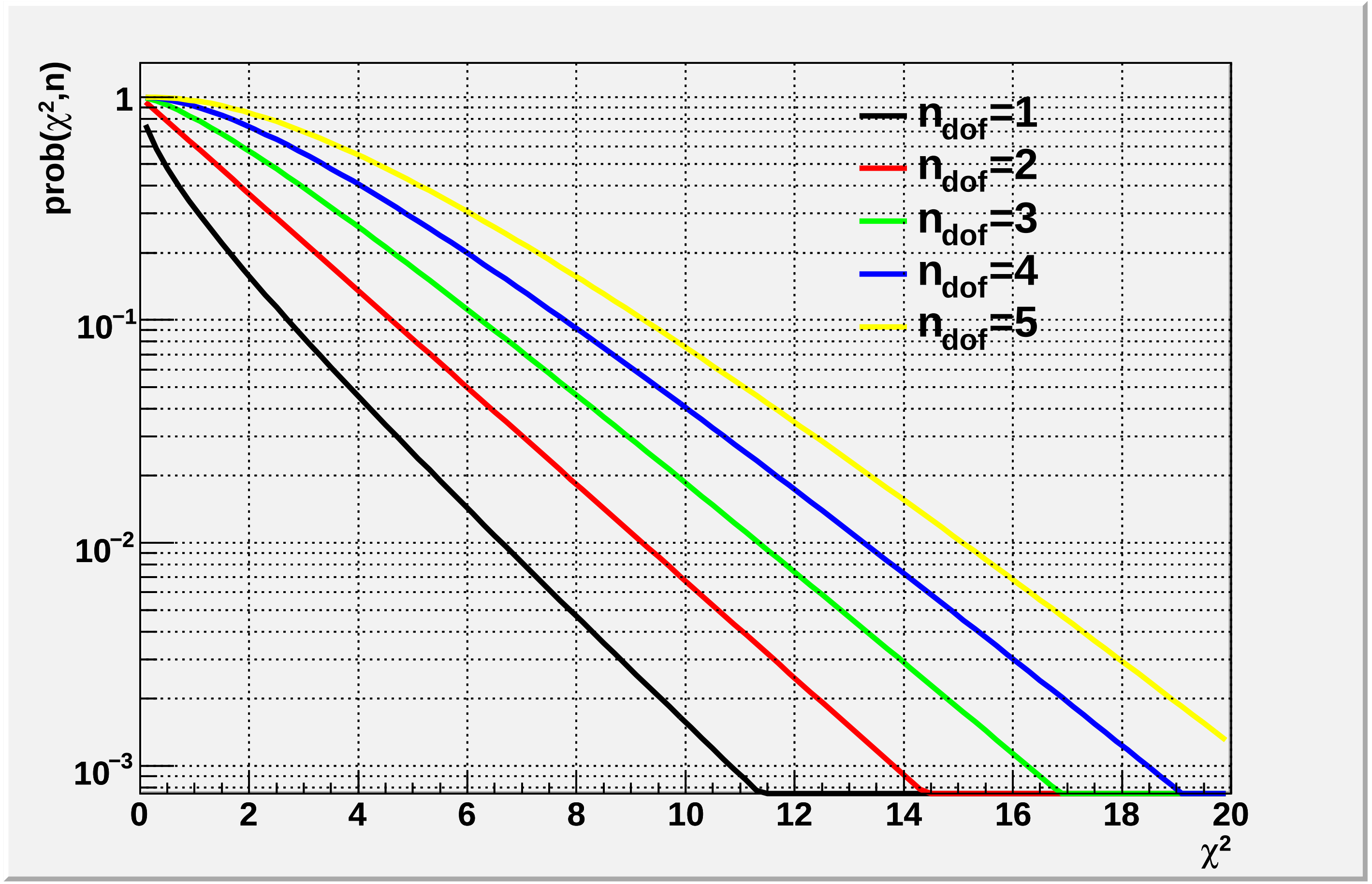
$$prob(\chi^2, 2) = \int_{\chi^2}^{\infty} f(\chi'^2, 2) d\chi'^2$$

Probability to observe for repeated experiments a χ^2 greater or equal than the current one is a common measure for consistency

Here $prob(\chi^2 = 3, 2) = \underbrace{TMath :: Prob(3, 2)}_{\text{ROOT Function}} = 22\%$

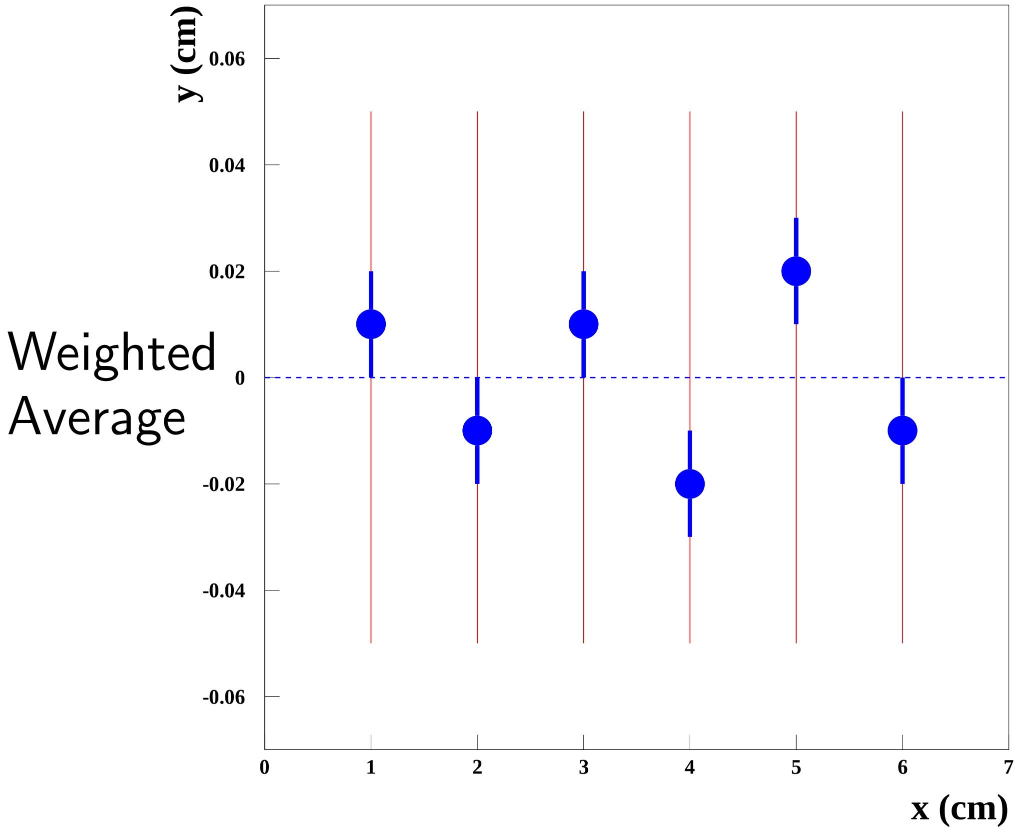
$prob(\chi^2, n)$-function for n degrees of freedom

$$prob(\chi^2, n) = \int_{\chi^2}^{\infty} f(\chi'^2, n) d\chi'^2 = \frac{1}{\Gamma(n/2)} \cdot \int_{\chi^2/2}^{\infty} dt\, e^{-t}\, t^{n/2-1}$$



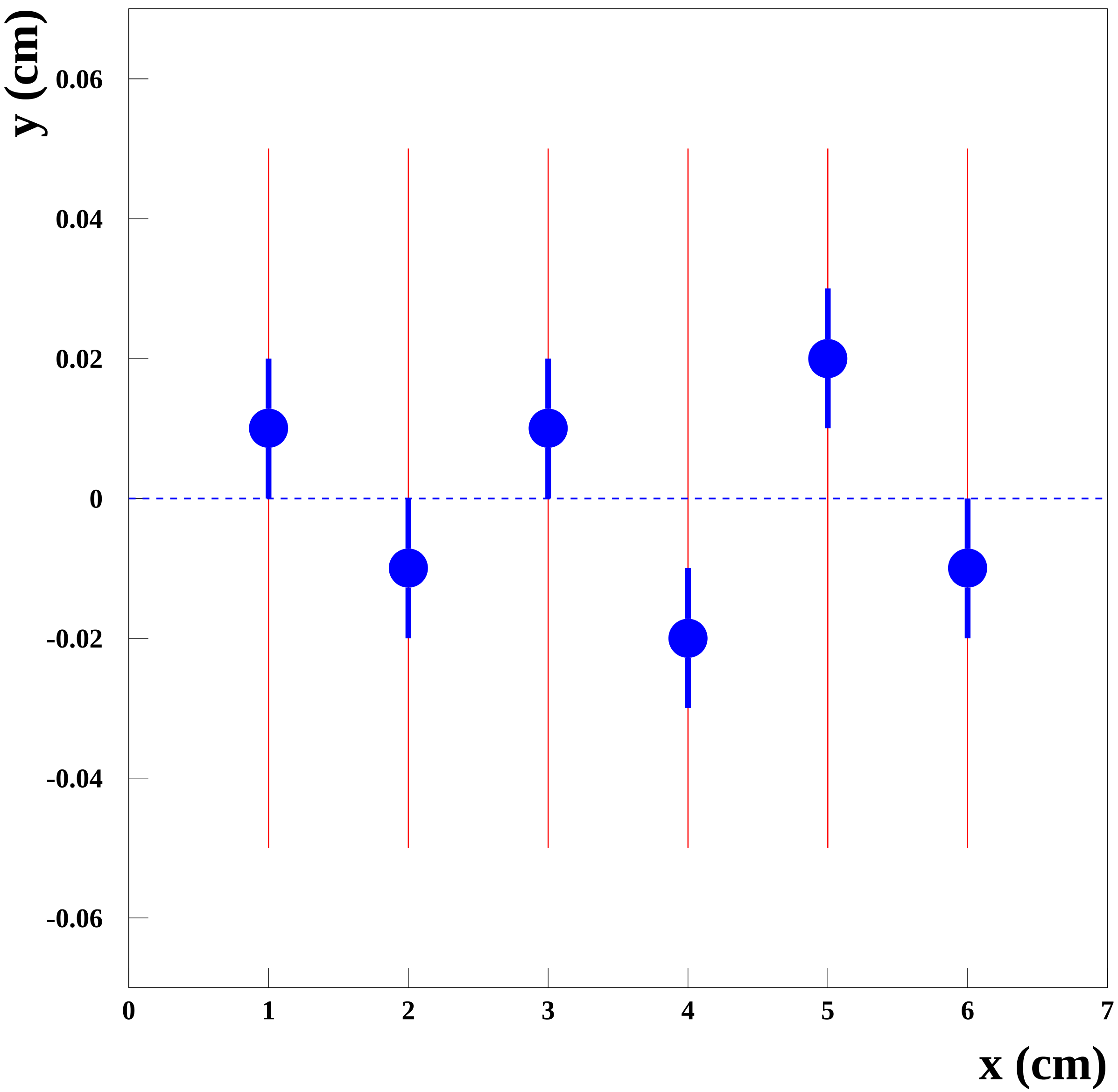
Note: for repeated experiments expect the observed values of $prob(\chi^2, n)$ to be evenly distributed over interval $[0, 1]$

χ^2 for averaging measurements



The figure shows the result of a fit of a constant using n measurements. When repeating the fit many times the resulting χ^2_{min} distribution should follow a χ^2 distribution with $n-1$ degrees of freedom. One degree of freedom is sacrificed to determine the weighted average. A prove for this (for $n=2$) is given in the appendix.

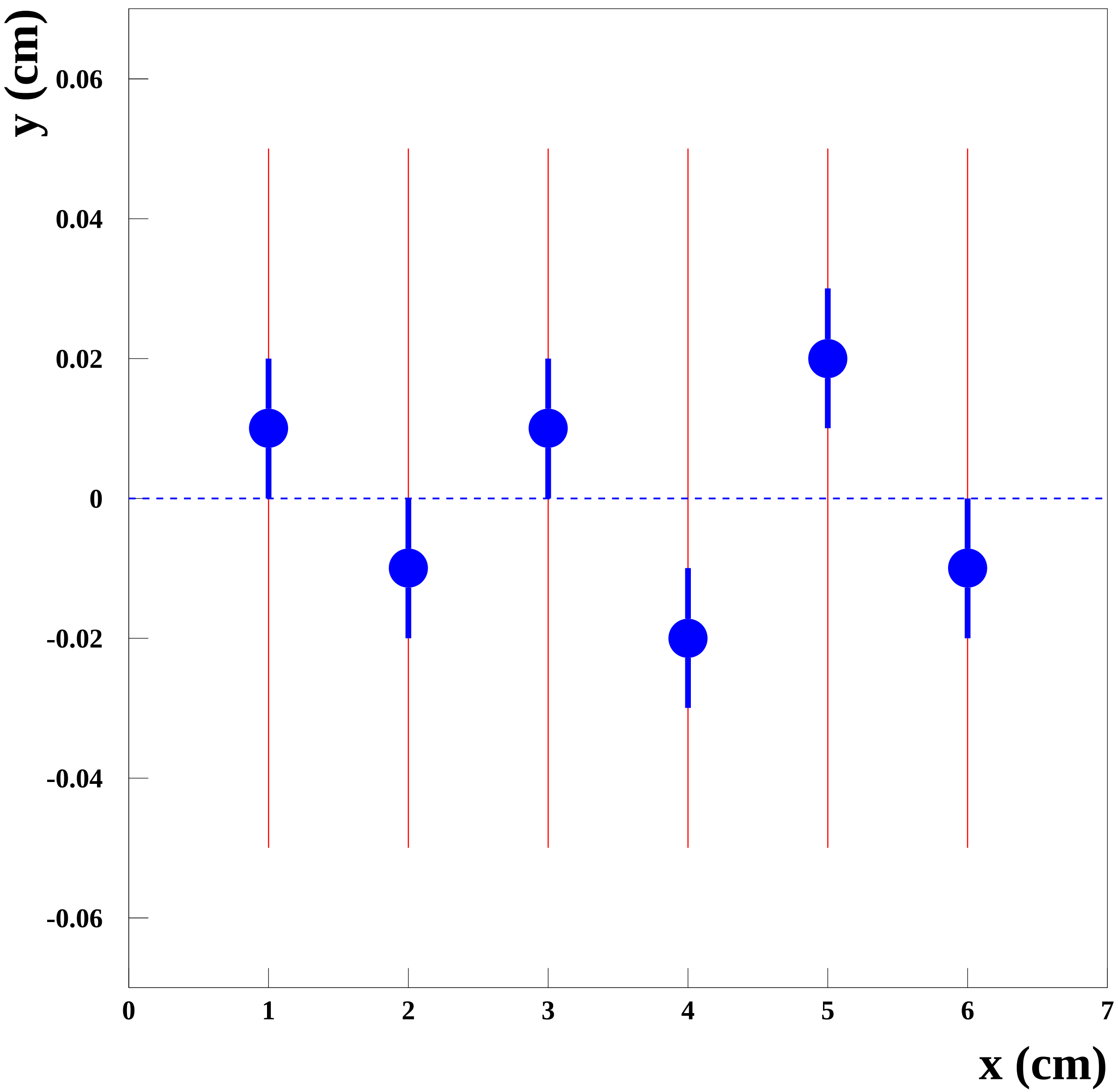
χ^2 and probability



The figure shows the result of a fit of a constant. Determine

- the total χ^2 from reading the figure
- the χ^2-probability with the ROOT command: **TMath::Prob(χ^2,ndf)**, where ndf is the numbers of degree of freedom.

Mini-exercise χ^2 and probability

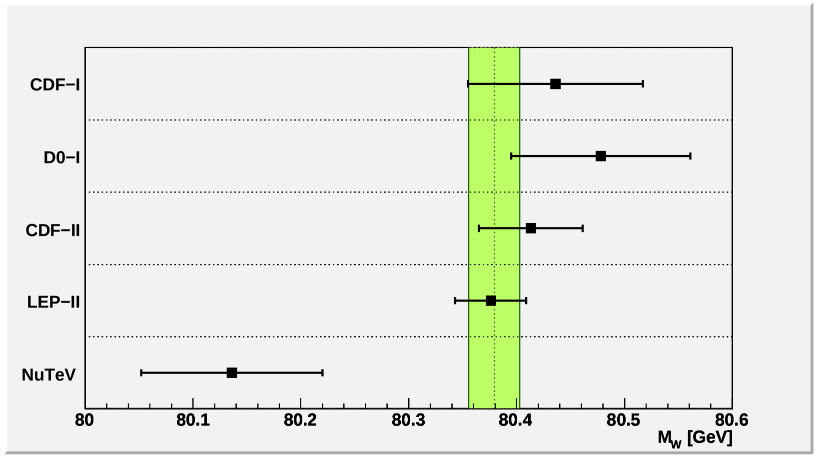


The figure shows the result of a fit of a constant. Determine

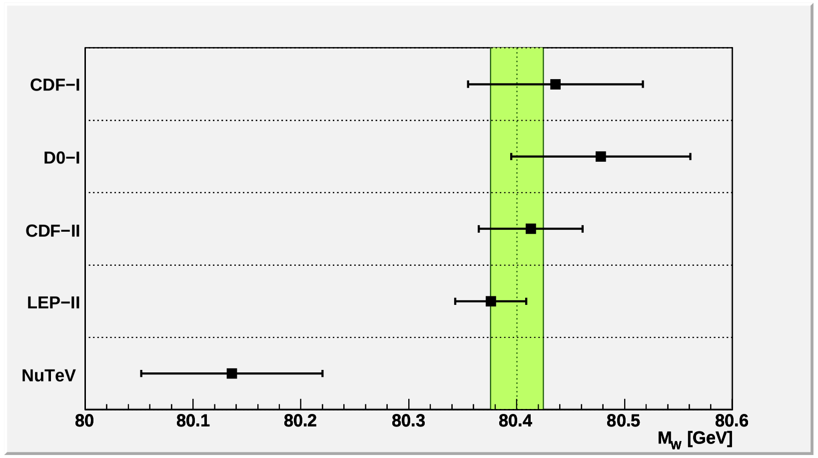
- the total χ^2 from reading the figure: $= 12$
- the χ^2-probability with the ROOT command: **TMath::Prob(12,5)** $=$ 3.8%.

World average of W boson mass

or how to arrive at a good χ^2

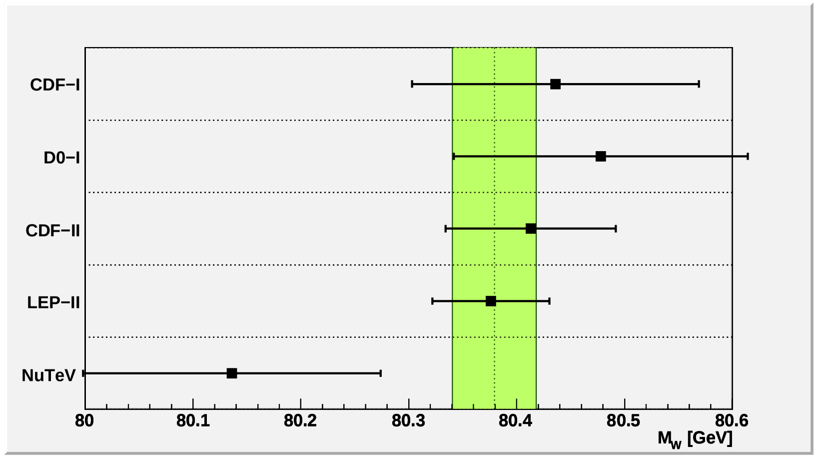


$\chi^2_{min} = 10.8$, $n_{dof} = 4$, probability = 0.029



Taking out NuTeV result:
$\chi^2_{min} = 1.7$, $n_{dof} = 3$, probability = 0.64

"Outlier rejection", is this allowed?



Scaling all errors by $S = \sqrt{\chi^2_{min}/n_{dof}} = 1.64$
$\chi^2_{min} = 4.$, $n_{dof} = 4$, probability = 0.4

Standard procedure by Particle Data group
$\rightarrow$ "destroying" the hard work of many experimentalists

Computer exercise: Average 10 measurements with noise:

Computer exercise: Average 10 measurements with noise: with Root Macro p0toyf.C

Note: Macro available at
www.desy.de/~obehnke/stat/school_apr14/p0toyf.C

Task instructions available at
www.desy.de/~obehnke/stat/school_apr14/compueb_p0toyf.pdf

Or you copy from afs:

- Code:
  cp ***/afs/desy.de/user/o/obehnke/www/stat/school_apr14/p0toyf.C*** **.**
- Instructions:
  cp ***/afs/desy.de/user/o/obehnke/www/stat/school_apr14/compueb_p0toyf.pdf*** **.**

Dont' forget to initialise module load root before starting root!

Fits with problems: outliers

Toy simulations of constant fits through 10 data points

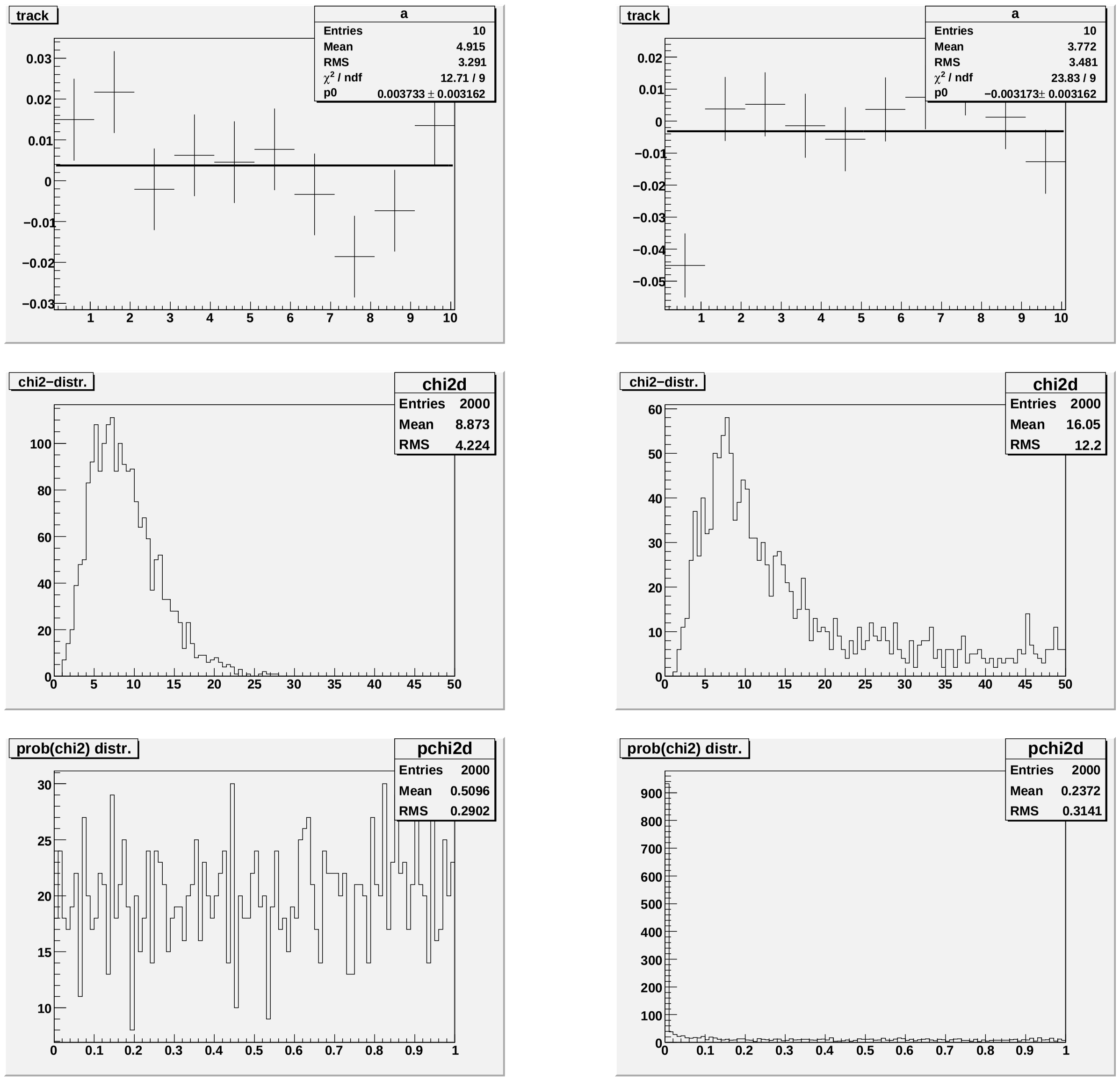
No outliers

10% random outliers (10σ)

Exemplary fit

χ^2_{min} distribution for 2000 experiments

$prob(\chi^2_{min}, 9)$ distribution for 2000 experiments



$\Rightarrow \chi^2_{min}$ and $prob(\chi^2_{min}, n_{dof})$ highly sensitive to wrong measurements

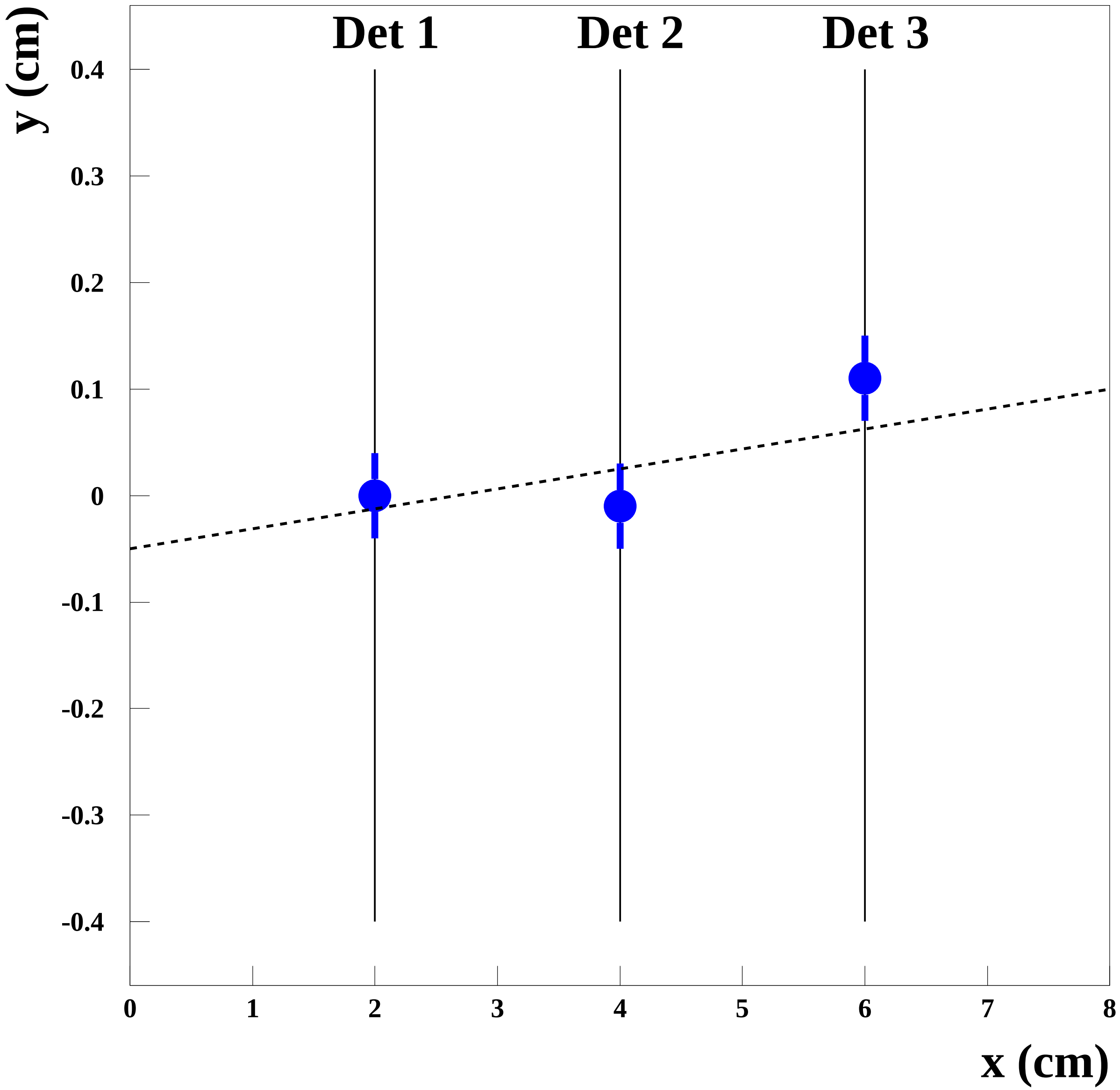
Mini summary of what we have learnt

- **The χ^2_{min} of a fit is a consistency check**
- **Expect $\chi^2_{min}/n_{dof} \sim 1$ for good fits**
- **if χ^2_{min}/n_{dof} significantly larger than one then suspect**
  - **data could contain outliers or errors are (generally) underestimated**
  - **the fitfunction might not be the correct model for the data**
- **for repeated experiments (e.g. many track fits) expect for good fits**
  - **mean value of χ^2_{min}/n_{dof} distribution $\rightarrow 1$**
  - **and flat $prob(\chi^2_{min}, n_{dof})$ distribution in interval [0,1]**

3. Linear least square fits

Think linear!

Example: Straight line track trajectory fit



$$y_i = a_0 + a_1 x_i$$

This is a classical linear least square fit problem.

Linear least square fits

$\vec{y}$ vector of n measurements $\begin{pmatrix} y_1(x_1) \\ . \\ y_n(x_n) \end{pmatrix}$ with cov-matrix V

Linear model $\vec{y} := A\,\vec{a}$, $\vec{a}$ vector of m fitparameters $\begin{pmatrix} a_1 \\ . \\ a_m \end{pmatrix}$

Example: $y = a_0$;

Linear least square fits

$\vec{y}$ vector of n measurements $\begin{pmatrix} y_1(x_1) \\ . \\ y_n(x_n) \end{pmatrix}$ with cov-matrix V

Linear model $\vec{y} := A\,\vec{a}$, $\quad$ $\vec{a}$ vector of m fitparameters $\begin{pmatrix} a_1 \\ . \\ a_m \end{pmatrix}$

Example: $y = a_0; \quad \rightarrow \vec{a} = (a_0); \quad A = \begin{pmatrix} 1 \\ .. \\ 1 \end{pmatrix}$

In general: $A = A(\vec{x})$, but no dependence on $\vec{a}$

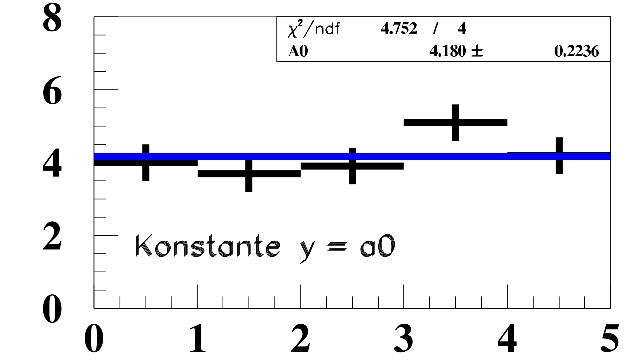
"Master formula": $$\chi^2 = (\vec{y} - A\,\vec{a})^t\, V^{-1}\, (\vec{y} - A\,\vec{a})$$

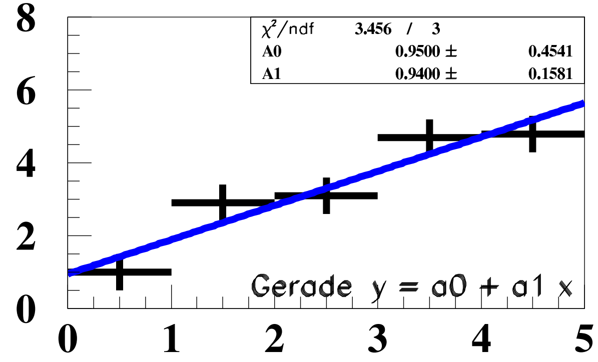
$\rightarrow$ to be minimised w.r.t $\vec{a}$

$\rightarrow$ obtain estimators $\hat{\vec{a}}$ and covariance matrix $V_{\hat{\vec{a}}}$

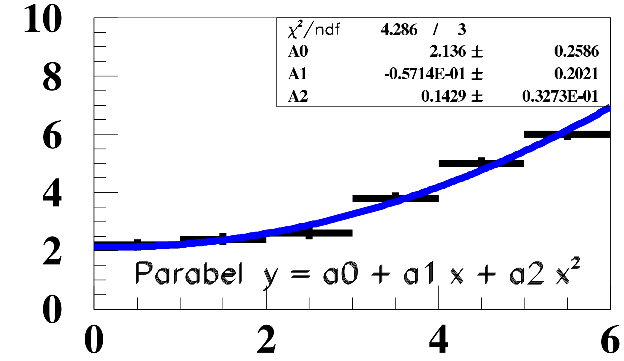
Examples for linear least square fits

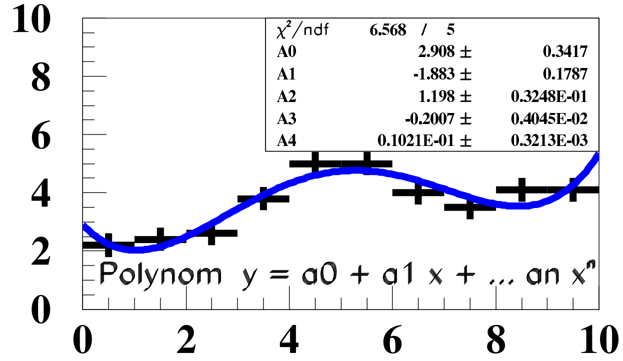
Linear means that y depends linearly on the fitparameters a_i.

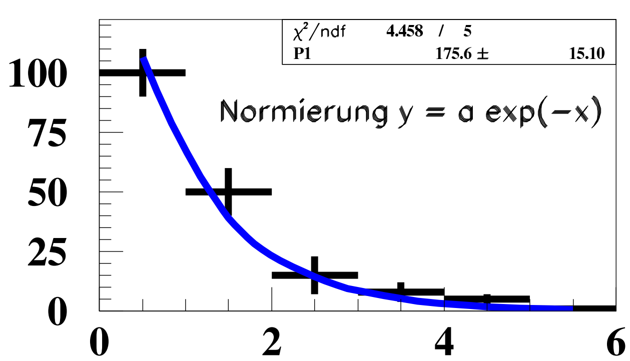




$$\vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}; \; A = \begin{pmatrix} 1 & x_1 \\ . & . \\ 1 & x_n \end{pmatrix}$$







$$\vec{a} = (a); \; A = \begin{pmatrix} e^{-x_1} \\ .. \\ e^{-x_n} \end{pmatrix}$$

← **Watch out: function can be highly non-linear in x**

General solution via normal equations

$$\chi^2 = (\vec{y} - A\vec{a})^t V^{-1} (\vec{y} - A\vec{a})$$
$$= \vec{y}^t V^{-1} \vec{y} - 2\vec{a}^t A V^{-1} \vec{y} + \vec{a}^t A^t V^{-1} A \vec{a}$$

Min. $\chi^2 \rightarrow \frac{d\chi^2}{d\vec{a}^t} = -2A^t V^{-1} \vec{y} + 2A^t V^{-1} A \vec{a} = 0$

General solution via normal equations

$$\chi^2 = (\vec{y} - A\vec{a})^t V^{-1} (\vec{y} - A\vec{a})$$
$$= \vec{y}^t V^{-1} \vec{y} - 2\vec{a}^t A V^{-1} \vec{y} + \vec{a}^t A^t V^{-1} A \vec{a}$$

Min. $\chi^2 \rightarrow \frac{d\chi^2}{d\vec{a}^t} = -2A^t V^{-1} \vec{y} + 2A^t V^{-1} A \vec{a} = 0$

$\Rightarrow$ Normal equation Solution:

$$\hat{\vec{a}} = (A^t V^{-1} A)^{-1} A^t V^{-1} \vec{y}$$
$$= H^{-1} A^t V^{-1} \vec{y}$$
$$= U A^t V^{-1} \vec{y} \text{ with } U = H^{-1} = Cov(\hat{\vec{a}})$$

Powerful & simple linear algebra

Note: $H = \frac{1}{2}\frac{d^2\chi^2}{d\vec{a}^2}$ is 'an old friend', the Hesse Matrix

Linear least square fits: Covariance Matrix

Proof that Covariance matrix U of fit parameters $\hat{\vec{a}}$ is given by $U = H^{-1}$

Use Normal Equations:

$$\hat{\vec{a}} = B\vec{y} \quad \text{with} \quad B = H^{-1}A^tV^{-1}$$

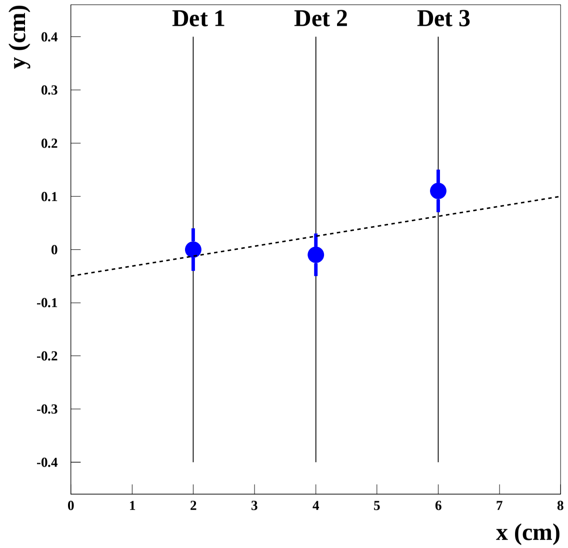
Then apply errorpropagation:

$$\begin{aligned} \rightarrow U &= BVB^t = H^{-1}A^tV^{-1}VV^{-1}AH^{-1} \\ &= H^{-1}A^tV^{-1}AH^{-1} = H^{-1}HH^{-1} = H^{-1} \end{aligned}$$

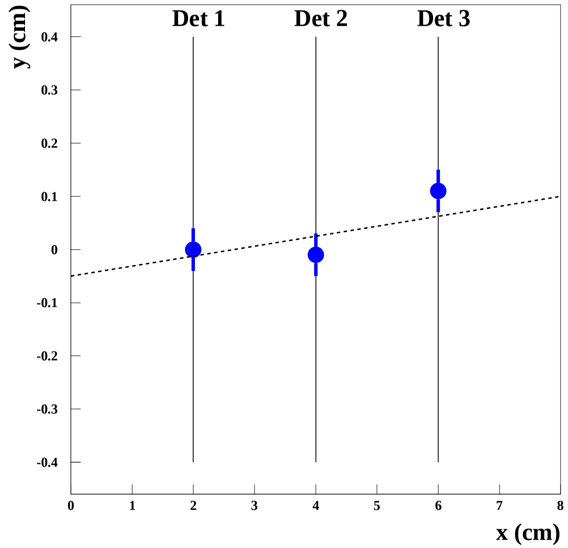
Straight line fit through n detector layers

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

$$\vec{y} = A\vec{a};\;\; \vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix};\; A = \begin{pmatrix} 1 & x_1 \\ . & . \\ 1 & x_n \end{pmatrix};\; A^t = \begin{pmatrix} 1 & .. & 1 \\ x_1 & .. & x_n \end{pmatrix};\; V = \begin{pmatrix} \sigma^2 & & 0 \\ & . & \\ 0 & & \sigma^2 \end{pmatrix}$$



Straight line fit through n detector layers



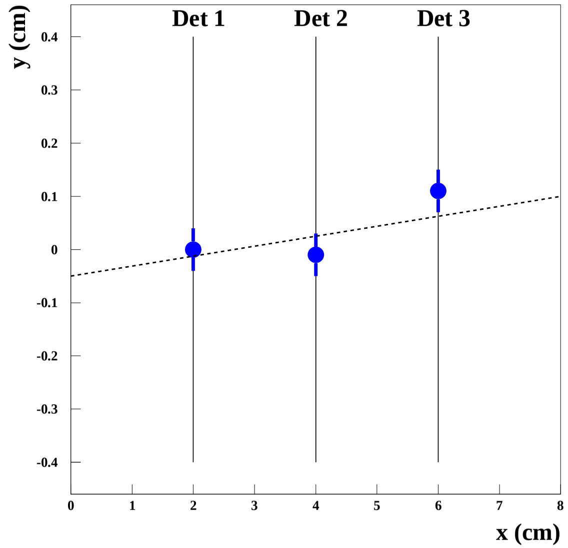
$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1\, x_i)^2}{\sigma^2}$$

$$\vec{y} = A\vec{a};\ \ \vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix};\ A = \begin{pmatrix} 1 & x_1 \\ . & . \\ 1 & x_n \end{pmatrix};\ A^t = \begin{pmatrix} 1 & .. & 1 \\ x_1 & .. & x_n \end{pmatrix};\ V = \begin{pmatrix} \sigma^2 & & 0 \\ & . & \\ 0 & & \sigma^2 \end{pmatrix}$$

Apply normal equations:

$$\hat{\vec{a}} \ = \ (A^t V^{-1} A)^{-1} A^t V^{-1} \vec{y} = \sigma^2 (A^t A)^{-1} \cdot \frac{1}{\sigma^2} A^t \cdot \vec{y} \ = \ (A^t A)^{-1} A^t \cdot \vec{y}$$

Straight line fit through n detector layers



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

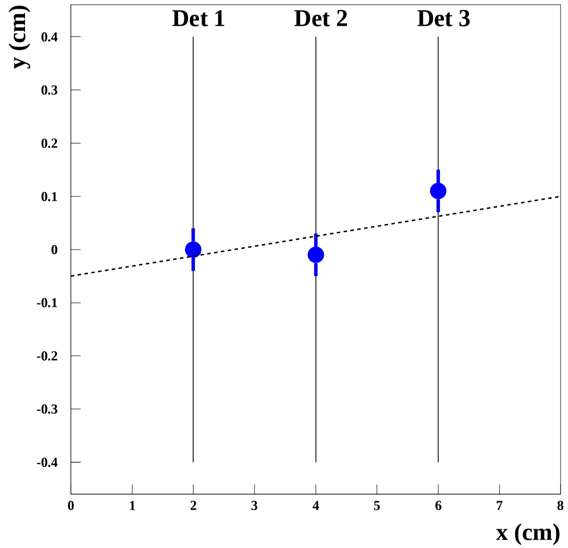
$$\vec{y} = A\vec{a};\ \ \vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix};\ A = \begin{pmatrix} 1 & x_1 \\ . & . \\ 1 & x_n \end{pmatrix};\ A^t = \begin{pmatrix} 1 & .. & 1 \\ x_1 & .. & x_n \end{pmatrix};\ V = \begin{pmatrix} \sigma^2 & 0 \\ & . \\ 0 & \sigma^2 \end{pmatrix}$$

Apply normal equations:

$$\hat{\vec{a}} \ = \ (A^tV^{-1}A)^{-1}A^tV^{-1}\vec{y} = \sigma^2(A^tA)^{-1}\cdot\frac{1}{\sigma^2}A^t\cdot\vec{y} \ = \ (A^tA)^{-1}A^t\cdot\vec{y}$$

$$= \begin{pmatrix} \sum_i 1 & \sum_i x_i \\ . & . \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} = \begin{pmatrix} N & N\overline{x} \\ N\overline{x} & N\overline{x^2} \end{pmatrix}^{-1} \cdot \begin{pmatrix} N\overline{y} \\ N\overline{xy} \end{pmatrix}$$

Straight line fit through n detector layers



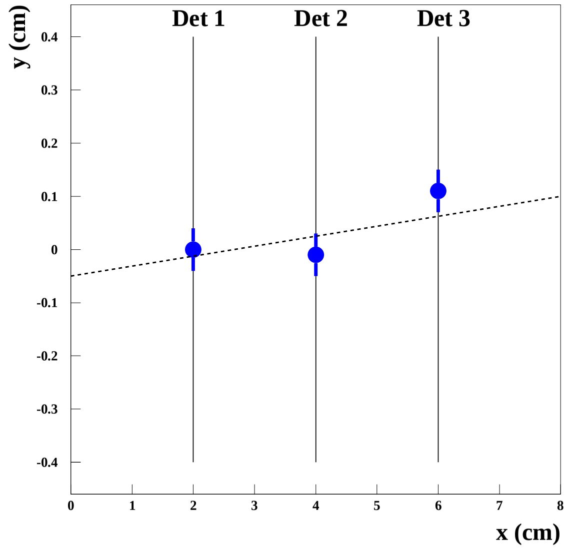
$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1\, x_i)^2}{\sigma^2}$$

$$\vec{y} = A\vec{a};\ \ \vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix};\ A = \begin{pmatrix} 1 & x_1 \\ . & . \\ 1 & x_n \end{pmatrix};\ A^t = \begin{pmatrix} 1 & .. & 1 \\ x_1 & .. & x_n \end{pmatrix};\ V = \begin{pmatrix} \sigma^2 & & 0 \\ & . & \\ 0 & & \sigma^2 \end{pmatrix}$$

Apply normal equations:

$$\begin{aligned}
\hat{\vec{a}} &= (A^tV^{-1}A)^{-1}A^tV^{-1}\vec{y} = \sigma^2(A^tA)^{-1}\cdot\frac{1}{\sigma^2}A^t\cdot\vec{y} \ = \ (A^tA)^{-1}A^t\cdot\vec{y} \\
&= \begin{pmatrix} \sum_i 1 & \sum_i x_i \\ . & . \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} = \begin{pmatrix} N & N\overline{x} \\ N\overline{x} & N\overline{x^2} \end{pmatrix}^{-1} \cdot \begin{pmatrix} N\overline{y} \\ N\overline{xy} \end{pmatrix} \\
&= \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix} = \frac{1}{\overline{x^2} - \overline{x}^2} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix} = \frac{1}{V[x]} \cdot \begin{pmatrix} \overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy} \\ -\overline{x}\,\overline{y} + \overline{xy} \end{pmatrix}
\end{aligned}$$

Straight line fit through n detector layers



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

$$\vec{y} = A\vec{a};\ \ \vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix};\ A = \begin{pmatrix} 1 & x_1 \\ . & . \\ 1 & x_n \end{pmatrix};\ A^t = \begin{pmatrix} 1 & .. & 1 \\ x_1 & .. & x_n \end{pmatrix};\ V = \begin{pmatrix} \sigma^2 & & 0 \\ & . & \\ 0 & & \sigma^2 \end{pmatrix}$$

Apply normal equations:

$$\hat{\vec{a}} = (A^tV^{-1}A)^{-1}A^tV^{-1}\vec{y} = \sigma^2(A^tA)^{-1}\cdot\frac{1}{\sigma^2}A^t\cdot\vec{y} = (A^tA)^{-1}A^t\cdot\vec{y}$$

$$= \begin{pmatrix} \sum_i 1 & \sum_i x_i \\ . & . \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} = \begin{pmatrix} N & N\overline{x} \\ N\overline{x} & N\overline{x^2} \end{pmatrix}^{-1} \cdot \begin{pmatrix} N\overline{y} \\ N\overline{xy} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix} = \frac{1}{\overline{x^2} - \overline{x}^2} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix} = \frac{1}{V[x]} \cdot \begin{pmatrix} \overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy} \\ -\overline{x}\,\overline{y} + \overline{xy} \end{pmatrix}$$

$$U = \begin{pmatrix} \sigma^2_{\hat{a}_0} & cov(\hat{a}_0, \hat{a}_1) \\ cov(\hat{a}_0, \hat{a}_1) & \sigma^2_{\hat{a}_1} \end{pmatrix} = (A^tV^{-1}A)^{-1} = \frac{\sigma^2}{NV[x]} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}$$

Mini-exercise: Straight line track-fit

The covariance formula

$$\begin{pmatrix} \sigma^2_{\hat{a}_0} & cov(\hat{a}_0, \hat{a}_1) \\ cov(\hat{a}_0, \hat{a}_1) & \sigma^2_{\hat{a}_1} \end{pmatrix} = \frac{\sigma^2}{NV[x]} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}$$

is valid for e.g. a straight line track fit in N detectors of resolution σ:

Determine the improvements on the slope error $\sigma_{\hat{a}_1}$ by:

a) Doubling the number of detector layers N within the same interval in x

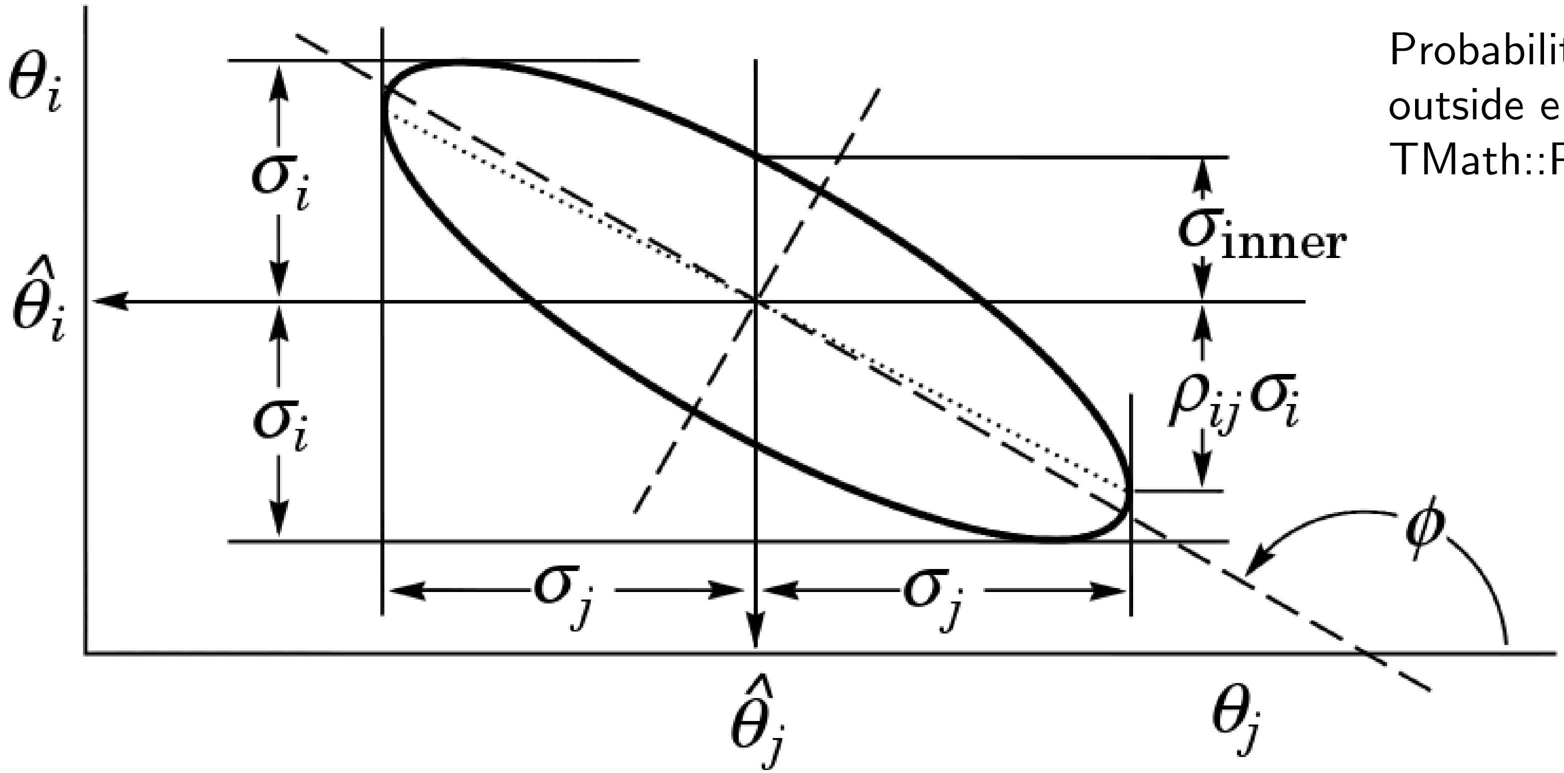
b) Distributing the detector layers over an interval in x twice as large

c) Buying detectors with measurement uncertainties σ reduced by a factor two

Covariance ellipses

The covariance matrix $V = \begin{pmatrix} \sigma_i^2 & \rho_{ij}\sigma_i\sigma_j \\ \rho_{ij}\sigma_i\sigma_j & \sigma_j^2 \end{pmatrix}$ of two parameters θ_i and θ_j can be represented by error ellipses (see Fig. from PDG below) The role of the correlation coefficient ρ_{ij}:

- If one shifts θ_i to $\hat{\theta}_i + \sigma_i$ one has to shift θ_j to $\hat{\theta}_j + \rho_{ij}\sigma_j$ to keep the χ^2 increase minimal (stay down in the χ^2 valley)
- When fixing θ_j the error on θ_i is reduced to $\sigma_{inner} = \sqrt{1-\rho^2}\,\sigma_i$



Probability for values outside ellipse: TMath::Prob(1,2)∼ 40%

Computer exercise: Straight line trajectory fit

with Root Macro StraightLineFit.C

Computer exercise: Straight line trajectory fit with Root Macro StraightLineFit.C

Note: Macro available at
www.desy.de/~obehnke/stat/school_apr14/StraightLineFit.C

Task instructions available at
www.desy.de/~obehnke/stat/school_apr14/compueb_StraightLineFit.pdf

Or you copy from afs:

- Code:
  cp */afs/desy.de/user/o/obehnke/www/stat/school_apr14/StraightLineFit.C* **.**
- Instructions:
  cp
  /afs/desy.de/user/o/obehnke/www/stat/school_apr14/compueb_StraightLineFit.pdf **.**

Dont' forget to initialise module load root before starting root!

Mini summary of what we have learnt

- **Linear least square problems:** $\vec{y} = A\vec{a}$,
  $\rightarrow y$ **is a linear function of the fitparameters** $\vec{a}$ **but can be a linear or nonlinear function of the continuous parameter** x.

- **The normal equations are a powerful tool to solve linear least square fit problems**
  $\hat{\vec{a}} = (A^tV^{-1}A)^{-1}A^tV^{-1}\vec{y}, \;\; cov(\hat{\vec{a}}) = (A^tV^{-1}A)^{-1}$

- **Straight line fits are a typical application and there are many others (e.g. parabolas, higher order polynoms, etc.)**

4. Nonlinear least square fits

Think linear but act nonlinear!

Reminder least square method (one parameter)

$\rightarrow$ $\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i,a))^2}{\sigma_i^2}}$ $\leftrightarrow$ Minimum w.r.t a

$\Rightarrow$ determine estimator $\hat{a}$ from $\frac{d\chi^2}{da} = 0$

Nonlinear case

Examples: $f(x,a) = tan(ax), \quad ln(ax), \quad a\, exp(-ax)$

Define $g = \frac{d\chi^2}{da}$ and $G = \frac{d^2\chi^2}{da^2} = g'$

$\hookrightarrow$ Newton steps to find root of g:

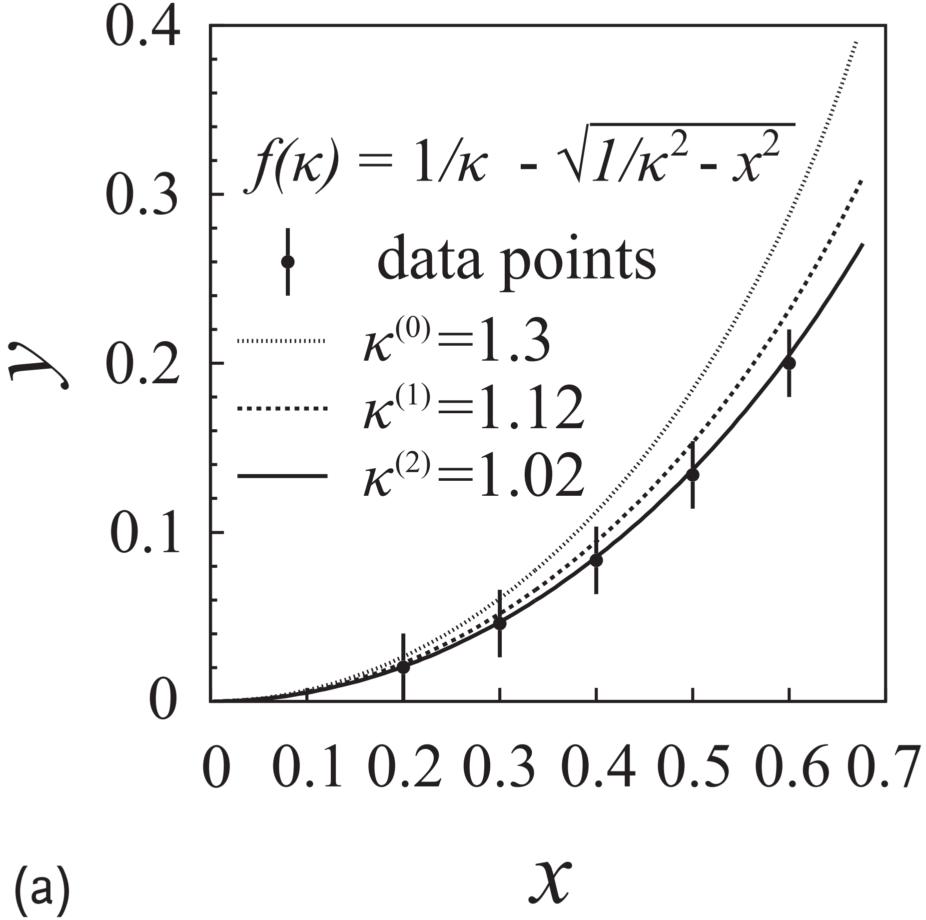
$a_{m+1} = a_m - \frac{g(a_m)}{g'(a_m)}$ (iteration index m)

For several parameters $\vec{a}$ generalise: $\vec{g} = \frac{d\chi^2}{d\vec{a}^t}$ and $G = \frac{d^2\chi^2}{d\vec{a}^2}$

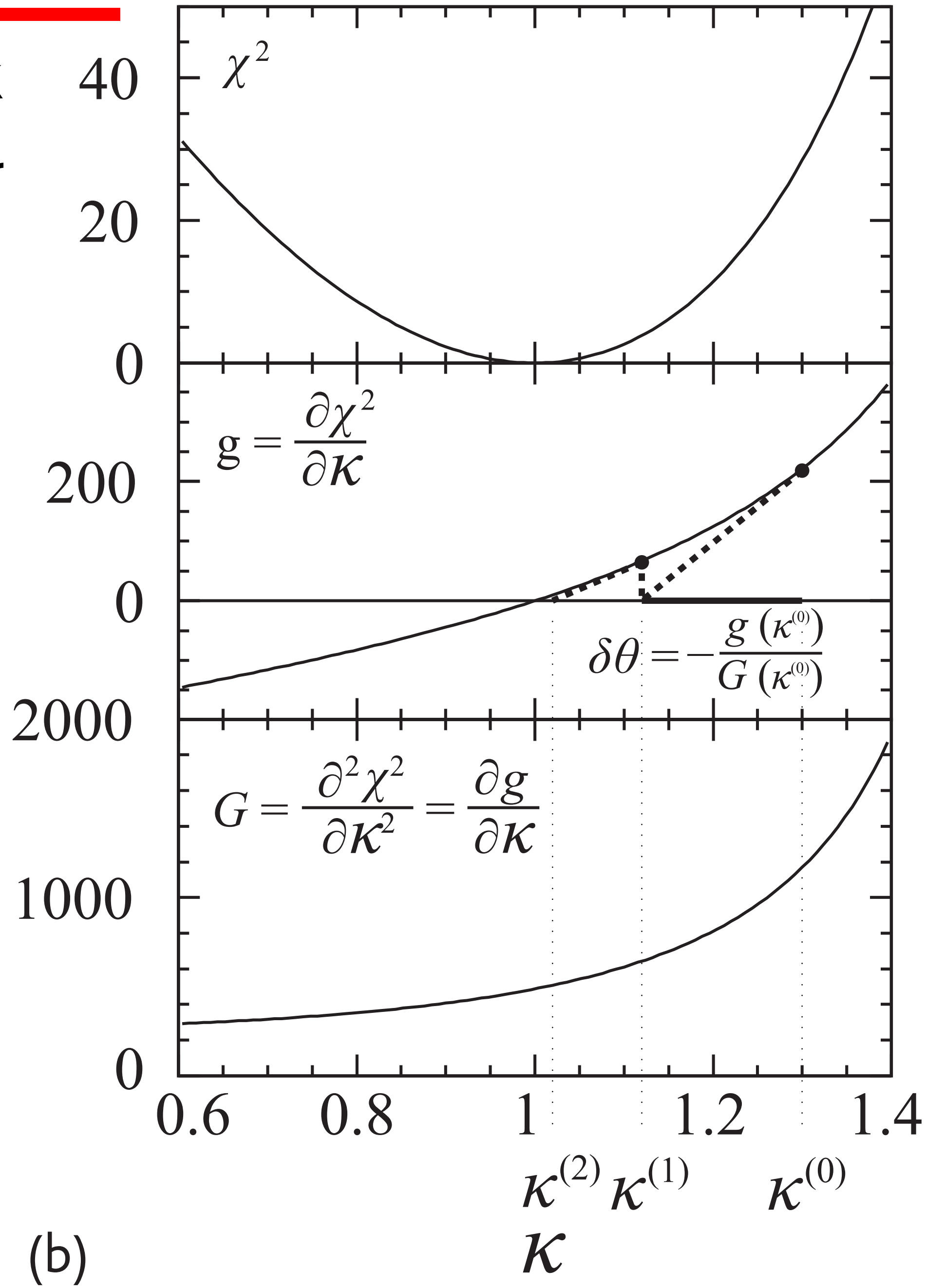
$\hookrightarrow$ Newton step: $\delta\vec{a} = -G^{-1}\vec{g}$

Illustration of Newton steps

Fit the curvature κ of a track flying through perpendicular magnetic field



(a)



(b)

Mini-exercise Nonlinear χ^2 Fit: determine particle mass from signal peak

Paper exercise

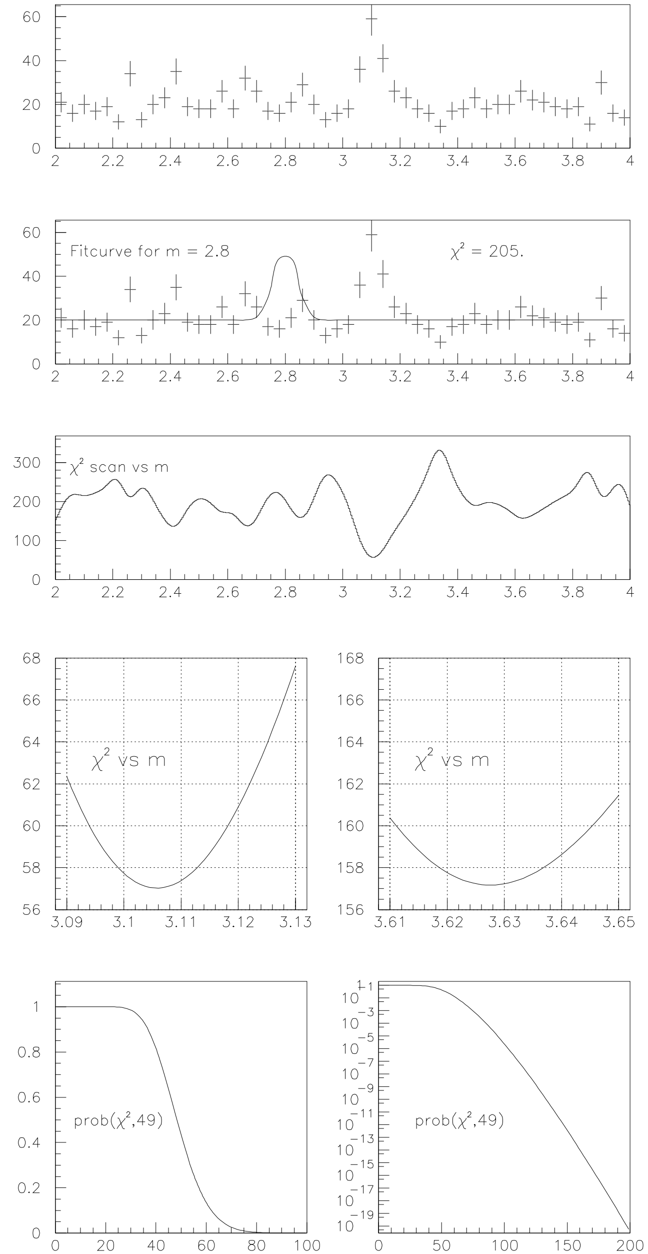
Mini-exercise Nonlinear χ^2 Fit: determine particle mass from signal peak

Fit function: $f(M) = U + S \cdot e^{-\dfrac{(M-m)^2}{2\sigma^2}}$ with

- U known constant background (noise)
- S: predicted Signalstrength
- σ: known detector resolution
- m: unknown mass of particle

$$\chi^2 = \sum_{bin_i} \frac{(k_i - f(M_i, m))^2}{k_i}$$

with k_i = #observed events in bin i



Tasks:

- Find in the $\chi^2(m)$-curve (third plot) the global χ^2-Minimum and the second deepest minimum. Read the χ^2-value in the respective minima and determine the 'Fit-*Probability*' (this you can read off in the lowest two plots).
- Determine for the two χ^2-Minima (4. and 5. plot) at which masses
  - the χ^2 is minimal
  - the χ^2 is $\chi^2_{min} + 1$, thus estimating the error for the mass

---

Please take note: We will see later in the maximum likelihood lecture that the χ^2 estimator used here (called 'Neyman-χ^2') is not recommended for such a binned mass peak fit, it is better to use a 'Poisson likelihood'.

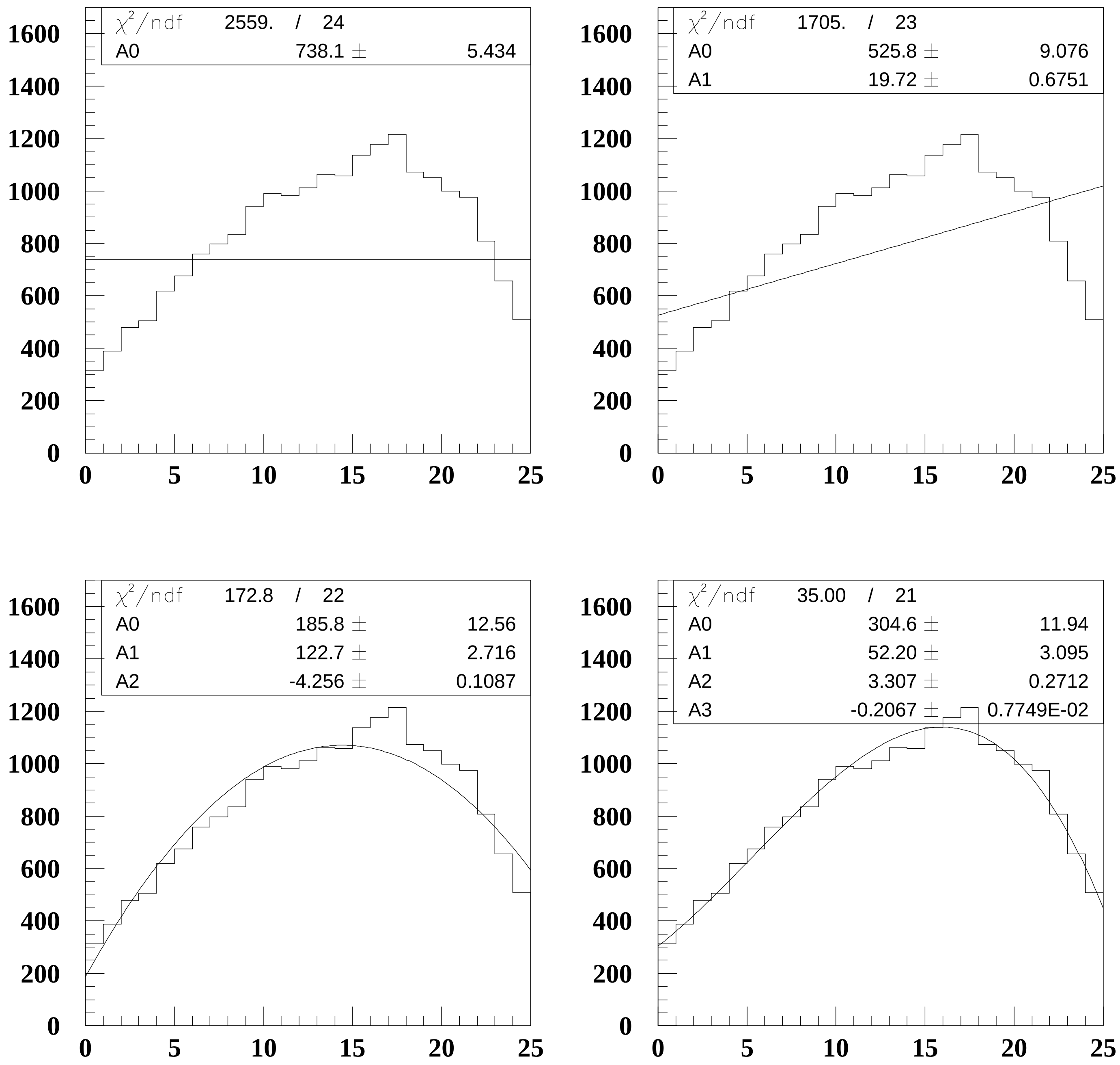
Mini-exercise: Unknown parameterisation - Find best one

The plots on the next two pages show χ^2-fits of the same data distribution with eight different parametrisations (polynomials of different orders). Which parametrisation is the most reasonable one? Try to judge using the following three criteria:

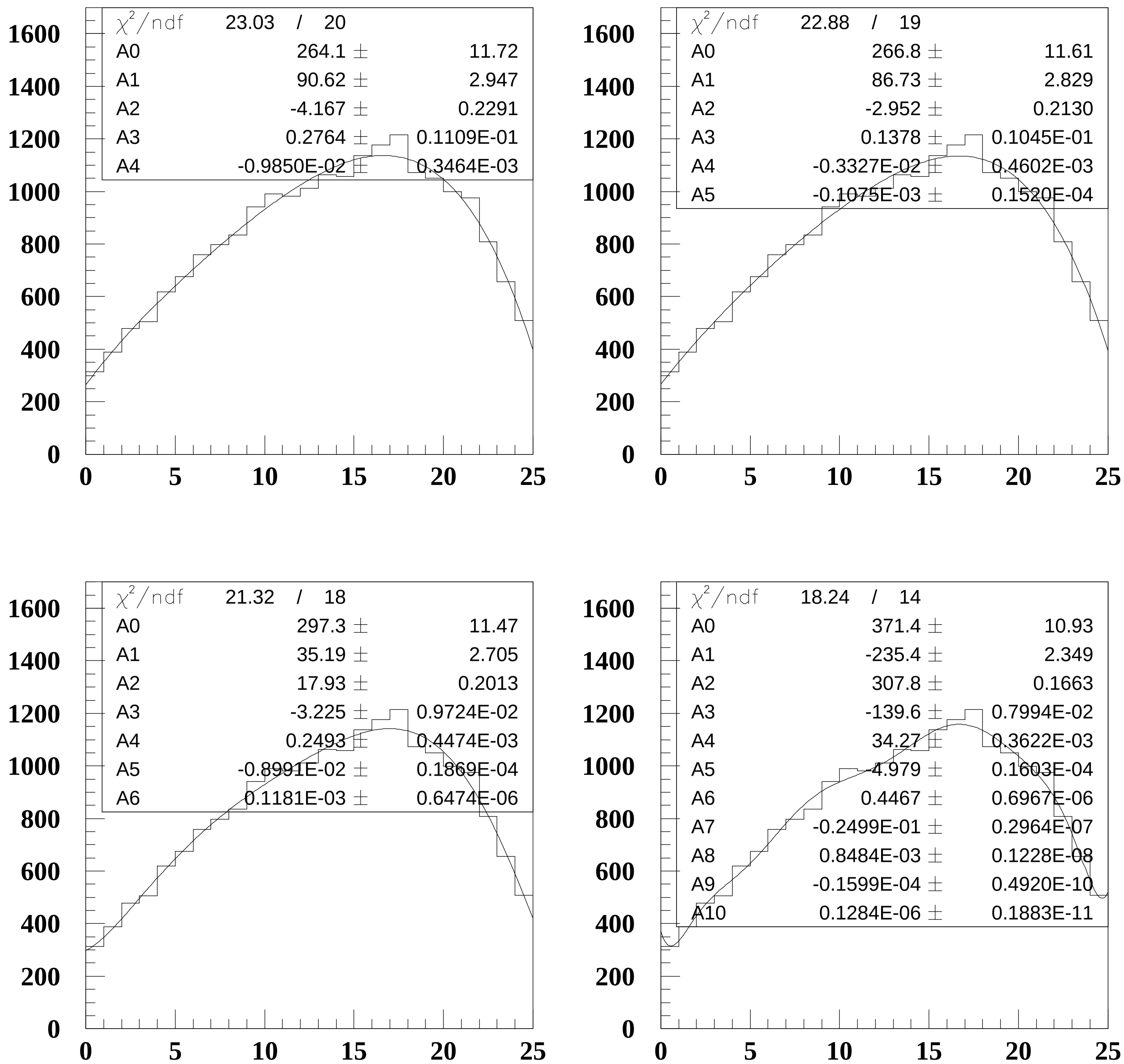
1. optically, how well the curves fit the data
2. a reasonable parametrisation should lead to $\chi^2/ndf \approx 1$.
3. choose only a more complicated parametrisation if a significant improvement of χ^2/ndf can be achieved

$\rightarrow$ Try to find your personal favorite.

Mini-exercise: Find best data parametrisation



Mini-exercise: **Find best data parametrisation**



Appendix

Content:

- Proof that χ^2_{min} for averaging two measurements follows χ^2-distribution with one degree of freedom

χ^2 for two measurements with unknown true value

$$\chi^2_{min} = \frac{(y_1-\hat{a})^2}{\sigma_1^2} + \frac{(y_2-\hat{a})^2}{\sigma_2^2}; \quad \hat{a} = \frac{1}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}} \cdot \left(\frac{y_1}{\sigma_1^2}+\frac{y_2}{\sigma_2^2}\right) = \frac{G_1 y_1 + G_2 y_2}{G_1+G_2} \quad \text{(with } G_i := 1/\sigma_i^2\text{))}$$

$$\begin{aligned}
\Rightarrow \chi^2_{min} &= G_1 \cdot \left(y_1 - \frac{(G_1 y_1 + G_2 y_2)}{G_1+G_2}\right)^2 + G_2 \cdot \left(y_2 - \frac{(G_1 y_1 + G_2 y_2)}{G_1+G_2}\right)^2 \\
&= G_1 \cdot \left(\frac{(G_2 y_1 - G_2 y_2)}{G_1+G_2}\right)^2 + G_2 \cdot \left(\frac{(G_1 y_2 - G_1 y_1)}{G_1+G_2}\right)^2 \\
&= \frac{G_1 G_2^2}{(G_1+G_2)^2}(y_1-y_2)^2 + \frac{G_2 G_1^2}{(G_1+G_2)^2}(y_1-y_2)^2 \\
&= \frac{G_1 G_2 (G_1+G_2)}{(G_1+G_2)^2} \cdot (y_1-y_2)^2 = \frac{G_1 \cdot G_2}{G_1+G_2} \cdot (y_1-y_2)^2 \\
&= \frac{1}{1/G_1+1/G_2} \cdot (y_1-y_2)^2 = \frac{1}{\sigma_1^2+\sigma_2^2} \cdot (y_1-y_2)^2
\end{aligned}$$

$\Delta = \frac{y_1-y_2}{\sqrt{\sigma_1^2+\sigma_2^2}}$ should follow *(errorpropagation!)* gauss distribution $\sim e^{-\frac{\Delta^2}{2}}$

$\rightarrow \chi^2 = \Delta^2$ follows 1-dim χ^2 distr.!

$\rightarrow$ One degree of freedom "sacrificed" for determination of $\hat{a}$.

General: n-measurements with one unknown a
$\rightarrow$ follows χ^2 distribution with $n-1$ degrees of freedom