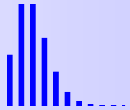


Disentanglement in Data Analysis

Michael Schmelling – MPI for Nuclear Physics

e-Mail: Michael.Schmelling@mpi-hd.mpg.de

Slides: http://www.mpi-hd.mpg.de/personalhomes/michaelt

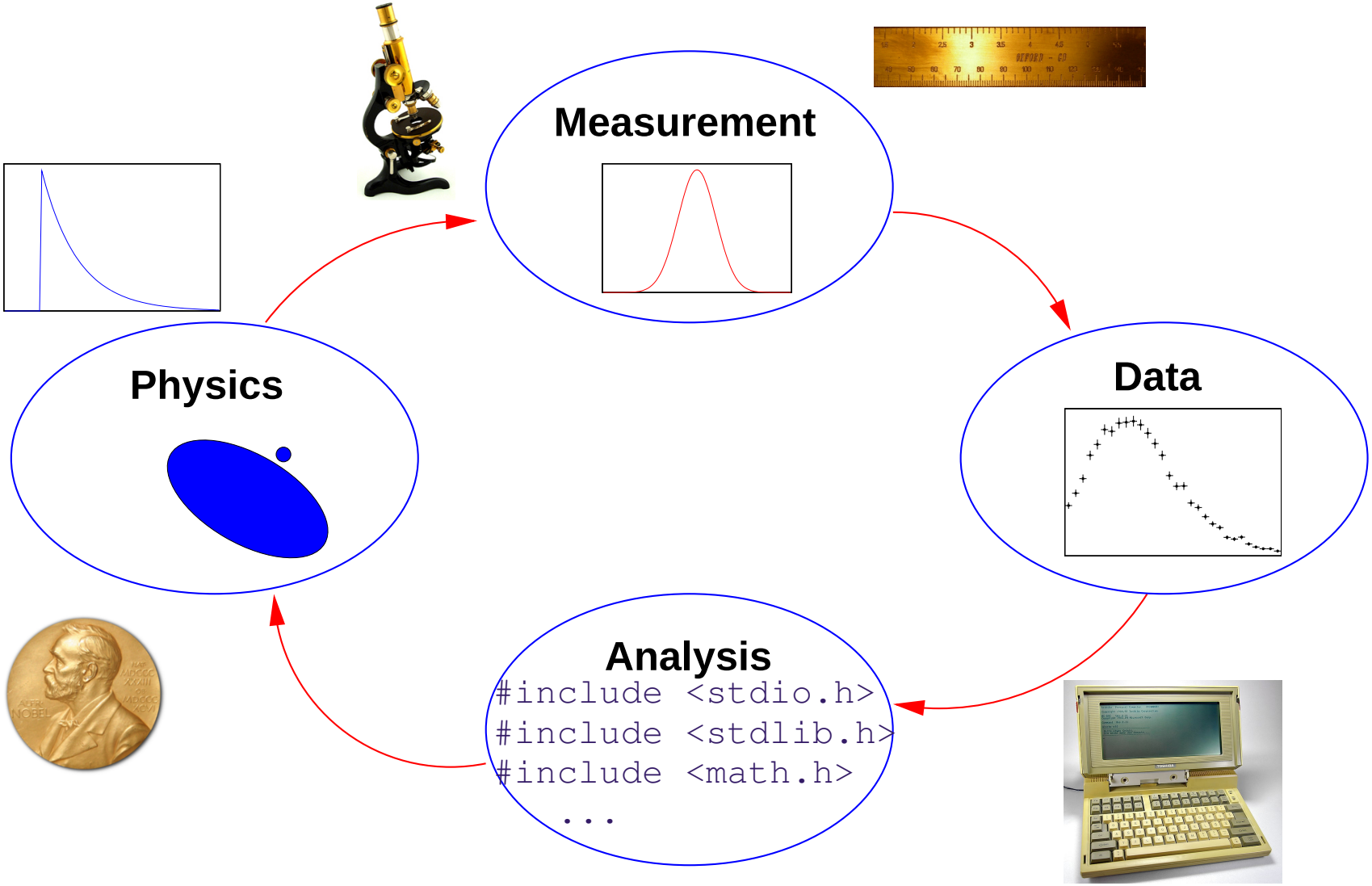
Access Password: 9885f3

Overview

- Introduction
- Basics
- Point Estimates
- Unfolding

➜ *in alphabetic order...*

- R.J. Barlow, Statistics (Wiley)
- S. Brand, Data Analysis (Springer)
- G.D. Cowan, Statistical Data Analysis (Oxford University Press)
- H.L. Harney, Bayesian Inference (Springer)
- F. James, Statistical Methods in Experimental Physics (World Scientific)
- F. James, Telling the Truth with Statistics (CERN Academic Training)
- D.E. Knuth, The Art of Computer Programming (Addison Wesley)
- W.T. Press et al., Numerical Recipes (Cambridge University Press)
- D.S. Sivia, Data Analysis - A Bayesian Tutorial (Oxford University Press)


Measurement
Data
Analysis
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
...
Physics

DISCLAIMER

➜ *do not start from scratch - assume. . .*

- intuitive understanding of probability
- familiarity with statistical fluctuations and measurement noise
- basic math
  - ➜ calculus
  - ➜ linear algebra
  - ➜ Dirac's delta-function
  - ➜ Einstein's convention to sum over repeated indices
- some knowledge about expectation value
- experience with fitting
- . . .

some of the above will be repeated in the following. . .

hopefully from a new and interesting angle. . .

❖ DEFINITION:

A function $f(x)$ is a "Probability Density Functions" (PDF) if

$$f(x) \geq 0 \quad \forall\, x \qquad \text{and} \qquad \int_{-\infty}^{+\infty} dx\, f(x) = 1\ .$$

❖ INTERPRETATION OF $f(x)$:

The probability $p(x, x+dx)$ that an event falls into the infinitesimal intervall $[x, x+dx]$ falls is given by:

$$p(x, x+dx) = f(x)dx\ .$$

❖ DISCRETE PROBABILITIES:

Discrete probabilities p_i can be written as a PDF using Dirac's delta-function:

$$f(x) = \sum_{i=1}^{n} p_i\, \delta(x-i) \qquad \text{where} \qquad \int dx\, f(x) = \sum_i p_i = 1$$

Histograms and PDFs

A histogram is a convenient tool to estimate the PDF $f(x)$ from a set of events drawn from this parent distribution. A 1-dim histogram is defined through a range $[x_{\mathrm{min}}, x_{\mathrm{max}}]$, the number of bins n_x over this range and the number of entries in each bin. With

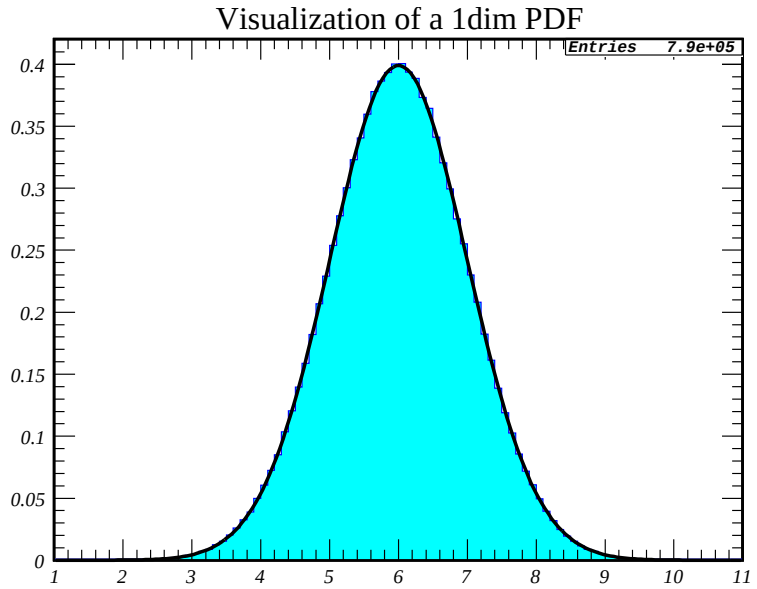
N : total number of entries in the histogram

h : bin width

n_k : number of entries in bin $[x_k - h/2, x_k + h/2]$

connection to the PDF $f(x)$:

$$
\begin{aligned}
n_k &= N \times p(x - h/2, x + h/2) \\
&= N \int_{x_k - h/2}^{x_k + h/2} dx\, f(x) \\
&\approx N f(x_k)\, h \\
\Longrightarrow \quad f(x_k) &\approx \frac{n_k}{h \cdot N}
\end{aligned}
$$



❖ General Definition:

Expectation values or "Moments" are mappings $f(x) \mapsto C$, from a PDF to a number, by means of an integral transformation of the PDF with an appropriate weight function $w(x)$:

$$\int_{-\infty}^{\infty} dx\, f(x)\, w(x) = \langle w(x) \rangle$$

The expectation value is a linear operation:

$$\langle a_1 \cdot w_1(x) + a_2 \cdot w_2(x) \rangle = a_1 \langle w_1(x) \rangle + a_2 \langle w_2(x) \rangle$$

Expectation values summarize the property of $f(x)$ by a single number. The term "expectation value" derives from the fact that for many distributions the expectation value with the special weight function $w(x) = x$ is a good estimator for the center and the most likely value of the PDF. Common weight functions include:

$w(x)$	$=$	x^n	algebraic moments, $n = 0, 1, 2, \ldots$
$w(x)$	$=$	$(x - \langle x \rangle)^n$	central moments, $n = 0, 1, 2, \ldots$
$w(x)$	$=$	e^{ikx}	Fourier transform, k any real number

Mean Value, Standard Deviation, Variance

A measure s for the scatter of x distributed according to $f(x)$ around a point a is given by

$$s^2 = \int dx\ (x-a)^2 f(x)$$

To characterize the distribution, a should be chosen such that the scatter is minimal, i.e. as that point around which the distribution is most concentrated. Miminizing s^2 yields:

$$\frac{\partial s^2}{\partial a} = -2 \int dx\ (x-a) f(x) \stackrel{!}{=} 0 \quad \text{i.e.} \quad a_{\min} = \int dx\ x\, f(x) = \langle x \rangle$$

Thus the "mean value" $\langle x \rangle$ is an estimate for the center of a PDF. For a symmetric distribution $f(a-x) = f(a+x)$ it is also the symmetry point

$$\langle x \rangle = \int dx\ x\, f(x) = \int dx\ (x-a)\, f(x) + a \int dx\ f(x) = 0 + a \times 1 = a$$

The scatter σ measured around $\langle x \rangle$ is also referred to as "standard deviation" or "rms"-scatter, its square as "variance". The following relation holds:

$$\sigma^2 = \int dx\ (x - \langle x \rangle)^2 f(x) = \int dx\ (x^2 - 2x\langle x \rangle + \langle x \rangle^2) f(x) = \langle x^2 \rangle - \langle x \rangle^2$$

❖ Definition and Interpretation:

A 2-dim PDF is a function $f(x, y) \geq 0$ with

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x, y) = 1$$

Given a 2-dim PDF $f(x, y)$, the probability to observe an event inside the (infinitesimal) rectangle $[x, x + dx] \times [y, y + dy]$ becomes

$$p(x, x + dx; y, y + dy) = f(x, y)\, dx dy$$

The histogram technique for estimating the density function discussed for 1-dim PDFs can equally be applied to higher dimensional cases.

➜ Relation between number of entries in a bin and PDF:

$$f(x, y) \approx \frac{\text{entries in bin}}{\text{total number of entries} \times \text{bin size}} = \frac{n}{N\ dx\ dy}$$

➜ generalization to higher dimensions is straightforward . . .

Correlations and Covariance

Simplest form of a 2-dim PDF $f(x, y)$: direct product of two 1-dim PDFs:

$$f(x, y) = g_1(x) \cdot g_2(y)$$

x and y are independent or "uncorrelated", i.e. the PDF of x does not depend on the value of y. In general there can be correlations, which can be detected by studying the moments of the distribution. The leading order ones are:

normalization	$\langle 1 \rangle$
1st moments	$\langle x \rangle$, $\langle y \rangle$
2nd moments	$\langle x^2 \rangle$, $\langle xy \rangle$, $\langle y^2 \rangle$
3rd moments	$\langle x^3 \rangle$, $\langle x^2 y \rangle$, $\langle xy^2 \rangle$, $\langle y^3 \rangle$ etc.

The lowerst order term sensitive to correlation is $\langle xy \rangle$. For uncorrelated variables it is:

$$\langle xy \rangle = \int dx \int dy \; x \cdot y \cdot g_1(x) \cdot g_2(y) = \left(\int dx \; x \cdot g_1(x) \right) \left(\int dy \; y \cdot g_2(y) \right) = \langle x \rangle \langle y \rangle$$

This suggests the following quantity as a measure for correlations:

$$C_{xy} = \langle xy \rangle - \langle x \rangle \langle y \rangle \qquad \text{"covariance" of } x \text{ and } y.$$

❖ GENERAL CASE OF N-VARIABLES

Set of covariances of all (ordered) pairs of variables (also pairings of a variable with itself):

$$C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

➜ *discussion:*

- the diagonal elements of C_{ij} are the variances of the individual variables
- off-diagonal elements are the covariances between all pairs of variables
- the covariance matrix is symmetric and positive definite
- it can be diagonalized by a rotation in the space of the random variables
- covariance matrix C_{ij} and expectations values $\langle x_i \rangle$ decribe (in leading order)
  - ➜ location,
  - ➜ extension and
  - ➜ orientation of a PDF
- C is also referred to as "error matrix"
  - ➜ simple and well defined concept to quantify uncertainties
  - ➜ probability content of n-σ-interval depends on shape of distribution

❖ LINEAR TRANSFORMATIONS: $y_k = A_{ki} x_i$

given: covariance matrix $C_{ij}(x)$; wanted: covariance matrix $C_{kl}(y)$

$$\begin{aligned} C_{kl}(y) &= \langle y_k y_l \rangle - \langle y_k \rangle \langle y_l \rangle \\ &= \langle (A_{ki} x_i)(A_{lj} x_j) \rangle - \langle A_{ki} x_i \rangle \langle A_{lj} x_j \rangle \\ &= A_{ki} A_{lj} (\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle) = A_{ki} A_{lj} C_{ij}(x) \end{aligned}$$

Matrix notation:

$$\vec{y} = A \cdot \vec{x} \quad \text{and} \quad C(y) = A \cdot C(x) \cdot A^T \; .$$

A not necessarily a square matrix!

❖ GENERAL (NONLINEAR) TRANSFORMATIONS: $y_k = F_k(x_1, x_2, \ldots, x_n)$

Exact treatment requires knowledge of PDF of $\vec{x}$. Linearization (leading order Taylor expansion) yields the (for non-linear F_k approximate) general solution

$$C_{kl}(y) = \frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} C_{ij}(x)$$

"Gaussian Error Propagation"

➜ general formulation

Given a n-dim PDF $f(x_1, x_2, \ldots, x_n)$, find the PDF $g(y)$, with $y = h(x_1, x_2, \ldots, x_n)$ when the $x_1, x_2, \ldots, x_n$ are distributed according to $f_i(x_1, x_2, \ldots, x_n)$.

- obvious to do numerically by Monte Carlo (especially for independent variables)
- formal solution via cumulative distribution $G(y)$ of $g(y)$

$$G(Y) = \int_{-\infty}^{Y} dy\ g(y)$$

$G(Y)$ obtained by summing all probability elements $dx_1\ dx_2 \cdots dx_n\, f(x_1, x_2, \ldots, x_n)$ which satisfy the boundary condition $h(x_1, x_2, \ldots, x_n) < Y$, i.e.

$$G(Y) = \int dx_1\ dx_2 \cdots dx_n\ f_1(x_1, x_2, \ldots, x_n)\ \Theta(Y - h(x_1, x_2, \ldots, x_n))$$

From $G(Y)$, the density $g(y)$ is obtained by differentiation with respect to the upper limit:

$$g(y) = \int dx_1\ dx_2 \cdots dx_n\ f_1(x_1, x_2, \ldots, x_n)\ \delta(y - h(x_1, x_2, \ldots, x_n))$$

general expression for the transformation of PDFs

➜ *the "textbook convolution"* $y = x_1 + x_2$

$$g(y) = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) \delta(y - x_1 - x_2) = \int dx_1\, f_1(x_1) f_2(y - x_1)$$

moments of $g(y)$:

$$\langle y^k \rangle = \int dy\; y^k\; g(y) = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2)^k$$

mean value and variance:

$$\langle y \rangle = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2) = \langle x_1 \rangle + \langle x_2 \rangle$$

$$\langle y^2 \rangle - \langle y \rangle^2 = \left[\int dx_1\, dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2)^2 \right] - \left[\langle x_1 \rangle + \langle x_2 \rangle \right]^2$$

$$= \left[\langle x_1^2 \rangle + 2 \langle x_1 \rangle \langle x_2 \rangle + \langle x_2^2 \rangle \right] - \left[\langle x_1 \rangle^2 + 2 \langle x_1 \rangle \langle x_2 \rangle + \langle x_2 \rangle^2 \right]$$

$$= \left[\langle x_1^2 \rangle - \langle x_1 \rangle^2 \right] + \left[\langle x_2^2 \rangle - \langle x_2 \rangle^2 \right]$$

Mean values and variances always add up under the convolution $y = x_1 + x_2$!

❖ ADDITIVE CONVOLUTION OF MANY RANDOM VARIABLES

For PDFs $f_i(x_i)$, $i = 1, \ldots, n$ with mean values $\mu_i = 0$ and finite variances σ_i^2, consider

$$y = \frac{1}{S} \sum_{i=1}^{n} x_i \qquad \text{with} \qquad S^2 = \sum_{i=1}^{n} \sigma_i^2 \; .$$

By construction one has $\langle y \rangle = 0$ and $\langle y^2 \rangle = 1$. In the limit $n \to \infty$ the PDF for y converges to the "normal" or "gaussian" distribution

$$g(y) = N(0,1) = \frac{1}{\sqrt{2\pi}} \; e^{-\frac{y^2}{2}}$$

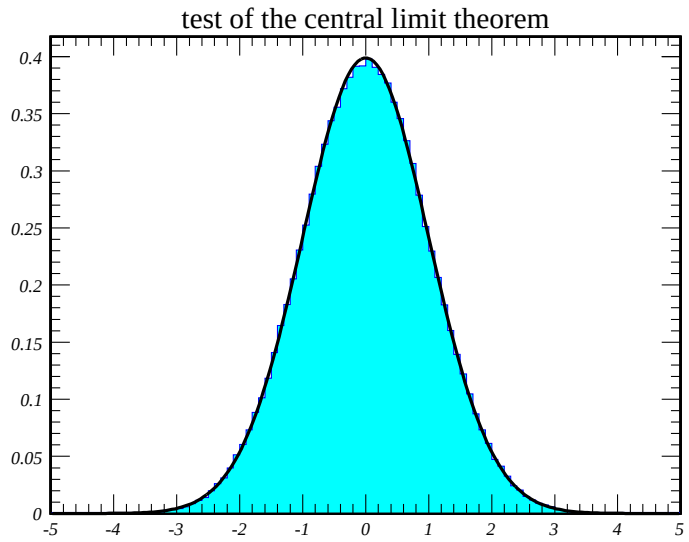
independent of the shape of the functions $f_i(x_i)$. The generalization to arbitrary mean value μ and variance σ^2 is given by:

$$g(y) = N(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

➜ *convergence towards a gaussian*

- generate n random numbers x_i from two types of parent distributions
  - ➜ uniform random numbers in $[-0.5, 0.5]$: $\sigma = 1/\sqrt{12}$ and $S^2 = n/12$
  - ➜ exponential random numbers $[-1, \infty]$: $\sigma = 1$ and $S^2 = n$
- calculate
  - ➜ $y = \sqrt{12/n} \sum_i x_i$ for the uniform random numbers
  - ➜ $y = \sqrt{1/n} \sum_i x_i$ for the exponential random numbers
- histogram y-values

A simple example how to do convolutions numerically . . .



➜ *disentangle true value(s) and measurement effects*

❖ SECONDARY SCHOOL EXAMPLE

In a grocery store, a box containing 12 apples and 18 oranges costs 15.60 EUR, another box with 20 apples and 10 oranges costs 14.00 EUR. What is the price per item?

➜ answer: solve the following system of equations

$$\begin{pmatrix} 12 & 18 \\ 20 & 10 \end{pmatrix} \cdot \begin{pmatrix} a_A \\ a_O \end{pmatrix} = \begin{pmatrix} 15.6 \\ 14.0 \end{pmatrix}$$

➜ complication: the shop assistant lost his pocket calculator

- 5 boxes need to be labelled with a price tag
- the shop assistent is not very good with numbers...
  - ➜ actual prices scatter around their nominal values
- what can be learned about the price per item
  - ➜ if the scatter of the prices is known
  - ➜ if the scatter of the prices is NOT known

. . .will come back to this problem later. . .

❖ PHYSICS EXAMPLE

The natural width σ_0 of a spectral line shall be determined. The spectrometer has a known resolution σ_s, the measured width of the line is σ.

➜ solution

$$\sigma_0 = \sqrt{\sigma^2 - \sigma_s^2}$$

➜ *discussion*

- removal of noise from measurement
- exploit that variances add when PDFs are convoluted
- knowledge of σ and σ_s only approximate
- unstable results if $\sigma \sim \sigma_s$
  - ➜ extract upper limit for σ_0

➜ *noise reduction from measurements*

Given n measurements x_i, $i = 1, \ldots n$ which all scatter with the same variance $\sigma_i^2 = \sigma^2$ around a common value μ, an estimate $\hat{\mu}$ for μ is the arithmetic average:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

expectation value:

$$\langle \hat{\mu} \rangle = \left\langle \frac{1}{n} \sum_{i=1}^{n} x_i \right\rangle = \frac{1}{n} \sum_{i=1}^{n} \langle x_i \rangle = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

variance:

$$\sigma^2(\hat{\mu}) = \sum_{i=1}^{n} \left(\frac{\partial \hat{\mu}}{\partial x_i} \right)^2 \cdot \sigma_i^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = \frac{\sigma^2}{n}$$

- the arithmetic average is unbiased
- the precision increases with $1/\sqrt{n}$
- the PDF of $\hat{\mu}$ converges towards a gaussian
  - ➜ independent of the PDF of the x_i (central limit theorem)!

➜ *optimal averaging*

Given n measurements x_i, $i = 1, \ldots n$ which scatter with the known variances σ_i^2 around a common value μ, find an optimal estimate $\hat{\mu}$ for μ.

❖ ANSATZ: LINEAR COMBINATION WITH MINIMAL VARIANCE

expectation value:

$$\langle \hat{\mu} \rangle = \left\langle \sum_{i=1}^{n} w_i x_i \right\rangle = \sum_{i=1}^{n} w_i \langle x_i \rangle = \sum_{i=1}^{n} w_i \mu = \mu \sum_{i=1}^{n} w_i \stackrel{!}{=} \mu$$

variance:

$$\sigma^2(\hat{\mu}) = \sum_{i=1}^{n} \left(\frac{\partial \hat{\mu}}{\partial x_i} \right)^2 \cdot \sigma_i^2 = \sum_{i=1}^{n} w_i^2 \sigma_i^2 \stackrel{!}{=} \min$$

solution: minimize variance subject to the constraint $\sum w_i = 1$

$$\hat{\mu} = \frac{1}{S} \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \pm \frac{1}{\sqrt{S}} \qquad \text{with} \qquad S = \sum_i \frac{1}{\sigma_i^2}$$

➜ *generalization of the previous discussion*

- averaging applies if each measurement estimates the parameter one is interested in
- In general a measurement $y = f(x; a)$ will be a function of unknown "physics" parameters a and control parameters x, e.g.:
  - ➜ $y = a_0 + a_1 x$ ("straight line fit")
  - ➜ $y = a_0 + a_1 x + a_2 x^2$ ("fit of a parabola")
  - ➜ $y = a_A x_A + a_O x_O$ ("apples and oranges")
- actual measurements y_i scatter with variances σ_i^2 around the true values
- for functions $f(x, a)$, linear in a one can construct unbiased estimators for a from linear combination of the measurements
- for known variances σ_i^2 the optimal linear combination is conveniently expressed as the minimum of a quadratic function, the χ^2-function:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i; a))^2}{\sigma_i^2} \quad \rightarrow \quad \left.\frac{\partial \chi^2}{\partial a}\right|_{a=\hat{a}} = 0$$

(reproduces all the previous results)

Properties of the Least Squares Method

- the Least Squares Method is a distribution-free way for parameter estimates
- input knowledge only variances (covariance matrix) of the data
  - ➜ variances must not be correlated with measurements!
- constructed for linear models, generalizes easily to non-linear case
  - ➜ properties of linear case approximately apply
- unbiased parameter estimates $\hat{a}$ for linear models
- estimates are linear combinations with minimum variance

$$\hat{a} = W \cdot y \qquad \text{and} \qquad C(\hat{a}) = W \cdot C(y) \cdot W^T$$

- for linear models: $W = W(x)$, i.e. only a function of control parameters
- PDF of $\hat{a}$ approximately gaussian (central limit theorem)
- $\langle \chi^2_{\min} \rangle = N_{data} - n_{par} = N_{df}$
- special case of equal size measurement errors $\sigma_i = \sigma$
  - ➜ identical parameter estimates when setting $\sigma = 1$ (unweighted fit)
  - ➜ σ can be estimated from the condition $\chi^2_{\min} = N_{df}$
  - ➜ estimate of parameter errors also in case of unknown measurement errors

➜ *uncorrelated measurements*

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_i^2} = S_{yy} + a_0^2 S_1 + a_1^2 S_{xx} - 2a_0 S_y - 2a_1 S_{xy} + 2a_0 a_1 S_x$$

with $$S_{\{1,\,x,\,xx,\,xy,\,y,\,yy\}} = \sum_{i=1}^{n} \frac{\{1,\, x_i,\, x_i^2,\, x_i y_i,\, y_i,\, y_i^2\}}{\sigma_i^2}$$

Minimization:

$$\frac{\partial \chi^2}{\partial a_0} = 2(a_0 S_1 - S_y + a_1 S_x) = 0 \quad \text{and} \quad \frac{\partial \chi^2}{\partial a_1} = 2(a_1 S_{xx} - S_{xy} + a_0 S_x) = 0$$

and thus

$$\begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} \quad \text{or more compact} \quad M \cdot \hat{a} = g$$

Results (after some algebra):

$$\hat{a} = M^{-1} \cdot g \qquad \text{and} \qquad C(\hat{a}) = M^{-1}$$

Since S_y und S_{xy} are linear in the measurements y_i also the estimates $\hat{a}$ are linear combinations of the y_i. The covariance matrix is only a function of the σ_i and x_i.

➜ *straight line fit*

- straight line $y = 1 + x$, i.e. $a_0 = a_1 = 1$
- 20 equidistant points in $0 < x < 2$
- fluctuate each measurement with $\sigma = 0.1$ around its expectation value, using a
  - ➜ gaussian
  - ➜ exponential distribution
  - ➜ distribution with two maxima
  - ➜ uniform distribution

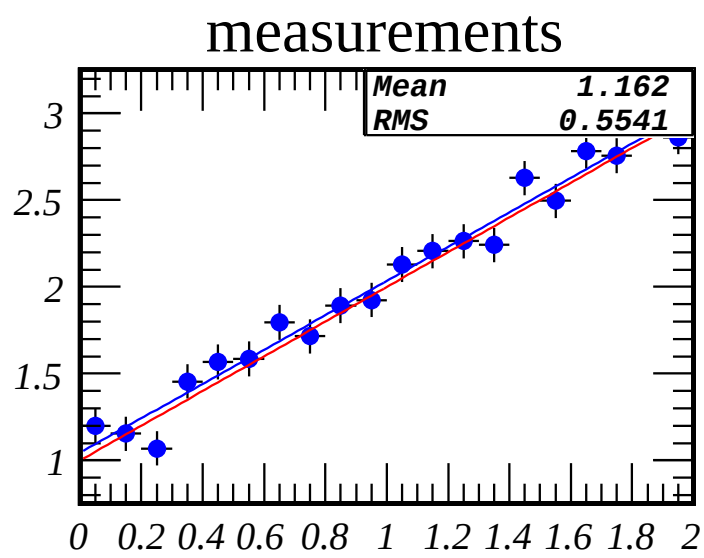
❖ NOTE:

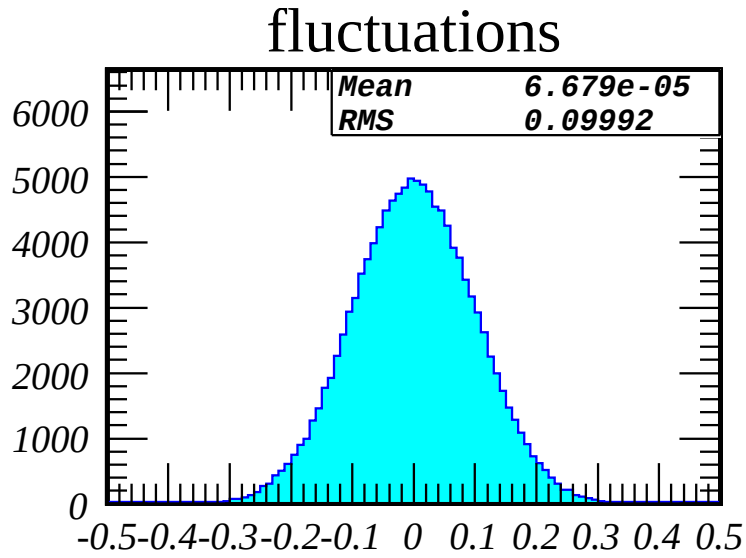
A least squares fit only uses the measurements and their errors. The PDF of the fluctuations does not enter. It follows that the covariance matrix of the fits is identical for all of the above cases.

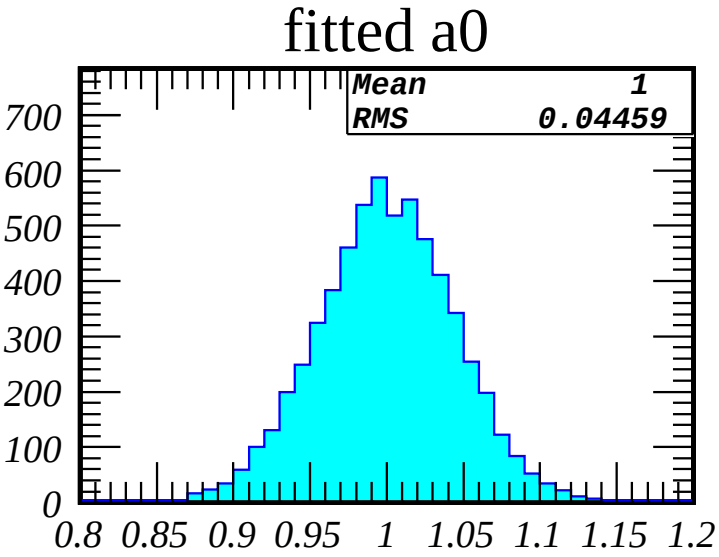
$$C(a) = \begin{pmatrix} 0.002004 & -0.001504 \\ -0.001504 & 0.001504 \end{pmatrix} \quad \begin{matrix} \sigma(a_0) = 0.044763 \\ \sigma(a_1) = 0.038778 \end{matrix} \quad \rho = -0.866296$$

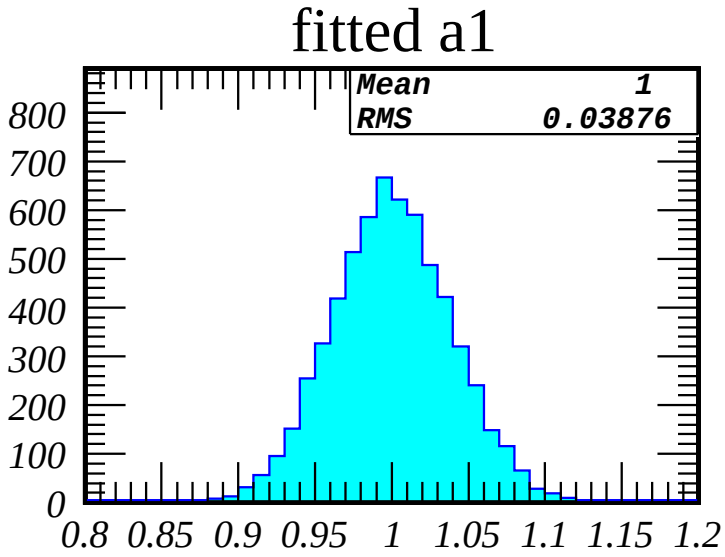
➜ verify . . .

Numerical Tests - central values

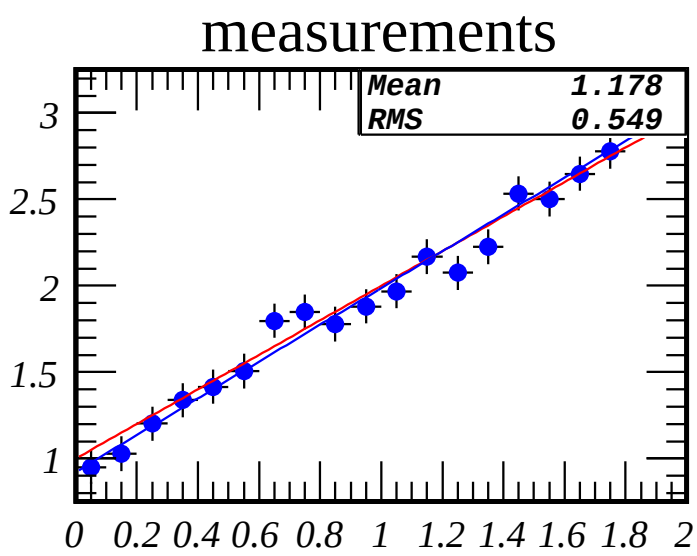


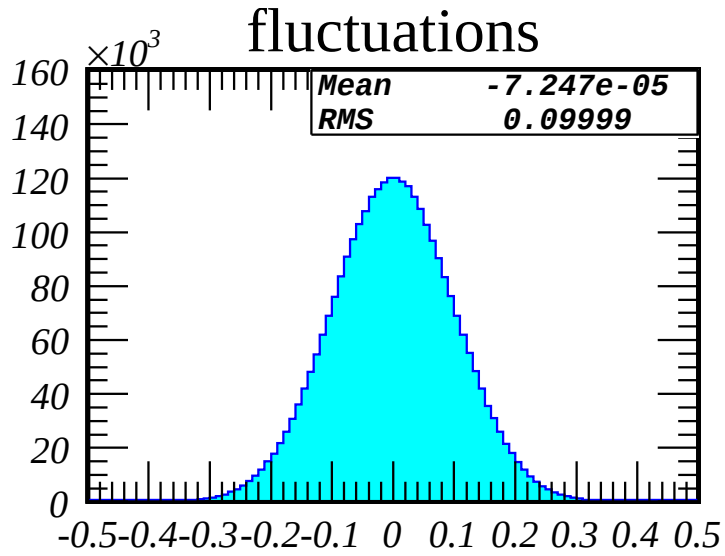


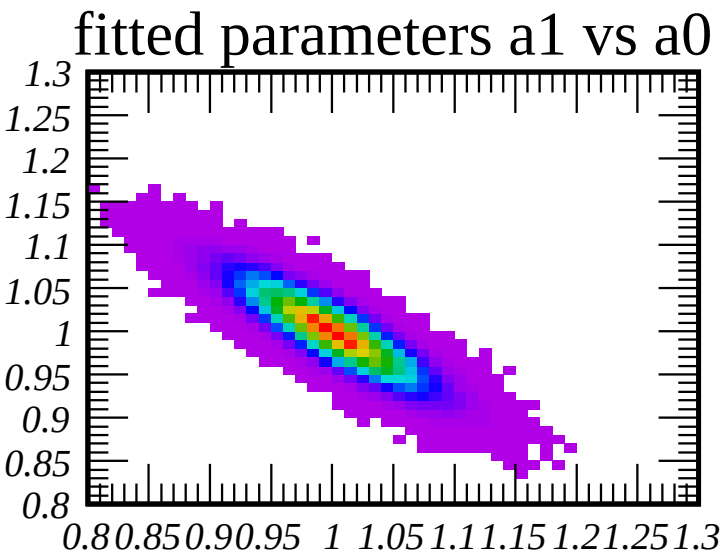


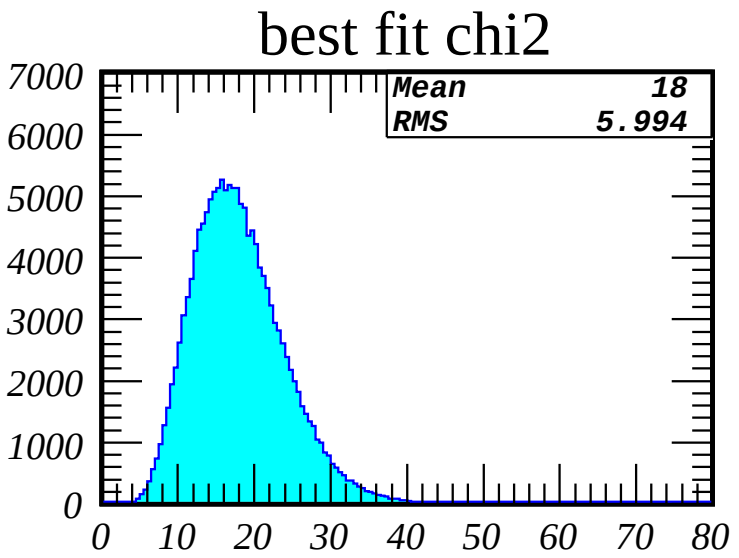


Numerical Tests - correlations and χ^2









➜ *the shop assistant's price tags*

number of apples	number of oranges	price of the box/EUR
12	18	15.6
20	10	14.0
8	22	16.0
10	20	15.4
15	15	16.0

❖ PERFORM A LEAST-SQUARES ANALYSIS AND EXTRACT. . .

- an estimate for the price per apple and the price per orange
- uncertainties for those estimates
- an estimate for the scatter of the price tags around their proper values

❖ SETTING THE STAGE

The distribution $b(y)$ of observable y is measured with an imperfect detector having inefficiencies, systematic shifts and finite resolution. It is described by a "response function" $g(x, y)$, the distribution of the measured x for every y. Alternative names for $g(x, y)$:

- response function (experimental physicists)
- point-spread function (astronomer)
- green's function (theorist)
- kernel (mathematician)

➜ relation between $b(y)$ and observable distribution $a(x)$:

$$a(x) = \int_{y_{\mathrm{min}}}^{y_{\mathrm{max}}} dy \; g(x, y) \, b(y)$$

➜ *the unfolding problem:*

❖ CONSTRUCT AN ESTIMATE FOR $b(y)$ GIVEN

- ☞ (estimate of) the response function $g(x, y)$
- ☞ a sample of n events drawn according to $a(x)$

Discretization

- In case a parametric model $b(y;\, a)$ with a small number of parameters a exists, unfolding can be done by extracting the parametes with e.g. a least squares fit.
- In practical applications the density $a(x)$ is sampled with a finite number of measurements x_i, $i = 1, \ldots, n$.
  - ➜ the available information is finite
  - ➜ a truly model-independent unfolding of $b(y)$ with continuous y is impossible
  - ➜ resort to a flexible description of $b(y)$ with a sufficiently large number of parameters. The problem has to be discretized.

➜ *expansion of PDFs into base-functions $\alpha_k(x)$ and $\beta_l(y)$*

$$a(x) = \sum_{k=1}^{n_a} a_k \; \alpha_k(x) \qquad \text{and} \qquad b(y) = \sum_{l=1}^{n_b} b_l \; \beta_l(y)$$

- for example . . .
  - ➜ harmonic functions (➜ Fourier-components)
  - ➜ orthogonal polynomial
  - ➜ histogram bins (0th order splines, orthogonal)
  - ➜ B-splines (not orthogonal)

Discretisation by Histogram Bins

- simple intuitive interpretation for coefficients a_k and b_i
- no assumptions about smoothness or curvature of distributions
- sufficiently large number of bins required for $b(y)$ to limit quantization errors

➜ base functions:

$$\alpha_k(x) = \begin{cases} 1/(x_k - x_{k-1}) & \text{if} \quad x_{k-1} \leq x < x_k \\ 0 & \text{else} \end{cases}$$

$$\beta_i(y) = \begin{cases} 1/(y_i - y_{i-1}) & \text{if} \quad y_{i-1} \leq y < y_i \\ 0 & \text{else} \end{cases}$$

➜ discretized Distributions:

$$a_k = \int_{x_{k-1}}^{x_k} dx \; a(x) \quad \text{and} \quad b_i = \int_{y_{i-1}}^{y_i} dy \; b(y)$$

➜ response matrix:

$$G_{ki} = \frac{1}{y_i - y_{i-1}} \int_{x_{k-1}}^{x_k} dx \int_{y_{i-1}}^{y_i} dy \; g(x,y)$$

➜ unfolding problem reduced to linear algebra:

$$a = G \cdot b$$

➜ *PDFs of true distributions on the intervall* $0 \leq y \leq 1$

- two Breit-Wigner peaks on a smooth background

$$b_1(y) = \frac{20.334}{100 + (10y - 2)^2} + \frac{2.0334}{1 + (10y - 4)^2} + \frac{4.0668}{4 + (20y - 15)^2}$$

- two narrow gaussian peaks

$$b_2(y) = 5.31923 \exp\left(-200(y - 0.35)^2\right) + 2.659615 \exp\left(-200(y - 0.65)^2\right)$$

- step function

$$b_3(y) = \begin{cases} 2 & \text{for} \quad 0.25 < y < 0.75 \\ 0 & \text{else} \end{cases}$$

➜ *parametrization of the response function*

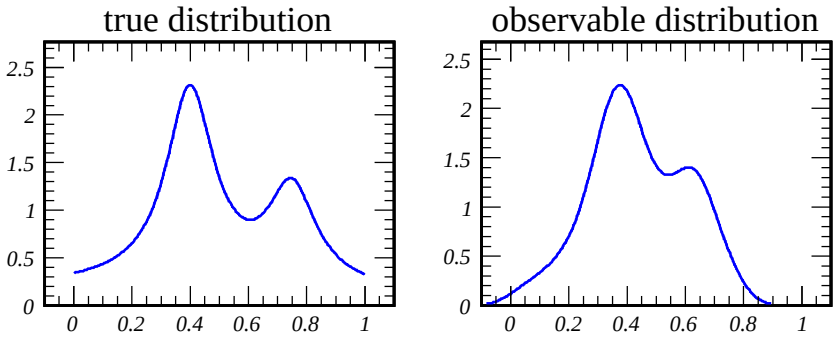
$$g(x, y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(x - \left[y - \beta y^2\right]\right)^2\right) \cdot \left(1 - 4\alpha\left(y - \frac{1}{2}\right)^2\right)$$

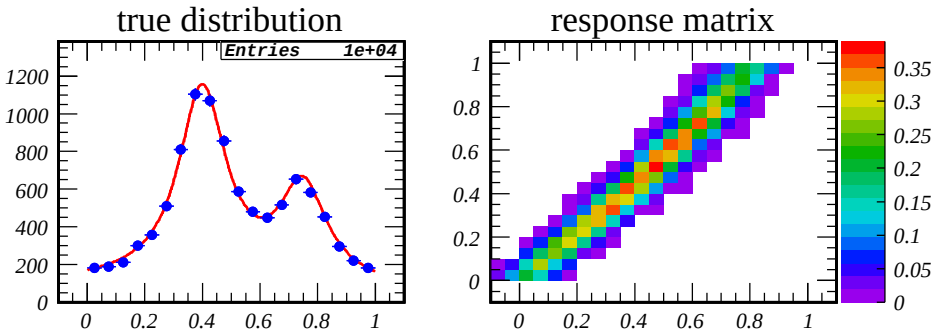
- gaussian resolution function (parameter σ)
- quadratic bias as a function of y (parameter β)
- parabolic shape of efficiency loss twoards phase space limits (parameter α)

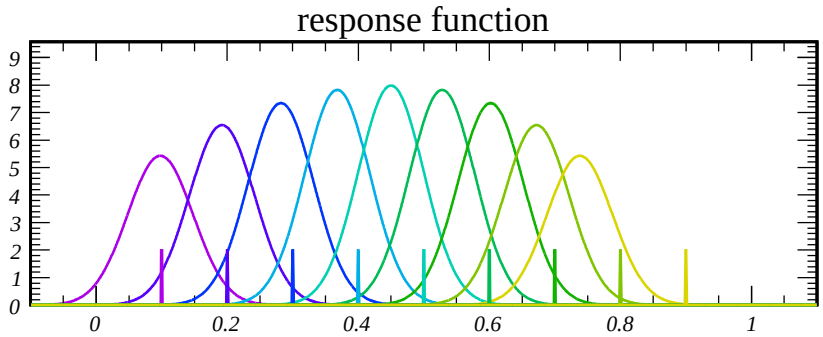
Illustration

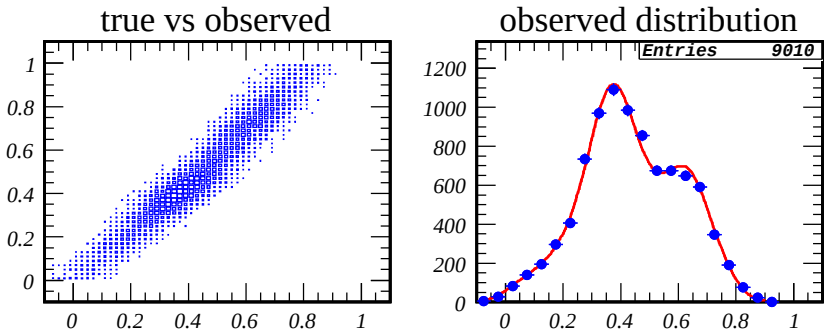
SHOWS

- properties of different models
- action of response function
- effect of finite statistics









IN THE FOLLOWING

- take response matrix to be exact, i.e. no quantization errors
- focus on the effect of finite statistics on unfolding methods

Specifics of the Toy Model Studies

- parameter settings for the response function

	$b(y)$	σ	α	β
Problem 1	$b_1(y)$	$1/20$	$1/2$	$1/10$
Problem 2	$b_2(y)$	$1/12$	0	0
Problem 3	$b_3(y)$	$1/8$	0	0

- histogram diskretisation with equidistant binning
  - ➜ restrict true and observed distribution to the range $x, y \in [0, 1]$
  - ➜ n_a bins for the observed distribution $a(x)$
  - ➜ n_b bins for the true distribution $b(y)$
  - ➜ statistical precision of N measurements, relative errors proportional to $1/\sqrt{N}$
- relation between observable and true distribution

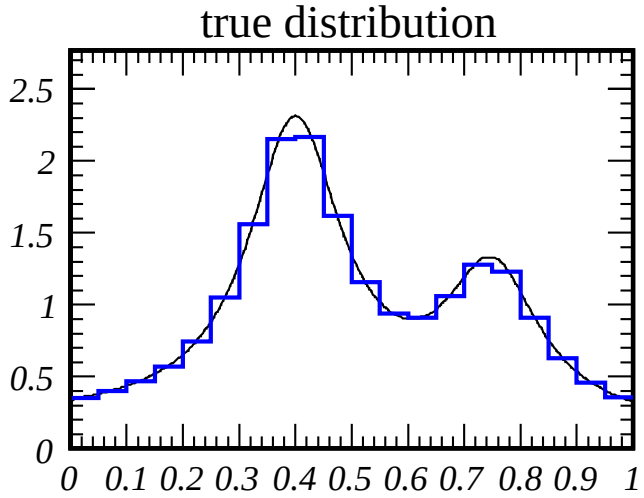
$$\langle a \rangle = G \cdot b$$

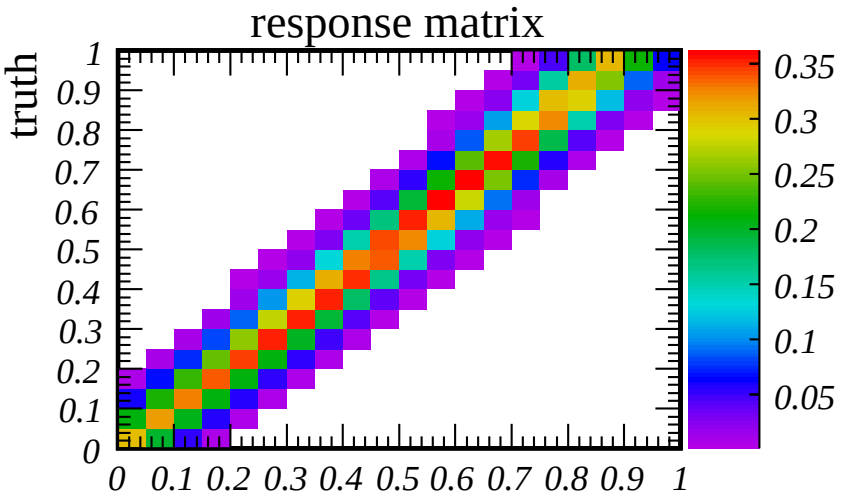
- actual measurements fluctuate around expectation values

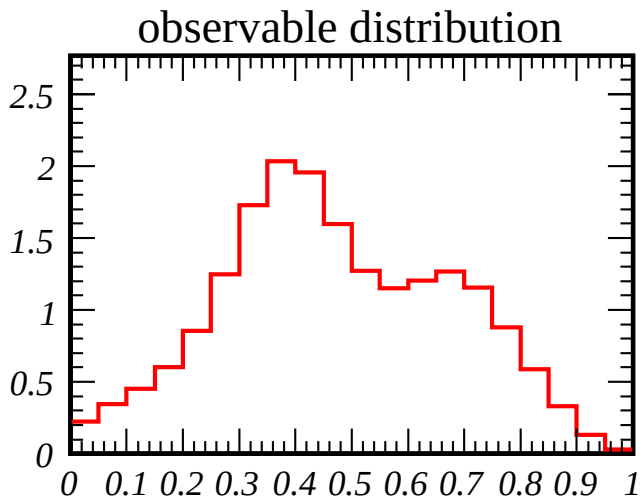
$$a = \langle a \rangle + r$$

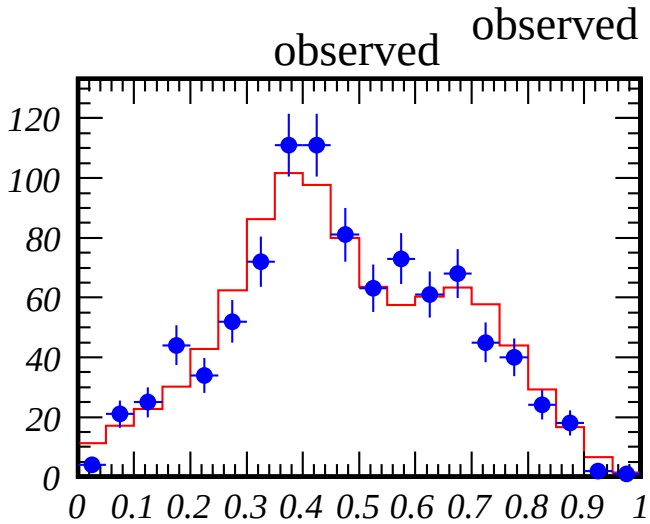
  - ➜ with statistics fluctuation r around zero, i.e. $\langle r \rangle = 0$
  - ➜ relative size of fluctuations according to assumed statistics

Illustration









➜ performance of different unfolding methods. . .

➜ *simplest and most widely used method*

- same binning for observed and true distribution
- bin-dependent correction factors c_k

$$b_k = a_k \cdot c_k$$

- determination of the correction factors
  - ➜ start with assumption for b_k
  - ➜ determina a_k by folding (multiplication) with response matrix
  - ➜ calculate $c_k = b_k / a_k$

$$c_k = \frac{b_k}{\sum_{l=1}^{n_b} G_{kl}\, b_l}$$

- Correction factors depend on the assumed distribution b_k. Possible choices:
  - ➜ (approximate/expecetd) true distribution (unknown)
  - ➜ uniform distribution (“objectiv”)
  - ➜ measured distribution (hopefully similar to truth . . .)
- correct result is garanteed only for $b_k = b_k^{true}$
- in general a partial correction should be achievable

Iterated Correction Factors

❖ GET INDEPENDENT OF SPECIFIC ASSUMPTION FOR b_k

➜ choose initial setting: (for example)

$$b_k^{(0)} = a_k$$

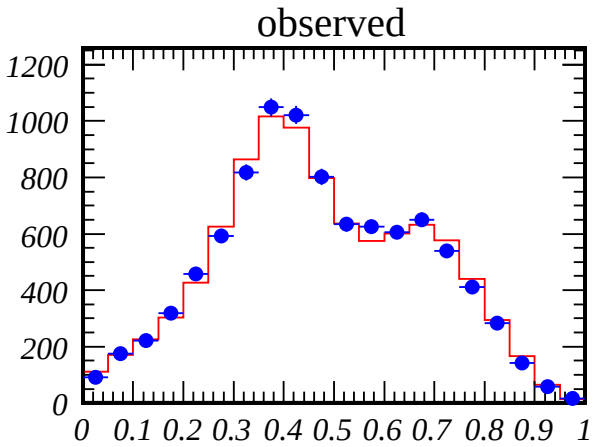
➜ iteration:

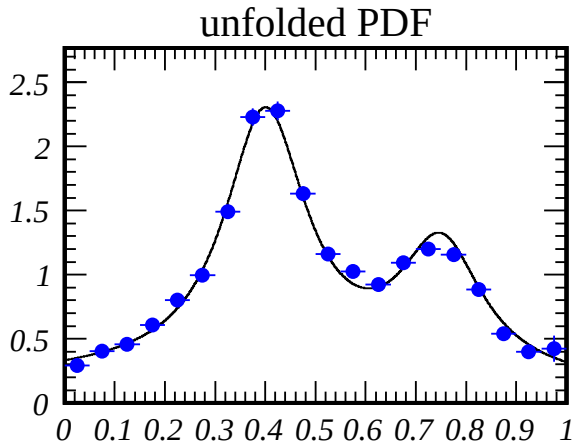
$$b_k^{(n+1)} = a_k \cdot c_k^{(n+1)} = a_k \cdot \frac{b_k^{(n)}}{\sum_{l=1}^{n_b} G_{kl}\, b_l^{(n)}}$$

error estimate for the unfolded distribution:

$$\sigma(b_k) = c_k \cdot \sigma(a_k) \qquad \text{correct for } C(\vec{c}) = 0$$

➜ *application to test-problems:*





Discussion

- correction factors work well if the true distribution is known
  - ➜ no iteration required
  - ➜ stable result
  - ➜ no new information from measurement
- conceptual problems
  - ➜ empty bins are corrected to zero
  - ➜ data from outside physics phase space are ignored
- first iteration step depends on assumed true distribution
- iteration removes dependence on unknown distribution, but . . .
  - ➜ results are unstable
  - ➜ naive error propagation evidently wrong
  - ➜ analytic error calculation not feasible: the iterated result is a highly non-linear function of the measurements

➜ do it properly. . .

Improved Correction Factors

- fluctuate measurements $\vec{a}$ according to their error
  - ➜ generate N pseudo-samples $\vec{a}_n$ with $n = 1, \ldots, N$
- for each pseudo-Sample $\vec{a}_n$ determine $\vec{b}_n$ using M-times iterated correction factors
- take average unfolded distribution as nominal result

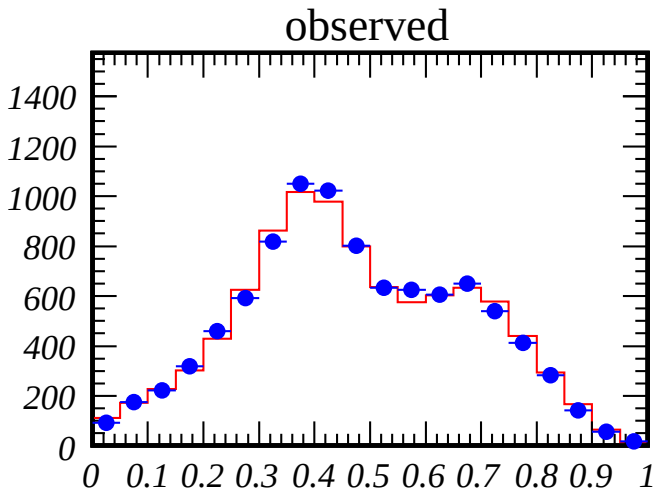
$$\vec{b} = \frac{1}{N} \sum_{n=1}^{N} \vec{b}_n$$

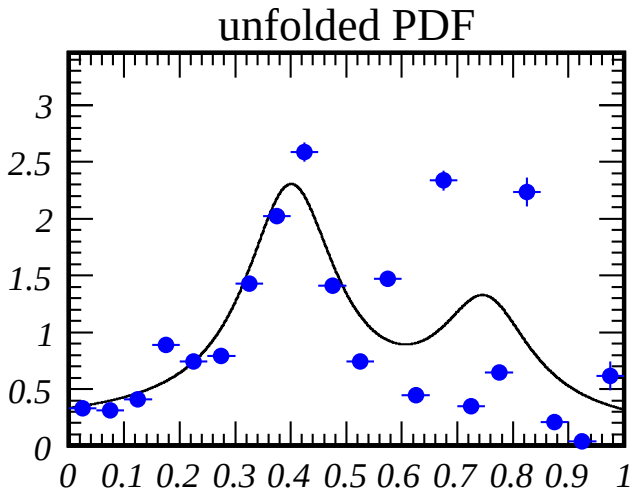
- estimate errors by the empirical covariance matrix of the results

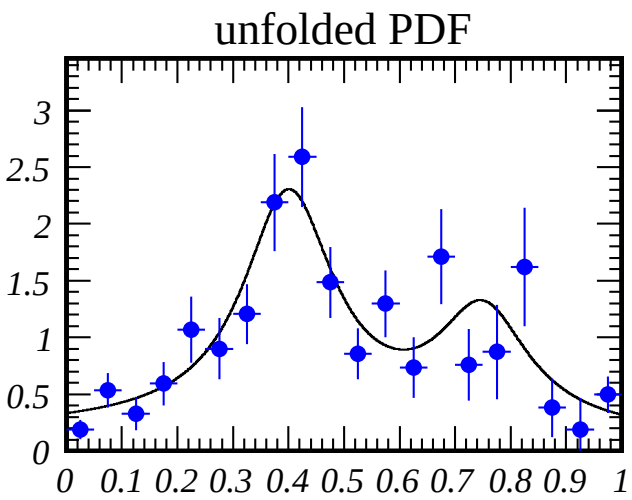
$$C(b) = \left(\frac{1}{N} \sum_{n=1}^{N} \vec{b}_n \cdot \vec{b}_n^T \right) - \vec{b} \cdot \vec{b}^T$$

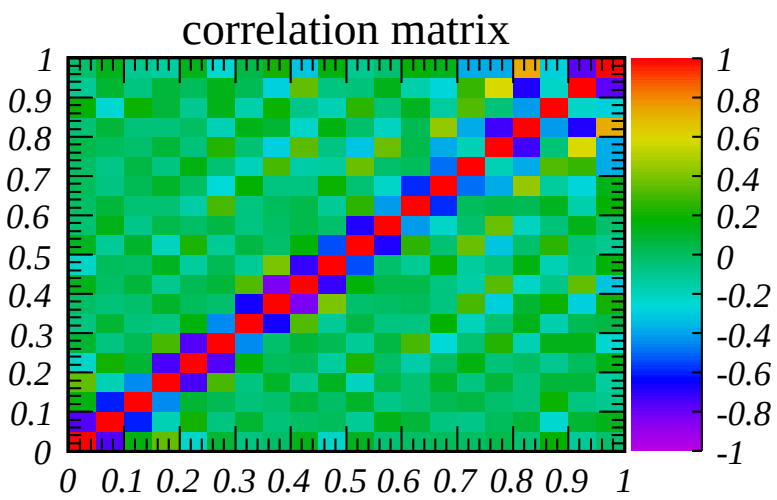
  - ➜ correlations between bins of the unfolded distribution handled properly
- numerical studies show
  - ➜ surprisingly large error in the unfolded distribution
  - ➜ strong correlations between neighboring bins
  - ➜ errors grow with the number M of iterations

Numerical Test









consider alternative methods. . .

➜ *Unfolding based on conditional probabilities*

introduce discrete probabilities p_i for the true distribution:

$$b_i = B \cdot p_i \quad \text{and} \quad a_i = \sum_{k=1}^{n_b} G_{ik}\, b_k = B \sum_{k=1}^{n_b} G_{ik}\, p_k$$

Interpretation of the response matrix G_{ik} as conditional probabilities

$$G_{ik} = p(\text{measurement } i | \text{true value } k)$$

exploit Bayes' theorem to construct an unfolding matrix H_{ik}:

$$\begin{aligned} H_{ik} &= p(\text{true value } k | \text{measurement } i) \\ &= \frac{p(\text{measurement } i | \text{true value } k) \cdot p(\text{true value } k)}{p(\text{measurement } i)} \\ &= \frac{p(\text{measurement } i | \text{true value } k) \cdot p(\text{true value } k)}{\sum_j p(\text{measurement } i | \text{true value } j) \cdot p(\text{true value } j)} = \frac{G_{ik} \cdot p_k}{\sum_{j=1}^{n_b} G_{ij}\, p_j} \end{aligned}$$

- H_{ik} depends on the unknown distribution b_k
- H_{ik} corrects smearing, no correction for inefficiencies

application ➜

➜ *bayesian unfolding*

- determination of unfolded distribution
  - ➜ use the unfolding matrix to correct for smearing
  - ➜ then correct efficiencies as described by the response matrix
  - ➜ if necessary determine the normalization
- synopsis

$$q_j = \frac{1}{\epsilon_j} \sum_{j=1}^{n_a} a_i \cdot H_{ij} \quad \text{with} \quad \epsilon_j = \sum_{k=1}^{n_a} G_{kj} \quad \text{and} \quad p_j = \frac{q_j}{\sum_{i=1}^{n_b} q_i}$$

- naive error propagation for q_j

$$C_{ij}(q) = \sum_{k,l=1}^{n_a} \frac{\partial q_i}{\partial a_k} \frac{\partial q_j}{\partial a_l} C_{kl}(a) = \frac{1}{\varepsilon_i \varepsilon_j} \sum_{k,l=1}^{n_a} H_{ki} H_{lj} C_{kl}(a)$$

  - ➜ correlated errors due to unfolding matrix
  - ➜ otherwise similar dependence on measurements as bin-by-bin corrections

❖ CHARAKTERISTICS OF BAYESIAN UNFOLDING

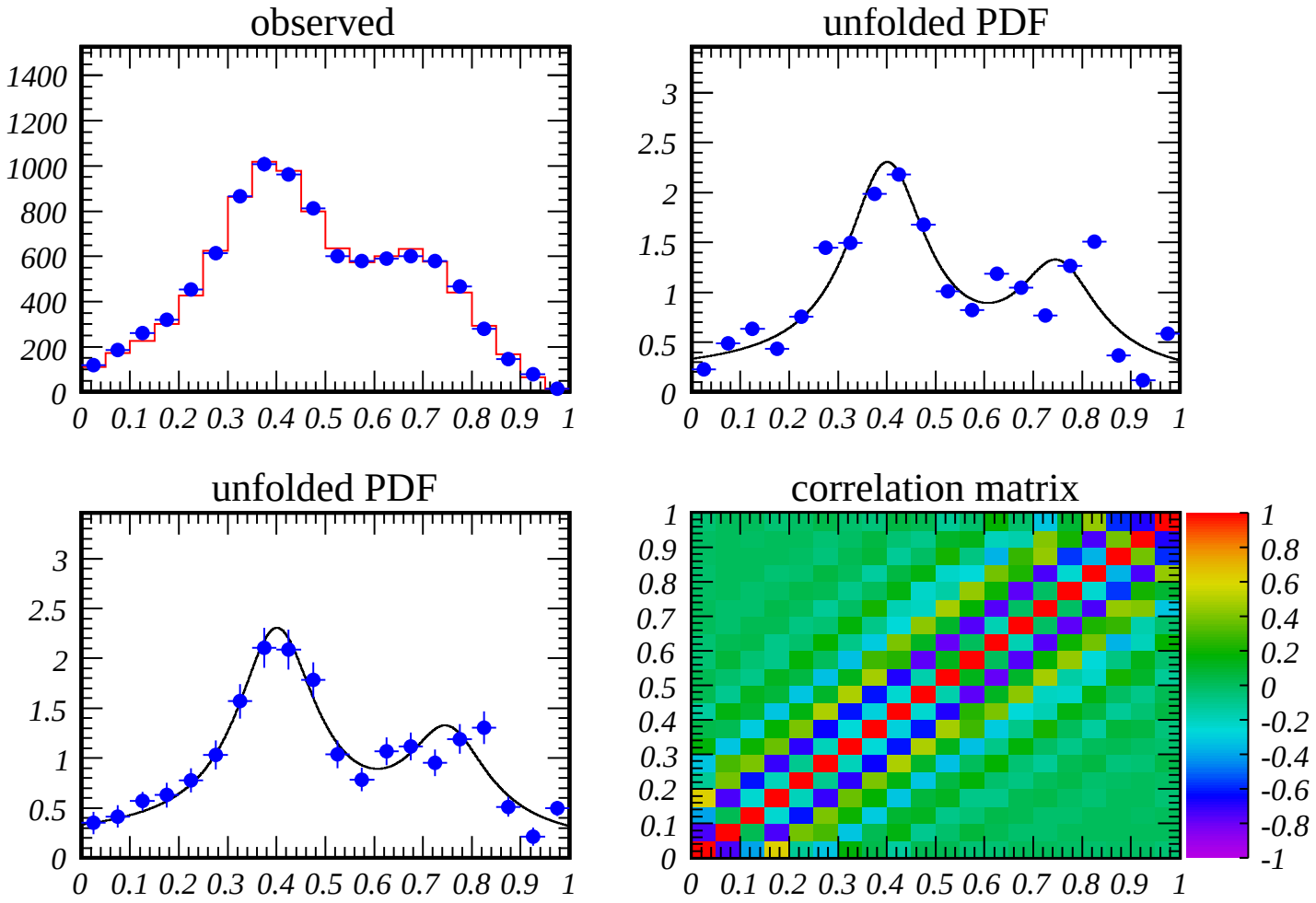
- mathematically sound aproach
- explicitly use positvity of probabilities
- can move measurements from unphysical region into allowed phase space
- no matrix inversion required
  - ➜ unfolding works also for non-square matrices G_{ik}
  - ➜ if needed the normalization B of $b_i = Bp_i$ is obtained from

$$\sum_i B \sum_k G_{ik} \cdot p_k = \sum_i a_i$$

- same problem with initial values as correction factors
- iteration makes H independent of initial values p_i
- error Monte Carlo is the method of choise to . . .
  - ➜ reliably determinate the covariance matrix of the unfolded distribution
  - ➜ stabilize the result against statistical fluctuations in the measurements

➜ test the method. . .

Numerical Tests



- slow convergence (if at all?) with the number of iterations
- number of iterations correlates structure in covariance matrix and size of errors

➜ *observation*

- too few iterations: result strongly correlated with initial values
- too many iterations:result becomes unstable

❖ CONCEPTAL APPROACH

The number of iteration can be chosen freely. Consequences of a particular choice can be quantified by means of the covariance matrix of the result. Schematically one has for the case of a square response matrix:

$$b_{unf} = H \cdot a = H \cdot (G \cdot G^{-1}) \cdot a = (H \cdot G) \cdot b_{true}$$

The unfolded distribution is a linear function of the measurements. The connection with the true distribution is given by a residual response matrix G_{res}:

$$G_{res} = (H \cdot G)$$

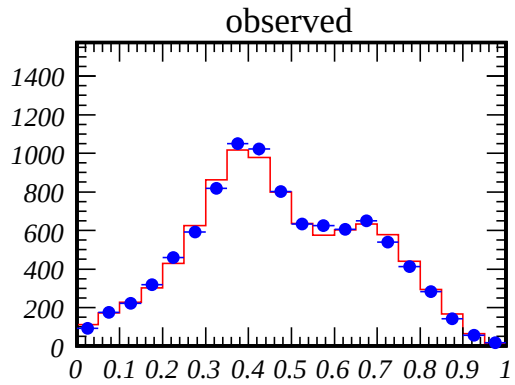
- $H = G^{-1}$ corresponds to full correction
- $H \neq G^{-1}$ implies residual distortions
  - ➜ the unfolding procedure did achieve a partial correction
  - ➜ improvement of resolution instead of full correction

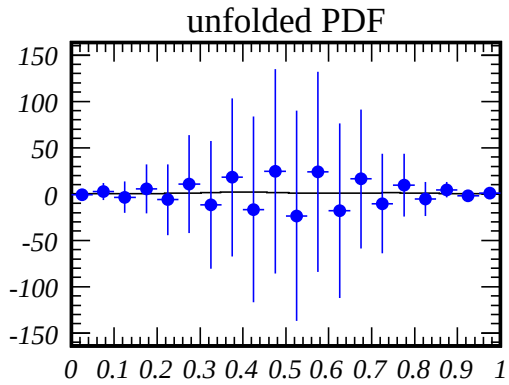
Unfolding by Matrix Inversion

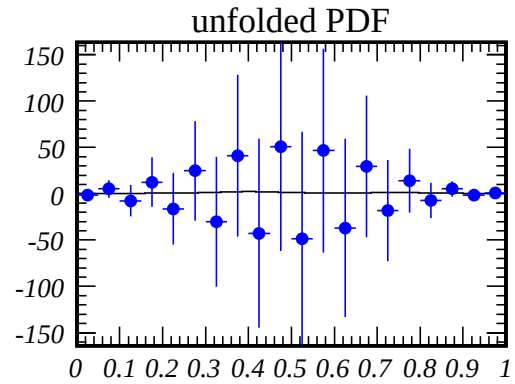
➜ *restricted to the case* $n_a = n_b$

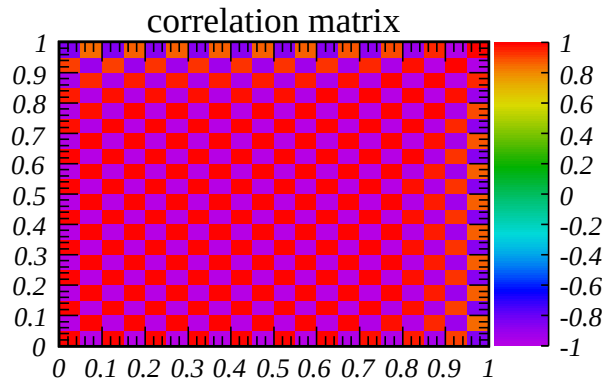
$$a = G \cdot b \quad \rightarrow \quad b = G^{-1} \cdot a \quad \text{with} \quad C(b) = G^{-1} C(a) (G^{-1})^T$$

- formally correct
- proper covariance matrix
- completely useless









diagonalize the unfolding problem to understand the strange behaviour . . .

❖ REMINDER: SVD FOR ANY MATRIX $A[m,n]$ $(m \geq n)$

$$A[m,n] = U[m,n] \cdot W[n,n] \cdot V[n,n]^T$$

with $U^T \cdot U = V^T \cdot V = V \cdot V^T = \mathbf{1}_n$ and positive definite diagonal matrix W

➜ *diagonalization of the unfolding problem*

- transform measurements $x = M \cdot a$ such that $C(x) = \mathbf{1}$
- the unfolding problem now reads $x = M \cdot a = M \cdot G \cdot b$
- apply singular value decomposition (SVD) to new response matrix $M \cdot G$

  $$x = M \cdot G \cdot b = U \cdot W \cdot V^T \cdot b \quad \text{or} \quad U^T x = W \cdot V^T \cdot b$$

- introduce "normalized moments" u and v

  $v = V^T \cdot b$ and $u = U^T \cdot x = U^T \cdot M \cdot a$ with $C(u) = U^T \cdot C(x) \cdot U = \mathbf{1}$

  which diagonalize the unfolding problem

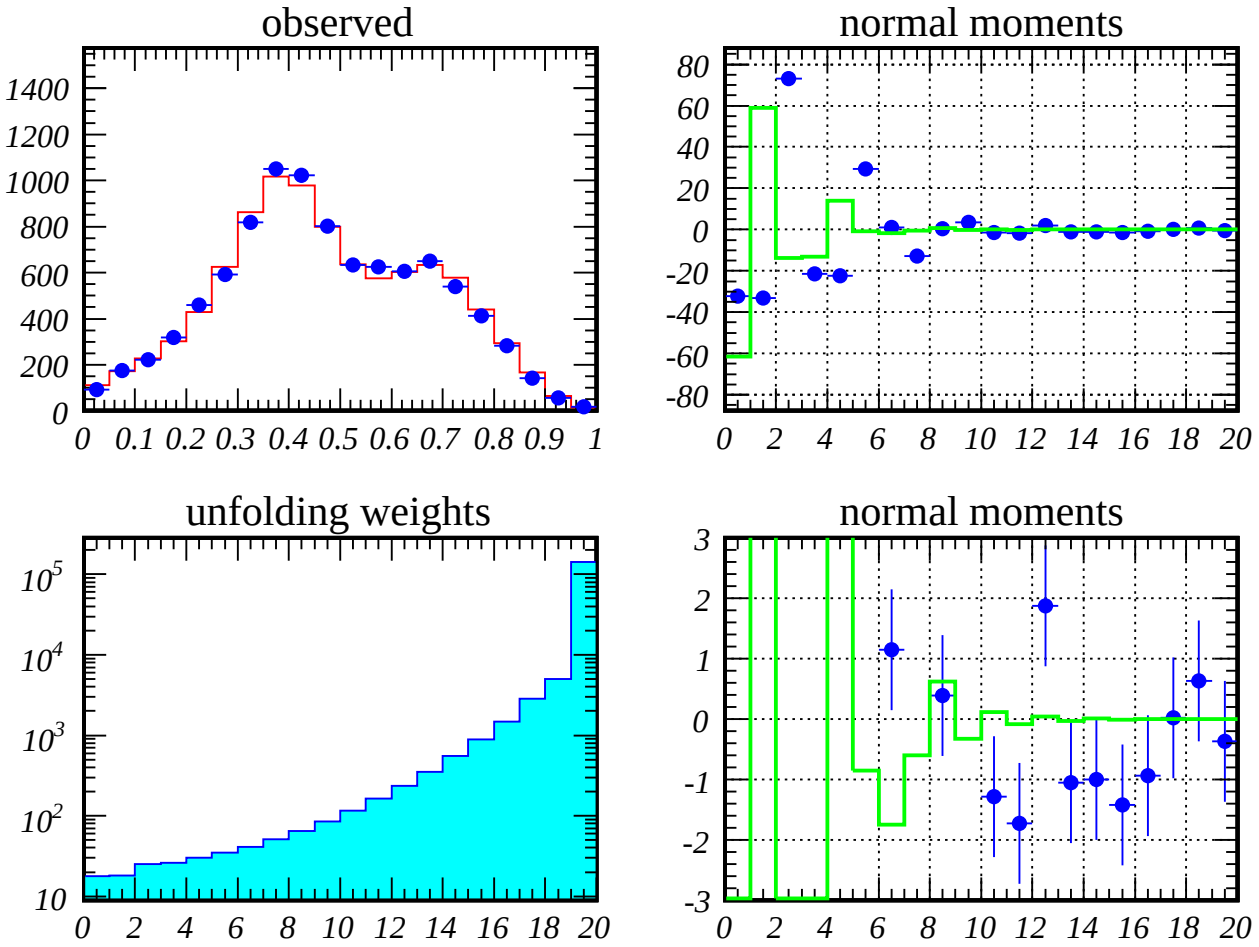
  $$u = W \cdot v$$

  ➜ a simple rotation now relates v to the unfolded distribution b

  ➜ the diagonal correction factors ("unfolding weights") are $1/W_{kk}$

➜ compare

- normalized moments for measurements and correction factors
- expectation for uniform unfolded distribution



➜ *diagonalization of the unfolding problem shows:*

The higher order moments u_k are exponentially suppressed by the response matrix. To measure them requires extrem large statistical precision. Accepting those components for the unfolding means exponential amplification of statistical fluctations.

❖ NOTE:

The higher order moments describe fine structur of $b(y)$. Using

$$b = V \cdot v$$

the eigen-functions (eigen-vectors) for the individual components v_i can be read off from the columns of V. Those

- eigenvectors are orthogonal
- the number of sign-changes grows with increasing order
- the highest order vector
  - ➜ has alternating signs
  - ➜ has the largest correction factors
  - ➜ dominates the matrix-inversion result

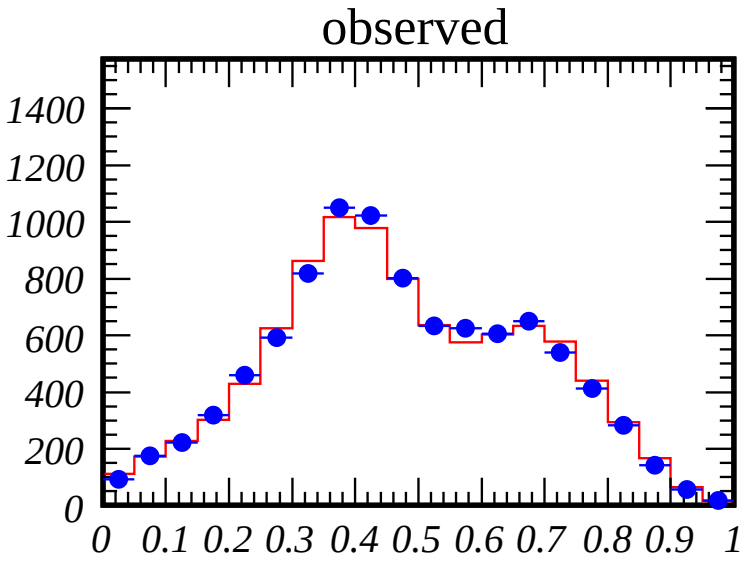
Regularization

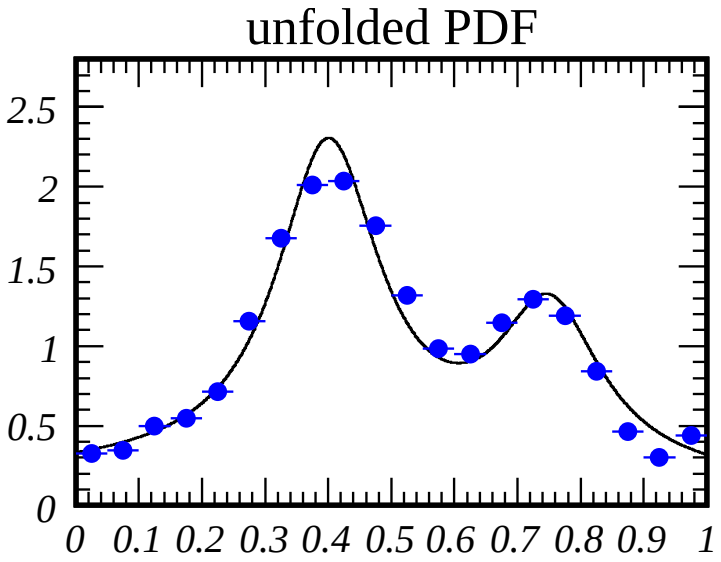
- unfolding requires n_b measurements u_k
- the effective number of measurements with information about $b(y)$ is smaller
- heuristic ansatz to count the effective number of measurements
  - ➜ compare measured values of normalized moments u_i with expectation $\tilde{u}_i$, from a (e.g. flat) prior distribution
  - ➜ count only those which are significantly different from the prior expectation
- Regularisation:
  - ➜ take those normalized moments from data which differ significanttly from the prior
  - ➜ take the others from the prior
  - ➜ construct the unfolded distribution

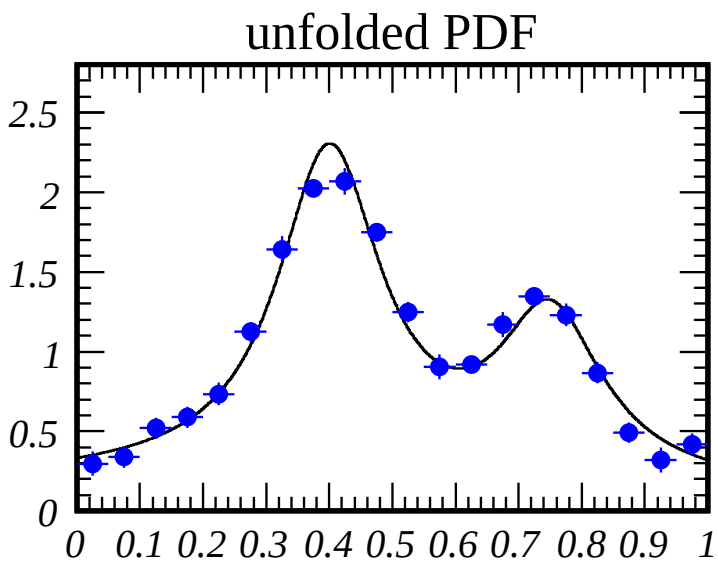
➜ *in general: replace missing information by assumptions*

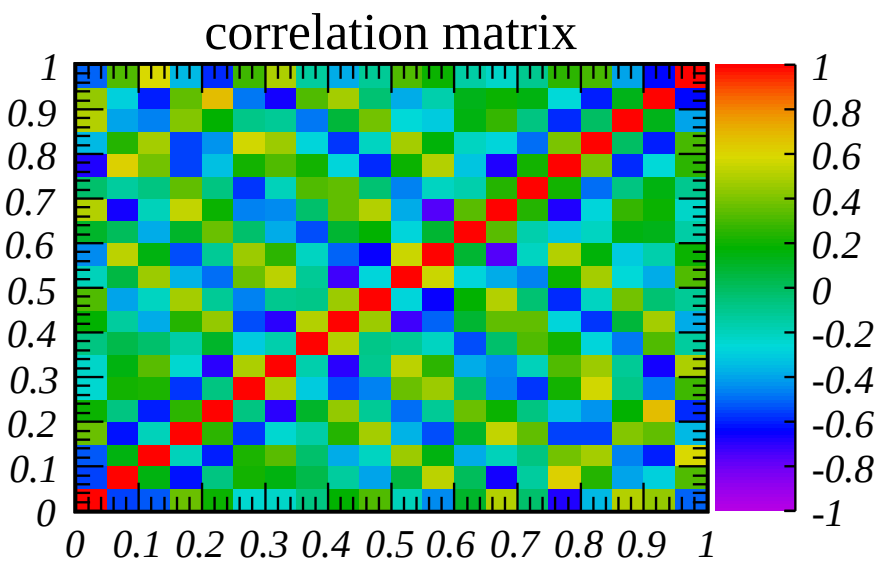
- many possibilities to add information
  - ➜ pure heuristics
  - ➜ assumption about smoothness or curvature of the result
  - ➜ information theoretical approaches (Maximum Entropy)
- many possibilities where to put the cut on the measurements
  - ➜ . . . “adjustment of the regularization parameter”

Numerical Examples









- significantly “better” results than straightforward matrix inversion
- almost perfect unfolding of smooth structures
- difficulties with discontinuities

❖ EXPLANANATION:

Describe regularization by a diagonal matrix R applied to the normalized moments:

$$
\begin{aligned}
b &= V \cdot W^{-1} \cdot R \cdot u = (V \cdot R \cdot W^{-1}) \cdot u \\
&= (V \cdot R \cdot W^{-1}) \cdot U^T \cdot M \cdot a \\
&= V \cdot R \cdot (V^T \cdot V) \cdot W^{-1} \cdot U^T \cdot M \cdot a \\
&= (V \cdot R \cdot V^T) \cdot (V \cdot W^{-1} \cdot U^T) \cdot M \cdot a \\
&= (V \cdot R \cdot V^T) \cdot (M \cdot G)^{-1} M \cdot a \\
&= (V \cdot R \cdot V^T) \cdot G^{-1} \cdot a = G_{\mathrm{res}} \cdot b(R=1) \quad \text{with} \quad G_{\mathrm{res}} = (V \cdot R \cdot V^T)
\end{aligned}
$$

The regularized result is equivalent to a smearing of the exaxt result! If the residual smearing G_{res} has a resolution better than typical structures in the true distribution the result appears undistorted. Regularization replaces complete unfolding by improvement of the resolution function.

➜ *analysis of the response matrix has shown*

- unfolding problem often are effectively under-constrained
- the response matrix is “ill conditioned”, i.e. the spectrum of the eigenvalues (singular values) varies over many orders of magnitude.
- matrix inversion or a least squares fit of bin contents usually unsatisfactory
- regularisation is a possible solution
  - ➜ suppress the not well measured components by some prior information
- effects of regularization
  - ➜ often acceptable unfolding results
  - ➜ residual smearing in the unfolded distribution
  - ➜ bias in normalized moments
    - ✘ negligible bias with respect to the measured distribution
    - ✘ unknown bias in the unfolding result
  - ➜ trade statistical errors against resolution
- bias can be controlled through the structure of the residual response matrix or the correlation lengths seen in the covariance matrix of the unfolded distribution.