Comments on χ^2 minimisation

Volker Blobel – University of Hamburg September 2003

1. Introduction

The determination of **parameters in fits to measured data** is a standard task of data analysis.

Examples are

- determination of calorimeter calibration constants,
- fit of parton densities in a global analysis of a wide range of deep inelastic and related scattering data (many different experiments),
- detector alignment and calibration procedures (many thousand parameters)

Standard statistical methods for parameter determination are

- Method of Least Squares
- Maximum Likelihood method

which have certain **optimal statistical properties**, which can be proven on the **basis of certain conditions**. In both methods a multidimensional objective function is constructed, taking into account the statistical properties of the data, and the minimum (or maximum) of the function w.r.t. the parameters has to be determined.

A popular method for parameter estimation is χ^2 minimisation $\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta}$ – is this identical to least squares?
The minimum value of the objective function in Least Squares follows often (not always) a χ^2 distribution.

In contrast to the well-defined standard methods

- in χ^2 minimisation a variety of different non-standard concepts is used,
- often apparently motivated by serious problems to handle the experimental data in a consistent way;
- especially for the error estimation there are non-standard concepts and methods.

From publications:

> To determine these parameters one must minimize a χ^2 which compares the measured values . . . to the calculated ones . . .

> Our analysis is based on an effective global chi-squared function that measures the quality of the fit between theory and experiment . . .

Two examples are given, which demonstrate that χ^2 minimisation can give **biased results**:

- Calorimeter calibration
- Averaging data with common normalisation error

Calorimeters for energy measurements in a particle detector require a calibration, usually based on test beam data (measured cell energies y_{ik}) with known energy E. A common method [1, 2, 3, 4, 5, 6, 7, 8] based on the χ^2 minimisation of

$$\chi^2 = \frac{1}{N}\sum_{k=1}^{N}\left(a_1 y_{1,k} + a_2 y_{2,k} + \ldots + a_n y_{n,k} - E\right)^2$$

for the determination of the a_j can produce biased results (D. Lincoln et al. [9]).

If there would be one cell only, one would have data y_k with standard deviation σ, with a mean value of $\overline{y} = \sum_k y_k/N$, and the intended result is simply $a = E/\overline{y}$

A one-cell version of the above χ^2 definition is

$$\chi^2 = \frac{1}{N}\sum_{k=1}^{N}\left(a \cdot y_k - E\right)^2 \qquad \text{with the biased result} \quad a = \frac{E\cdot\overline{y}}{\left(\sum_k y_k^2\right)/N} = \frac{E\cdot\overline{y}}{\overline{y}^2+\sigma^2} \quad \neq E/\overline{y}$$

There would be no bias, if the inverse constant a_{inv} would have been determined from

$$\chi^2 = \frac{1}{N}\sum_{k=1}^{N}\left(y_k - a_{\text{inv}} E\right)^2$$

In a χ^2 expression the measured values y_k should not be modified; instead the expectation has to take into account all known effects.

Given N data x_k with different standard deviations σ_k and a common relative normalisation error of ε. Apparently the mean value $\overline{y}$ can not be affected by the normalisation error, but its standard deviation is.
One method is to use the full covariance matrix for the correlated data, e.g. in the case $N = 2$:

$$\boldsymbol{V}_a = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} + \varepsilon^2 \cdot \begin{pmatrix} y_1^2 & y_1 y_2 \\ y_1 y_2 & y_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2 y_1^2 & \varepsilon^2 y_1 y_2 \\ \varepsilon^2 y_1 y_2 & \sigma_2^2 + \varepsilon^2 y_2^2 \end{pmatrix}$$

and minimising

$$\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta} \qquad \text{with} \quad \boldsymbol{\Delta} = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \end{pmatrix}$$

Example (from [10]): Data are
$y_1 = 8.0 \pm 2\%$ and $y_2 = 8.5 \pm 2\%$, with a common (relative) normalisation error of $\varepsilon = 10\%$.
The mean value resulting from χ^2 minimisation is:

$$7.87 \pm 0.81 \qquad \text{i.e. } < y_1 \text{ and } < y_2$$

- this is apparently wrong.

> . . . that including normalisation errors in the correlation matrix will produce a fit which is biased towards smaller values . . . [11]

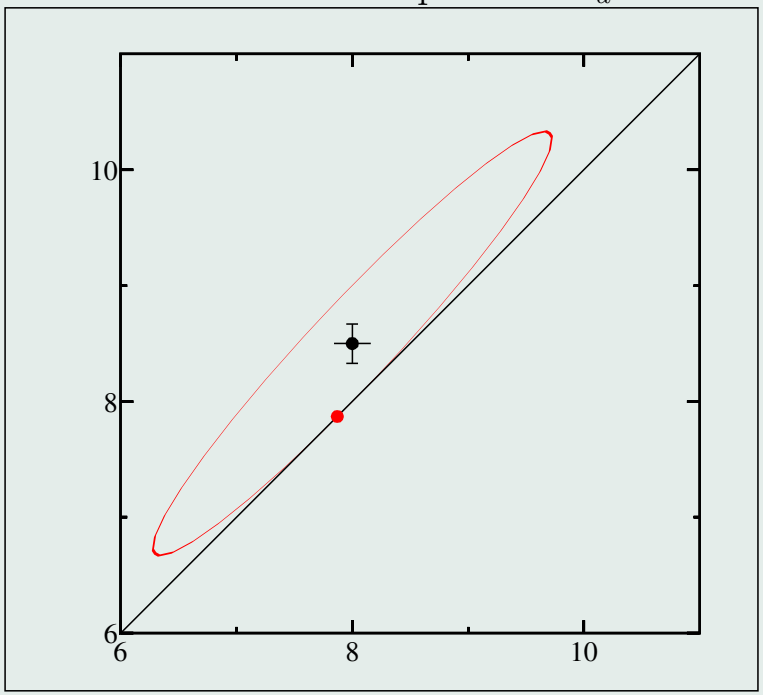
> . . . the effect is a direct consequence of the hypothesis to estimate the empirical covariance matrix, namely the linearisation on which the usual error propagation relies. [10, 12]

Origin of the apparent problem ... the used covariance matrix!

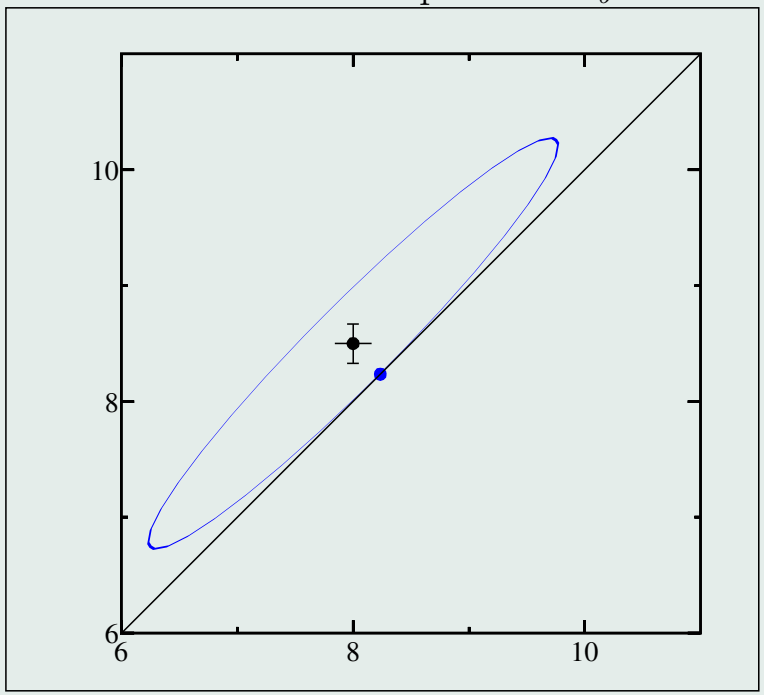
Correct model: true y_1 and y_2 and the normalisation errors $\varepsilon \cdot$ value are identical:

$$\boldsymbol{V}_b = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} + \varepsilon^2 \cdot \begin{pmatrix} \overline{y}^2 & \overline{y}^2 \\ \overline{y}^2 & \overline{y}^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2 \overline{y}^2 & \varepsilon^2 \overline{y}^2 \\ \varepsilon^2 \overline{y} & \sigma_2^2 + \varepsilon^2 \overline{y}^2 \end{pmatrix}$$

Covariance ellipse for $\boldsymbol{V}_a$



Covariance ellipse for $\boldsymbol{V}_b$



Axis of ellipse is tilted w.r.t. the diagonal and ellipse touches the diagonal at a biased point.

Axis of the ellipse is $\approx 45°$ and ellipse touches the diagonal at the correct point.

The result of χ^2 minimisation may depend critically on details of the model implementation!

2. Standard methods

Standard statistical methods for parameter determination are

- Method of Least Squares $S(\boldsymbol{a})$ • χ^2 minimisation is equivalent: $\chi^2 \equiv S(\boldsymbol{a})$
- Maximum Likelihood method $F(\boldsymbol{a})$
  ...improves the parameter estimation if the detailed probability density is known.

Least squares and Maximum Likelihood can be combined, e.g

$$F_{\text{total}}(\boldsymbol{a}) = \frac{1}{2} S(\boldsymbol{a}) + F_{\text{special}}(\boldsymbol{a})$$

Doubts about justification of χ^2 minimisation from publications:

> The justification for using least squares lies in the assumption that the measurement errors are Gaussian distributed. [13]

> However it is doubtful that Gaussian errors are realistic.

> A bad χ^2 ...Finally the data may very well not be Gaussian distributed.

Solution (from $\partial F/\partial \boldsymbol{a} = 0$) is linear transformation of the data vector $\boldsymbol{y}$:

$$\hat{\boldsymbol{a}} = \left[\left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} \right)^{-1} \boldsymbol{A}^T \boldsymbol{W} \right] \boldsymbol{y} \qquad = \boldsymbol{B}\,\boldsymbol{y}$$

Starting from **Principles** properties of the solution are derived, which are valid under certain conditions:

- Data are unbiased: $E[\boldsymbol{y}] = \boldsymbol{A}\,\bar{\boldsymbol{a}}$ ($\bar{\boldsymbol{a}}$ = true parameter vector)
- Covariance matrix $\boldsymbol{V}_y$ of the data is known (and correct).

Distribution-free properties of least squares estimates in linear problems are:

- Estimated parameters are unbiased: $\boldsymbol{W} = \boldsymbol{V}_y^{-1}$

$$E[\hat{\boldsymbol{a}}] = \left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} \right)^{-1} \boldsymbol{A}^T \boldsymbol{W}\, E[\boldsymbol{y}] = \bar{\boldsymbol{a}}$$

- In the class of unbiased estimates, which are linear in the data, the Least Squares estimates $\hat{\boldsymbol{a}}$ have the smallest variance (Gauß-Markoff theorem).
- The expectation of the sum of squares of the residuals is $\hat{S} = (n - p)$.

Special case of Gaussian distributed measurement errors:

$$\hat{S} \quad \text{distributed according to the} \quad \chi^2_{n-p} \quad \text{distribution}$$

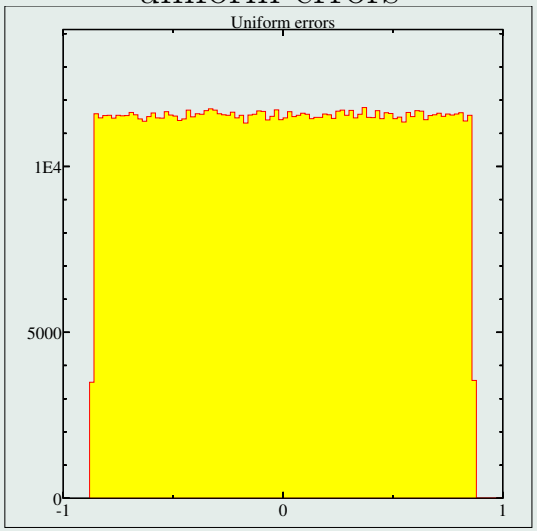
to be used for goodness-of-fit test. Properties are not valid, if conditions violated.
Covariance matrix of $\boldsymbol{a}$ by "error" propagation

$$\boldsymbol{V}[\hat{\boldsymbol{a}}] = \boldsymbol{B}\,\boldsymbol{V}[\boldsymbol{y}]\,\boldsymbol{B}^T = \left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} \right)^{-1} \qquad = \text{inverse of second der. matrix of } F(\boldsymbol{a})$$

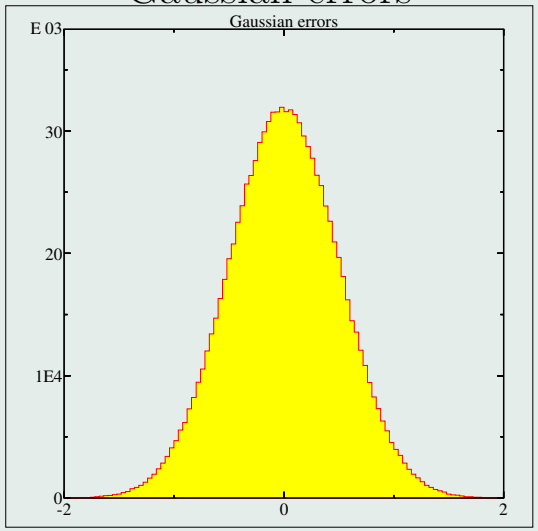
Test of non-Gaussian data

MC test of least squares fit of 20 data points to straight line (two parameters), generated with data errors from different distributions, but always mean = 0 and same standard deviation $\sigma = 0.5$.

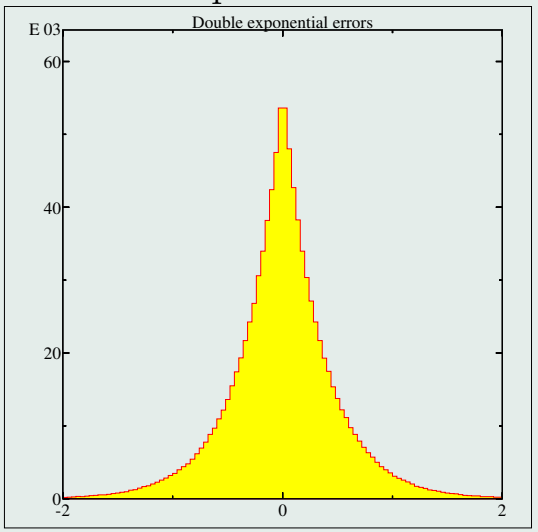
uniform errors



Gaussian errors



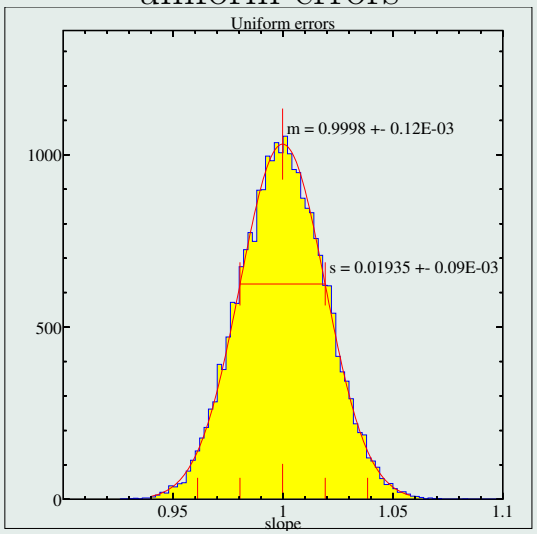
double exponential errors



Results for slope parameters

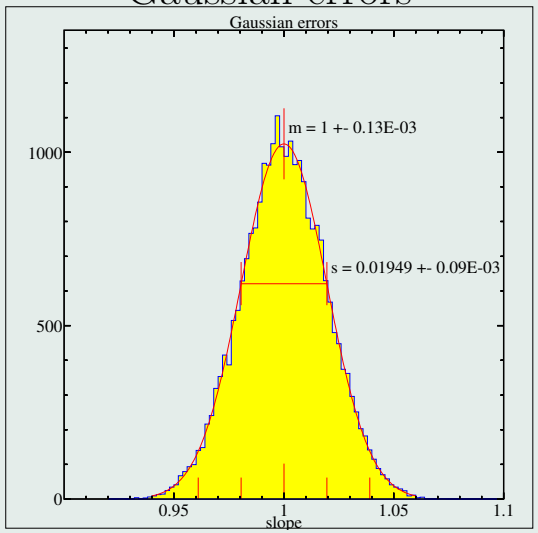
25000 entries

uniform errors



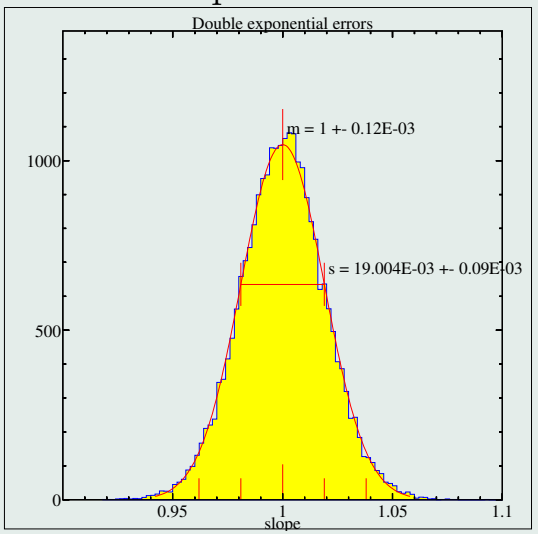
$\sigma = 0.0194$

Gaussian errors



$\sigma = 0.0195$

double exponential errors



$\sigma = 0.0190$

- All parameter distributions are Gaussian, and of the width, expected from the standard error calculation.
- This is valid for both fitted parameters.
- Mean χ^2-values are all equal to $n_{\mathrm{df}} = 20 - 2 = 18$, as expected, but
- χ^2-probabilities have different distributions, as expected.

Likelihood function and information

Case of m variables $a_1, \ldots, a_j, \ldots, a_m$: the information I is a m-by-m symmetric matrix $\boldsymbol{I}$ with elements

$$I_{jk} = E\left[\frac{\partial \ln \mathcal{L}}{\partial a_j}\frac{\partial \ln \mathcal{L}}{\partial a_k}\right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a_j \partial a_k}\right]$$

The minimal variance $\boldsymbol{V}\left[\hat{\boldsymbol{a}}\right]$ of an estimate $\hat{\boldsymbol{a}}$ is given by inverse of the information matrix $\boldsymbol{I}$:

minimal variance	$\boldsymbol{V}\left[\hat{\boldsymbol{a}}\right] = \boldsymbol{I}^{-1}$

Define the negative log likehood function as objective function and find minimum

$$F(\boldsymbol{a}) = -\ln \mathcal{L}(\boldsymbol{a}) \qquad\qquad \boldsymbol{g} = \frac{\partial F}{\partial a_j} = 0\ .$$

In case of good statistic the Hessian is almost constant in the region around the minimum and the inverse $\boldsymbol{H}^{-1}$ is a good estimate of the covariance matrix $\boldsymbol{V}_a$ of the parameters $\boldsymbol{a}$.

$$\boldsymbol{V}_a = \boldsymbol{H}^{-1}$$

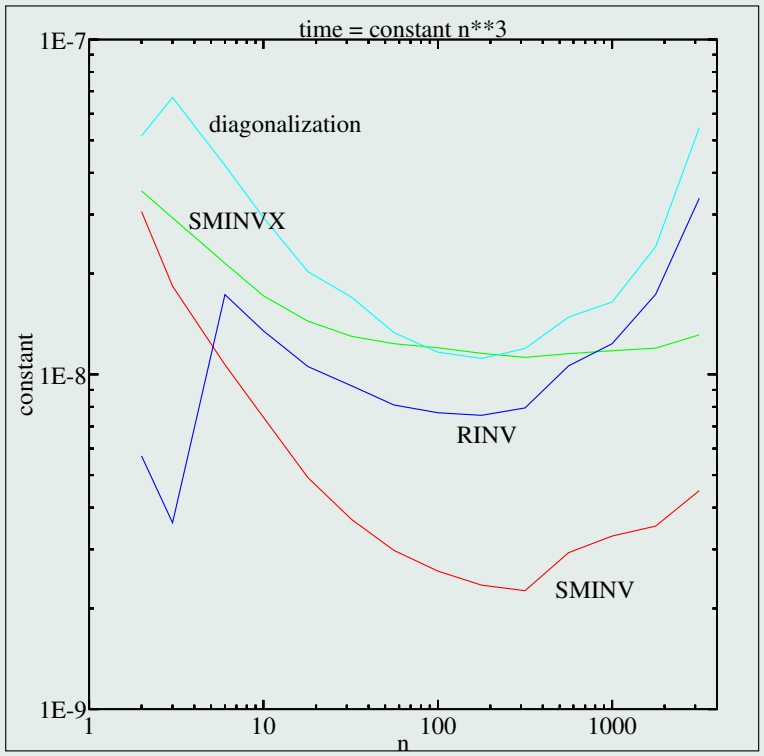
This corresponds to standard error propagation from the data errors to the parameter errors. The curvature (second derivative) of $F(\boldsymbol{a})$ determines the covariance matrix; this is essentially **error propagation** from the input (data) errors to the parameter errors; it does **not** depend on the goodness-of-fit.

Matrix inversion – timing

$n =$	**RINV**	**SMINV**	**HHLROT**	unit
10	13.5	7.4	29.4	μsec
100	7.7	2.6	11.6	msec
1000	12.4	3.3	16.4	sec
3162	17.9	2.4	28.7	min
words	n^2	$1/2\ n^2$	$3/2\ n^2$	

Inversion with $n = 25000$ will take $\approx$ one day (SMINV), but would require 1.25 GB.

time $=$ constant $\times\ n^3$

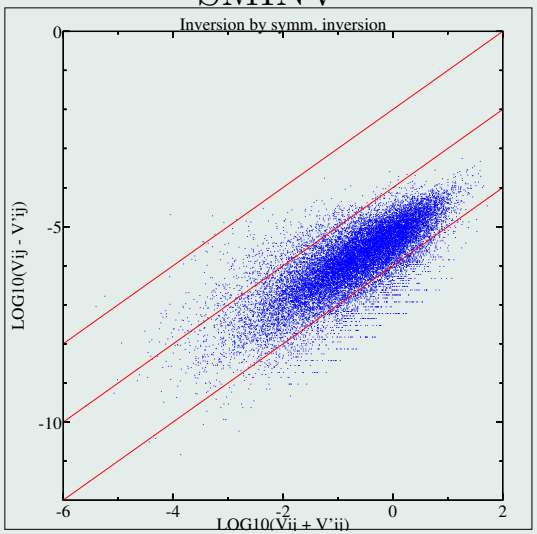


SMINV: Special Gauss-Jordan algorithm for **symmetric matrices** with pivot selection on diagonal. **Detects singularity by check of diagonal elements**, making use of the global correlation coefficient, and inverts a submatrix in case of singular matrix.

2.6 MHz Pentium with 512 MB; single precison computation.

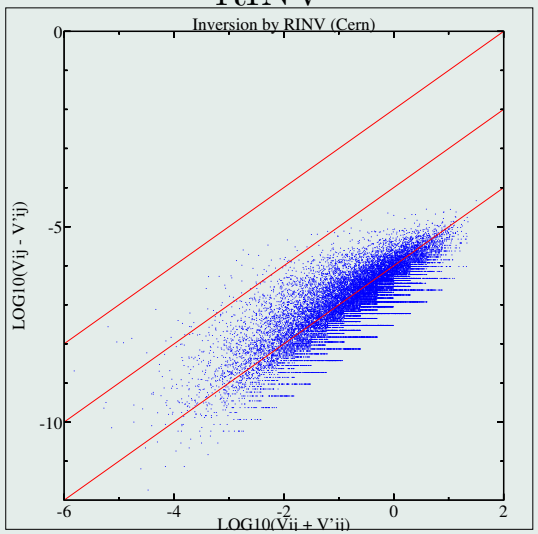
Check of accuracy based on $\boldsymbol{V}' = \left(\boldsymbol{V}^{-1}\right)^{-1}$. Plots show $\log_{10}$ of *difference* versus *sum* of elements; lines correspond to 10^{-2} 10^{-4} 10^{-6}.

SMINV



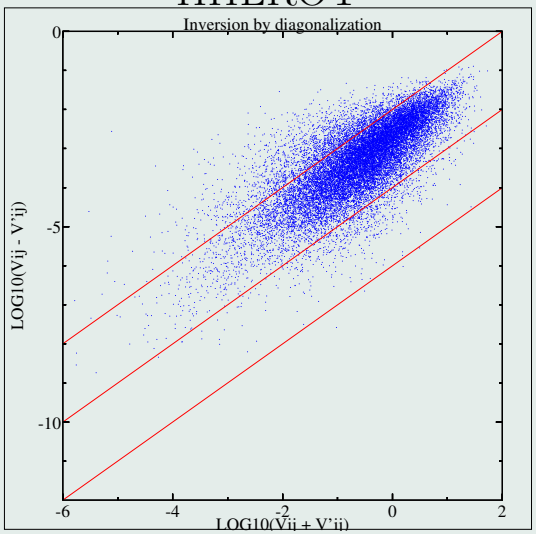
$\varepsilon \approx 10^{-5}$

RINV



$\varepsilon \approx 10^{-6}$

HHLROT



$\varepsilon \approx 10^{-3}$

Highest precision by RINV (Cern), but fails without result for a rank defect of 1. Accuracy of SMINV not changed by rank defect of 1.

3. Data and parameter errors

Data errors: Statistical and systematic uncertainties can only be correctly taken into account in a fit, if there is a clear **model** describing all aspects of the uncertainties.

Statistical data errors: described either

- by ("uncorrelated") errors – standard deviation σ_i for data point y_i (origin is usually counts – Poisson distribution),
- by a covariance matrix $\boldsymbol{V}_y$.

Two alternative models for **systematic errors**:

- **multiplicative effects** – normalisation errors
- **additive effects** – offset errors

that had to be accounted for in *different* ways in a fit.

Data y_i in Particle Physics are often (positive) cross sections, obtained from counts and several factors (Luminosity, detector acceptance, efficiency).

In general there is a normalisation error, given by a relative error ε. If data from > 1 experiment are combined, the normalisation error ε has to be taken into account.

Method: Introduce one additional factor α, which has been measured to be $\alpha = 1 \pm \varepsilon$, modify expectation according to

$$f_i = \alpha \cdot f(x_i, \boldsymbol{a})$$

and make fit with

$$S(\boldsymbol{a}) = \sum_i \frac{\left(y_i - \alpha \cdot f(x_i, \boldsymbol{a})\right)^2}{\sigma_i^2} + \Delta S^{\text{norm}} \qquad \text{with} \quad \Delta S^{\text{norm}} = \frac{(\alpha - 1)^2}{\varepsilon^2}$$

$$\text{or} \quad \Delta S^{\text{norm}} = \ln \alpha \left(3 + \frac{\ln \alpha}{\ln\left(1 + \varepsilon^2\right)}\right) \qquad \text{lognormal distribution}$$

One factor α_k has to be introduced for each experiment, if data from more than one experiment are fitted.

Example: error of calorimeter constant – a change of the constant will change *all* data values y_i – events are moved between bins.
Determine **shifts** s_i of data values y_i, for a one-standard deviation change of the calorimeter constant – the shifts s_i will carry a relative sign.

1. Method: Modify covariance matrix to include contribution(s) due to systematic errors

$$\boldsymbol{V}_a = \boldsymbol{V}_{\text{stat}} + \boldsymbol{V}_{\text{syst}} \qquad \text{with} \quad \boldsymbol{V}_{\text{syst}} = \boldsymbol{s}\boldsymbol{s}^T \quad \text{(rank=1 matrix)}$$

2. Method: Introduce one additional parameter β, which has been measured to be 0 ± 1, for each systematic error source, modify expectation according to

$$f_i = f(x_i, \boldsymbol{a}) + \beta \cdot s_i$$

and make fit with

$$S(\boldsymbol{a}) = \sum_i \frac{\left(y_i - (f(x_i, \boldsymbol{a}) + \beta s_i)\right)^2}{\sigma_i^2} \; + \; \beta^2$$

Advantage of additional parameter β:

- Allows to test the pull $= \hat{\beta}/\sqrt{1 - V_{\beta\beta}}$ due to the systematic error.
- Allows to test the effect of the fit model on the systematic effect from the global correlation coefficient $\rho_\beta^{\text{global}}$.
- Allows more insight into systematic effect by inspection of the correlation coefficients ρ_{β,a_j} between β and the other parameters.
- First derivative of expectation (for fits) is trivial: $\partial f_i/\partial\beta = s_i$.

The parameter(s) β can be eliminated in a modified χ^2 definition. [14]

Single parameter a_j: Calculate, for many fixed values of a_j, the function value $S(\boldsymbol{a})$, which requires always a minimisation with $(m-1)$ parameters (MINOS feature of MINUIT).

Function $g(\boldsymbol{a})$: Calculate, for many fixed values of g, the function value $S(\boldsymbol{a})$, which requires always a function minimisation. The standard method of constraining, in a fit, the $g(\boldsymbol{a})$ to a fixed value g_{fix} is by the method of Lagrange multipliers, minimizing

$$F(\boldsymbol{a}) + \lambda \cdot (g(\boldsymbol{a}) - g_{\text{fix}})$$

w.r.t. the parameters $\boldsymbol{a}$ and the Lagrange multiplier λ. This defines an $(m-1)$-dimensional subspace.

Note that the extremum is a saddle point: F is minimal w.r.t. $\boldsymbol{a}$ and maximal w.r.t. λ, and standard minimisation programs (like MINUIT) cannot be used.

An alternative is to assume, by trial-and-error, fixed values of the Lagrange multiplier λ and to minimize

$$F(\boldsymbol{a}) + \lambda \cdot g(\boldsymbol{a})$$

and, after minization, to calculate the corresponding fixed $g(\boldsymbol{a})$ (allows to use MINUIT). [14]

Systematic errors in χ^2 expressions

There is a variety of methods:

$$\chi^2 = \sum_i \frac{(\alpha \cdot f_i - y_i)^2}{\sigma_i^2} + \frac{(\alpha - 1)^2}{\varepsilon^2} \qquad \chi^2 = \sum_i \frac{(f_i - \alpha \cdot y_i)^2}{\sigma_i^2} + \frac{(\alpha - 1)^2}{\varepsilon^2}$$

$$\chi^2 = \sum_i \frac{(f_i/(1+\beta s_i) - y_i)^2}{\sigma_i^2} + \beta^2 \qquad \chi^2 = \sum_i \frac{(f_i \cdot (1+\beta s_i) - y_i)^2}{\sigma_i^2} + \beta^2$$

. . . in my nomenclatur.

"Offset method": Systematic errors are ignored in the fit ("forces the theory prediction to be as close as possible to the data"), but later added in quadrature. [13]

The fit result must be biased, if incomplete error information is used.

Parameter errors in χ^2 minimisation

Notice that the covariance matrix

$$V_{ij}^p = \langle \Delta_i \Delta_j \rangle = \Delta\chi^2 \cdot H_{ij}^{-1}$$

depends on the choice of $\Delta\chi^2$ which usually, but not always, is taken to be $\Delta\chi^2 = 1$. This choice ... corresponds to the definition of the width of a Gaussian distribution. [13]

In full global fit art in choosing "correct" $\Delta\chi^2$ given complication of errors. Ideally $\Delta\chi^2 = 1$, but unrealistic. [15]

... and $\Delta\chi^2$ is the allowed variation in χ^2. ... and a suitable choice of $\Delta\chi^2$ ... and $\Delta\chi^2$ is the allowed deterioration in fit quality for the error determination. [16]

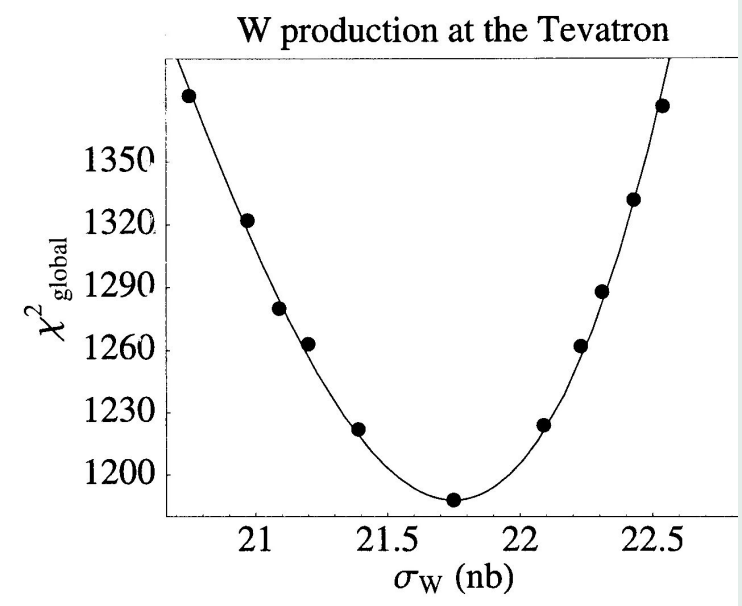
Group	$\Delta\chi^2$	Ref.	#	Value of $\alpha_S(M_Z^2)$		
H1	1	[17]	2	0.115 ± 0.0017 (exp)	$^{+0.0009}_{-0.0005}$ (model)	± 0.005 (theory)
GKK	1	[18, 19]	3	0.112 ± 0.001 (exp)		
MRST02	20	[16]	many	0.1195 ± 0.002 (exp)	± 0.003 (theory)	
ZEUS	50	[20, 21]	several	0.1166 ± 0.0040 (exp)	± 0.0081 (model)	± 0.004 (theory)
CTEQ6	100	[22]	several	0.1165 ± 0.0065 (exp)		

Also the errors obtained for functions of the parameters by error propagation are multiplied by a $\Delta\chi^2$.

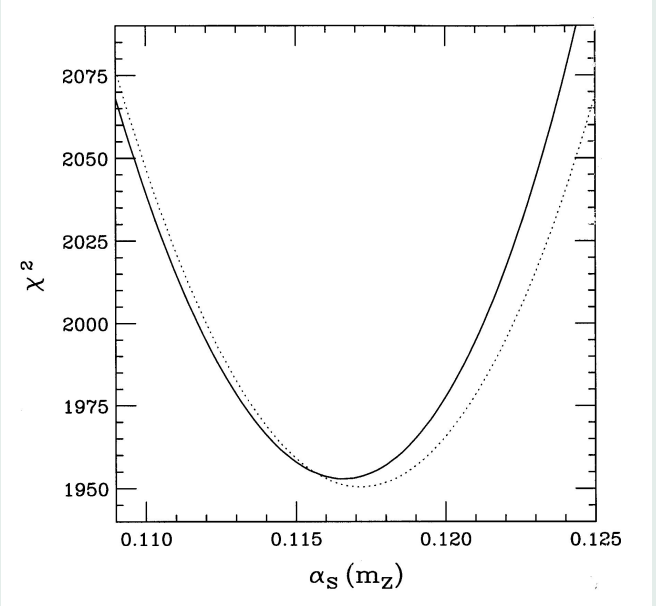
Examples with large $\Delta\chi^2$

The large, artificial and arbitrary magnification of errors is hardly acceptable – the procedure points to a deep problem in the whole data analysis. Two examples from parton distribution fits: [14, 15]

W production at the Tevatron



$\alpha_S(M_Z^2)$



Both curves are parabolas to a very good approximation over a range of $\Delta\chi^2 > 100$ ...

... while usually one would consider only a range of $\Delta\chi^2 \approx 4$, corresponding to two standard deviations.

4. Statistical properties of the data

A theorists view:

> Indeed, we have always believed the theory, rather than experiment, will provide the dominant source of error. [16]

But let us look at the statistical and systematic properties of the data.

- Are the data points (highly) correlated?

  > The main problem of the analysis is the correction for measurement errors (unsmearing corrections), which are ... [later more]

  $$g(y) = \int_\Omega A(y,x) f(x) \mathrm{d}x \qquad \text{or short} \qquad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

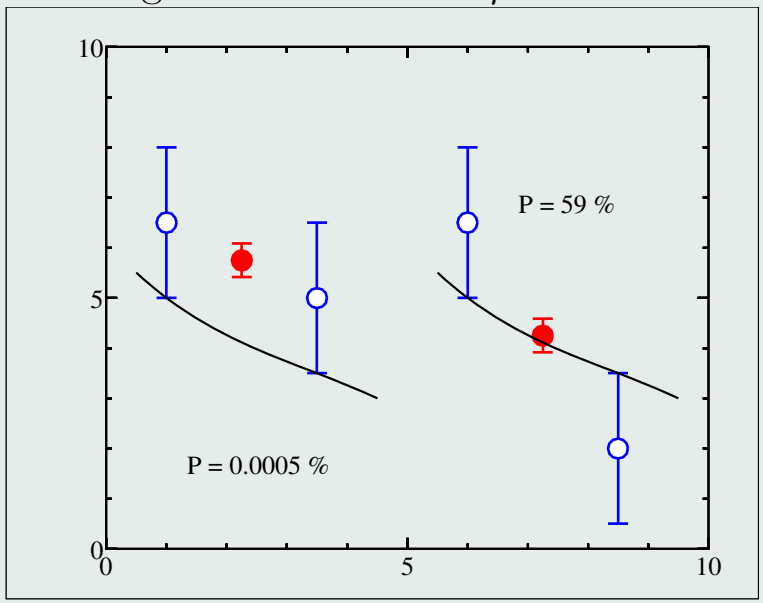
  $f(x)$ = true distribution, $g(y)$ = measured distribution, $A(y,x)$ = resolution function.

- Ist there enough information available on correlations/systematic errors, to be used in a fit?

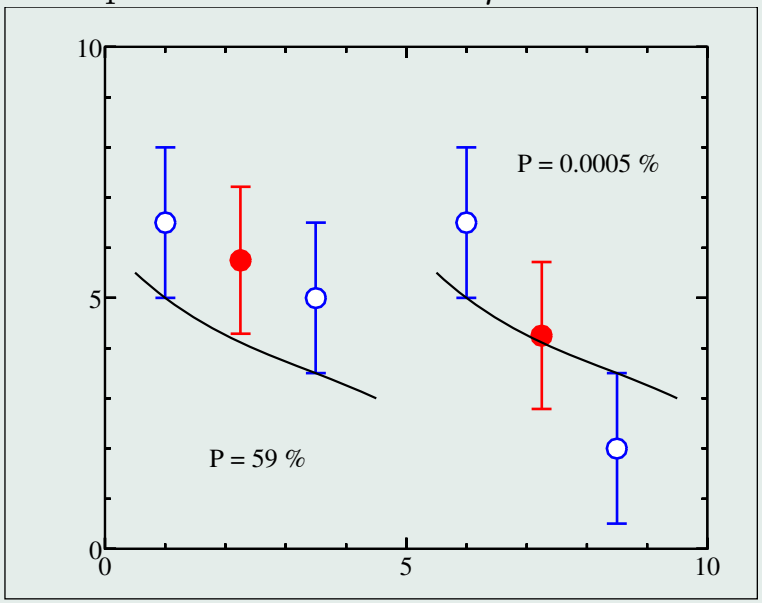
Comparing correlated data points

The two blue points with high negative/positive correlation are compared to a theoretical curve.

negative correlations $\rho = -0.9$



positive correlations $\rho = +0.9$

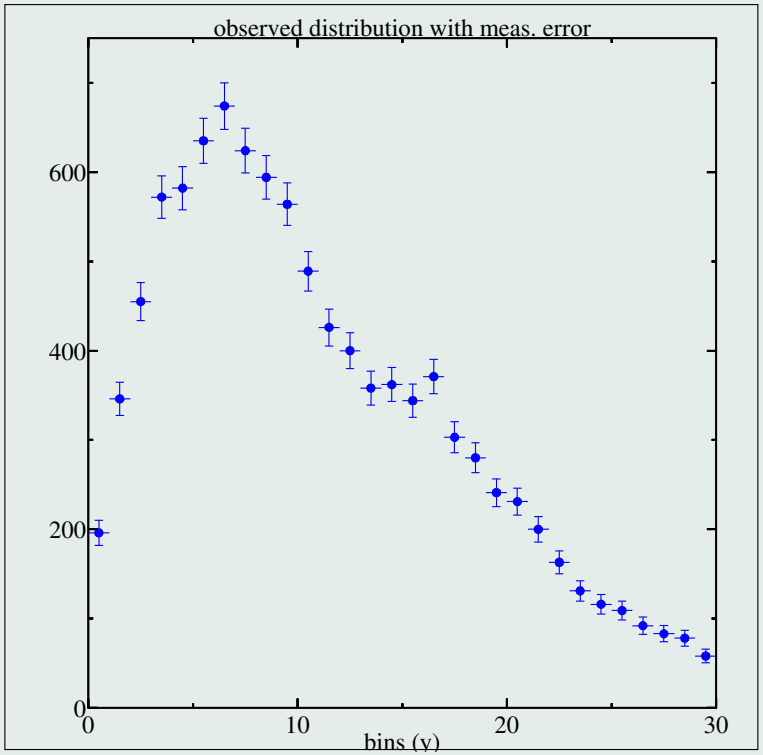


The χ^2-probabilities P are quite different for same sign/opposite sign deviations to the theoretical curve.

The average data point (red) of the two blue data points is very precise for negative correlations, but of almost the same precision as both single points for positive correlations.

An example for a measured histogram

... using migration parameter $\varepsilon = 0.24$, i.e. 52 % of true events remain in the same bin, and for 10 000 events.



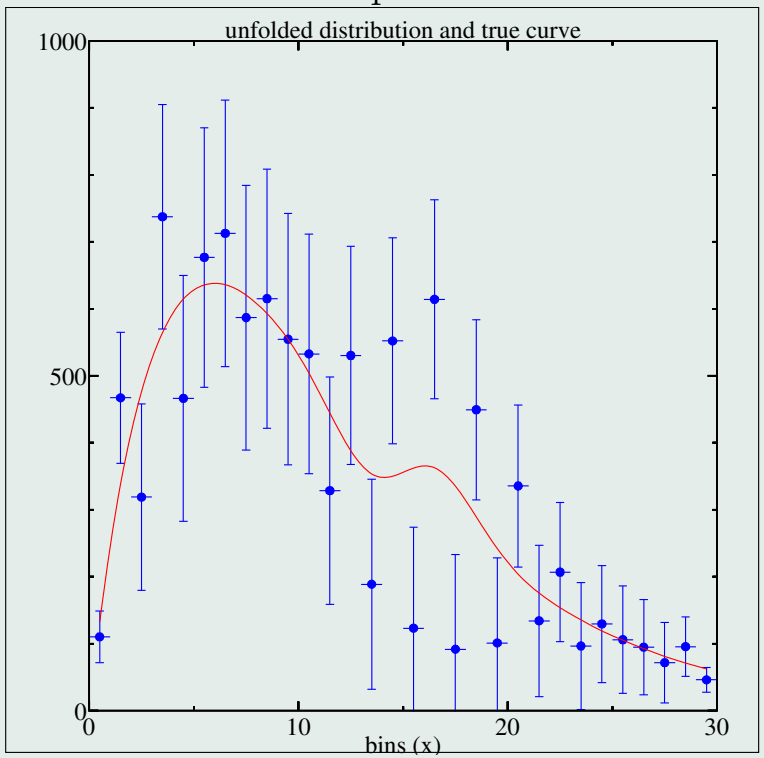
Note the small structure in the center:

- It may be just a statistical fluctuation ($\rightarrow$ smooth after unfolding)
- If it is a real structure in the distribution, then the true peak has to be higher!

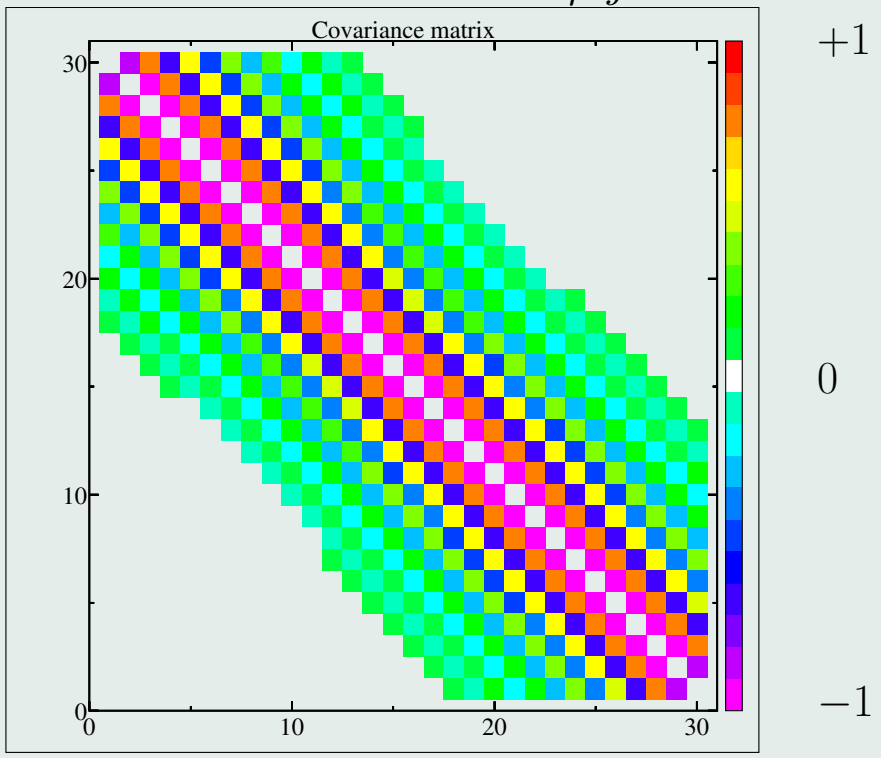
Result of solution by inversion

$\widehat{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{y}$

Reconstructed data points and true curve



Correlation coefficients ρij

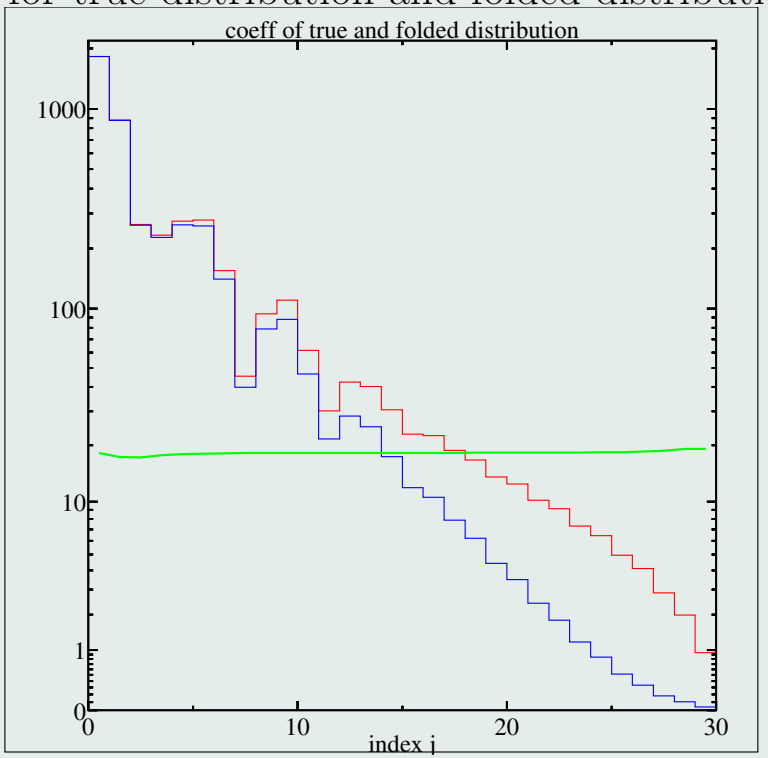


Highly fluctuation data points due to large negative correlations, caused by limited resolution. Correlation coefficients ρ_{ij} with $|\rho_{ij}| > 0.05$ are shown by colour boxes: here the coefficients $\rho_{i,i+1}$ between neighbour bins are ≈ -0.95.

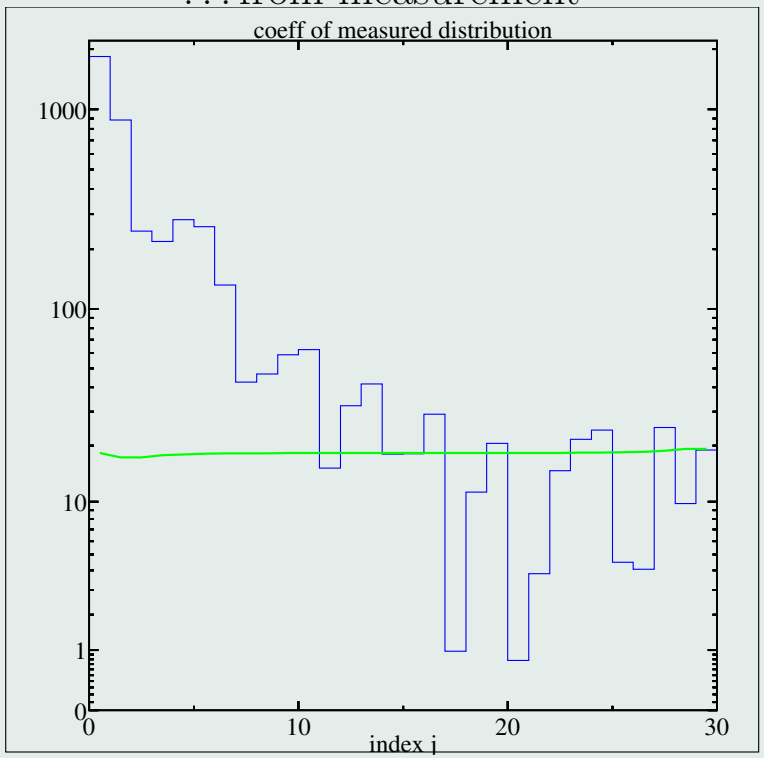
$$\boldsymbol{U}^T \cdot \;|\qquad \boldsymbol{y} \cong \boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{x}$$

$$\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y} \cong \boldsymbol{D}\left(\boldsymbol{U}^T\boldsymbol{x}\right) = \boldsymbol{D}\boldsymbol{b} \qquad \boldsymbol{b} = \boldsymbol{D}^{-1}\boldsymbol{c}$$

. . . for true distribution and folded distribution

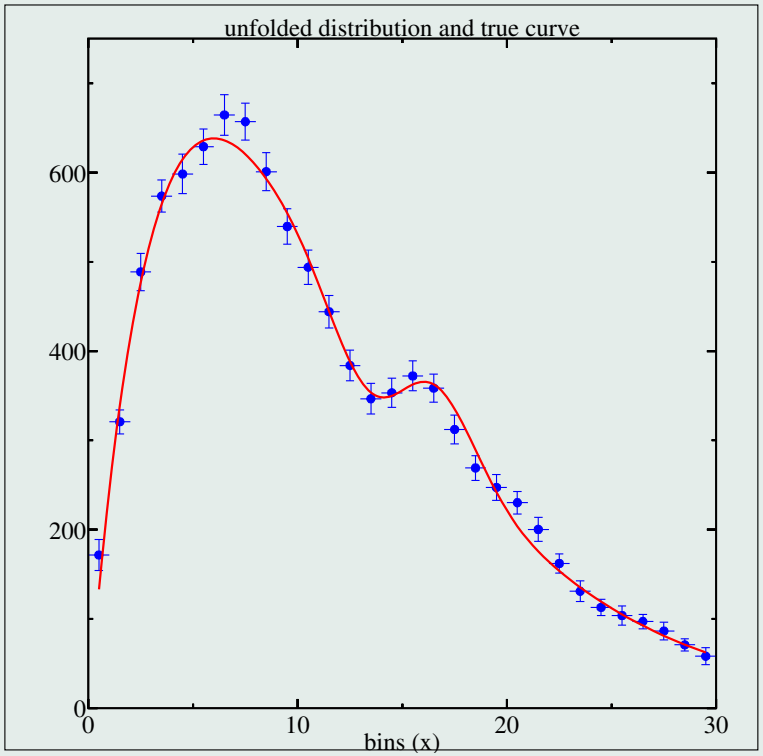


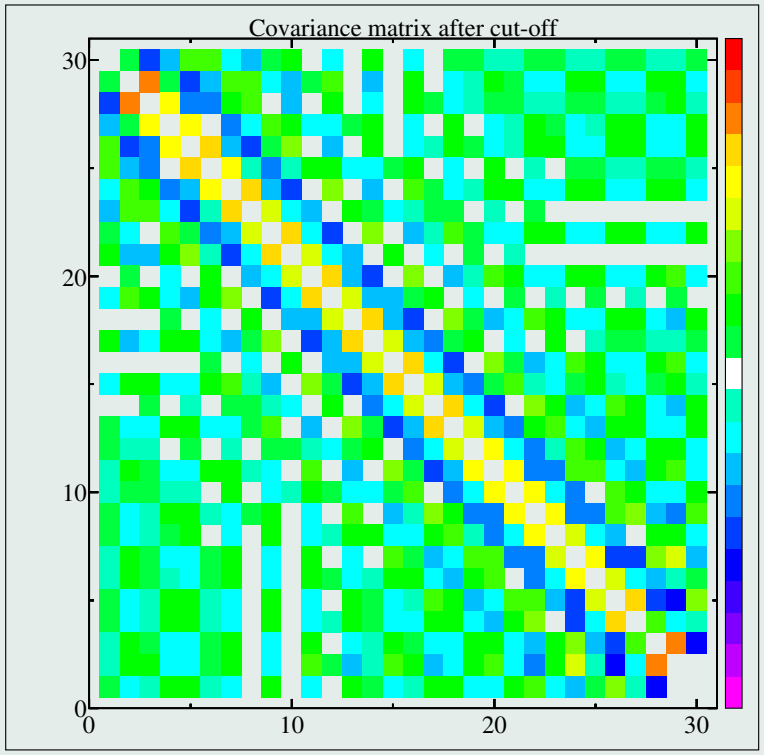
. . . from measurement



Folded amplitudes are measured and can be transformed to reconstruct the true amplitudes. Green line represents statistical errors (noise level). True and folded amplitudes below the noise level can not be reconstructed.

Take only the significant first 15 amplitudes to reconstruct the distribution with 30 data points.

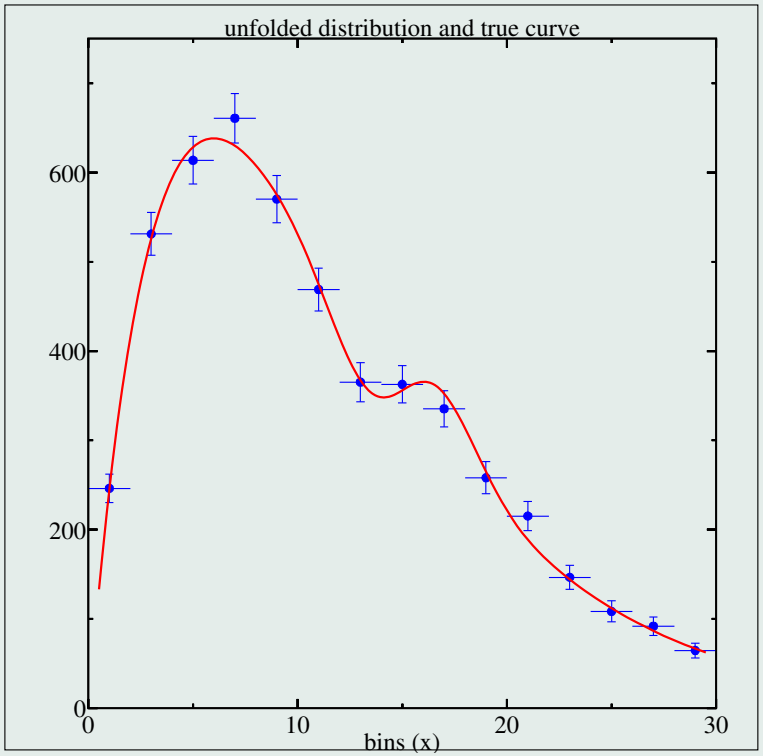


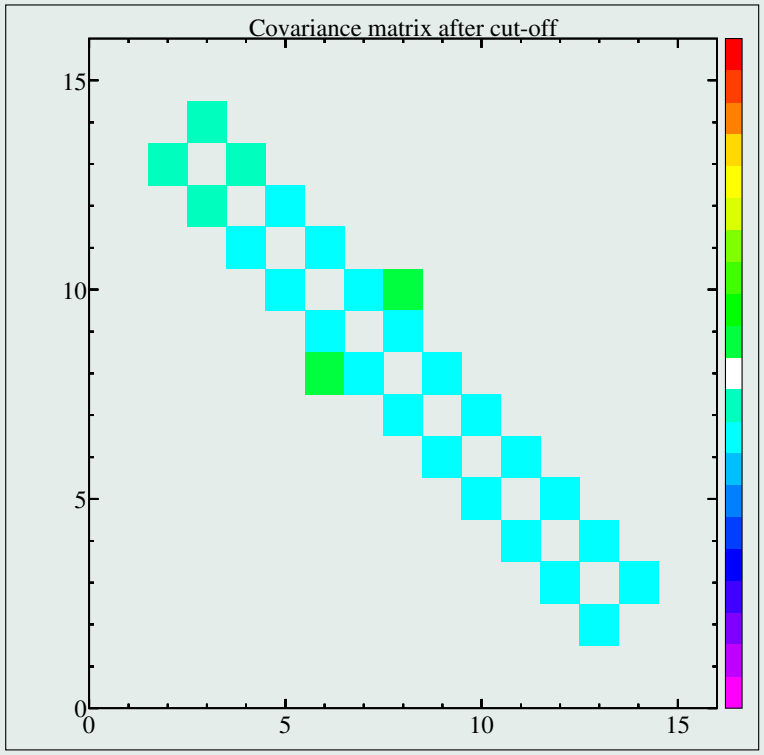


Covariance matrix has rank 15 and coefficients $\rho_{i,i+1}$ between neighbour bins are large and positive ($\approx +0.6$): "statistical" errors are smaller than original errors!

Solution with $N/2$ data points

If two bins are combined to one, the distribution has a covariance matrix with full rank.





All correlation coefficients are small, even between neighbour bins ($|\rho_{i,i+1}| < 0.2$) – but at the cost of a reduced number of points.

The standard method in particle physics to correct for the limited resolution is explained *in words* (no mathematical formula):

> ...The main problem of the analysis is the correction for measurement errors (unsmearing corrections), which are large at large x where the structure functions vary rapidly with x. We proceed by assuming a "true" structure function and calculate by Monte Carlo simulation, on the basis of the known experimental resolution functions, the result to be expected in the apparatus. **By iteration a "true" distribution which reproduces the experimental result is found.** The "unsmearing factor" is the ratio of Monte Carlo events for any particular (x, Q^2) bin in the "true" distribution divided by those in the resolution smeared distribution. If this factor differs from unity by more than 30 %, the bin is not retained. ... [23]

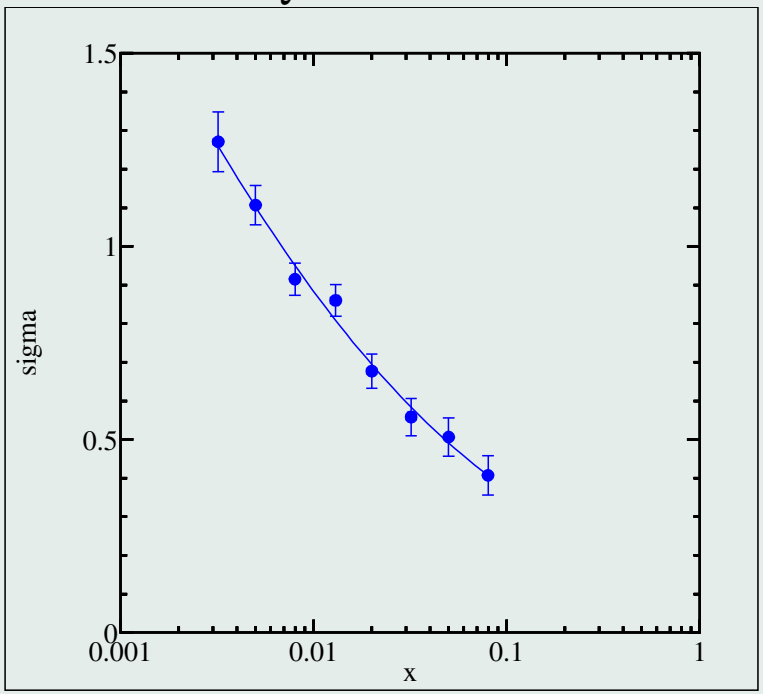
The method above is correct, if the "true" distribution is found without error. One could stop the procedure once the "true" distribution is found, but what about the measurement errors?

Any "true" distribution assumed to be very smooth may result in **positive correlations** between neighbour bins – the data have weights in fits which are too large.

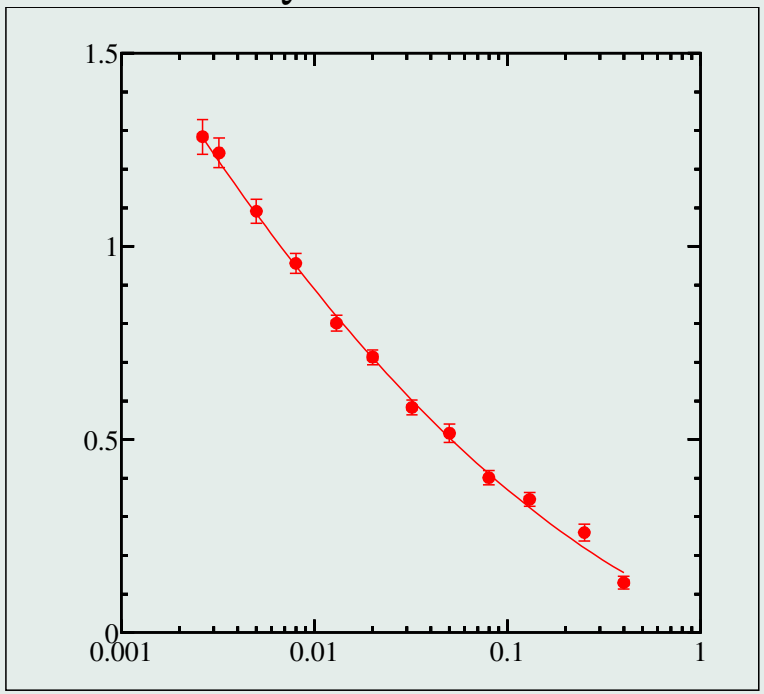
Published data on deep inelastic scattering

Shown are the total (statistical and "uncorrelated" systematic) errors. In addition there are "correlated" systematic errors and a normalisation error of 1.8 % and 1.5 %, resp. The curve is a fitted parabola, with a χ^2, that is better than expected (the data are rather smooth).

1998/1999 data (16.4 pb^{-1})
at $Q^2 = 200$ GeV^2



1999/2000 data (65.2 pb^{-1})
at $Q^2 = 200$ GeV^2

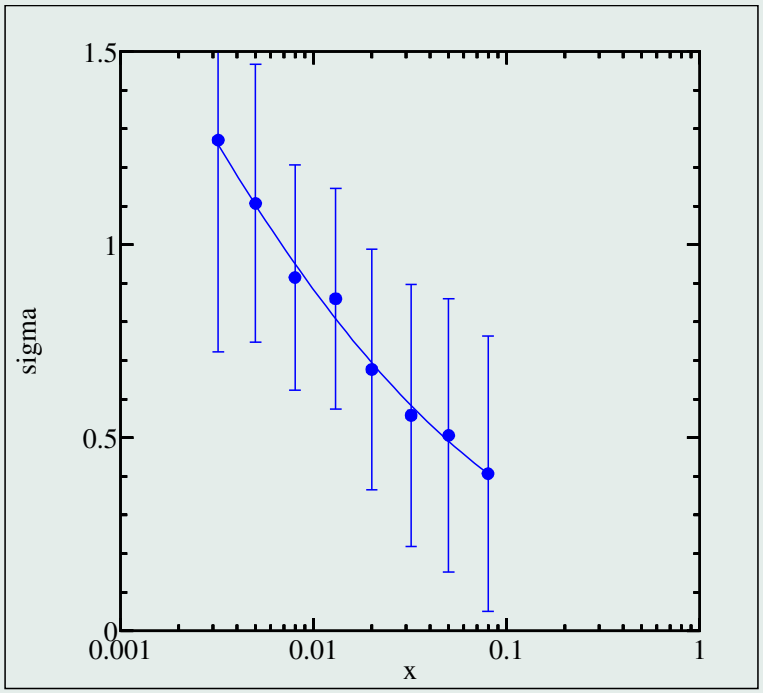


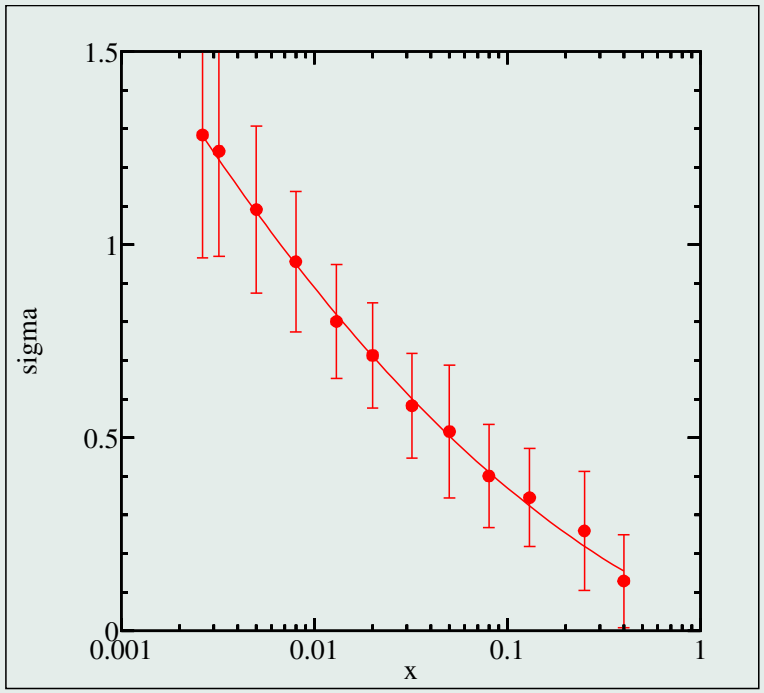
Unsmearing corrections are done based on earlier fits; bins are required to have stability and purity of $> 30\%$.

Same data

Errors multiplied by $\sqrt{50}$

Taking $\Delta\chi^2$ of 50 to calculate 1 standard deviation errors is equivalent to multiplication of all input errors by $\sqrt{50}$.





This looks strange indeed!

Summary

Objective functions to be minimized should follow statistical principles in order to avoid biased results:

- input data should have no bias, and correct statistical errors or covariance matrix, resp.
- complete systematic error contributions specified with a clear model of their meaning in fits
- the introduction of additional parameters for systematic error contributions allows insight into the interdependence (correlation coefficients, global correlations, pulls)
- parameter errors should be based only on error propagation from the input errors, not on arbitrary $\Delta\chi^2$ factors.

1. Introduction

The determination of **parameters in fits to measured data** is a standard task of data analysis.

Examples are

- determination of calorimeter calibration constants,
- fit of parton densities in a global analysis of a wide range of deep inelastic and related scattering data (many different experiments),
- detector alignment and calibration procedures (many thousand parameters)

Standard statistical methods for parameter determination are

- Method of Least Squares
- Maximum Likelihood method

which have certain **optimal statistical properties**, which can be proven on the **basis of certain conditions**. In both methods a multidimensional objective function is constructed, taking into account the statistical properties of the data, and the minimum (or maximum) of the function w.r.t. the parameters has to be determined.

A popular method for parameter estimation is χ^2 minimisation $\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta}$ – is this identical to least squares?
The minimum value of the objective function in Least Squares follows often (not always) a χ^2 distribution.

In contrast to the well-defined standard methods

- in χ^2 minimisation a variety of different non-standard concepts is used,
- often apparently motivated by serious problems to handle the experimental data in a consistent way;
- especially for the error estimation there are non-standard concepts and methods.

From publications:

> To determine these parameters one must minimize a χ^2 which compares the measured values . . . to the calculated ones . . .

> Our analysis is based on an effective global chi-squared function that measures the quality of the fit between theory and experiment . . .

Two examples are given, which demonstrate that χ^2 minimisation can give **biased results**:

- Calorimeter calibration
- Averaging data with common normalisation error

Calorimeters for energy measurements in a particle detector require a calibration, usually based on test beam data (measured cell energies y_{ik}) with known energy E. A common method [1, 2, 3, 4, 5, 6, 7, 8] based on the χ^2 minimisation of

$$\chi^2 = \frac{1}{N} \sum_{k=1}^{N} \left(a_1 y_{1,k} + a_2 y_{2,k} + \ldots + a_n y_{n,k} - E \right)^2$$

for the determination of the a_j can produce biased results, as pointed out by D. Lincoln et al. [9].

If there would be one cell only, one would have data y_k with standard deviation σ, with a mean value of $\overline{y} = \sum_k y_k / N$, and the intended result is simply $a = E/\overline{y}$

A one-cell version of the above χ^2 definition is

$$\chi^2 = \frac{1}{N} \sum_{k=1}^{N} \left(a \cdot y_k - E \right)^2$$

and minimizing this χ^2 has the biased result

$$a = \frac{E \cdot \overline{y}}{\left(\sum_k y_k^2 \right) / N} = \frac{E \cdot \overline{y}}{\overline{y}^2 + \sigma^2} \quad \neq E/\overline{y}$$

The bias mimics a non-linear response of the calorimeter.
A known bias in fitted parameters is easily corrected for.

Example: for a hadronic calorimeter one may have

$$\text{Energy resolution } \frac{\sigma}{E} = \frac{0.7}{\sqrt{E}} \qquad \text{which will result in a biased ratio} = \frac{E}{E + 0.7^2}$$

(at $E = 10$ GeV the resolution is 22 % and the bias is 5 %).

There would be no bias, if the inverse constant a_{inv} would have been determined from

$$\chi^2 = \frac{1}{N} \sum_{k=1}^{N} (y_k - a_{\text{inv}} E)^2$$

General principle: In a χ^2 expression the measured values y_k should not be modified; instead the expectation has to take into account all known effects.

There are N data x_k with different standard deviations σ_k and a common relative normalisation error of ε. Apparently the mean value $\overline{y}$ can not be affected by the normalisation error, but its standard deviation is.
One method is to use the full covariance matrix for the correlated data, e.g. in the case $N = 2$:

$$\boldsymbol{V}_a = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} + \varepsilon^2 \cdot \begin{pmatrix} y_1^2 & y_1 y_2 \\ y_1 y_2 & y_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2 y_1^2 & \varepsilon^2 y_1 y_2 \\ \varepsilon^2 y_1 y_2 & \sigma_2^2 + \varepsilon^2 y_2^2 \end{pmatrix}$$

and minimising

$$\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta} \qquad \text{with} \quad \boldsymbol{\Delta} = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \end{pmatrix}$$

Example (from [10]): Data are
$y_1 = 8.0 \pm 2\%$ and $y_2 = 8.5 \pm 2\%$, with a common (relative) normalisation error of $\varepsilon = 10\%$.
The mean value resulting from χ^2 minimisation is:

$$7.87 \pm 0.81 \qquad \text{i.e. } < y_1 \text{ and } < y_2$$

- this is apparently wrong.

> . . . that including normalisation errors in the correlation matrix will produce a fit which is biased towards smaller values . . . [11]

> . . . the effect is a direct consequence of the hypothesis to estimate the empirical covariance matrix, namely the linearisation on which the usual error propagation relies. [10, 12]

The contribution to $\boldsymbol{V}$ from the normalisation error was calculated from the measured values, which were different; the result is a covariance ellipse with axis different from 45° and this produces a biased mean value.

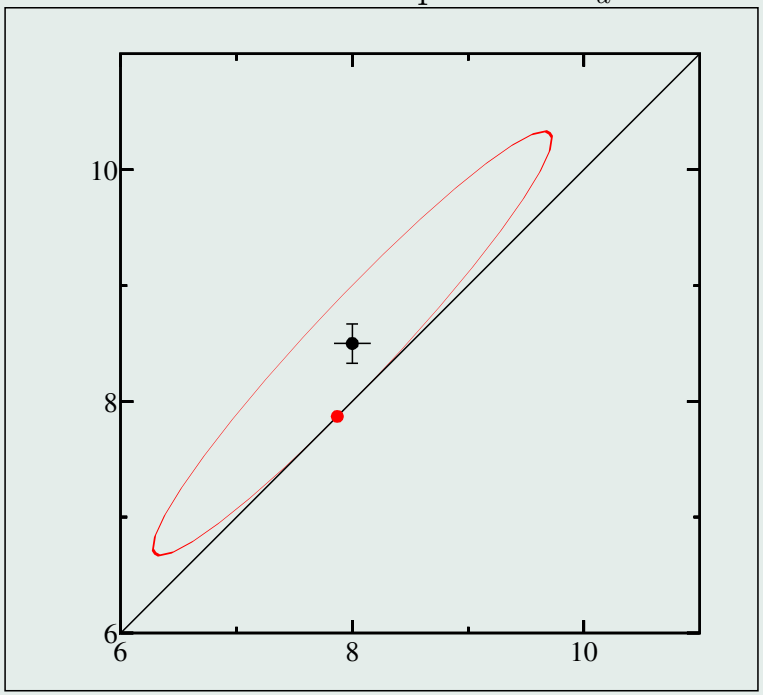
The correct model is: y_1 and y_2 have the same true value, then the normalisation errors $\varepsilon \cdot$ value are identical, with

$$\boldsymbol{V}_b = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} + \varepsilon^2 \cdot \begin{pmatrix} \overline{y}^2 & \overline{y}^2 \\ \overline{y}^2 & \overline{y}^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2\overline{y}^2 & \varepsilon^2\overline{y}^2 \\ \varepsilon^2\overline{y} & \sigma_2^2 + \varepsilon^2\overline{y}^2 \end{pmatrix}$$

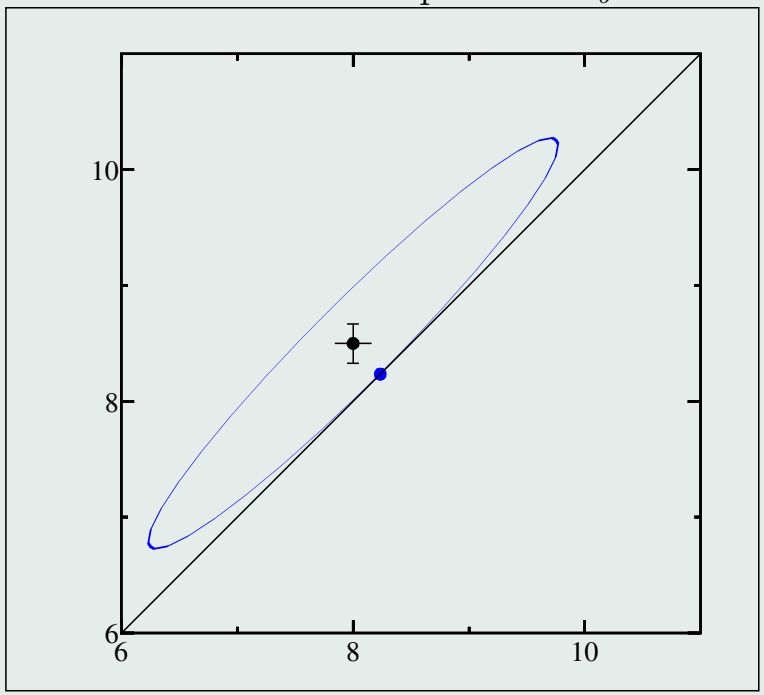
i.e. the covariance matrix depends on the resulting parameter.

Ellipses

Covariance ellipse for $\boldsymbol{V}_a$



Covariance ellipse for $\boldsymbol{V}_b$



Axis of ellipse is tilted w.r.t. the diagonal and ellipse touches the diagonal at a biased point.

Axis of the ellipse is $\approx 45°$ and ellipse touches the diagonal at the correct point.

The result may depend critically on certain details of the model implementation.

The method with one additional parameter ...

Another method often used is to define

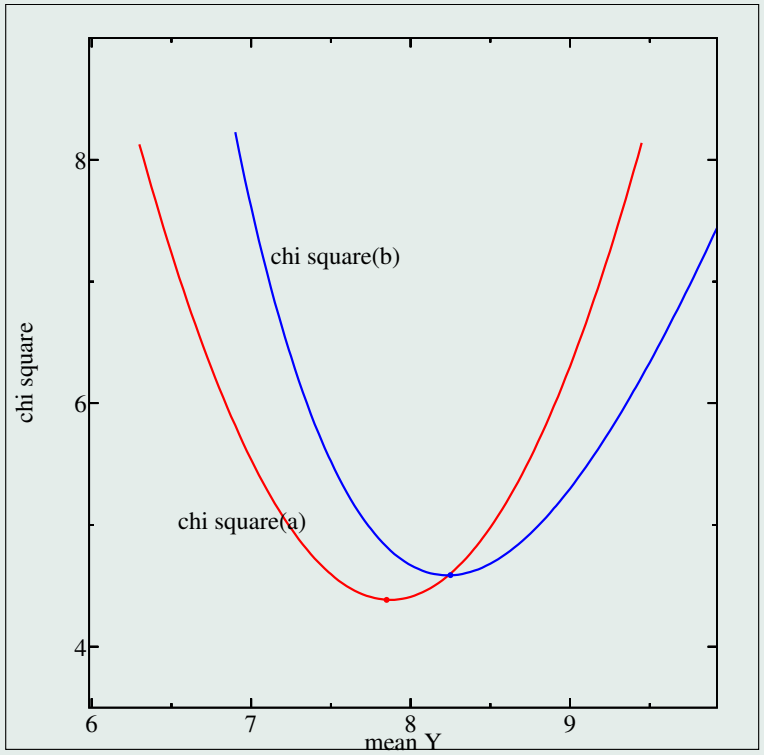
$$\chi^2_a = \sum_k \frac{(f \cdot y_k - \overline{y})^2}{\sigma_k^2} + \frac{(f-1)^2}{\varepsilon^2} \ ,$$

which will also produce a biased result.

The χ^2 definition for this problem

$$\chi^2_b = \sum_k \frac{(y_k - f \cdot \overline{y})^2}{\sigma_k^2} + \frac{(f-1)^2}{\varepsilon^2}$$

will give the correct result (data unchanged and fitted value according to the model), as seen by blue curve.



2. Standard methods

Standard statistical methods for parameter determination are

- Method of Least Squares $S(\boldsymbol{a})$ • χ^2 minimisation is equivalent: $\chi^2 \equiv S(\boldsymbol{a})$
- Maximum Likelihood method $F(\boldsymbol{a})$
  ...improves the parameter estimation if the detailed probability density is known.

Least squares and Maximum Likelihood can be combined, e.g

$$F_{\text{total}}(\boldsymbol{a}) = \frac{1}{2} S(\boldsymbol{a}) + F_{\text{special}}(\boldsymbol{a})$$

Doubts about justification of χ^2 minimisation from publications:

> The justification for using least squares lies in the assumption that the measurement errors are Gaussian distributed. [13]

> However it is doubtful that Gaussian errors are realistic.

> A bad χ^2 ...Finally the data may very well not be Gaussian distributed.

The standard linear least squares method

The model of Linear Least Squares: $\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{a}$

$\boldsymbol{y}=$ measured data $\quad \boldsymbol{A}=$ matrix (fixed) $\quad \boldsymbol{a}=$ parameters $\quad \boldsymbol{V}_y=$ covariance matrix of $\boldsymbol{y}$

Least Squares Principle: minimize the expression $\quad (\boldsymbol{W} = \boldsymbol{V}_y^{-1})$

$$S(\boldsymbol{a}) = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a})^T\ \boldsymbol{W}\ (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a}) \qquad \text{or} \quad F(\boldsymbol{a}) = \frac{1}{2}S(\boldsymbol{a})$$

with respect to $\boldsymbol{a}$.

Derivatives of expression $F(\boldsymbol{a})$:

$$\boldsymbol{g} = \frac{\partial F}{\partial \boldsymbol{a}} = -\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} + \left(\boldsymbol{A}^T\boldsymbol{W}\,\boldsymbol{A}\right)\boldsymbol{a}$$

$$\boldsymbol{H} = \frac{\partial^2 F}{\partial a_j a_k} = \left(\boldsymbol{A}^T\boldsymbol{W}\,\boldsymbol{A}\right) \qquad = \text{constant}$$

Solution (from $\partial F/\partial \boldsymbol{a} = 0$) is linear transformation of the data vector $\boldsymbol{y}$:

$$\hat{\boldsymbol{a}} = \left[\left(\boldsymbol{A}^T\boldsymbol{W}\,\boldsymbol{A}\right)^{-1}\ \boldsymbol{A}^T\boldsymbol{W}\right]\ \boldsymbol{y} \qquad = \boldsymbol{B}\,\boldsymbol{y}$$

Covariance matrix of $\boldsymbol{a}$ by "error" propagation

$$\boldsymbol{V}[\hat{\boldsymbol{a}}] = \boldsymbol{B}\,\boldsymbol{V}[\boldsymbol{y}]\,\boldsymbol{B}^T = \left(\boldsymbol{A}^T\boldsymbol{W}\,\boldsymbol{A}\right)^{-1} \qquad = \text{inverse of } \boldsymbol{H}$$

Properties of the solution

Starting from **Principles** properties of the solution are derived, which are valid under certain conditions:

- Data are unbiased: $E[\boldsymbol{y}] = \boldsymbol{A}\,\bar{\boldsymbol{a}}$ ($\bar{\boldsymbol{a}}$ = true parameter vector)
- Covariance matrix $\boldsymbol{V}_y$ of the data is known (and correct).

Distribution-free properties of least squares estimates in linear problems are:

- Estimated parameters are unbiased:

$$E[\hat{\boldsymbol{a}}] = \left(\boldsymbol{A}^T\boldsymbol{W}\,\boldsymbol{A}\right)^{-1}\;\boldsymbol{A}^T\boldsymbol{W}\;E[\boldsymbol{y}] = \bar{\boldsymbol{a}}$$

- In the class of unbiased estimates, which are linear in the data, the Least Squares estimates $\hat{\boldsymbol{a}}$ have the smallest variance (Gauß-Markoff theorem).
- The expectation of the sum of squares of the residuals is $\hat{S} = (n - p)$.

Special case of Gaussian distributed measurement errors:

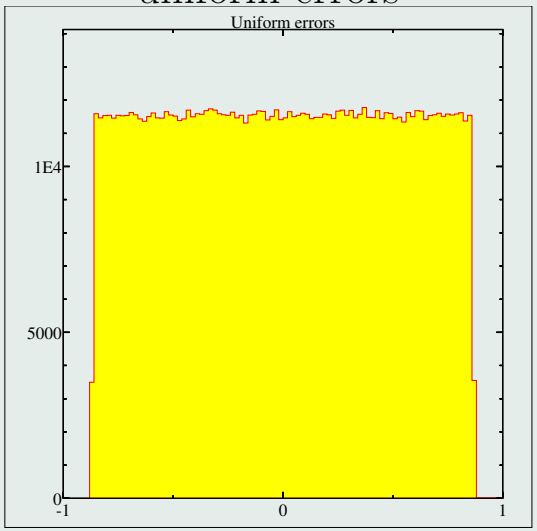
$$\hat{S}/\sigma^2 \quad \text{distributed according to the} \quad \chi^2_{n-p} \quad \text{distribution}$$

to be used for goodness-of-fit test. Properties are not valid, if conditions violated.

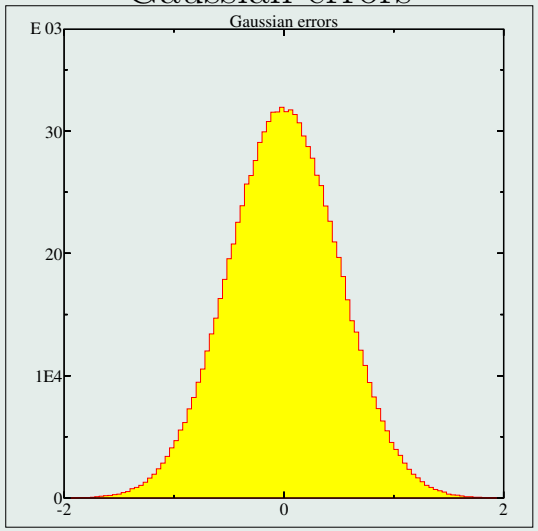
Test of non-Gaussian data

MC test of least squares fit of 20 data points to straight line (two parameters), generated with data errors from different distributions, but always mean = 0 and same standard deviation $\sigma = 0.5$.

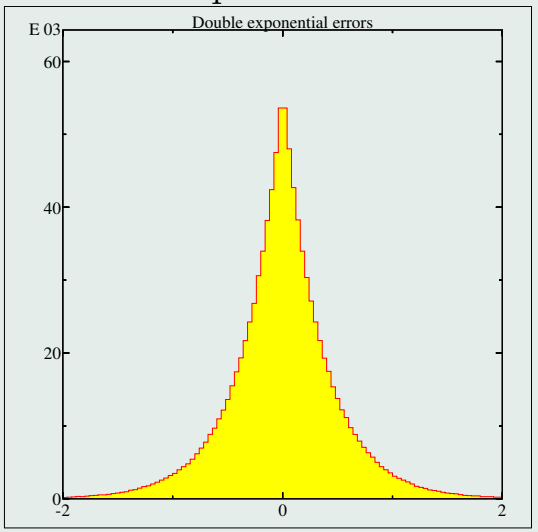
uniform errors



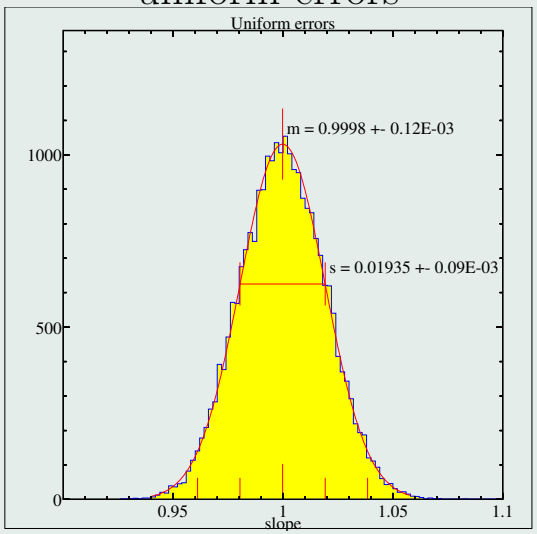
Gaussian errors



double exponential errors

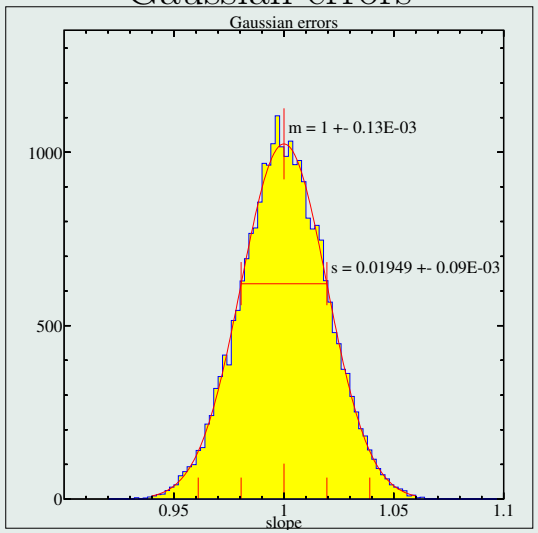


uniform errors



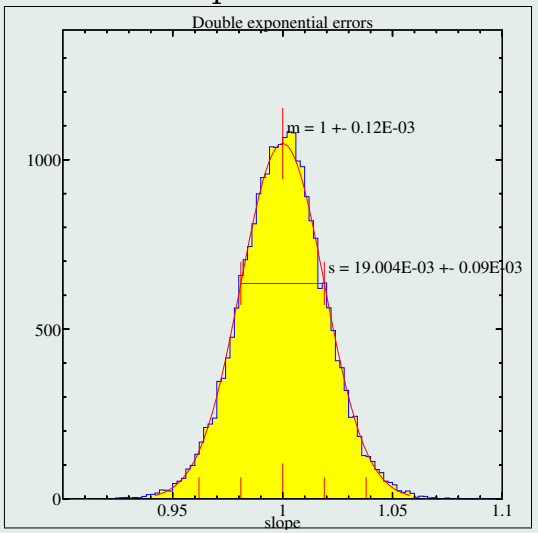
$\sigma = 0.0194$

Gaussian errors



$\sigma = 0.0195$

double exponential errors

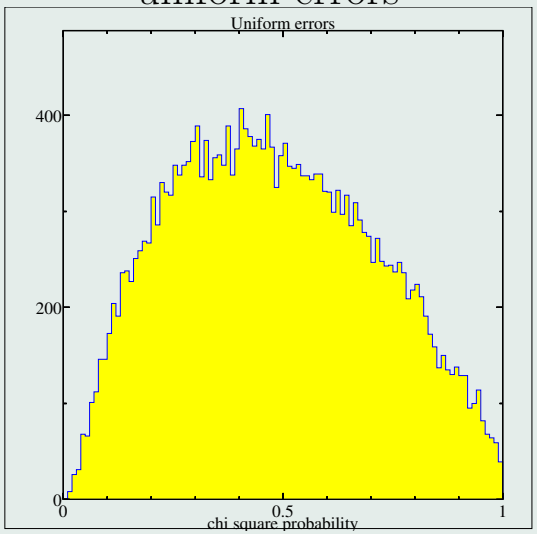


$\sigma = 0.0190$

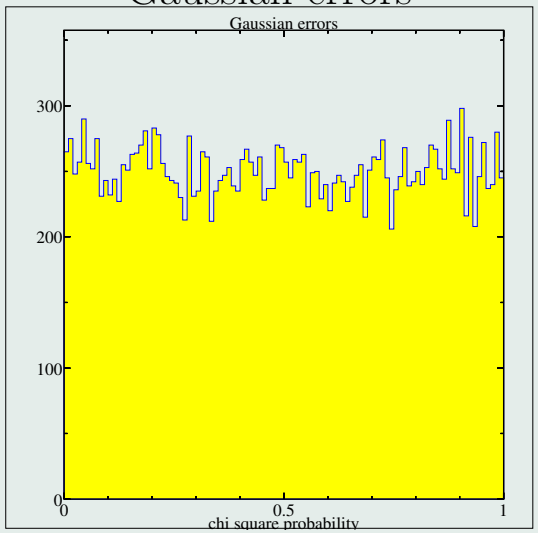
- All parameter distributions are Gaussian, and of the width, expected from the standard error calculation.
- This is valid for both fitted parameters.

- Mean χ^2-values are all equal to $n_{\rm df} = 20 - 2 = 18$, as expected, but
- χ^2-probabilities have different distributions, as expected.

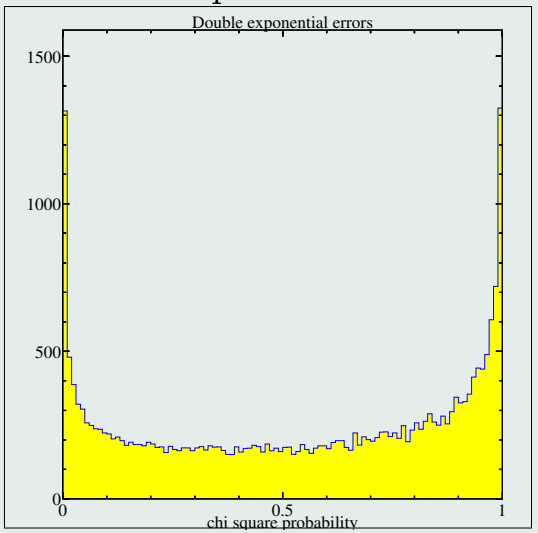
uniform errors



Gaussian errors



double exponential errors



Conclusion: Least squares works fine and as expected, also for non-Gaussian data,
if . . . and only if

- data are unbiased and covariance matrix is complete and correct.

Likelihood function and information

Given a sample $x_1, \ldots, x_i, \ldots, x_n$ (or short $\{\boldsymbol{x}\}$) of measured values from a distribution $p(x; a)$ (i.e. normalized density).

What is the information in the sample about the parameter(s) a?

The likelihood function as *joint density* of the observed values of the random variable x:

$$\boxed{\text{Likelihood function} \qquad \mathcal{L}(a) = \prod_{i=1}^{n} p(x_i; a)}$$

with normalisation

$$\int_\Omega \mathcal{L}(a)\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 1$$

Case of m variables $a_1, \ldots, a_j, \ldots, a_m$: information I becomes a m-by-m symmetric matrix $\boldsymbol{I}$ with elements

$$I_{jk} = E\left[\frac{\partial \ln \mathcal{L}}{\partial a_j}\frac{\partial \ln \mathcal{L}}{\partial a_k}\right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a_j \partial a_k}\right]$$

The minimal variance $\boldsymbol{V}\left[\hat{\boldsymbol{a}}\right]$ of an estimate $\hat{\boldsymbol{a}}$ is given by the inverse of the information matrix $\boldsymbol{I}$:

$$\boxed{\text{minimal variance} \qquad \boldsymbol{V}\left[\hat{\boldsymbol{a}}\right] = \boldsymbol{I}^{-1}}$$

Maximum likelihood method in practice

Define the negative log likehood function as objective function and find minimum

$$F(\boldsymbol{a}) = -\ln \mathcal{L}(\boldsymbol{a}) \qquad\qquad \boldsymbol{g} = \frac{\partial F}{\partial a_j} = 0 \ .$$

In case of good statistic the Hessian is almost constant in the region around the minimum and the inverse $\boldsymbol{H}^{-1}$ is a good estimate of the covariance matrix $\boldsymbol{V}_a$ of the parameters $\boldsymbol{a}$.

$$\boldsymbol{V}_a = \boldsymbol{H}^{-1}$$

This corresponds to standard error propagation from the data errors to the parameter errors.

The covariance matrix

- the function value $F(\hat{\boldsymbol{a}})$ determines the **goodness-of-fit** (not always); the goodness-of-fit has to be acceptable.
- the curvature (second derivative) of $F(\boldsymbol{a})$ determines the covariance matrix; this is essentially **error propagation** from the input (data) errors to the parameter errors; it does **not** depend on the goodness-of-fit.

Minimisation of objective function

$$F(\boldsymbol{a}+\boldsymbol{\Delta a}) = F(\boldsymbol{a}) + \boldsymbol{g}^T \cdot \boldsymbol{\Delta a} + \frac{1}{2}\boldsymbol{\Delta a}^T \boldsymbol{H} \boldsymbol{\Delta a} + \ldots$$

with

$$\text{gradient} \quad g_j = \frac{\partial F}{\partial a_j} \qquad\qquad \text{Hessian} \quad H_{jk} = \frac{\partial^2 F}{\partial a_j \partial a_k}$$

$$\text{Newton step} \qquad\qquad \boldsymbol{\Delta a} = -\boldsymbol{H}^{-1}\boldsymbol{g}$$

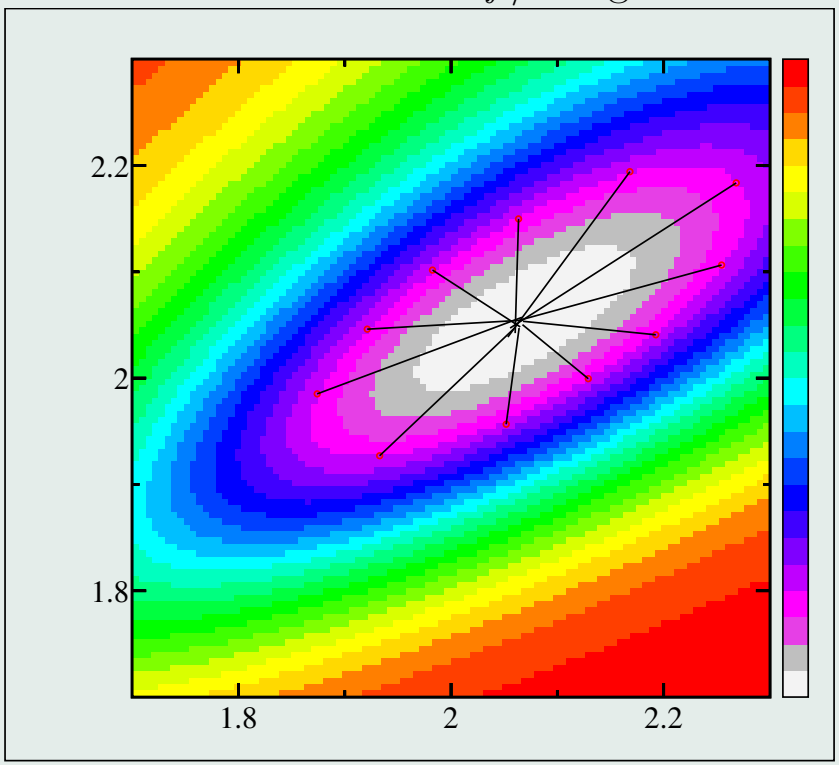
Least squares contributions

$$\begin{aligned}
F(\boldsymbol{a}) &= \frac{1}{2}\sum_i \frac{1}{\sigma_i^2}\left(y_i - f(x_i,\boldsymbol{a})\right)^2 \\
\frac{\partial F}{\partial a_j} &= \sum_i \frac{1}{\sigma_i^2}\frac{\partial f}{\partial a_j}\left(y_i - f(x_i,\boldsymbol{a})\right) \\
\frac{\partial^2 F}{\partial a_j \partial a_k} &= \sum_i \frac{1}{\sigma_i^2}\left(\frac{\partial f}{\partial a_j}\frac{\partial f}{\partial a_k} - \frac{\partial^2 f}{\partial a_j \partial a_k}\left(y_i - f(x_i,\boldsymbol{a})\right)\right)
\end{aligned}$$

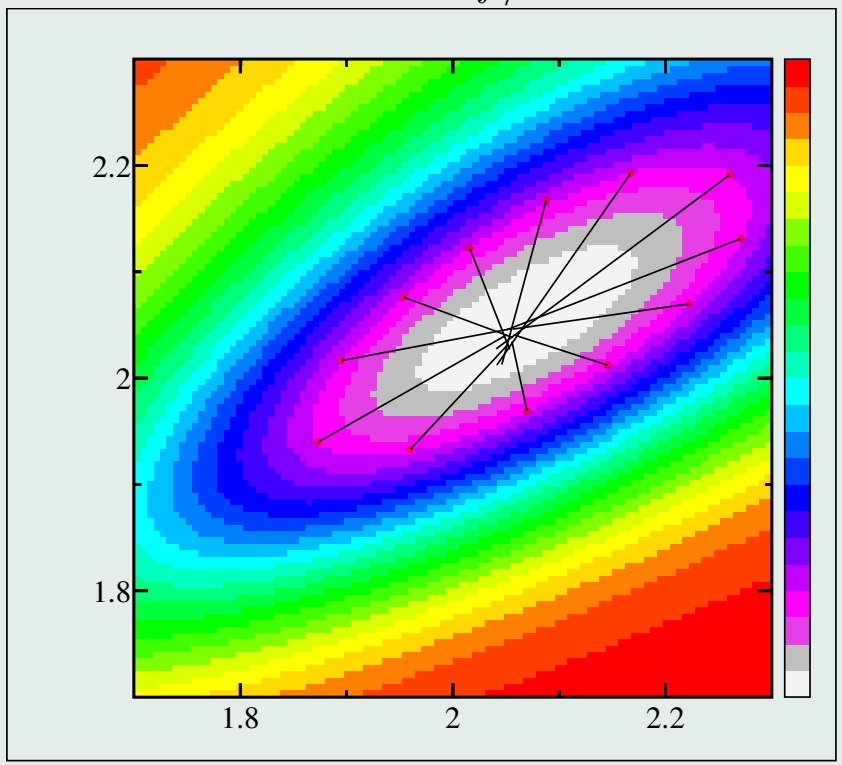
Ignoring second derivatives *improves* the Newton step!

Colour contours of objective function $S(\boldsymbol{a})$: steps correspond to $\Delta\chi^2 \approx 50$

Second derivatives $\partial f/\partial a$ ignored



Second derivatives $\partial f/\partial a$ included



Ignoring second derivatives *improves* the Newton step!

Minimisation

MINUIT is standard, general, well documented and "easy" to use, but it requires to code the objective function (χ^2 function) – which is not always simple and straightforward.

Often the objective function could have standard form – a fit of function $f(x_i, \boldsymbol{a})$ to data sets y_i – and would allow:

- Standardized handling for function integration over bins (histograms) and of the systematic additive and multiplicative (normalisation) errors.
- strategy to have a reduced number of parameter changes for those parameters, where this is expensive,
- use of first derivative (analytical or numerical) of $f(x, \boldsymbol{a})$ to construct first (gradient) and second derivative (Hessian) of objective function,
- making use of the structure of the equations e.g. for the systematic error contributions,
- allow equality constraints between parameters (Lagrange).

Matrix inversion

Matrix inversion is an essential part of minimisation and covariance matrix calculation.

Inversion is a n^3 process and can be time consuming for large matrix dimension. Matrix inversion fails for singular matrices and is inaccurate for almost singular matrices.

> For experiments with many data points, the inversion of such large matrices may lead to numerical instabilities, in addition to being time-consuming. [14]

> Minimizing χ^2 ...is impractical because it involves the inversion of the measurement covariance matrix which, in global fits, tends to be very large.

Matrices to be inverted in statistical compuation are *symmetric* and represent covariance/Hessian matrices. The storage and computation can make use of the symmetry.

Matrices with highly correlated parameters are almost singular.
Strategy: if parameters are highly correlated – invert submatrix and treat dependent parameters as fixed (zero correction).

Matrix programs

RINV Cern-Library program for full matrices, using triangular factorization with row interchange (special code for $n \leq 3$). Returns flag for singularity, but singularity will often go undetected.

SMINVX Special Gauss-Jordan algorithm for **symmetric matrices** with pivot selection on diagonal. **Detects singularity by check of diagonal elements** and inverts a submatrix in case of singular matrix.

SMINV Same as **SMINVX**, but with index calculation avoiding integer multiply and up to a factor of 3 faster.

HHLROT Diagonalization (eigenvalues + eigenvectors) by Householder transformation followed by diagonalization of tridiagonal matrices. [24] Allows to recognize insignificant components of the solution.

Global correlation and pivot selection

The **global correlation coefficient**, ρ_k is a measure of the total amount of correlation betwen the k-th parameter and *all* the other variables. It is the largest correlation between the k-th parameter and every possible linear combination of all the other variables.

$$\rho_k = \sqrt{1 - \frac{1}{(\boldsymbol{V})_{kk} \cdot \left(\boldsymbol{V}^{-1}\right)_{kk}}} \qquad \text{and} \qquad (\boldsymbol{V})_{kk} \cdot \left(\boldsymbol{V}^{-1}\right)_{kk} = \frac{1}{1-\rho_k^2}$$

Rule in SMINV: use largest pivot element (on diagonal), but avoid elements with $(\boldsymbol{V})_{kk} \cdot \left(\boldsymbol{V}^{-1}\right)_{kk} > 1/\varepsilon$. Stop inversion if no acceptable pivot can be found and clear corresponding matrix elements.

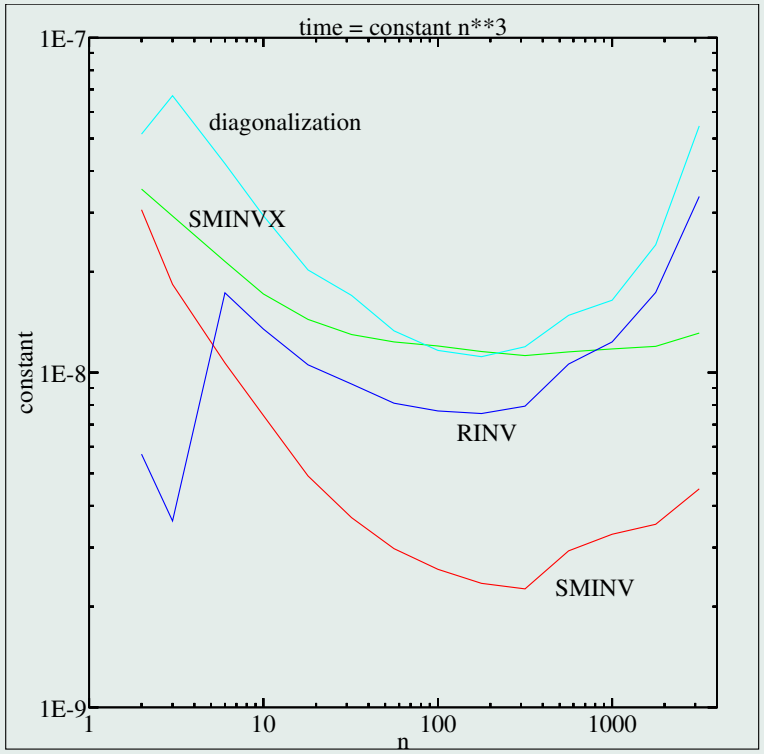
i.e. invert the largest possible submatrix if complete matrix is singular.

Matrix inversion – timing

$n =$	**RINV**	**SMINV**	**HHLROT**	unit
10	13.5	7.4	29.4	μsec
100	7.7	2.6	11.6	msec
1000	12.4	3.3	16.4	sec
3162	17.9	2.4	28.7	min
words	n^2	$1/2\ n^2$	$3/2\ n^2$	

Inversion with $n = 25000$ will take $\approx$ one day (SMINV), but would require 1.25 GB.

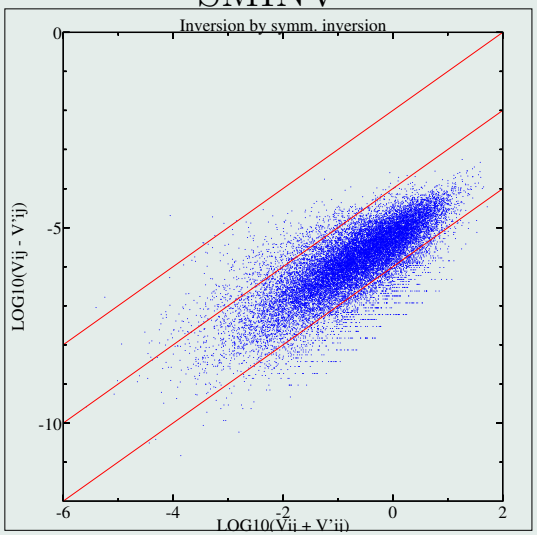
time $=$ constant $\times\ n^3$



2.6 MHz Pentium with 512 MB; single precison computation.

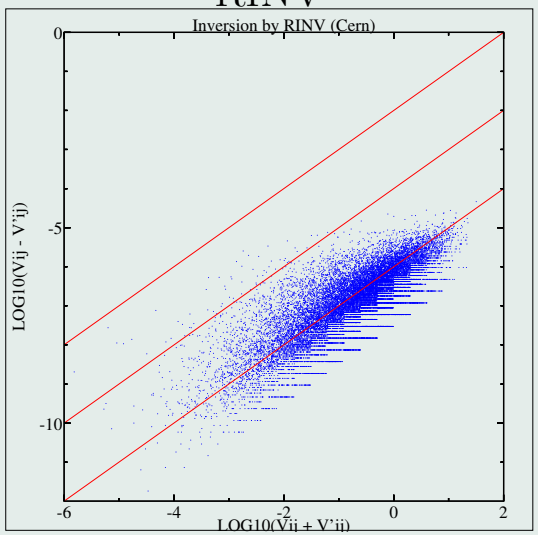
Check of accuracy based on $\boldsymbol{V}' = \left(\boldsymbol{V}^{-1}\right)^{-1}$. Plots show $\log_{10}$ of *difference* versus *sum* of elements; lines correspond to 10^{-2} 10^{-4} 10^{-6}.

SMINV



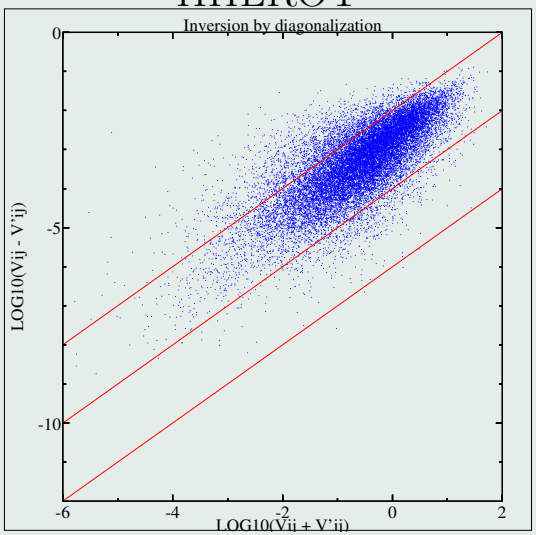
$\varepsilon \approx 10^{-5}$

RINV



$\varepsilon \approx 10^{-6}$

HHLROT

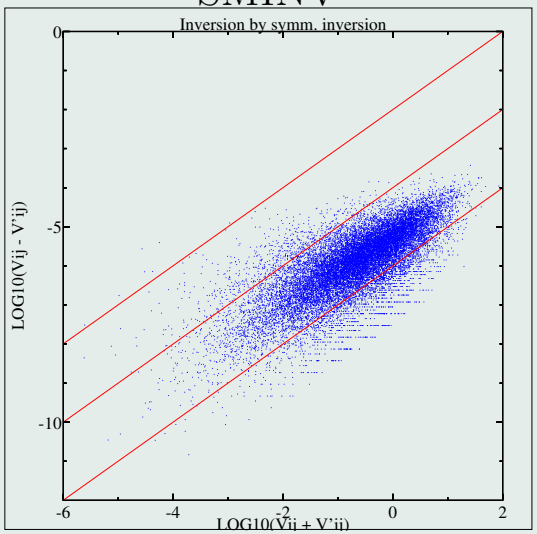


$\varepsilon \approx 10^{-3}$

Highest precision by RINV (Cern); lowest precision for inversion with diagonalisation (also more sensitive).

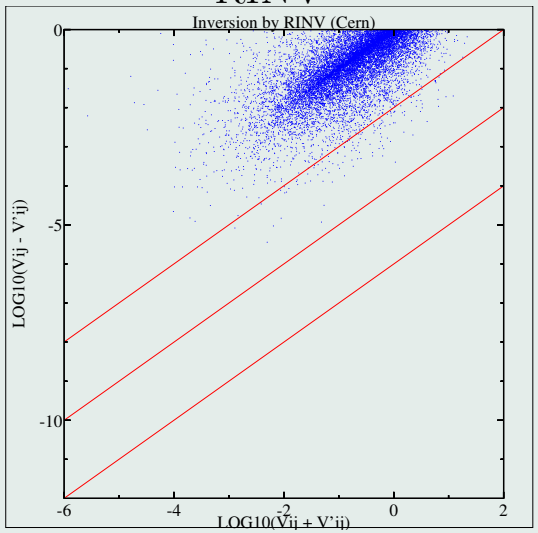
Result for matrix made singular with rank defect of 1.

SMINV



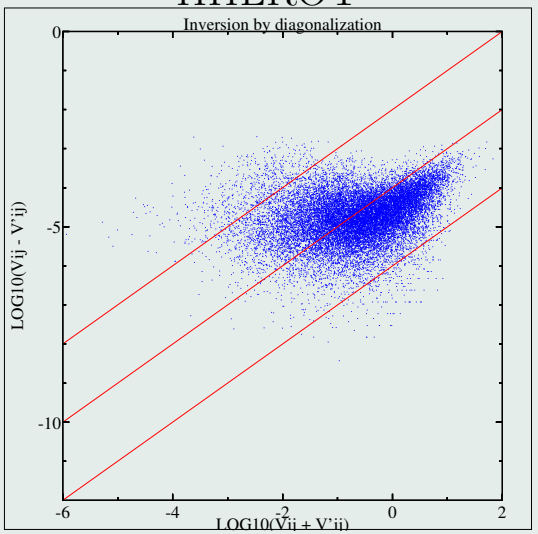
$\varepsilon \approx 10^{-5}$

RINV



IFAIL $= -1$

HHLROT



$\varepsilon \approx 10^{-3}$

RINV (Cern) fails without result; other algorithms have still useful result for 999 by 999 submatrix with unchanged precision.

Assume a Gaussian density function with 3 parameters N, μ and σ

$$f(x) = N \cdot \Delta x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

is fitted to a histogram (bin size Δx) using the Poisson maximum likelihood method. All three parameters are (almost) uncorrelated. The result for N will be the true value with an error of $\sqrt{N}$ because of the Poisson model (and error propagation).

If however the density is expressed by

$$f(x) = N \cdot \Delta x \quad \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

(e.g. `PAW` and `root`), then N is (negatively) correlated with σ and the relative error of N is enlarged due to the correlation.
After a proper full matrix error propagation $\boldsymbol{AVA}^T$ of course the previous error expression is obtained.

An example from parton density fits: the gluon parametrization is

$$xg(x, Q_0^2) = \quad \ldots \quad - A_- \, (1-x)^{\eta_-} \, x^{-\delta_-}$$

where $A_- \sim 0.2$, $\delta_- \sim 0.3$ and η_- fixed at ~ 10. A change of δ_- changes both shape *and* normalisation.

> . . . very small changes in the value of δ_- can be compensated almost exactly by a change in A_- and (to a lesser extent) in the other gluon parameters . . . [16]

> . . . we notice that a certain amount of redundancy in parameters leads to potentially disatrous departures . . . For example, in the negative term in the gluon parameterization very small changes in the value of δ_- can be compensated almost exactly by a change in A_- and in the other gluon parameters . . . [16]

> We found our input parameterization was sufficiently flexible to accomodate data, and indeed there is a certain redundancy evident. [25]

In that case the Hessian will be (almost) singular, inversion is impossible and the convergence of the fit is doubtful.

3. Data and parameter errors

Data errors: Statistical and systematic uncertainties can only be correctly taken into account in a fit, if there is a clear **model** describing all aspects of the uncertainties.

Statistical data errors: described either

- by ("uncorrelated") errors – standard deviation σ_i for data point y_i (origin is usually counts – Poisson distribution),
- by a covariance matrix $\boldsymbol{V}_y$.

Two alternative models for **systematic errors**:

- **multiplicative effects** – normalisation errors
- **additive effects** – offset errors

that had to be accounted for in *different* ways in a fit.

Data y_i in Particle Physics are often (positive) cross sections, obtained from counts and several factors (Luminosity, detector acceptance, efficiency).

In general there is a normalisation error, given by a relative error ε. If data from > 1 experiment are combined, the normalisation error ε has to be taken into account.

Method: Introduce one additional factor α, which has been measured to be $\alpha = 1 \pm \varepsilon$, modify expectation according to

$$f_i = \alpha \cdot f(x_i, \boldsymbol{a})$$

and make fit with

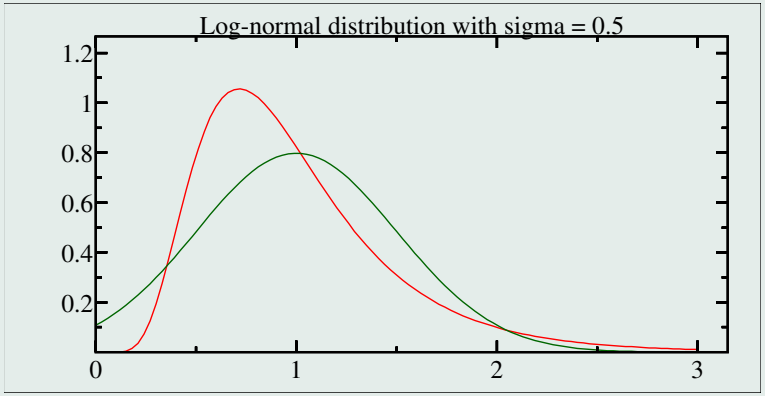
$$S(\boldsymbol{a}) = \sum_i \frac{\left(y_i - \alpha \cdot f(x_i, \boldsymbol{a})\right)^2}{\sigma_i^2} + \Delta S^{\text{norm}} \qquad \text{with} \quad \Delta S^{\text{norm}} = \frac{(\alpha - 1)^2}{\varepsilon^2}$$

One factor α_k has to be introduced for each experiment, if data from more than one experiment are fitted.

The normalisation factor determined in an experiment is more the product than the sum of random variables. According to the *multiplicative* central limit theorem the product of positive random variables follows the log-normal distribution, i.e. the logarithm of the normalisation factor follows the normal distribution.

For a log-normal distribution of a random variable α with $E[\alpha] = 1$ and standard deviation of ε the contribution to $S(\boldsymbol{a}, \alpha)$ is

$$\begin{aligned} \Delta S^{\text{norm}} &= \ln\alpha \left(3 + \frac{\ln\alpha}{\ln\left(1+\varepsilon^2\right)}\right) \\ &\to \frac{(\alpha-1)^2}{\varepsilon^2} \quad \text{for small } \varepsilon \end{aligned}$$



The normal and the log-normal distribution, both with mean 1 and standard deviation $\varepsilon = 0.5$.

Example: error of calorimeter constant – a change of the constant will change *all* data values y_i – events are moved between bins.

Determine **shifts** s_i of data values y_i, for a one-standard deviation change of the calorimeter constant – the shifts s_i will carry a relative sign.

1. Method: Modify covariance matrix to include contribution(s) due to systematic errors

$$\boldsymbol{V}_a = \boldsymbol{V}_{\text{stat}} + \boldsymbol{V}_{\text{syst}} \qquad \text{with} \quad \boldsymbol{V}_{\text{syst}} = \boldsymbol{s}\boldsymbol{s}^T \quad \text{(rank=1 matrix)}$$

e.g. $V_{ij}^{\text{stat}} = s_i s_j$, and use modified matrix in fit with $S(\boldsymbol{a}) = \boldsymbol{\Delta}^T \boldsymbol{V}_a^{-1} \boldsymbol{\Delta}$

- Requires inversion (once) of the $n \times n$ matrix of the data.
- Otherwise no change of formalism necessary.
- Used e.g. by LEP Electroweak Heavy Flavour WG.[26]

2. Method: Introduce one additional parameter β, which has been measured to be 0 ± 1, for each systematic error source, modify expectation according to

$$f_i = f(x_i, \boldsymbol{a}) + \beta \cdot s_i$$

and make fit with

$$S(\boldsymbol{a}) = \sum_i \frac{\left(y_i - \left(f(x_i, \boldsymbol{a}) + \beta s_i\right)\right)^2}{\sigma_i^2} \; + \; \beta^2$$

Advantage of additional parameter β:

- Allows to test the pull $= \hat{\beta}/\sqrt{1 - V_{\beta\beta}}$ due to the systematic error.
- Allows to test the effect of the fit model on the systematic effect from the global correlation coefficient $\rho_\beta^{\text{global}}$.
- Allows more insight into systematic effect by inspection of the correlation coefficients ρ_{β,a_j} between β and the other parameters.
- First derivative of expectation (for fits) is trivial: $\partial f_i/\partial\beta = s_i$.

The parameter(s) β can be eliminated in a modified χ^2 definition. [14]

Several systematic error sources

When using the 2. method one parameter β_ℓ has to be introduced for each systematic effect. The corresponding subvector of the gradient and submatrix of the Hessian is easily defined:

$$\text{Hessian} \qquad \boldsymbol{H} = \begin{pmatrix} \boldsymbol{H}_{aa} & \boldsymbol{H}_{a\beta} \\ \boldsymbol{H}^T_{a\beta} & \boldsymbol{H}_{\beta\beta} \end{pmatrix} \qquad \text{with} \quad (\boldsymbol{H}_{\beta\beta})_{\ell k} = \delta_{\ell k} + \sum_i \frac{s_{\ell i}\, s_{k i}}{\sigma_i^2}$$

The matrix $\boldsymbol{H}_{\beta\beta}$ is a constant and has to be inverted only once.

$$\text{Inverse:} \qquad \boldsymbol{H}^{-1} = \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{aa} & \boldsymbol{V}_{a\beta} \\ \boldsymbol{V}^T_{a\beta} & \boldsymbol{V}_{\beta\beta} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{H}_{aa} - \boldsymbol{H}_{a\beta}\boldsymbol{H}^{-1}_{\beta\beta}\boldsymbol{H}^T_{a\beta}\right)^{-1} & -\boldsymbol{V}_{aa}\boldsymbol{H}_{a\beta}\boldsymbol{H}^{-1}_{\beta\beta} \\ -\boldsymbol{H}^{-1}_{\beta\beta}\boldsymbol{H}^T_{a\beta}\boldsymbol{V}_{aa} & \boldsymbol{H}^{-1}_{\beta\beta} - \boldsymbol{H}^{-1}_{\beta\beta}\boldsymbol{H}^T_{a\beta}\boldsymbol{V}_{a\beta} \end{pmatrix}$$

$\boldsymbol{V}_{aa}$ is the covariance matrix of the parameters $\boldsymbol{a}$; the contribution from systematic errors ($\boldsymbol{H}_{\beta\beta}$) is *not* separated. $\boldsymbol{H}^{-1}_{aa} + \boldsymbol{H}^{-1}_{aa}\boldsymbol{H}^T_{a\beta}\boldsymbol{H}_{a\beta}\boldsymbol{H}^{-1}_{aa}$ is probably not correct.

See Press et al. [27]: Inversion can be done by an $n^{\log_2 7} = n^{2.807}$ algorithm.

Covariance matrix of parameters and error propagation

Parameter errors: The parameter errors are given by the inverse of the Hessian $\boldsymbol{H}_{aa}$ (the parameter submatrix)

$$\boldsymbol{V}_{aa}$$

If the approximation is good, then the function difference should follow a parabola in any direction $\boldsymbol{\Delta a}$:

$$F(\boldsymbol{a}+\boldsymbol{\Delta a}) - F(\boldsymbol{a}) = \frac{1}{2}\boldsymbol{\Delta a}^T \boldsymbol{H} \boldsymbol{\Delta a} \qquad \boldsymbol{H} = \boldsymbol{V}_{aa}^{-1}$$

This is usually an excellent approximation, also for "non-Gaussian" data errors.

Errors of a function $\boldsymbol{g}(\boldsymbol{a})$ of the fitted parameters are calculated by standard error propagation:

$$\boldsymbol{V}_g = \boldsymbol{T}\, \boldsymbol{V}_{aa}\, \boldsymbol{T}^T$$

where $\boldsymbol{T}$ is the matrix of derivatives.

Both aspects have to be checked, especially in case of poor statistic.

Single parameter a_j: Calculate, for many fixed values of a_j, the function value $S(\boldsymbol{a})$, which requires always a minimisation with $(m-1)$ parameters (MINOS feature of MINUIT).

Function $g(\boldsymbol{a})$: Calculate, for many fixed values of g, the function value $S(\boldsymbol{a})$, which requires always a function minimisation. The standard method of constraining, in a fit, the $g(\boldsymbol{a})$ to a fixed value g_{fix} is by the method of Lagrange multipliers, minimizing

$$F(\boldsymbol{a}) + \lambda \cdot (g(\boldsymbol{a}) - g_{\text{fix}})$$

w.r.t. the parameters $\boldsymbol{a}$ and the Lagrange multiplier λ. This defines an $(m-1)$-dimensional subspace.

Note that the extremum is a saddle point: F is minimal w.r.t. $\boldsymbol{a}$ and maximal w.r.t. λ, and standard minimisation programs (like MINUIT) cannot be used.

An alternative is to assume, by trial-and-error, fixed values of the Lagrange multiplier λ and to minimize

$$F(\boldsymbol{a}) + \lambda \cdot g(\boldsymbol{a})$$

and, after minization, to calculate the corresponding fixed $g(\boldsymbol{a})$ (allows to use MINUIT). [14]

Systematic errors in χ^2 expressions

There is a variety of methods:

$$\chi^2 = \sum_i \frac{(\alpha \cdot f_i - y_i)^2}{\sigma_i^2} + \frac{(\alpha - 1)^2}{\varepsilon^2} \qquad \chi^2 = \sum_i \frac{(f_i - \alpha \cdot y_i)^2}{\sigma_i^2} + \frac{(\alpha - 1)^2}{\varepsilon^2}$$

$$\chi^2 = \sum_i \frac{(f_i/(1+\beta s_i) - y_i)^2}{\sigma_i^2} + \beta^2 \qquad \chi^2 = \sum_i \frac{(f_i \cdot (1+\beta s_i) - y_i)^2}{\sigma_i^2} + \beta^2$$

. . . in my nomenclatur.

"Offset method": Systematic errors are ignored in the fit ("forces the theory prediction to be as close as possible to the data"), but later added in quadrature. [13]

The fit result must be biased, if incomplete error information is used.

An example from χ^2 minimisation

> . . . in the global χ^2, which has a value $\chi^2 = 2328$ for 2097 data points, usually signifies that the fit to one or more data sets is becoming unacceptable poor. . . . Overall, this gives 24 free parameters. . . . In fact we finish up with 15 free parameters in total . . . [16]

The χ^2-probability for a $\chi^2 = 2328$ at $2097 - 15 = 2082$ degrees of freedom is indeed very small: $P = 1.16 \times 10^{-4}$.

The standard procedure (PDG) is to assume that the data errors are too small by a factor of $\sqrt{2082/2328} = 0.946$ and to increase the parameter errors by a factor of $\sqrt{2328/2082} = 1.057$.

This would be only a small magnification of the standard errors.

Parameter errors in χ^2 minimisation

Notice that the covariance matrix

$$V_{ij}^p = \langle \Delta_i \Delta_j \rangle = \Delta\chi^2 \cdot H_{ij}^{-1}$$

depends on the choice of $\Delta\chi^2$ which usually, but not always, is taken to be $\Delta\chi^2 = 1$. This choice . . . corresponds to the definition of the width of a Gaussian distribution. [13]

In full global fit art in choosing "correct" $\Delta\chi^2$ given complication of errors. Ideally $\Delta\chi^2 = 1$, but unrealistic. [15]

. . . and $\Delta\chi^2$ is the allowed variation in χ^2. . . . and a suitable choice of $\Delta\chi^2$ . . . and $\Delta\chi^2$ is the allowed deterioration in fit quality for the error determination. [16]

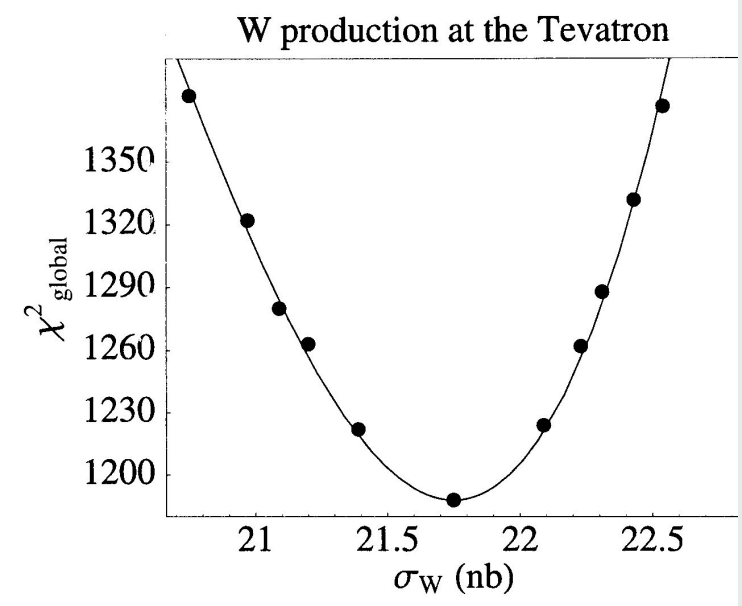
Group	$\Delta\chi^2$	Ref.	#	Value of $\alpha_S(M_Z^2)$		
H1	1	[17]	2	0.115 ± 0.0017 (exp)	$^{+0.0009}_{-0.0005}$ (model)	± 0.005 (theory)
GKK	1	[18, 19]	3	0.112 ± 0.001 (exp)		
MRST02	20	[16]	many	0.1195 ± 0.002 (exp)	± 0.003 (theory)	
ZEUS	50	[20, 21]	several	0.1166 ± 0.0040 (exp)	± 0.0081 (model)	± 0.004 (theory)
CTEQ6	100	[22]	several	0.1165 ± 0.0065 (exp)		

Also the errors obtained for functions of the parameters by error propagation are multiplied by a $\Delta\chi^2$.

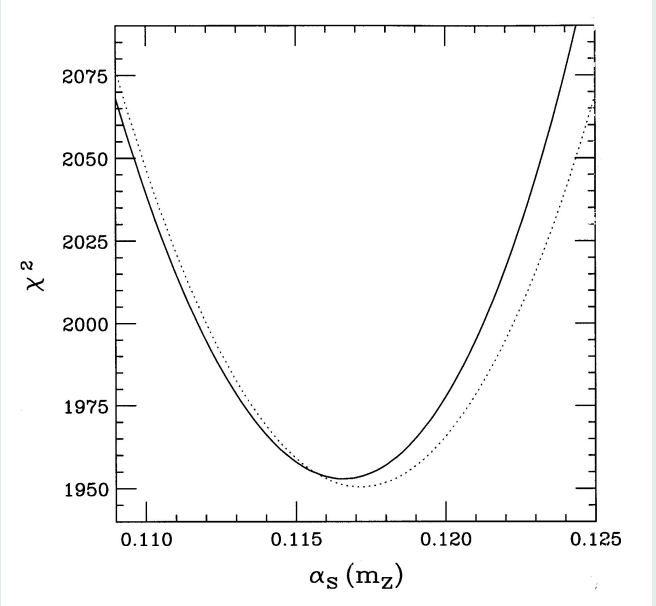
Examples with large $\Delta\chi^2$

The large, artificial and arbitrary magnification of errors is hardly acceptable – the procedure points to a deep problem in the whole data analysis. Two examples from parton distribution fits: [14, 15]

W production at the Tevatron



$\alpha_S(M_Z^2)$



Both curves are parabolas to a very good approximation over a range of $\Delta\chi^2 > 100$ ...

... while usually one would consider only a range of $\Delta\chi^2 \approx 4$, corresponding to two standard deviations.

4. Statistical properties of the data

A theorists view:

> Indeed, we have always believed the theory, rather than experiment, will provide the dominant source of error. [16]

But let us look at the statistical and systematic properties of the data.

- Are the data points (highly) correlated?

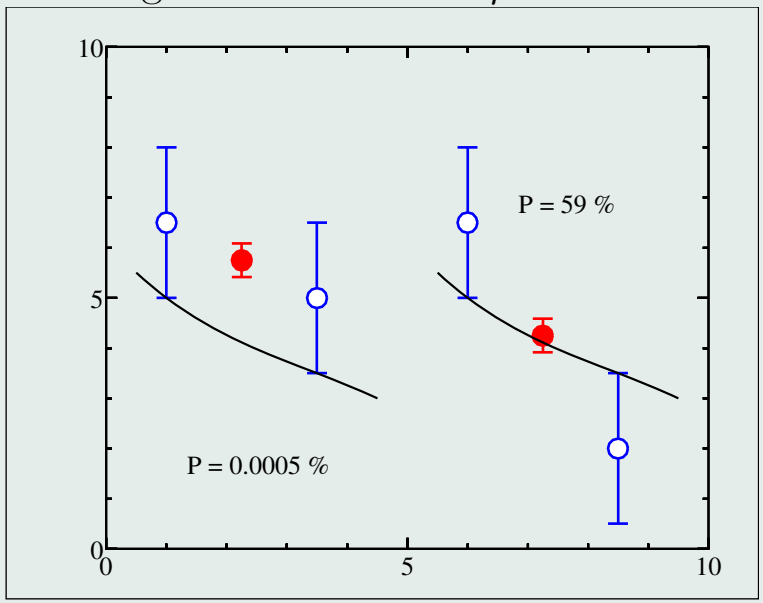
  > The main problem of the analysis is the correction for measurement errors (unsmearing corrections), which are ... [later more]

- Ist there enough information available on correlations/systematic errors, to be used in a fit?

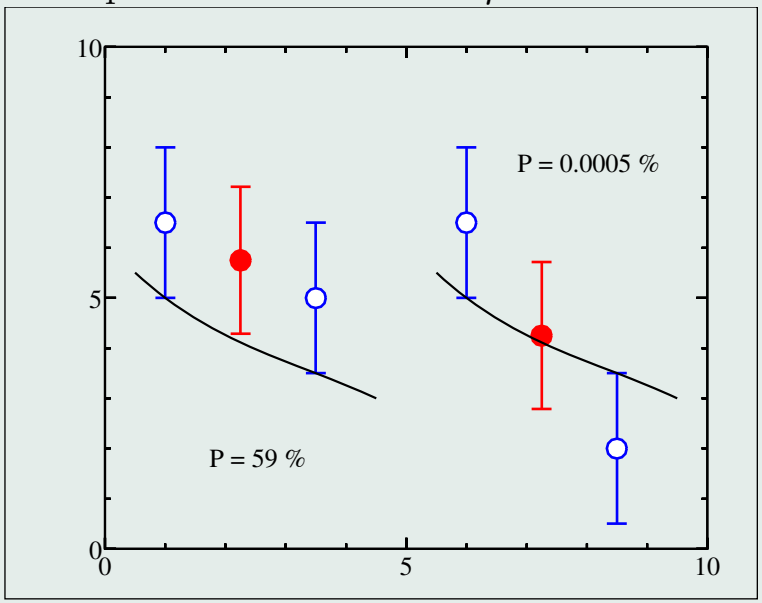
Comparing correlated data points

The two blue points with high negative/positive correlation are compared to a theoretical curve.

negative correlations $\rho = -0.9$



positive correlations $\rho = +0.9$



The χ^2-probabilities P are quite different for same sign/opposite sign deviations to the theoretical curve.

The average data point (red) of the two blue data points is very precise for negative correlations, but of almost the same precision as both single points for positive correlations.

The unfolding problem

Reconstruct the distribution $f(x)$ of the true variable x from the measured distribution $g(y)$ of the quantity y, which is related to the true variable x, based on the knowledge of the resolution (or migration) function $A(y, x)$.

$$g(y) = \int_\Omega A(y,x) f(x) \mathrm{d}x \qquad \text{or short} \qquad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

where

$g(y)$	=	measured distr.	$\boldsymbol{y}$	=	measured histogram
$f(x)$	=	ideal (true) distr.	$\boldsymbol{x}$	=	true histogram
$A(y,x)$	=	resolution fcn.	$\boldsymbol{A}$	=	resolution/migration matrix

The resolution matrix $\boldsymbol{A}$ has usually be determined from the x/y-pairs of a sample of MC events.

In matrix notation: determine vector $\boldsymbol{x}$ from the measured vector $\boldsymbol{y}$, with a known matrix $\boldsymbol{A}$ - i.e. solve a linear equation.

Examples for a migration matrix

A Matrix $\boldsymbol{A}$, depending on a single parameter ε (= migration parameter):

$$\boldsymbol{A} = \begin{pmatrix} 1-\varepsilon & \varepsilon & 0 & 0 & 0 \\ \varepsilon & 1-2\varepsilon & \varepsilon & 0 & 0 \\ 0 & \varepsilon & 1-2\varepsilon & \varepsilon & 0 \\ 0 & 0 & \varepsilon & 1-2\varepsilon & \varepsilon \\ 0 & 0 & 0 & \varepsilon & 1-\varepsilon \end{pmatrix}$$

Red elements show migration probability from second true bin into three bins of the measured distribution.

The matrix is symmetric and therefore has real eigenvalues.

A direct solution is possible with inversion of the matrix $\boldsymbol{A}$:

$$\begin{array}{ll} \text{estimate} & \widehat{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{y} \\ \text{error propagation} & \boldsymbol{V}(\widehat{\boldsymbol{x}}) = \boldsymbol{A}^{-1}\boldsymbol{V}_y\left(\boldsymbol{A}^{-1}\right)^T \end{array}$$

The method has good statistical properties – no bias:

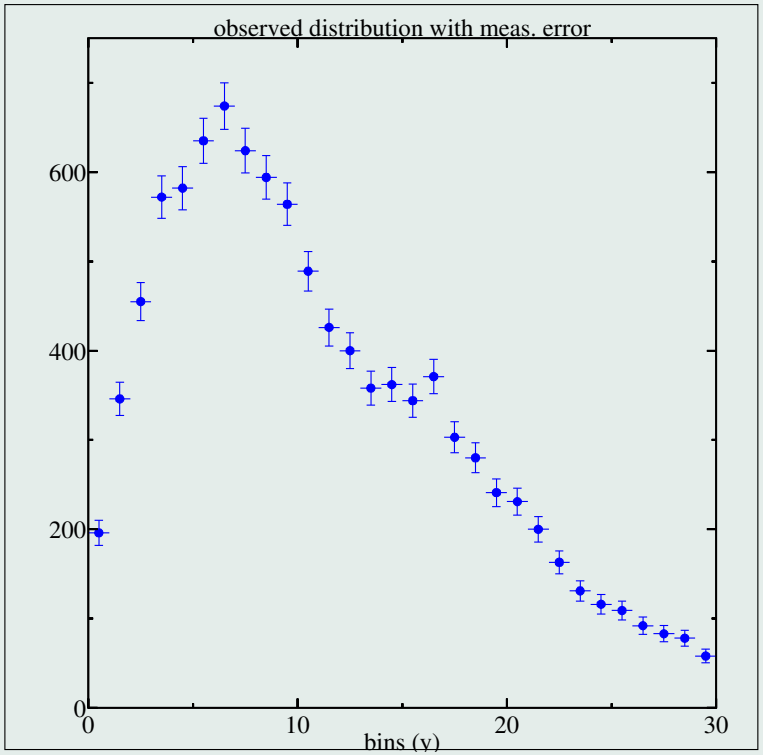
$$E[\boldsymbol{x}] = \boldsymbol{A}^{-1}E[\boldsymbol{y}] = \boldsymbol{A}^{-1}\boldsymbol{A}E[\boldsymbol{x}] = \boldsymbol{x}$$

In practice the result is satisfactory for a matrix $\boldsymbol{A}$ with dominating diagonal.

However: the results looks terrible if the matrix $\boldsymbol{A}$ describes a large migration to neighbour bins.

An example for a measured histogram

... using migration parameter $\varepsilon = 0.24$, i.e. 52 % of true events remain in the same bin, and for 10 000 events.

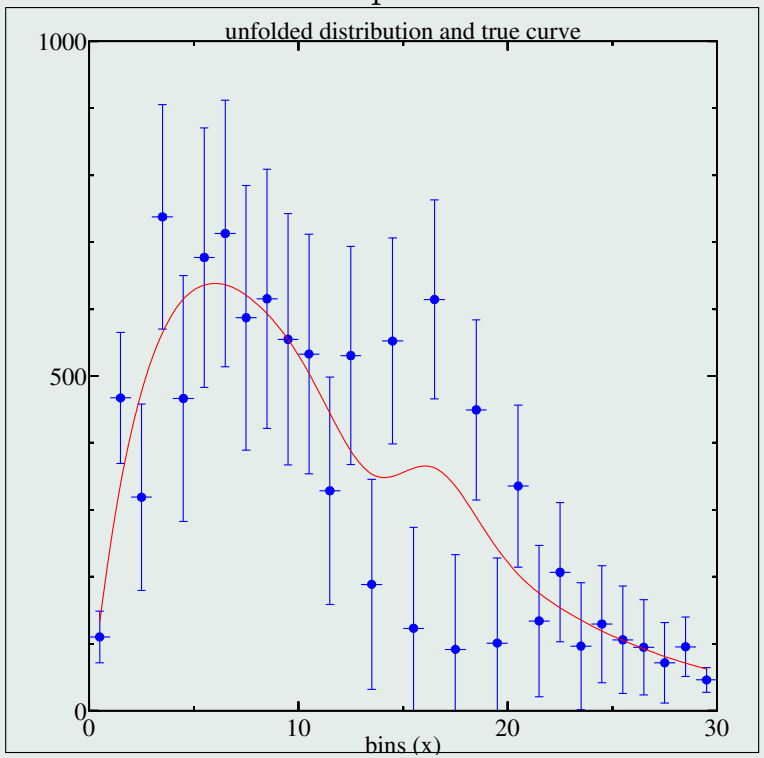


Note the small structure in the center:

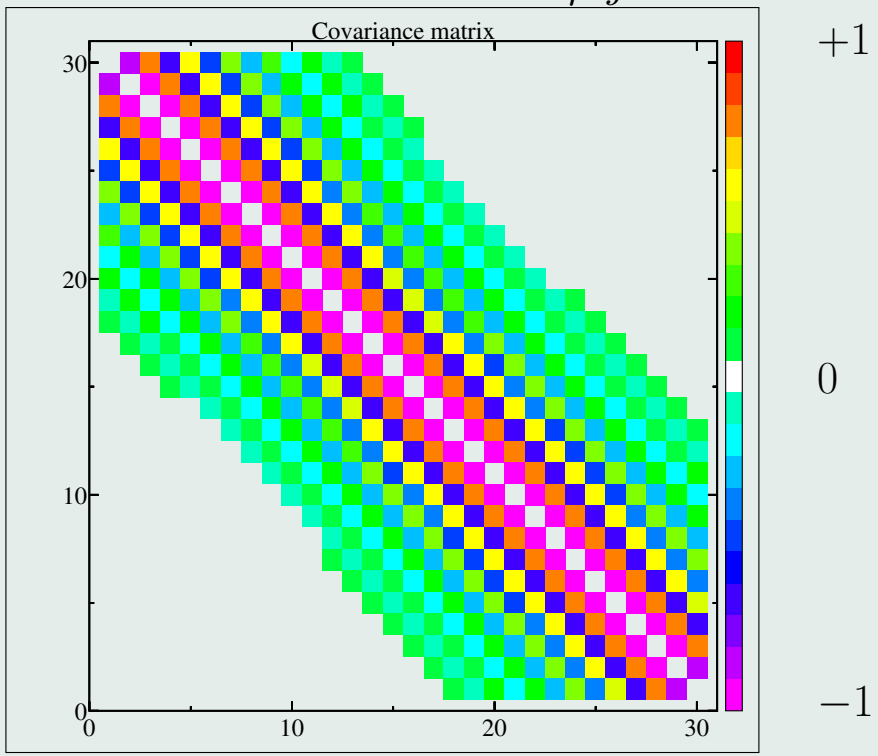
- It may be just a statistical fluctuation ($\rightarrow$ smooth after unfolding)
- If it is a real structure in the distribution, then the true peak has to be higher!

Result of solution by inversion

Reconstructed data points and true curve



Correlation coefficients ρij



Highly fluctuation data points due to large negative correlations, caused by limited resolution. Correlation coefficients ρ_{ij} with $|\rho_{ij}| > 0.05$ are shown by colour boxes: here the coefficients $\rho_{i,i+1}$ between neighbour bins are ≈ -0.95.

Decomposition of symmetric matrix $\boldsymbol{A}$: $\boldsymbol{A = U\,D\,U^T}$ with diagonal matrix $\boldsymbol{D}$ of eigenvalues λ and $\boldsymbol{U^T U = 1}$. Matrix $\boldsymbol{U}$ contains eigenvector $\boldsymbol{u}_j$ in j-th column.

Transformation to new basis:

$$\begin{array}{lrcll} \boldsymbol{U^T} \cdot \;\; | & \boldsymbol{y} & \cong & \boldsymbol{Ax = UDU^T x} & \\ & \boldsymbol{c = U^T y} & \cong & \boldsymbol{D\left(U^T x\right) = Db} & \boldsymbol{b = D^{-1} c} \end{array}$$

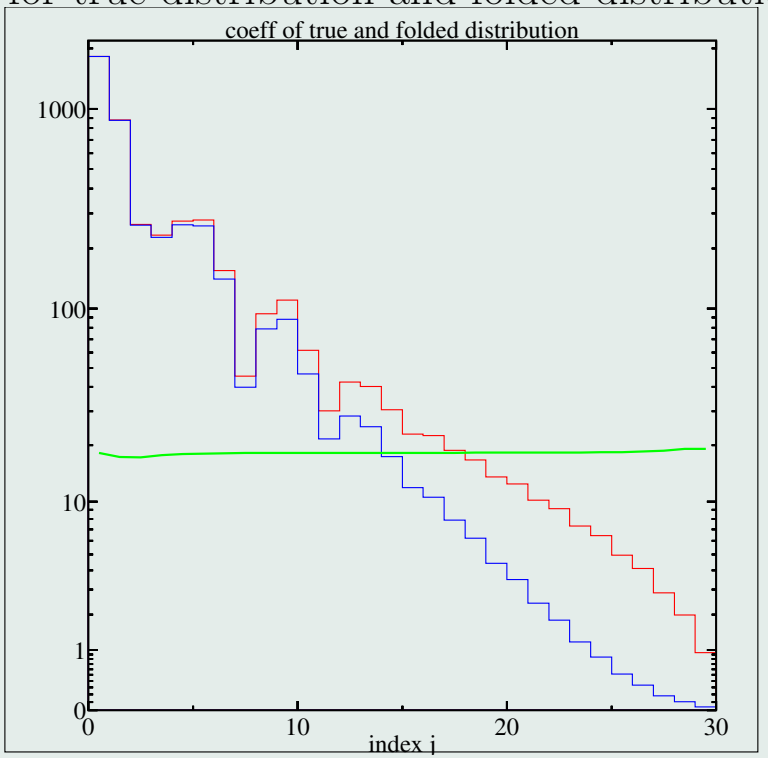
Vector $\boldsymbol{y}$ is transformed to vector $\boldsymbol{c}$ using matrix U (eigenvectors $\boldsymbol{u}_j$):

$$\boldsymbol{c \;=\; U^T y} \qquad \text{or} \qquad c_j = \boldsymbol{u}_j^T \boldsymbol{y} \qquad j = 1 \ldots n$$

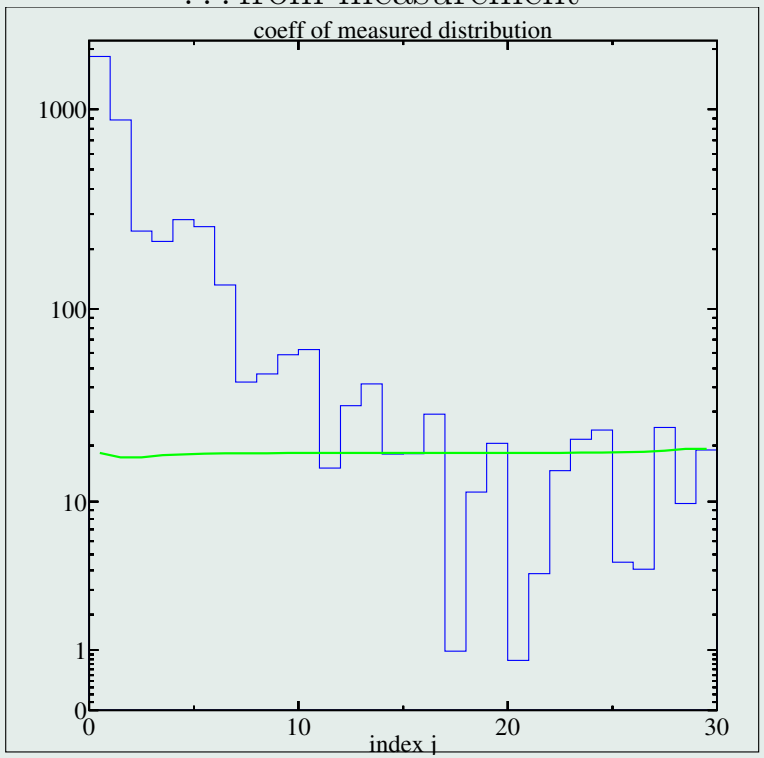
Unfolding is simply multiplication/division of coefficients by eigenvalues (and depends only an matrix $\boldsymbol{A}$):

$$\text{Unfolding} \qquad \boldsymbol{y \to x} \qquad c_j \to \frac{c_j}{\lambda_j} = b_j = \boldsymbol{u}^T \boldsymbol{x} \qquad j = 1 \ldots n$$

. . . for true distribution and folded distribution

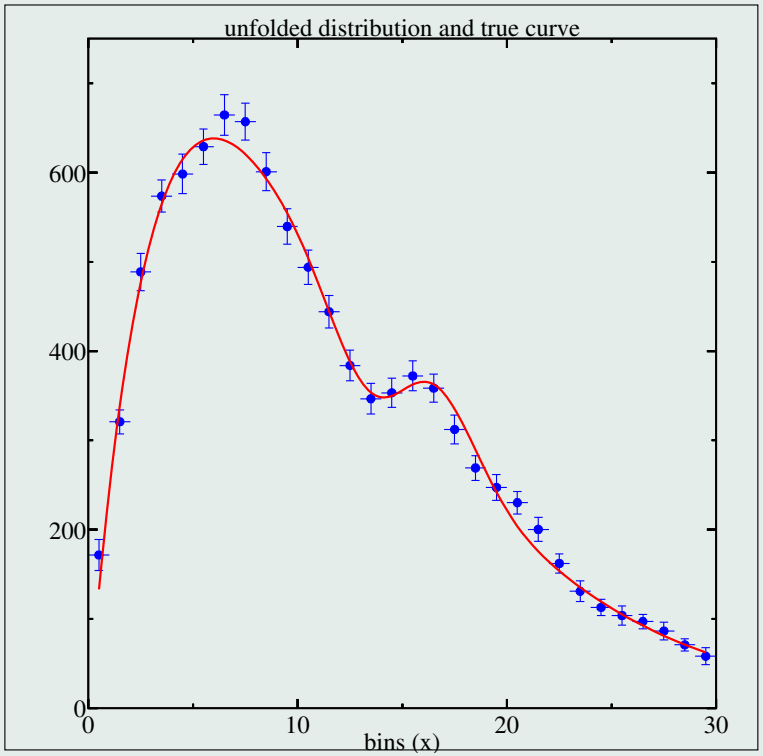


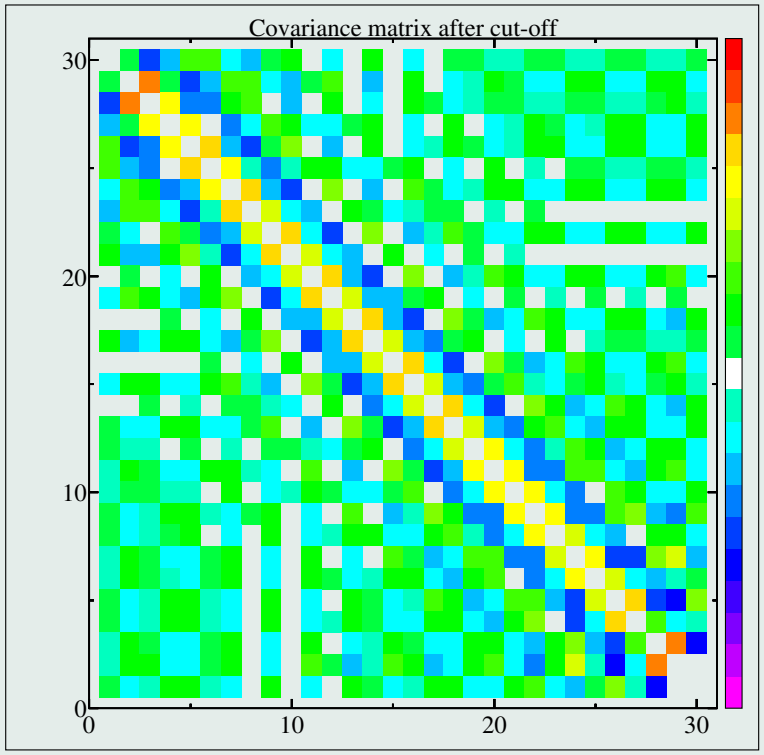
. . . from measurement



Folded amplitudes are measured and can be transformed to reconstruct the true amplitudes. Green line represents statistical errors (noise level). True and folded amplitudes below the noise level can not be reconstructed.

Take only the significant first 15 amplitudes to reconstruct the distribution with 30 data points.

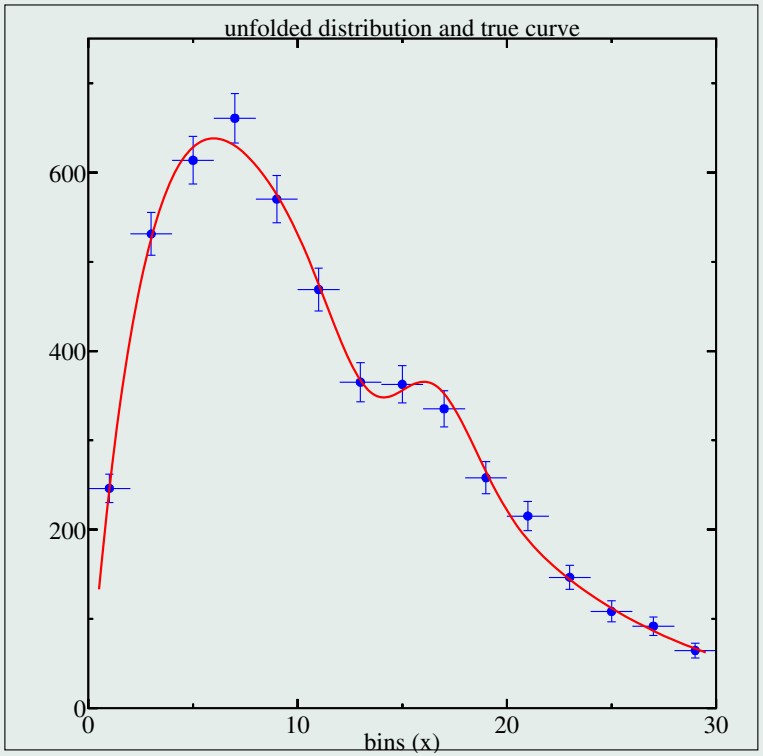


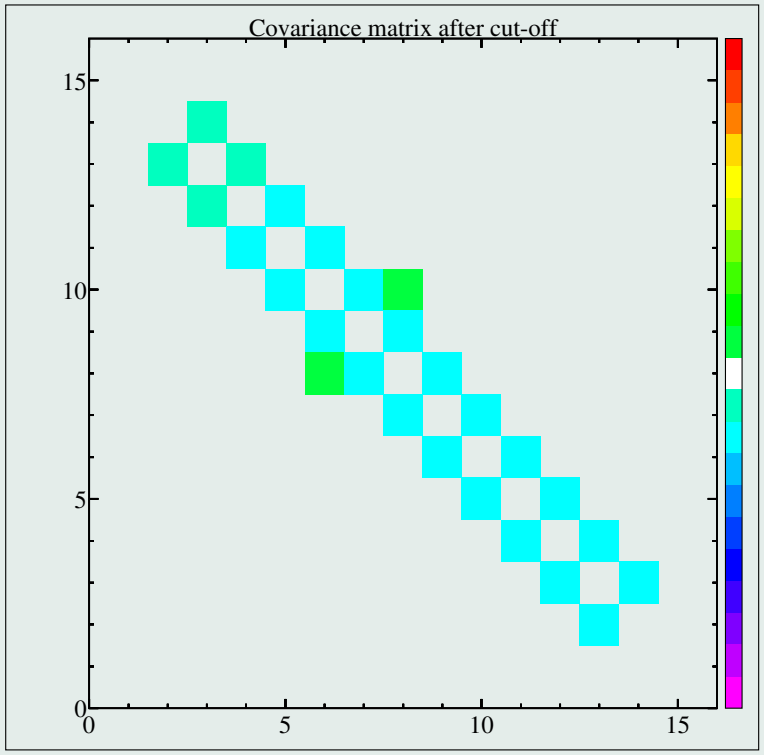


Covariance matrix has rank 15 and coefficients $\rho_{i,i+1}$ between neighbour bins are large and positive ($\approx +0.6$): "statistical" errors are smaller than original errors!

Solution with $N/2$ data points

If two bins are combined to one, the distribution has a covariance matrix with full rank.





All correlation coefficients are small, even between neighbour bins ($|\rho_{i,i+1}| < 0.2$) – but at the cost of a reduced number of points.

The standard method in particle physics to correct for the limited resolution is explained *in words* (no mathematical formula):

> ...The main problem of the analysis is the correction for measurement errors (unsmearing corrections), which are large at large x where the structure functions vary rapidly with x. We proceed by assuming a "true" structure function and calculate by Monte Carlo simulation, on the basis of the known experimental resolution functions, the result to be expected in the apparatus. **By iteration a "true" distribution which reproduces the experimental result is found.** The "unsmearing factor" is the ratio of Monte Carlo events for any particular (x, Q^2) bin in the "true" distribution divided by those in the resolution smeared distribution. If this factor differs from unity by more than 30 %, the bin is not retained. ... [23]

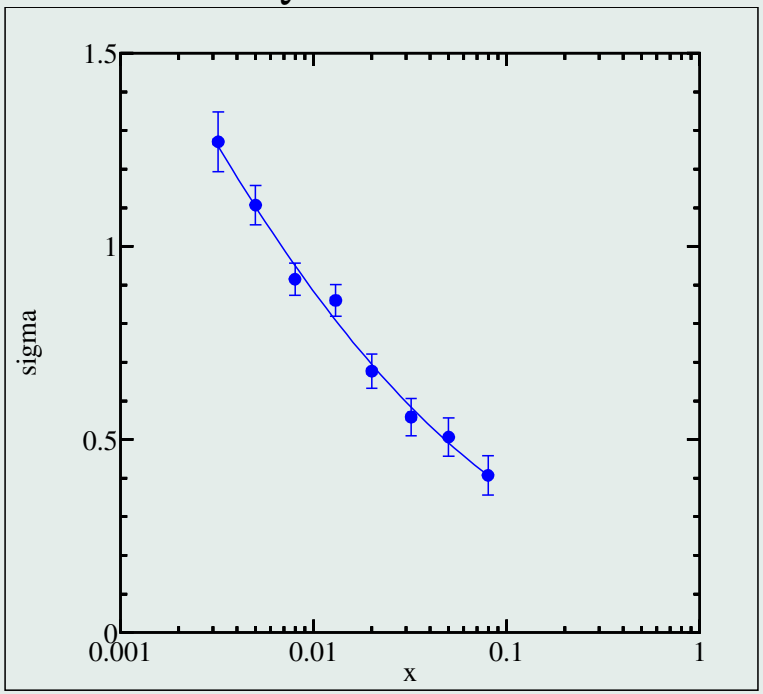
The method above is correct, if the "true" distribution is found without error. One could stop the procedure once the "true" distribution is found, but what about the measurement errors?

Any "true" distribution assumed to be very smooth may result in **positive correlations** between neighbour bins – the data have weights in fits which are too large.

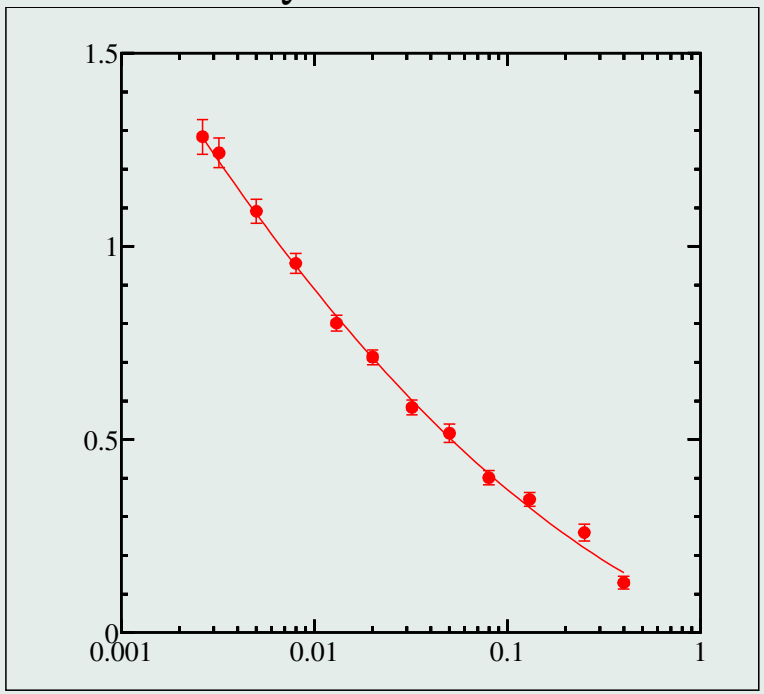
Published data on deep inelastic scattering

Shown are the total (statistical and "uncorrelated" systematic) errors. In addition there are "correlated" systematic errors and a normalisation error of 1.8 % and 1.5 %, resp. The curve is a fitted parabola, with a χ^2, that is better than expected (the data are rather smooth).

1998/1999 data (16.4 pb^{-1})
at $Q^2 = 200$ GeV^2



1999/2000 data (65.2 pb^{-1})
at $Q^2 = 200$ GeV^2

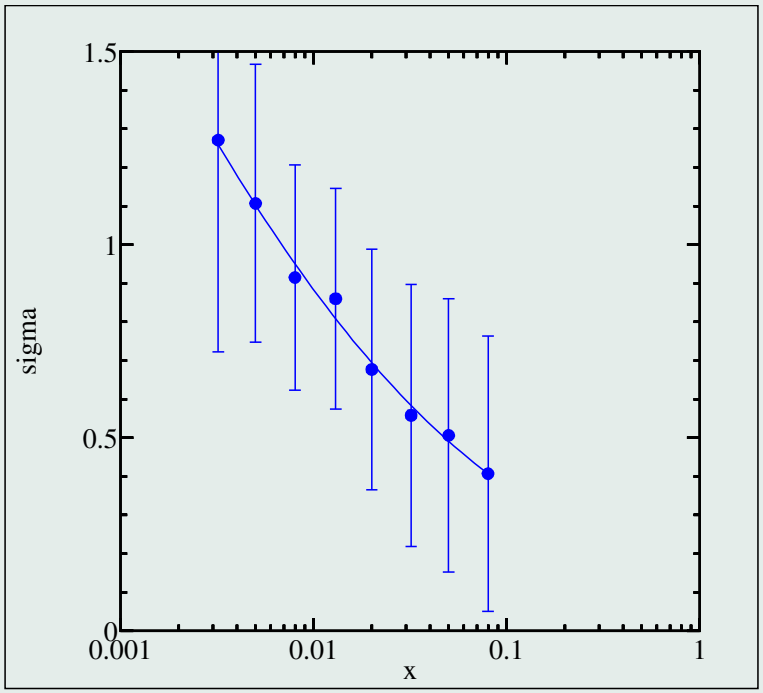


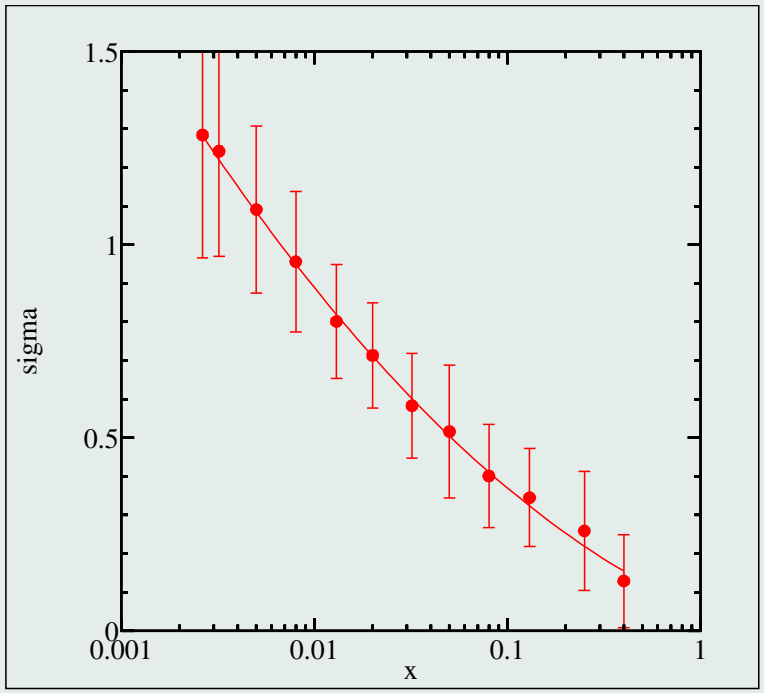
Unsmearing corrections are done based on earlier fits; bins are required to have stability and purity of $> 30\%$.

Same data

Errors multiplied by $\sqrt{50}$

Taking $\Delta\chi^2$ of 50 to calculate 1 standard deviation errors is equivalent to multiplication of all input errors by $\sqrt{50}$.





This looks strange indeed!

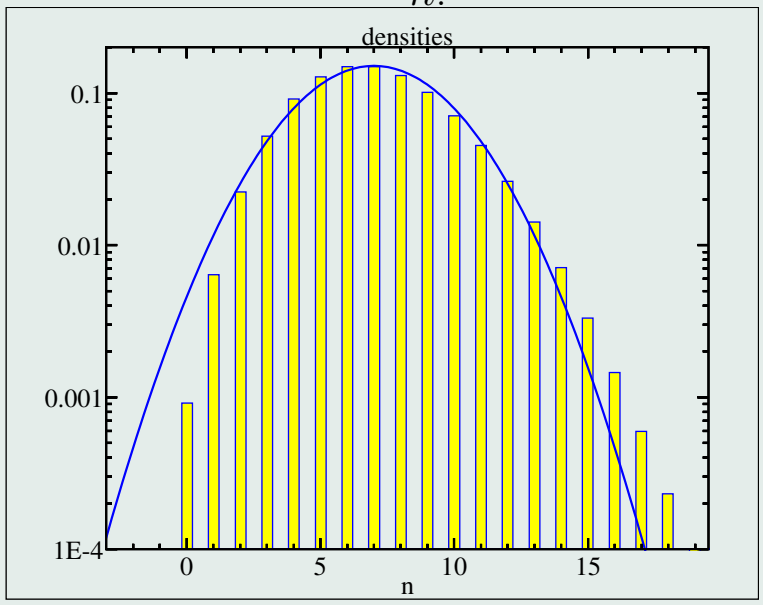
Selected topics

- Poisson approximation by Gaussian
- Histogram fits
- Poisson contribution to objective function
- Contaminated normal distribution
- Three eigenvectors in unfolding

Blue curve is Gaussian approximation with $\mu = \sigma^2 = 7$ in both figures.

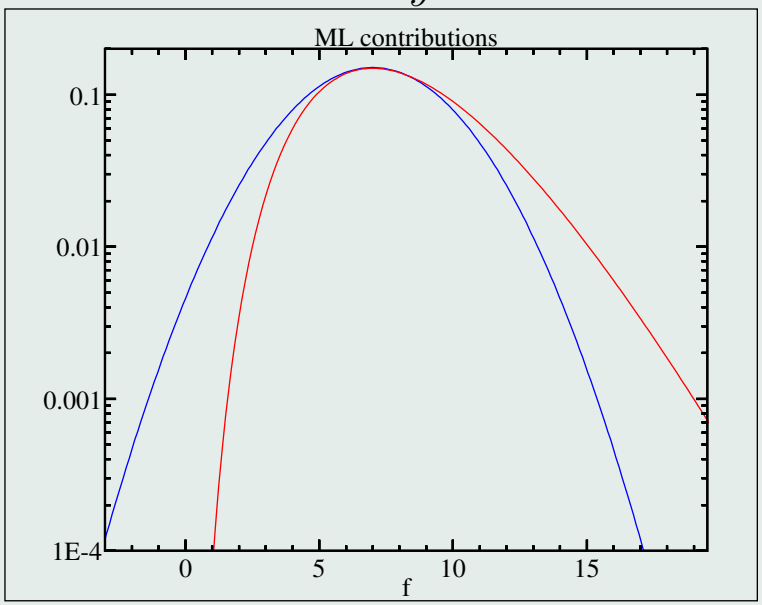
Poisson density for $\mu = 7$:

$$P = \frac{\mu^n e^{-\mu}}{n!}$$



Poisson ML contribution:

$$P = \frac{f^y e^{-f}}{y!}$$



Histogram fits

Should on use a least squares fit (χ^2 minimization) of Poisson maximum likelihood in a fit to histogram data?

> Some people put the requirement as low as $\lambda = 5$, but 10 is probably safer. [28]

> It is undesibale to have less than five events in any bin. [**?**]

Just excluding bins with no entries will introduce a bias.

Poisson contribution to objective function

$$
\begin{aligned}
F(\boldsymbol{a}) &= \sum_i f(x_i, \boldsymbol{a}) - y_i \ln f(x_i, \boldsymbol{a}) \\
\text{or better} \quad F(\boldsymbol{a}) &= \sum_i (f(x_i, \boldsymbol{a}) - y_i) + y_i \ln \frac{y_i}{f(x_i, \boldsymbol{a})} \\
\frac{\partial F}{\partial a_j} &= \sum_i y_i \frac{\frac{\partial f}{\partial a_j}}{f(x_i, \boldsymbol{a})} - \frac{\partial f}{\partial a_j} \\
\frac{\partial^2 F}{\partial a_j \partial a_k} &= \sum_i y_i \frac{\frac{\partial f}{\partial a_j}\frac{\partial f}{\partial a_k} - \frac{\partial^2 f}{\partial a_j \partial a_k} f(x_i, \boldsymbol{a})}{f^2(x_i, \boldsymbol{a})} - \sum_i \frac{\partial^2 f}{\partial a_j \partial a_k}
\end{aligned}
$$

> Everyone believes in the normal law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact. [Poincaré]

Outliers – single unusual large or small values among a sample – are dangerous and will usually introduce a bias in the result.

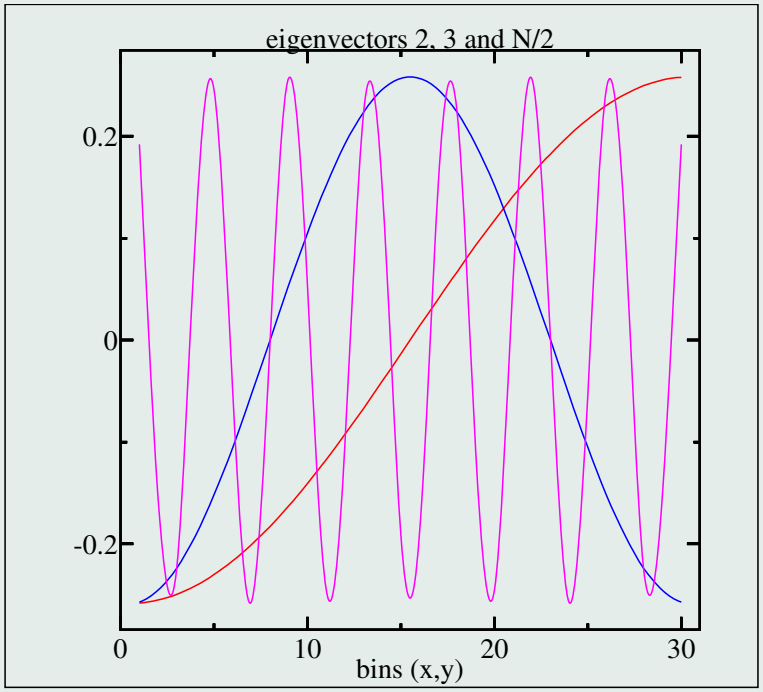
Modifications of the standard least squares procedure with

- recognition and
- special treatment of outliers

may be useful to reduce the unwanted bias in fitted parameters.

Three eigenvectors in unfolding

Condition number of matrix is $\lambda_{\max}/\lambda_{\min} = 24$. The migration parameter is $\varepsilon = 0.24$, i.e. 52 % of the entries after folding remain in the correct bin.



References

[1] T. Akesson et al. (HELIOS Collaboration). *Nuclear Instruments and Methods*, A(262):243 – 263, 1987.

[2] H. Tiecke et al. (ZEUS Calorimeter group). *Nuclear Instruments and Methods*, A(263):94 – 101, 1988.

[3] W. Braunschweig et al. (H1 Collaboration). *Nuclear Instruments and Methods*, A(275):246 – 257, 1989.

[4] M. Krammer et al. (UA1 Collaboration). *Nuclear Instruments and Methods*, A(283):630 – 634, 1989.

[5] F. Lacava et al. (UA1 Collaboration). *Nuclear Instruments and Methods*, A(289):482 – 489, 1990.

[6] R. Apsimon et al. (UA1 Collaboration). *Nuclear Instruments and Methods*, A(325):331 – 343, 1991.

[7] H. Aihara et al. (D0 Collaboration). *Nuclear Instruments and Methods*, A(325):393 – 416, 1993.

[8] B. Aubert et al. (RD3 Collaboration). *Nuclear Instruments and Methods*, A(330):405 – 415, 1993.

[9] D. Lincoln, G. Morrow, and P. Kaspar. A hidden bias in a common calorimeter calibration scheme. *Nuclear Instruments and Methods*, A(345):449 – 452, 1994.

[10] G.D'Agostini. On the use of the covariance matrix to fit correlated data. *Nuclear Instruments and Methods*, A(346):306 – 311, 1994.

[11] T. Takeuchi. The status of the determination of $\alpha(m_z)$ and $\alpha_s(m_z)$. Research Note CERN-TH/96-79, Theorie Division CERN, 1996. hep-ph/9603415.

[12] Giulio D'Agistini. *Bayesian Reasoning in Data Analysis*. World Scientific, 2003.

[13] M. Botje. Error estimates on parton density distributions. Research Note NIKHEF-01-014, NIKHEF, PO Box 41882, 1009DB Amsterdam, The Netherlands, 2001. hep-ph/0110123.

[14] D. Stump, J. Pumplin, P. Brock, D. Casey, J. Huston, J. Kalk, H.L. Lai, and W.K. Tung. Uncertainties of predictions from parton distibution functions i: the lagrange multiplier method (revised version). Research Note MSU-HEP-07102, CERN-TH/2000-359, Dept. of Physics and Astronomy, Michigan State University; Ming-Hsin Institute of Technology; Theory Division, CERN, 2001. hep-ph/0101051.

[15] Robert Thorne. Parton distribution functions. In *Lepton Photon Symposium, Fermilab*, August 2003.

[16] A.D. Martin, R.G. Roberts, W.J. Stirling, and R.S. Thorne. Uncertainties of predictions from parton distributions i: Experimental errors. Research Note IPPP/02/49, DCPT/02/98, Cavendish-HEP-2002/10, Dept. of Physics and Institute for Particle Physics Phenomenology, University of Durham; Theory Divison CERN; Cavendish Laboratory, University of Cambridge, 2002. hep-ph/0211080.

[17] C. Adloff et al. (H1 Collaboration). *Eur. Phys. J.*, 2001.

[18] W. .T. Giele and S. Keller. *Physical Review*, D 58:33?, 1998.

[19] S. Keller W. .T. Giele and D. A. Kosower. Technical report, 2001. hep-ph/01 04052.

[20] A.M. Cooper-Sakar. *J. Phys.*, G 28:2669, 2002.

[21] S. Chekanov et al. (ZEUS Collaboration). Technical report, 2002. hep-ph/0208023.

[22] J. Pumplin et al. *JHEP*, 0207:012, 2002.

[23] J. G. H. de Groot et al. Inclusive interactions of high-energy neutrinos and antineutrinos in iron. *Zeitschrift für Physik*, (C 1):143 ff., 1979.

[24] H. Rutishauser H. R.Schwarz and E. Stiefel. *Numerik symmetrischer Matrizen*. B. G. Teubner, Stuttgart, 1968.

[25] A.D. Martin, R.G. Roberts, W.J. Stirling, and R.S. Thorne. Uncertainties of predictions from parton distributions ii: Theoretical errors. Research Note IPPP/03/45, DCPT/03/90, Cavendish-HEP-2003/14, Dept. of Physics and Institute for Particle Physics Phenomenology, University of Durham; Theory Divison CERN; Cavendish Laboratory, University of Cambridge, 2003. hep-ph/0308087.

[26] L3 The LEP Experiments: ALEPH, DELPHI and OPAL. Combining heavy flavour electroweak measurements at lep. Technical report, 1996. CERN-PPE/96-017.

[27] Willaim T. Vetterling William H. Press, Saul A. Teukolsky and Brian P. Flannery. *Numerical Recipes in* FORTRAN – *The Art of Scientific Computing*. Cambridge University Press, 1986.

[28] R. Barlow and Chr. Beeston. Fitting using finite monte carlo samples. *Computer Physics Communications*, (77):219 – 228, 1993.

Comments on χ^2 minimisation