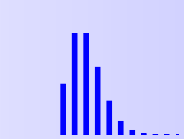
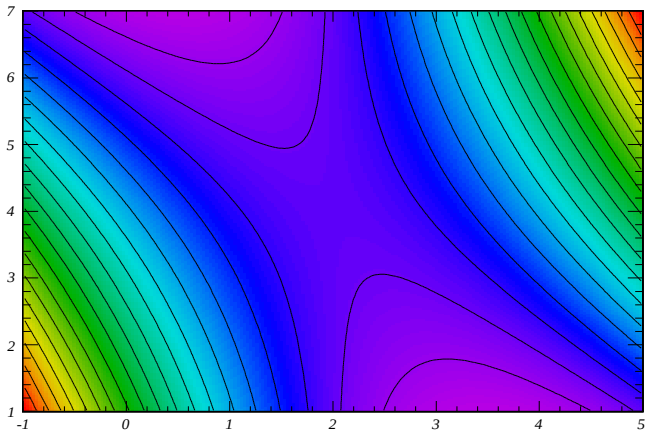


Statistische Methoden der Datenanalyse II

Michael Schmelling – MPI für Kernphysik

- *Einführung*
- *Fehler und Fehlerfortpflanzung*
- *Kleinste Quadrate & Maximum Likelihood*
- *Multivariate Analyse*
- *sWeights*
- *Markov Chain Monte Carlo*
- *Entfaltung und Parametrisierung*
- *Harmonische Analyse*



Literature

➜ *selected books and papers in alphabetical order. . .*

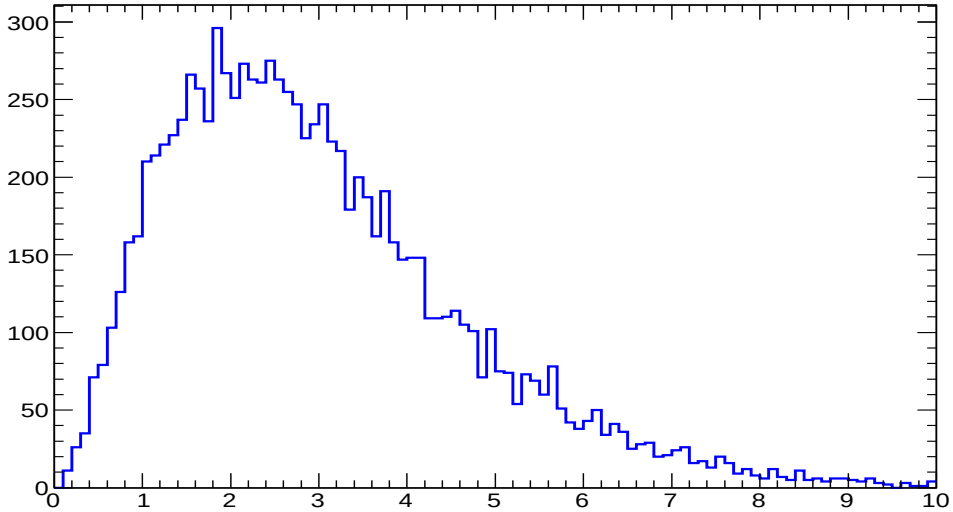
- R.J. Barlow,*Statistics*, Wiley
- S. Brand, *Data Analysis*, Springer
- G.D. Cowan,*Statistical Data Analysis*, Oxford University Press
- H.L. Harney, *Bayesian Inference*, Springer
- A. Hoecker et al., *TMVA 4 Users Guide*, http://tmva.sourceforge.net
- F. James, *Statistical Methods in Experimental Physics*, World Scientific
- D.E. Knuth, *The Art of Computer Programming*, Addison Wesley
- M. Pivk, F. R. Le Diberder. *sPlot*, NIM A555(2005)356, physics/0402083
- W.T. Press et al., *Numerical Recipes*, Cambridge University Press
- D.S. Sivia, *Data Analysis - A Bayesian Tutorial*, Oxford University Press

➜ *What are statistical methods?*

- recipes for data reduction: large data set ➜ single number e.g.. . .
  - ➜ md5sum: fingerprint characterizing the data set
  - ➜ particle lifetime from decay time measurements
  - ➜ CP violating phase from reconstructed B decays
- statistical methods are constructed
  - ➜ neither “right” nor “wrong” – characterized by properties
  - ➜ properties of a method need to be understood
    to judge the applicability and to interpret the results
- example: “central value” and “spread” of a set of measurements
  - ➜ different people will associate different things with those terms
  - ➜ usually no problem for qualitative discussions
  - ➜ quantitative science requires an exact definition

➜ how to characterize a data set

Example for a distribution of measurements



- "central value"
  - ➜ maximum value (after smoothing the distribution?)
  - ➜ median value - same number of measurements above and below
  - ➜ arithmetic average
- "spread"
  - ➜ Full-Width-at-Half-Maximum (FWHM) - but how to define the maximum
  - ➜ central 68% quantile
  - ➜ rms - average quadratic deviation from the mean

➜ start at the beginning

➜ *"probability" of an event: what does this mean?*

- probability $p = 0$: the event will not happen
- probability $p = 1$: the event will happen
- probability $p = 1/3$: suggestions?
  - ➜ the event will happen every third try
    - not consistent: equivalent to a sequence of p=0,p=0,p=1
  - ➜ the event will happen in 1/3 of infinitely many tries?
    - OK if the next result cannot be predicted from previous ones
    - provides a measurement prescription for repeatable tries
    - only approximate realization possible in practice
  - ➜ I should get paid 3 EUR if I invest 1 EUR and the event happens
    - OK - applicable also for non-repeatable events
    - basis of the world's financial system

Kolmogorov's axioms on probability

➜ *define properties of probabilities - don't care what they are!*

Build probability theory on a mapping of sets ➜ real numbers.

❖ Definitions:

Ω : the entire set

E : partial set of Ω

$p(E)$: probability of E

❖ Axioms:

1. $0 \leq p(E) \leq 1$
2. $p(\Omega) = 1$
3. $p(E_1 \cup E_2) = p(E_1) + p(E_2)$ if $E_1 \cap E_2 = 0$

Math of probabilities follows unambiguously - interpretation is left open.

Consider the probability of an event B occurring together with another one from a set of disjoint events $A_i, i = 1, \ldots, n$.

$$P(A_i, B) = p(B|A_i)\, p(A_i) = p(A_i|B)\, p(B)$$

It follows:

$$p(A_i|B) = \frac{p(B|A_i)\, p(A_i)}{p(B)}$$

Bayes' theorem

Having seen B, the prior $p(A_i)$ for A_i is updated to $p(A_i|B)$.

Bayes' theorem is at the heart of statistical inference based on empirical input.
If the probabilities of the A_i sum up to unity, then one has

$$p(B) = \sum_i p(B|A_i) p(A_i)$$

and thus:

$$p(A_k|B) = \frac{p(B|A_k) p(A_k)}{\sum_i p(B|A_i) p(A_i)}$$

Applying Bayes' theorem

Consider a test that detects the common cold in the early stages of an infection, where an efficient cure is available. The probability to test positive in case of an infection is $p(+|I) = 0.98$, the probability for a negative result on a healthy subject is $p(-|H) = 0.97$. In summer, the a priori probability for infection is $p(I) = 0.001$.

What's the probability for a person tested positive to be infected?

the probabilities are:

$$\begin{array}{llllll} p(I) & = & 0.001 & p(H) & = & 0.999 \\ p(+|I) & = & 0.980 & p(-|I) & = & 0.020 \\ p(+|H) & = & 0.030 & p(-|H) & = & 0.970 \end{array}$$

The rows sum up to unity. Application of Bayes' theorem yields

$$p(I|+) = \frac{p(+|I)p(I)}{p(+|I)p(I) + p(+|H)p(H)} \approx 0.032$$

Sweets for all patients diagnosed "infected" will yield 97% "healing rate"!

➜ *Kolmogorov's axiom for discrete sets: discrete probabilities*

Enumerate discrete probabilities by $i = 0, 1, 2, \ldots$

$$\sum_i p_i = 1 \quad \text{with} \quad p_i = \text{probability to find state } i$$

➜ *continuous sets: probability density functions (PDFs)*

A function $f(x)$ can be interpreted as a PDF if

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

The PDF gives the probability to observe an event in $[x, x + dx]$:

$$p(x, x + dx) = \int_x^{x+dx} dx\; f(x) \approx f(x)\, dx$$

➜ *the uniform distribution*

The probability density inside a range $[a, b]$ is constant.

- (most) fundamental, simple PDF
- convenient starting point to derive more complex PDFs
- core of numerical random generators

➜ *modelling measurement errors*

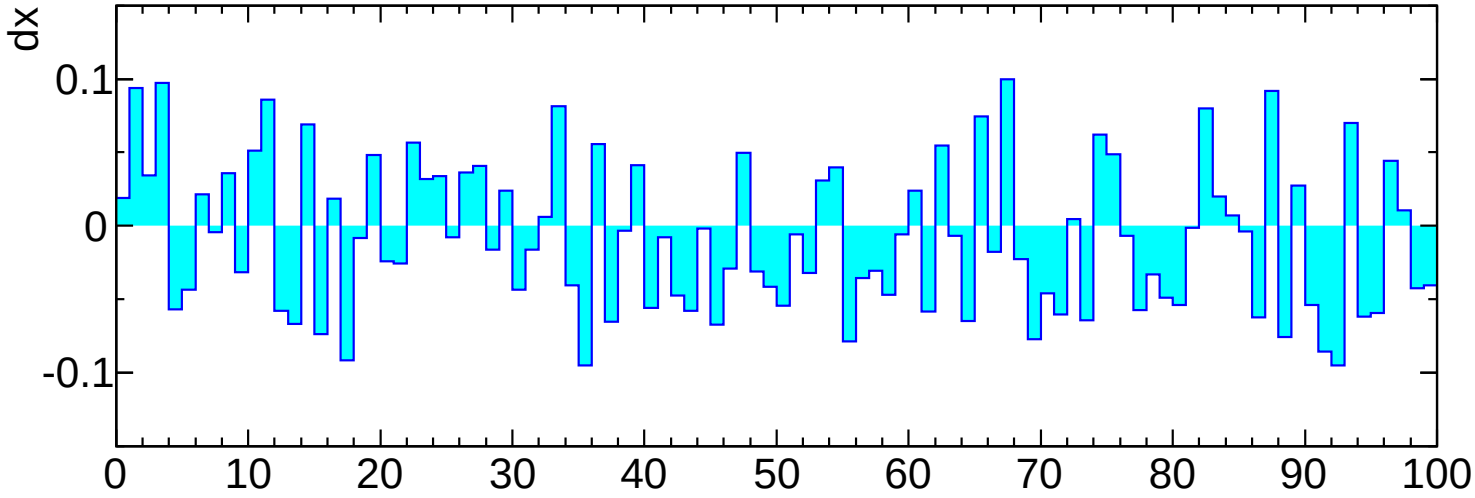
- example 1: astronomical observations
  - ➜ light rays are scattered at density variations in the atmosphere
- example 2: current over a resistor
  - ➜ current variations from thermal motion of many electrons

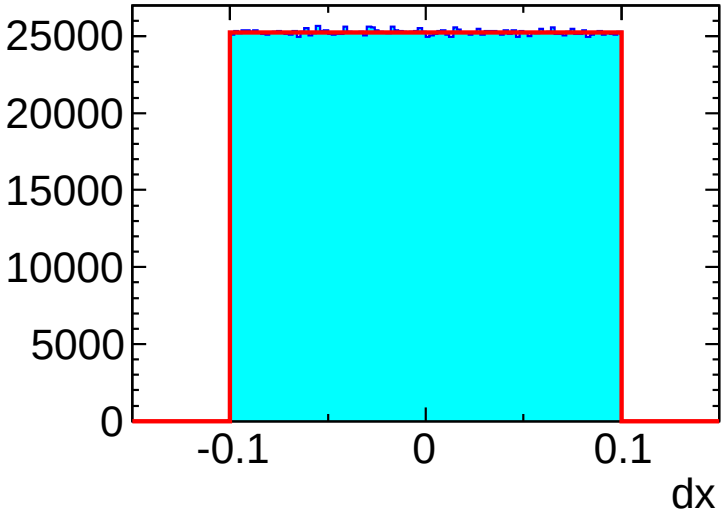
❖ common feature

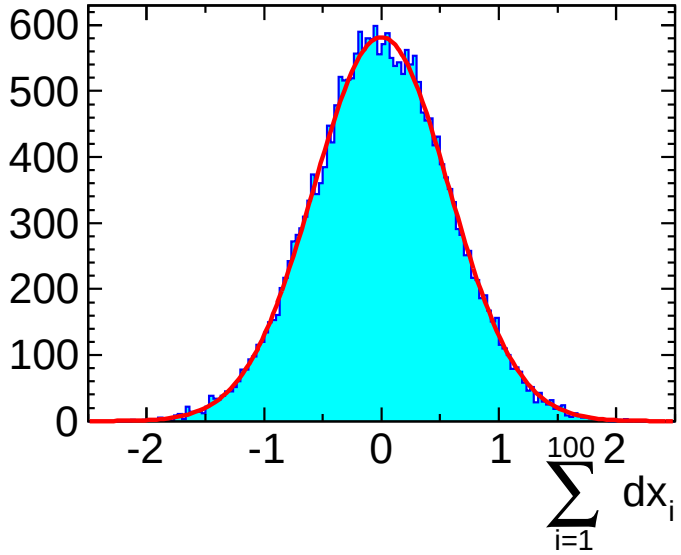
Many small variations add up to deviations between measurement und true value. Do a numerical study with uniform PDF for the variations.

➜

Simulation study







➜ *observation*

The sum of many random fluctuations is described by a Gaussian PDF

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$$

- symmetric around zero
- one parameter σ describing the width
- first published in by C.F. 1809 Gauss in "Theoria motus corporum coelestium in sectionibus conicis solem ambientium" (with Least-Squares and Maximum-Likelihood method)
- the exact conditions for convergence to a Gaussian are formally described by the central limit theorem
- due to its fundamental nature also referred to as “normal” distribution

➜ *statistics of counting experiments*

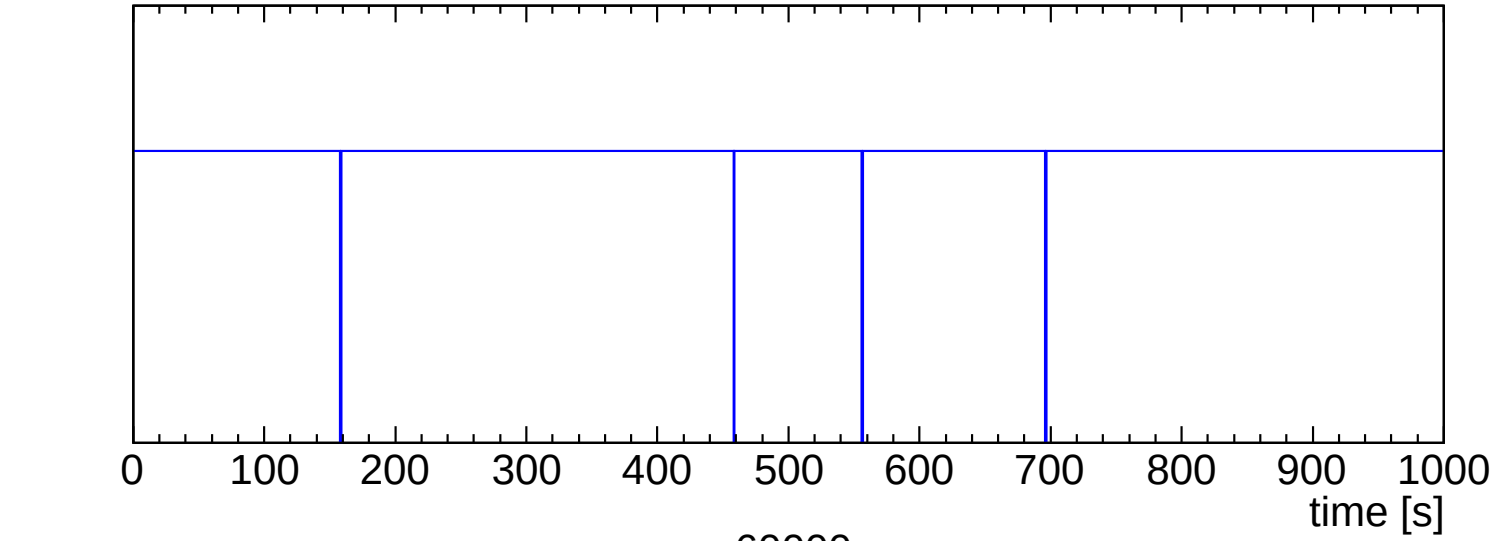
- examples:
  - ➜ decays in a radiactive source
  - ➜ cosmic muons observed at surface level on earth
  - ➜ number of soldiers in the Prussian army killed accidentally by horse kicks (Ladislaus Bortkiewicz, 1898)
- quantities of interest
  - ➜ time differences between subsequent events
  - ➜ number of events in time interval T

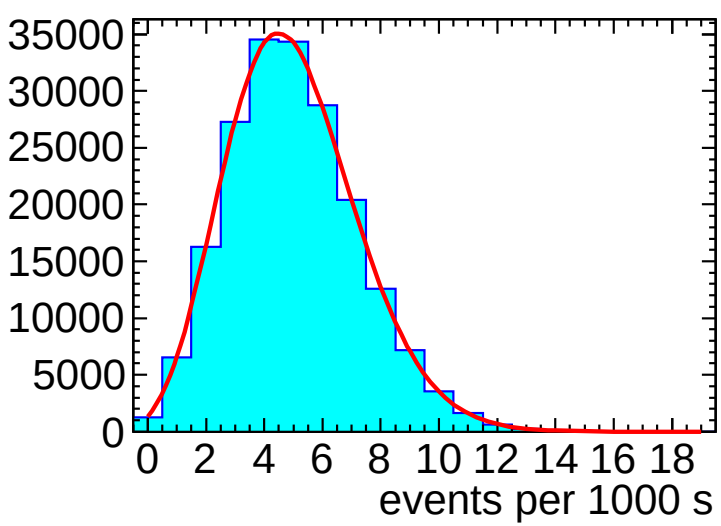
❖ numerical simulation

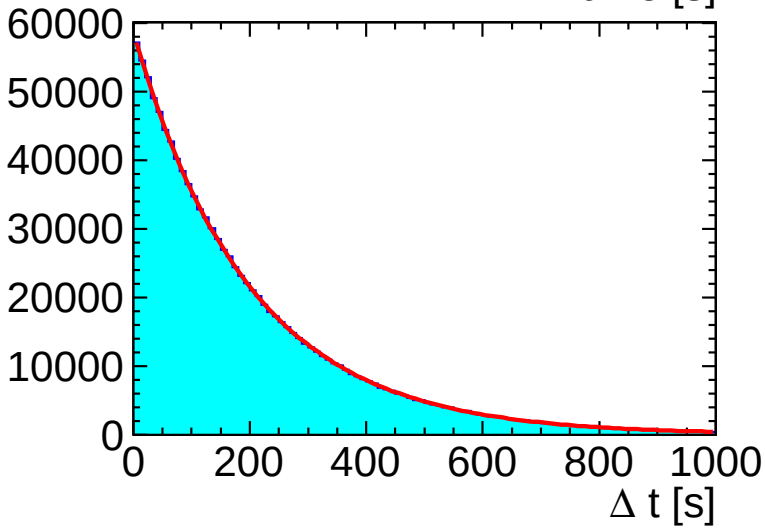
- split T into (many) subsequent time slices
- assume a probability to observe an event in a time slice $p \ll 1$

see what happens ➜

Simulation study

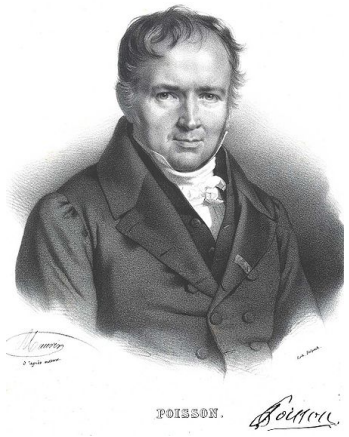






➜ *observation*

- results are described by simple functions of a single parameter (consequence of the single probability for an event per time slice)
- event counts per time interval: Poisson distribution
  - ➜ first published by Simèon Denis Poisson 1837 in "Recherches sur la probabilité des jugements en matière criminelle et en matière civile"



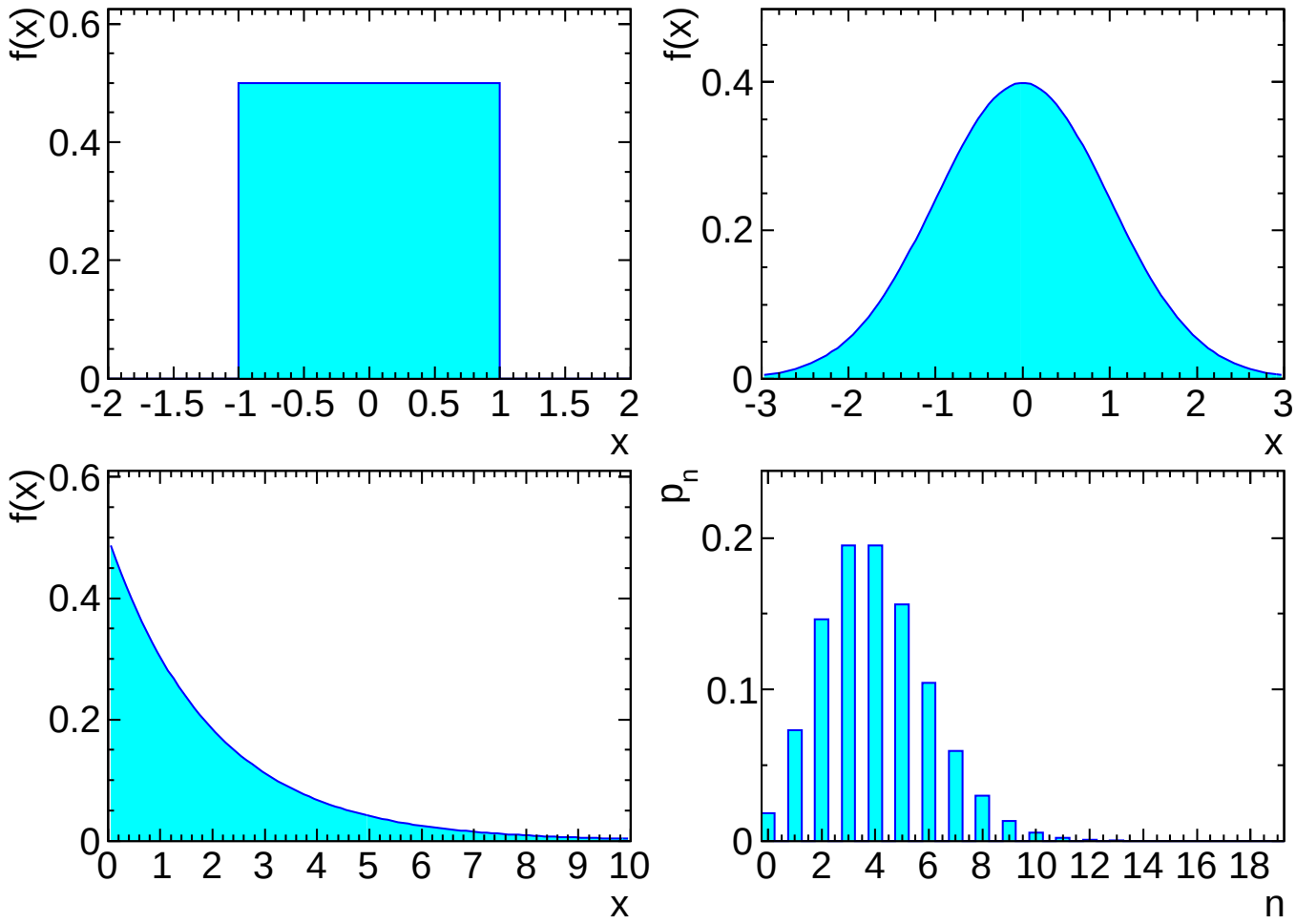
$$p_n = e^{-\mu} \frac{\mu^n}{n!}$$

- time difference between events: exponential distribution

$$f(t) = \frac{1}{\tau} e^{-t/\tau}$$

Characterizing probability distributions

➜ *consider the distributions introduced before*



❖ wanted: location and width

➜ *"typical" x-values*

- maximum of the distribution
  - ➜ not always well defined
  - ➜ can be at one edge of the distribution
- median value m

$$\int_{-\infty}^{m} dx\ f(x) = \int_{m}^{\infty} dx\ f(x)$$

  - ➜ not obvious for discrete distributions; insensitive to tails
- center-of-gravity $\langle \cdots \rangle$ – usually referred to as mean value

$$\langle x \rangle = \int dx\ x\, f(x) \quad \text{or} \quad \langle n \rangle = \sum_{n=0}^{\infty} n\, p_n$$

  - ➜ well defined for continuous and discrete distributions
  - ➜ sensitive to asymmetric tails; may even diverge
- median and center-of-gravity coincide for symmetric distributions
- all three "typical values" coincide for symmetric uni-modal distributions

Characterizing the width of a distribution

➜ *"typical" range covered by x-values*

- FWHM: full width at half the maximum value
  - ➜ not obvious for discrete distributions; insensitive to tails
- central q% quantile $[a, b]$, with, e.g., q=68.3%, 90% or 95%

$$\int_{-\infty}^{a} dx\, f(x) = \int_{b}^{\infty} dx\, f(x) = \frac{1}{2}(1-q)$$

  - ➜ not obvious for discrete distributions; insensitive to tails
- standard deviation σ

$$\sigma^2 = \int_{-\infty}^{\infty} dx\; f(x)\,(x - \langle x \rangle)^2 \quad \text{or} \quad \sigma^2 = \sum_{n=0}^{\infty} (n - \langle n \rangle)^2\, p_n$$

  - ➜ well defined for continuous and discrete distributions
  - ➜ sensitive to the functional form of the tails - may even diverge
  - ➜ simple linear operation on the PDF

"unified" approach ➜

A measure for the scatter s of x with PDF $f(x)$ around a point a is:

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

For s to characterize $f(x)$, a should be chosen to minimize s:

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

The mean value $\langle x \rangle$ is the location parameter that minimizes the scatter s. The minimal scatter s is called standard deviation, σ. Its square is called variance, σ^2.

➜ *note:*

- for symmetric PDFs $\langle x \rangle$ is the symmetry point
- the scatter around $\langle x \rangle$ is called "standard deviation" σ
- σ is also referred to as "rms"-width

➜ *uniform, gaussian, exponential and poisson distributions*

$$\frac{1}{2w}\Theta(x+w)\Theta(w-x) \quad , \quad \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \quad , \quad \frac{e^{-x/\tau}}{\tau} \quad , \quad e^{-\mu}\,\frac{\mu^n}{n!}$$

	median	mean	FWHM	68.3% quant.	stdev
uniform	0	0	$2w$	$1.366\,w$	$w/\sqrt{3}$
gaussian	0	0	$\sqrt{8\ln 2}\,\sigma$	$2\,\sigma$	σ
exponential	$\tau\ln 2$	τ	$\tau\ln 2$	$-\tau\ln 0.317$	τ
poisson		μ			$\sqrt{\mu}$

- ratios of different width or location estimators are $O(1)$
- analytically most convenient: mean value and standard deviation
  - ➜ simple integrals over the entire distributions
  - ➜ most commonly used estimators for location and width

➜ *generalization of concepts introduced before:*

Given a PDF $f(x)$ and a function $a(x)$, the expectation value $\langle a \rangle$ is:

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

- mapping of functions $f(x)$ to a real numbers – if the integral exists
- important property: linearity i.e. $\langle \alpha A + \beta B \rangle = \alpha \langle A \rangle + \beta \langle B \rangle$

❖ examples:

$$\langle x \rangle \quad : \quad \text{mean value}$$

$$\left\langle (x - \langle x \rangle)^2 \right\rangle : \quad \text{variance}$$

➜ *note:*

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

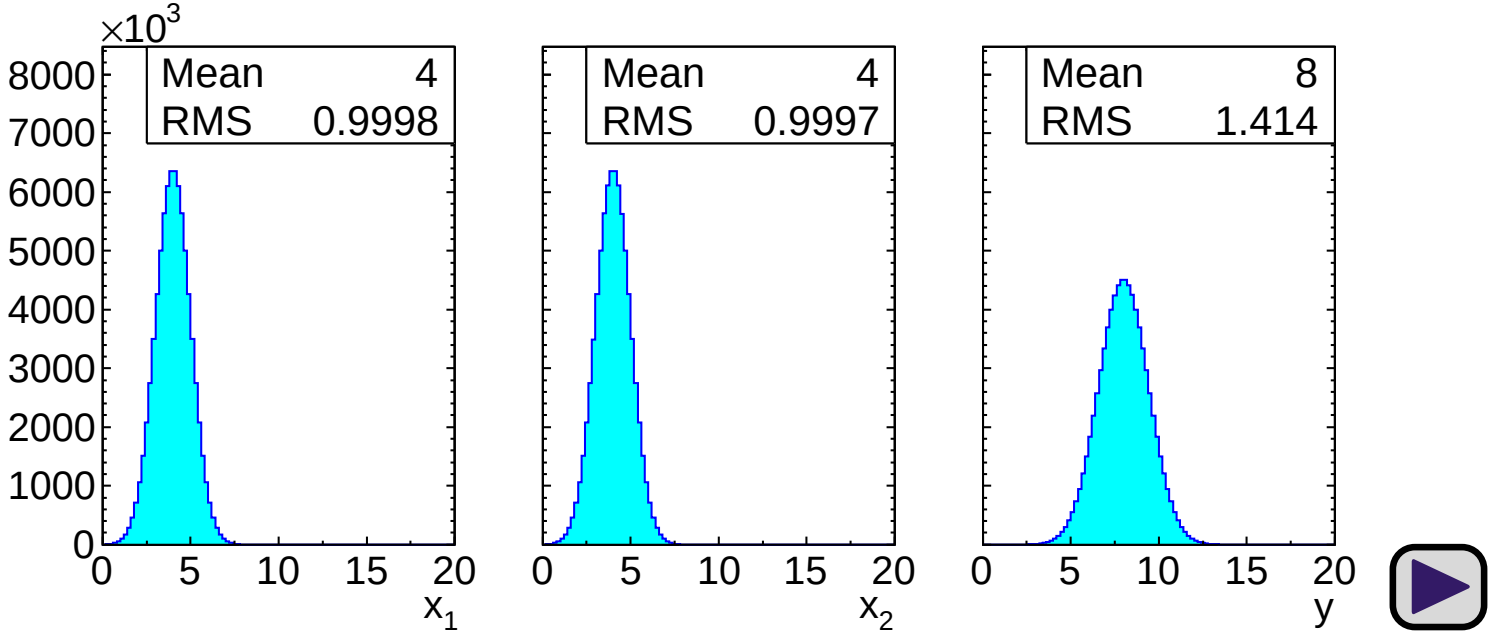
$$= \left(\int dx \; x^2 f(x) \right) - \langle x \rangle^2 = \langle x^2 \rangle - \langle x \rangle^2$$

➜ *the problem:*

Given PDFs $f_1(x_1)$ and $f_2(x_2)$, determine the PDF $g(y)$ of $y = h(x_1, x_2)$.

➜ *solution by Monte Carlo*

- generate x_1 and x_2 according to $f_1(x_1)$ and $f_2(x_2)$
- calculate and histogram $y = h(x_1, x_2)$



➜ *the problem:*

Given PDFs $f_1(x_1)$ and $f_2(x_2)$, determine the PDF $g(y)$ of $y = h(x_1, x_2)$.

➜ *analytic solution*

For the cumulative distribution $G(Y)$ one has:

$$G(Y) \equiv \int_{-\infty}^{Y} dy\ g(y) = \int dx_1\, dx_2 f_1(x_1) f_2(x_2)\ \Theta(Y - h(x_1, x_2))$$

Sum all probability elements $dp_1\, dp_2$, with $dp_i = dx_i f_i(x_i)$, which satisfy the constraint $h(x_1, x_2) < Y$. Differentiation with respect to the upper limit Y then yields the solution:

$$g(y) = \left.\frac{d}{dY} G(Y)\right|_{Y=y} = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) \delta(y - h(x_1, x_2))$$

Example: convolution of random variables

➜ *normalization, mean value and variance of* $y = x_1 + x_2$

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

$$= \int dx_1\, dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2)^k$$

expectation values:

$$\langle y^0 \rangle = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) = 1$$

$$\langle y^1 \rangle = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2) = \langle x_1 \rangle + \langle x_2 \rangle$$

$$\langle y^2 \rangle = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2)^2 = \langle x_1^2 \rangle + 2 \langle x_1 \rangle \langle x_2 \rangle + \langle x_2^2 \rangle$$

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

❖ *convolutions are normalized, mean values and variances are added!*

➜ *generalization of 1-dim PDFs*

- non-negative, normalizable functions in n dimensions
- discuss the most important concepts with 2-dim PDFs

❖ 2-dim PDF:

$$f(x,y) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x,y) = 1$$

❖ interpretation:

the Probability for (x,y) in the rectangle $[x, x+dx] \times [y, y+dy]$ is

$$p(x, x+dx; y, y+dy) = \int_{x}^{x+dx} dx \int_{y}^{y+dy} dy\, f(x,y) \approx f(x,y)\, dx\, dy$$

❖ independence of variables:

x and y are independent if the PDF factorizes: $f(x,y) = g_1(x) \cdot g_2(y)$

➜ *look for expectation values that are sensitive to dependencies*

0th order	$\langle 1 \rangle$
1st order	$\langle x \rangle \ , \langle y \rangle$
2nd order	$\langle x^2 \rangle \ , \langle xy \rangle \ , \langle y^2 \rangle$

The lowest order term probing dependencies between x and y is $\langle xy \rangle$.
For independent variables with $f(x,y) = g_1(x)\, g_2(y)$ one finds

$$\begin{aligned} \langle xy \rangle &= \int dx \int dy\, (x\, y)\, g_1(x)\, g_2(y) \\ &= \left(\int dx\, x\, g_1(x) \right) \left(\int dy\, y\, g_2(y) \right) = \langle x \rangle \langle y \rangle \end{aligned}$$

❖ measure of correlation: the "covariance" of x and y

$$C_{xy} = \langle x\, y \rangle - \langle x \rangle \langle y \rangle$$

not the only possibility, but simple and useful . . .

➜ *dimensionless measures of correlation between two variables*

$$\rho = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{C_{xy}}{\sqrt{C_{xx}\, C_{yy}}}$$

❖ properties

- $-1 \leq \rho \leq 1$
- $y = a\,x + b$ ➜ $\rho = \mathrm{sign}(a)$ (“100% (anti)correlation”)
- $\rho = 0$ necessary, but not sufficient for independence of x and y

❖ example: function $y = a\,x^2 + b\,x + c$ with gaussian distributed x

$$\rho = \frac{b}{\sqrt{2a^2\sigma_x^2 + b^2}}$$

➜ $|\rho| < 1$ if a parabolic term is present

➜ $\rho = 0$ if the linear term is absent

The covariance matrix

➜ *array of covariances between all variable-pairs of an n-dim PDF:*

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

Expressed through standard deviations and correlation coefficients it is

$$C_{ij} = \rho_{ij} \cdot \sigma_i \sigma_j \quad \text{with} \quad \rho_{ii} = 1 \ .$$

➜ *note:*

- the diagonal terms C_{ii} are the variances of the individual variables
- off-diagonal terms are covariances
- the covariance matrix is symmetric and positive definite
- it can be diagonalized by rotation in the space of the variables
- C also is referred to as "error matrix"
- C describes the extension and orientation of an n-dim PDF

➜ *exploit the linearity of expectation values*

Consider a linear transformation $y_k = \sum_i M_{ki} x_i$. Given the covariance matrix $C_{ij}(x)$, the covariance matrix $C_{kl}(y)$ is

$$\begin{aligned} C_{kl}(y) &= \langle y_k y_l \rangle - \langle y_k \rangle \langle y_l \rangle \\ &= \left\langle \left(\sum_i M_{ki} x_i \right) \left(\sum_j M_{lj} x_j \right) \right\rangle - \left\langle \sum_i M_{ki} x_i \right\rangle \left\langle \sum_j M_{lj} x_j \right\rangle \\ &= \sum_{ij} M_{ki} M_{lj} \left(\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \right) = \sum_{ij} M_{ki} M_{lj} C_{ij}(x) \end{aligned}$$

or in matrix notation:

$$\vec{y} = M \cdot \vec{x} \qquad \text{and} \qquad C(y) = M \cdot C(x) \cdot M^T$$

- ➜ if $C(x)$ is positive definite, so is $C(y)$
- ➜ M need not be a square matrix - the number of rows is arbitrary

➜ *what are errors?*

- "errors" are uncertainties - not to be confused with "mistakes"
- quantify how well one knows e.g. a constant of nature - but how?
- engineer: tolerance = maximum possible deviation
- physicist: many different conventions. . .
  - ➜ standard deviation σ
  - ➜ 3-σ uncertainties
  - ➜ confidence level intervals containing the true value. . .
    - in a certain fraction of experiments (frequentist)
    - with a certain probability (bayesian)

➜ ask the professionals. . .

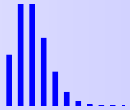
WG 1 (JCGM 100:2008, Recommendation INC-1 (1980)

➜ *Expression of experimental uncertainties*

1 The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:

A those which are evaluated by statistical methods,
B those which are evaluated by other means.

There is not always a simple correspondence between the classification into categories A or B and the previously used classification into“random” and “systematic” uncertainties. The term “systematic uncertainty” can be misleading and should be avoided. Any detailed report of the uncertainty should consist of a complete list of the components, specifying for each the method used to obtain its numerical value.




2 The components in category A are characterized by the estimated variances s_i^2 (or the estimated "standard deviations" s_i) and the number of degrees of freedom ν_i. Where appropriate, the covariances should be given.

3 The components in category B should be characterized by quantities u_j^2, which may be considered as approximations to the corresponding variances, the existence of which is assumed. The quantities u_j^2 may be treated like variances and the quantities u_j like standard deviations. Where appropriate, the covariances should be treated in a similar way.

4 The combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances. The combined uncertainty and its components should be expressed in the form of "standard deviations".

5 If, for particular applications, it is necessary to multiply the combined uncertainty by a factor to obtain an overall uncertainty, the multiplying factor used must always be stated.

(end of quote)

Discussion

➜ *why define uncertainties by variances and standard deviations*

- well defined procedures how to handle them
  - ➜ when propagating uncertainties into derived variables
  - ➜ for the combination of independent measurements
- rigorous limits on probability contents in the tails
- often asymptotically gaussian behaviour (central limit theorem)
- no (little) danger of mis-interpretation
- confidence level intervals . . .
  - ➜ not always obvious how they are defined
  - ➜ not obvious how to combine them
- warning: many physics papers actually mix concepts, combining "one-sided" variances in quadrature with confidence level intervals . . .

❖ focus first on variances/standard deviations!

➜ *probability content in the tails of a distribution*

Take any PDF $f(x)$, function $w(x) \geq 0$ and x-region with $w(x) \geq C$:

$$\langle w \rangle = \int dx\, f(x)\, w(x) \geq \int\limits_{w(x) \geq C} dx\; f(x) w(x) \geq C \int\limits_{w(x) \geq C} dx\; f(x) = C\, p(w(x) \geq C)$$

it follows $$p(w(x) \geq C) \;\leq\; \frac{\langle w \rangle}{C}\;.$$

For the special choice $w(x) = (x - \langle x \rangle)^2$ and $C = k^2\sigma^2$ one finds:

$$p_k \equiv p\left((x - \langle x \rangle)^2 > k^2\sigma^2\right) \leq \frac{1}{k^2}$$

- the probability beyond $\pm k\,\sigma$ around $\langle x \rangle$ is at most $1/k^2$
- actual probability contents for most PDFs are much lower
  - ➜ e.g. gaussian: $\{p_1, p_2, p_3\} \approx \{0.317, 0.0555, 0.0027\}$

Error propagation - setting the stage

➜ *definitions:*

- $\vec{x}$: vector of observed quantities
- $\langle \vec{x} \rangle$: expectation values of $\vec{x}$ - assumed to be the true values $\vec{x}^t$
- $C(x)$: covariance matrix of $\vec{x}$ - assumed to be known
- $\vec{y} = \vec{g}(\vec{x})$: vector of derived quantities
- $\vec{y}^t = \vec{g}(\langle \vec{x} \rangle)$: true vector of derived quantities
- $C(y)$: covariance matrix of $\vec{y}$ - to be determined

❖ study properties of the transition $\vec{x} \rightarrow \vec{y}$

➜ expectation values

➜ uncertainties

➜ *the expectation value of $\vec{y}$ is biased: $\langle \vec{y} \rangle \neq \vec{y}^{\,t}$*

Taylor expansion for a single component around $\langle x \rangle$ shows

$$y_k = g_k(\langle \vec{x} \rangle) + \sum_i \frac{\partial g_k(\langle \vec{x} \rangle)}{\partial x_i}(x_i - \langle x_i \rangle) + \frac{1}{2}\sum_{i,j} \frac{\partial^2 g_k(\langle \vec{x} \rangle)}{\partial x_i \partial x_j}(x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) + \ldots$$

and taking the expectation value yields:

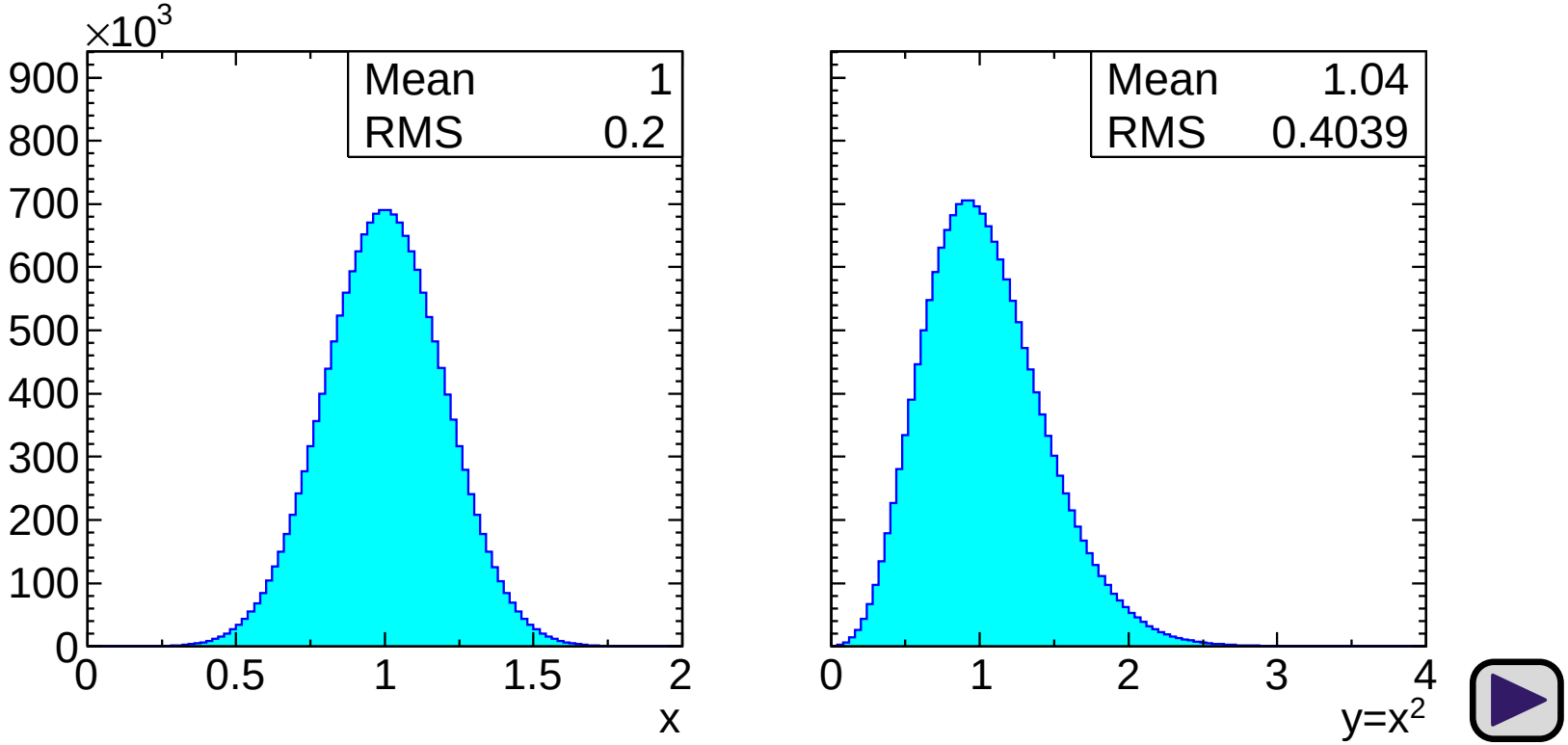
$$\langle y_k \rangle = y_k^t + \frac{1}{2}\sum_{i,j} \frac{\partial^2 g_k(\langle \vec{x} \rangle)}{\partial x_i \partial x_j} C_{ij}(x) + \ldots$$

❖ discussion

- in many cases the bias is small and can be neglected
- the leading order correction in principle is known
- don't average biased estimates of $\vec{y}$ - average the unbiased $\vec{x}$

Numerical example

➜ *transformation of a gaussian distributed* $x \rightarrow y = x^n$



- small non-linearities or small σ are uncritical
- biases are usually small compared to standard deviations
- bias correction is needed when averaging transformed values

➜ *leading order treatment in n dimensions*

$$y_k \approx g_k(\langle\vec{x}\rangle) + \sum_{i=1}^{n} \frac{\partial g_k(\langle\vec{x}\rangle)}{\partial x_i}(x_i - \langle x_i\rangle) \qquad \text{expansion around } \langle\vec{x}\rangle$$

$$\approx g_k(\langle\vec{x}\rangle) + \sum_{i=1}^{n} \frac{\partial g_k(\vec{x})}{\partial x_i}(x_i - \langle x_i\rangle) \qquad \text{derivatives taken at } \vec{x}$$

$$\approx \langle y_k\rangle + \sum_{i=1}^{n} \frac{\partial g_k(\vec{x})}{\partial x_i}(x_i - \langle x_i\rangle) \qquad \text{assume } \vec{y}^{\,t} = \langle\vec{y}\rangle$$

then calculate the covariance matrix $C_{kl}(y) = \langle(y_k - \langle y_k\rangle)(y_l - \langle y_l\rangle)\rangle$:

$$C_{kl}(y) \approx \sum_{i,j=1}^{n} \frac{\partial g_k}{\partial x_i}\frac{\partial g_l}{\partial x_j}\langle(x_i - \langle x_i\rangle)(x_j - \langle x_j\rangle)\rangle = \sum_{i,j=1}^{n} \frac{\partial g_k}{\partial x_i}\frac{\partial g_l}{\partial x_j} C_{ij}(x)$$

(note: derivatives are taken at the measured $\vec{x}$.)

➜ *matrix notation*

Under a transformation $\vec{y} = \vec{g}(\vec{x})$ the covariance matrix transforms as

$$C(y) = M(x) \cdot C(x) \cdot M^T(x)$$

with jacobian $M(x)$ and matrix elements $M_{ij} = \dfrac{\partial g_i}{\partial x_j}$.

The argument to M indicates that the derivatives are with respect to $\vec{x}$.

If $M(x)$ can be inverted then no information is lost in the transformation.

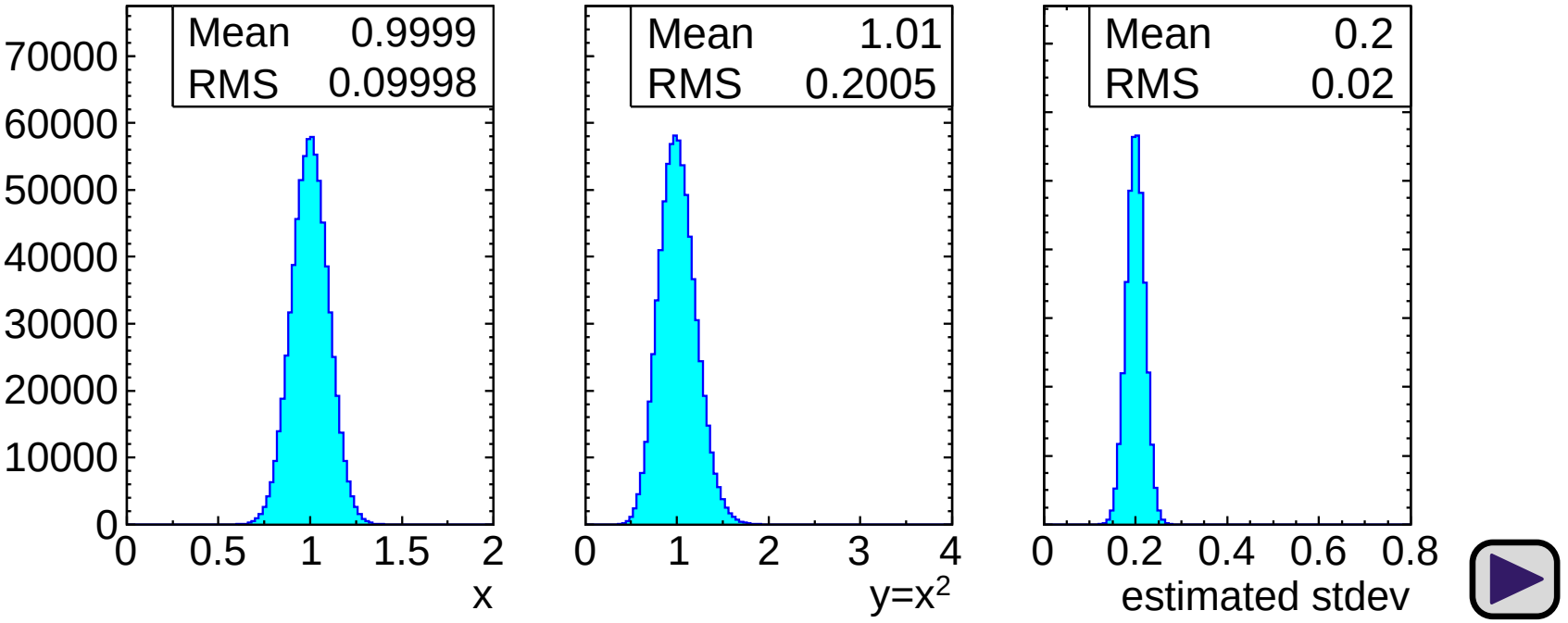
When chaining transformations one has:

$$\vec{y} = \vec{h}(\vec{g}(\vec{x})) \quad \text{and} \quad M_{ij} = \sum_{k=1}^{n} \frac{\partial h_i}{\partial g_k} \frac{\partial g_k}{\partial x_j} \quad \text{or} \quad M = M(g) \cdot M(x) \; .$$

Gaussian error propagation is consistent. The final covariance matrix is the same, if a transformation is done in one or in several steps.

Numerical studies (i)

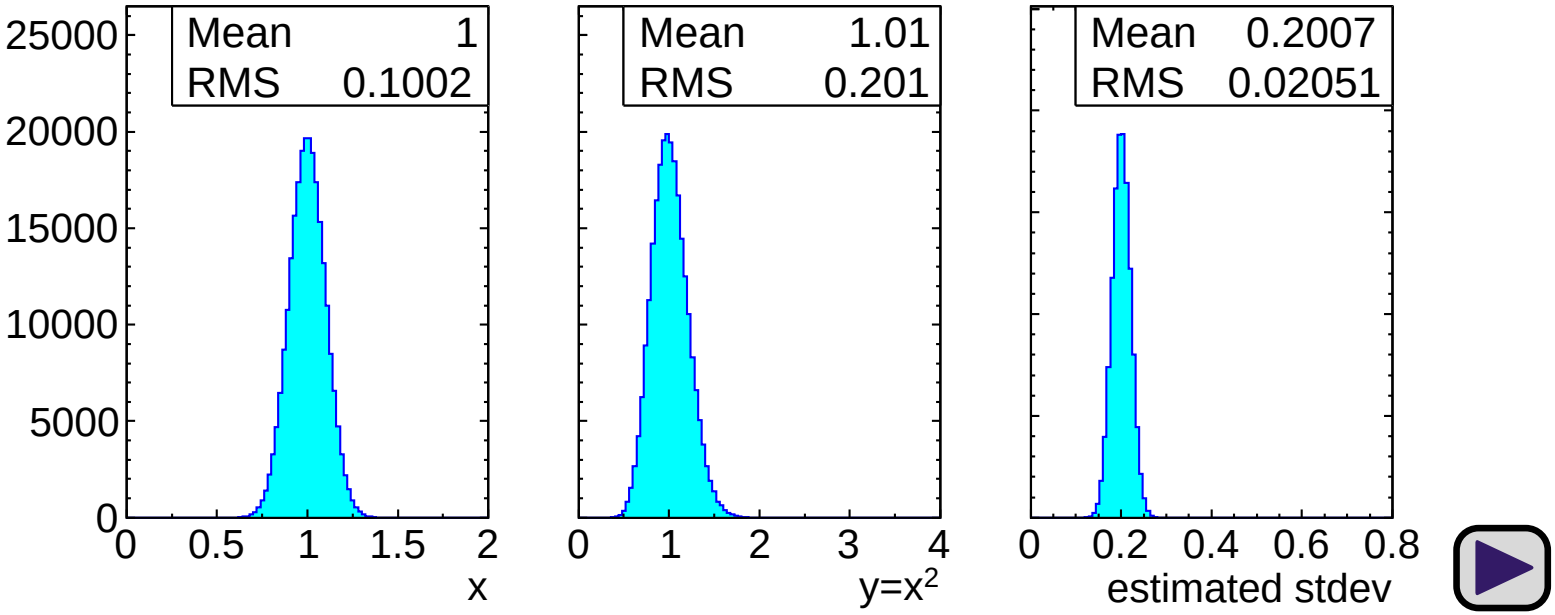
➜ *estimated and exact standard deviations for* $x \to y = x^n$



- average error estimates are OK
- actual values scatter proportional to relative errors of x

➜ *error MC and exact standard deviations for $x \to y = x^n$*

Fluctuate every measured value x by its known variance and estimate the standard deviation of y from the transformed x-values.



- similar behaviour as analytical results (slightly larger scatter)
- easy to implement as no derivatives are required
- small sensitivity to PDF of fluctuations

Summary up to this point

➜ *when uncertainties are quantified by the covariance matrix. . .*

- gaussian error propagation is . . .
  - ➜ consistent when chaining transformations
  - ➜ exact for linear transformations
  - ➜ approximate for non-linear transformation
- error propagation via MC is . . .
  - ➜ easy to implement
  - ➜ approximately the same accuracy as gaussian error propagation
- error estimates for non-linear transformation can have relative uncertainties of the same order of magnitude as the measured quantities - even if the variances of the measurements are known!
- non-linear transformation induces a bias
- leading order bias correction is recommended before averaging

Confidence level intervals

➜ *alternative ways to quantify uncertainties*

- no longer distribution-free – the underlying PDFs need to be known
- propagation of uncertainties usually not possible
  - ➜ requires full PDFs or likelihood functions
  - ➜ usually only the intervals are provided
- combination of uncertainties not well defined
  - ➜ common practice:

$$a = 42 \pm^{8}_{3} \pm^{6}_{4} = 42 \pm^{10}_{5}$$

  - ➜ little or no theoretical backing
  - ➜ implies the concept of asymmetric variance
  - ➜ implies that confidence level intervals behave like variances
- different concepts in bayesian and frequentist frameworks

a simple case study ➜

➜ *setting the scene:*

A counting experiment has observed n events. The experiment did counted independent random processes with a constant probability per time interval to happen, such as e.g. radiocative decays. It thus is known that n is a poissonian distributed random variable, i.e. the probability P_n to observe n events is:

$$P_n = P(n;\mu) = e^{-\mu}\ \frac{\mu^n}{n!}$$

➜ *question:*

What can be inferred about the expectation value μ?

Example: $n = 2$

➜ *quick check of a few hypotheses . . .*

- ➜ $P(2; \mu = 0.1) \approx 0.0045$
- ➜ $P(2; \mu = 1.0) \approx 0.1839$
- ➜ $P(2; \mu = 10.) \approx 0.0023$

- in principle any value for μ is possible
- a value $\mu = O(1)$ seems more plausible

❖ try to be quantitative about a certain range of μ

- discuss
  - ➜ the Bayesian approach
  - ➜ the frequentist approach

➜ *treat μ as a random variable*

- formally possible even if μ has a well defined true physical value
- interpret the PDF of μ as encoding the knowledge about μ
- use Bayes' theorem to improve the knowledge by the measurement:

$$P(\mu|n)\,P(n) = P(n|\mu)\,P(\mu)$$

  - ➜ $P(\mu)$: prior PDF of μ - to be defined
  - ➜ $P(n|\mu)$: Likelihood function
  - ➜ $P(n)$: probability for n, unknown constant
  - ➜ $P(\mu|n)$: posterior PDF for μ after the measurement

❖ it follows

$$P(\mu|n) \propto P(n|\mu)\,P(\mu) \quad \text{and thus} \quad P(\mu|n) = \frac{P(n|\mu)\,P(\mu)}{\int d\mu\; P(n|\mu)\,P(\mu)}$$

➜ *choice of prior distribution*

$$P(\mu) = \mu^k$$

- ad hoc - but allows to test sensitivity to prior, special cases:
- $k = 0$: equal probability for all possible values
- $k = -1$ Jeffries prior: invariance w.r.t scale-transformations $\mu \to \alpha\,\mu$

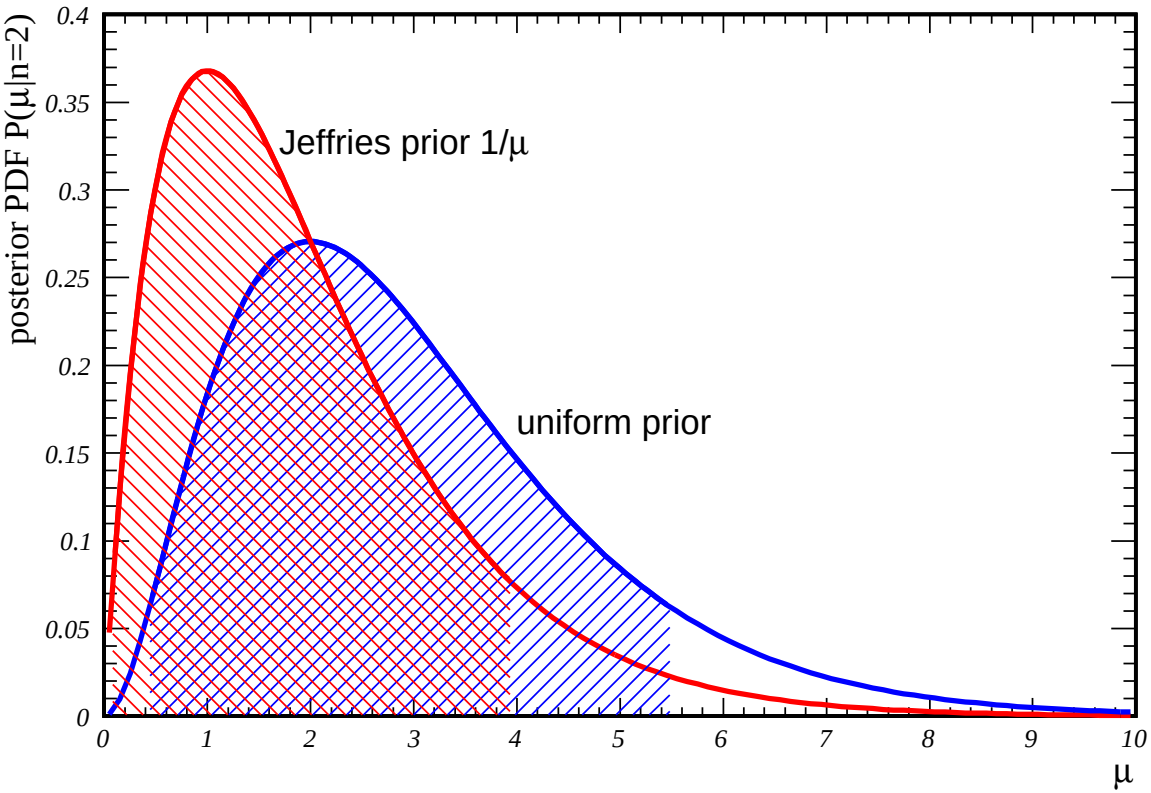
$$P(\mu|n) = \frac{e^{-\mu}\,\mu^{n+k}}{\int_0^\infty d\mu\; e^{-\mu}\,\mu^{n+k}} = e^{-\mu}\,\frac{\mu^{n+k}}{(n+k)!}$$

➜ equal to Poisson-likelihood to observe $n + k$ for given μ

results ➜

Bayesian result

➜ *posterior distributions and 90% CL intervals*



- $X\%$ confidence intervals are regions with $X\%$ probability content
  - ➜ many possibilities - usually take the smallest interval
- most probable values and confidence intervals depend on the prior

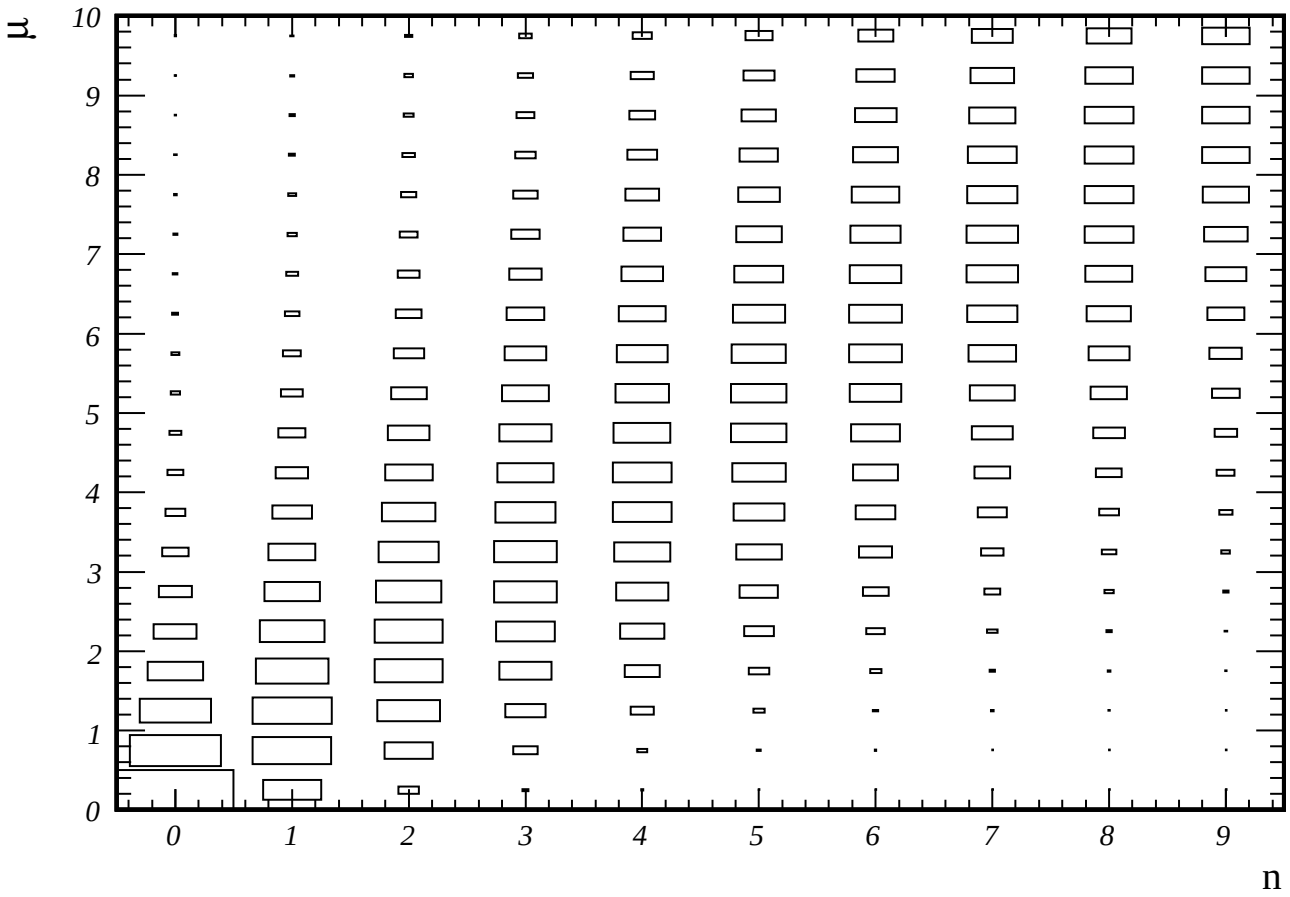
- bayesian approach formalizes gain of knowledge by measurement
  - ➜ posterior of first measurement can be prior of second, etc.

$$\begin{aligned} P(\mu|n_2, n_1) &\propto P(n_2|\mu)P(n_1|\mu)P(\mu) \\ &= P(n_2, n_1|\mu)P(\mu) \\ &= P(n_2|\mu)P_1(\mu) \quad \text{with} \quad P_1(\mu) = P(n_1|\mu)P(\mu) \end{aligned}$$

- consistent if a non-uniform prior (e.g. Jeffries') is used only once
  - ➜ avoid non-uniform priors for single measurements
  - ➜ if needed, use a non-uniform prior once when combining results
  - ➜ possible if likelihood functions are published
- use of uniform priors corresponds to maximum likelihood approach
- caveat: uniformity depends on the definition of the parameter
  - ➜ example: uniform in μ is non-uniform in $\sqrt{\mu}$

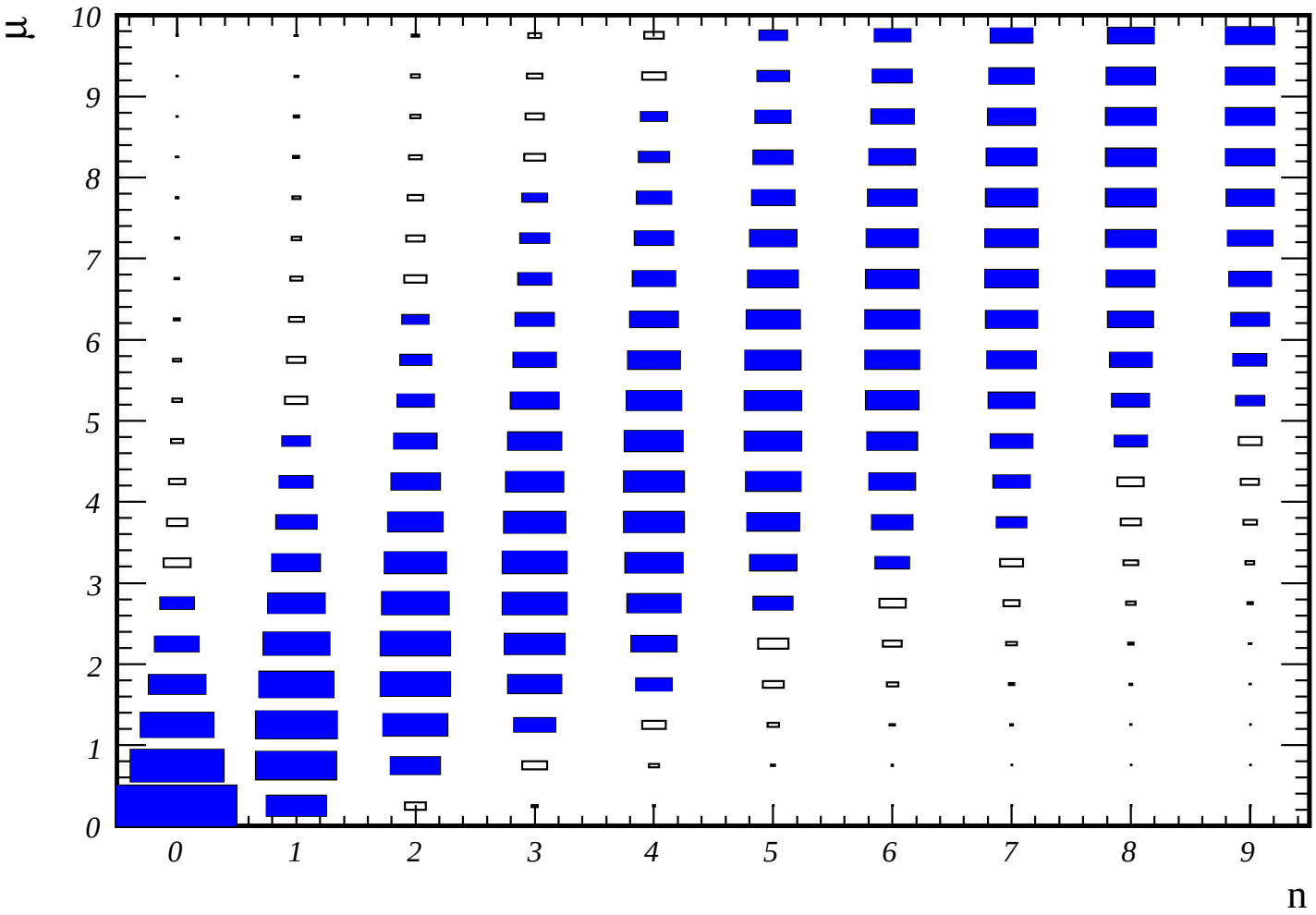
➜ *Likelihood-function-only based "Neyman construction"*

- start from table of probabilities for any observation and any vaue μ



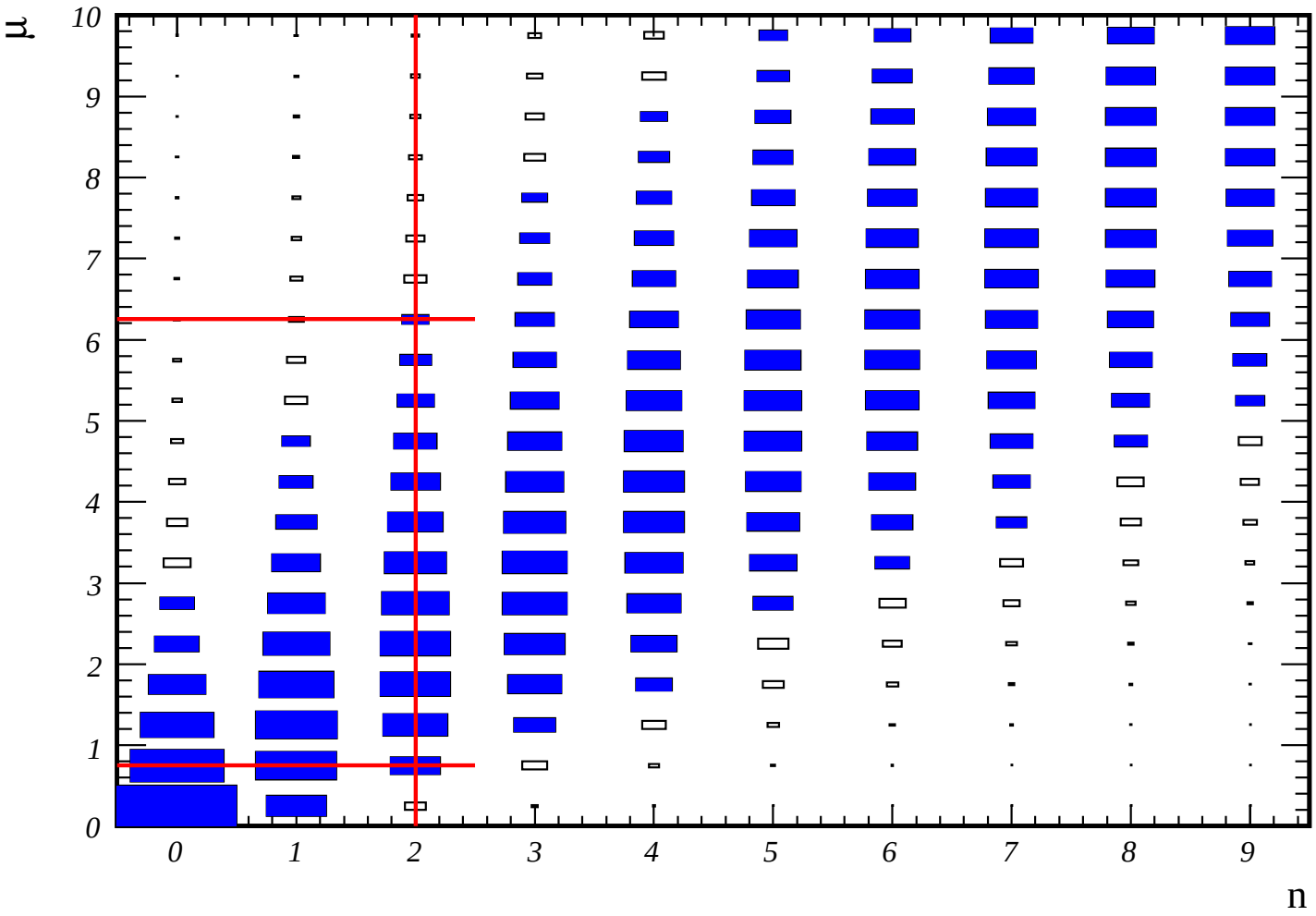
The Neyman construction (i)

- determine the shortest $\geq 90\%$ horizontal range for each μ



The Neyman construction (ii)

- given n, take the range of μ with n in the $\geq 90\%$ probability range



Properties of the Neyman construction

- a fixed interval for μ is assigned to every measurement n
- every interval contains the true value with 90% probability
  - ➜ false from the frequentist point of view – the true value μ is either inside or outside; it is a fixed value and does not depend on n.
- from an ensemble of measurements (at least) 90% of the confidence level intervals are expected to contain the true value
  - ➜ true – for any true μ, different measurements will find different values n and thus will quote different intervals. Take for example $\mu = 4.25$. It is contained in the intervals of $n = 1, \ldots, 7$, and by construction, (at least) 90% of the measurements are in that range. Analogous reasoning holds for all μ.
- the interval contains no information about preferred values!

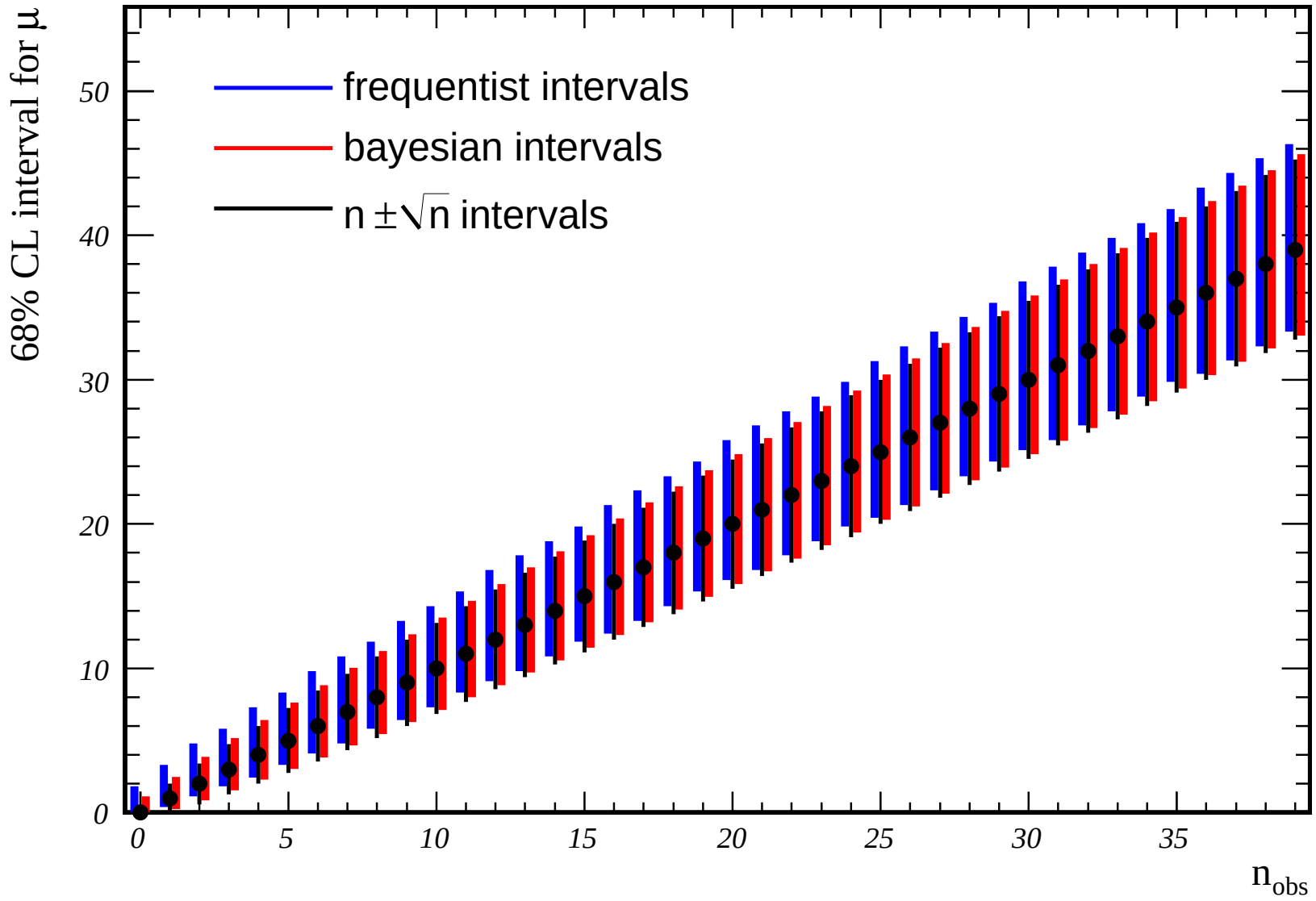
➜ *some common themes. . .*

- bayesian and frequentist methods define regions $[\mu_l(n), \mu_h(n)]$
- for each observation n there is a well defined interval
- another commonly used interval is $n \pm \sqrt{n}$
  - ➜ estimate for the standard deviation of the measurement
  - ➜ often taken also as approximate 68.3% confidence level interval

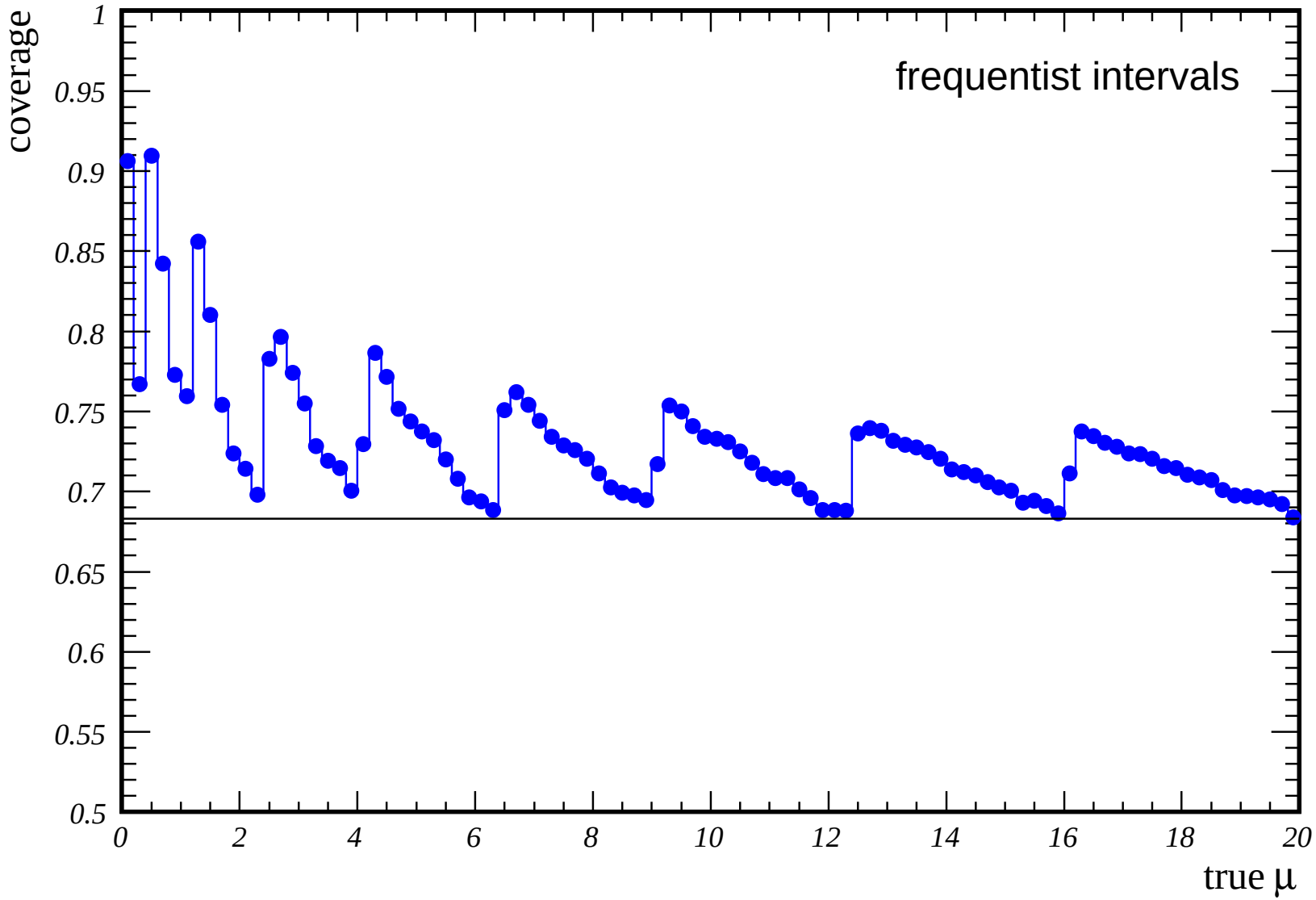
➜ *further studies. . .*

- compare the intervals defined by the different schemes
- MC check which fraction of intervals contain the true value
  - ➜ do the check as a function of the unknown true μ
  - ➜ check that the frequentists intervals have coverage
  - ➜ calculate coverage also for bayesian intervals
    - even if bayesians do not care about coverage . . .

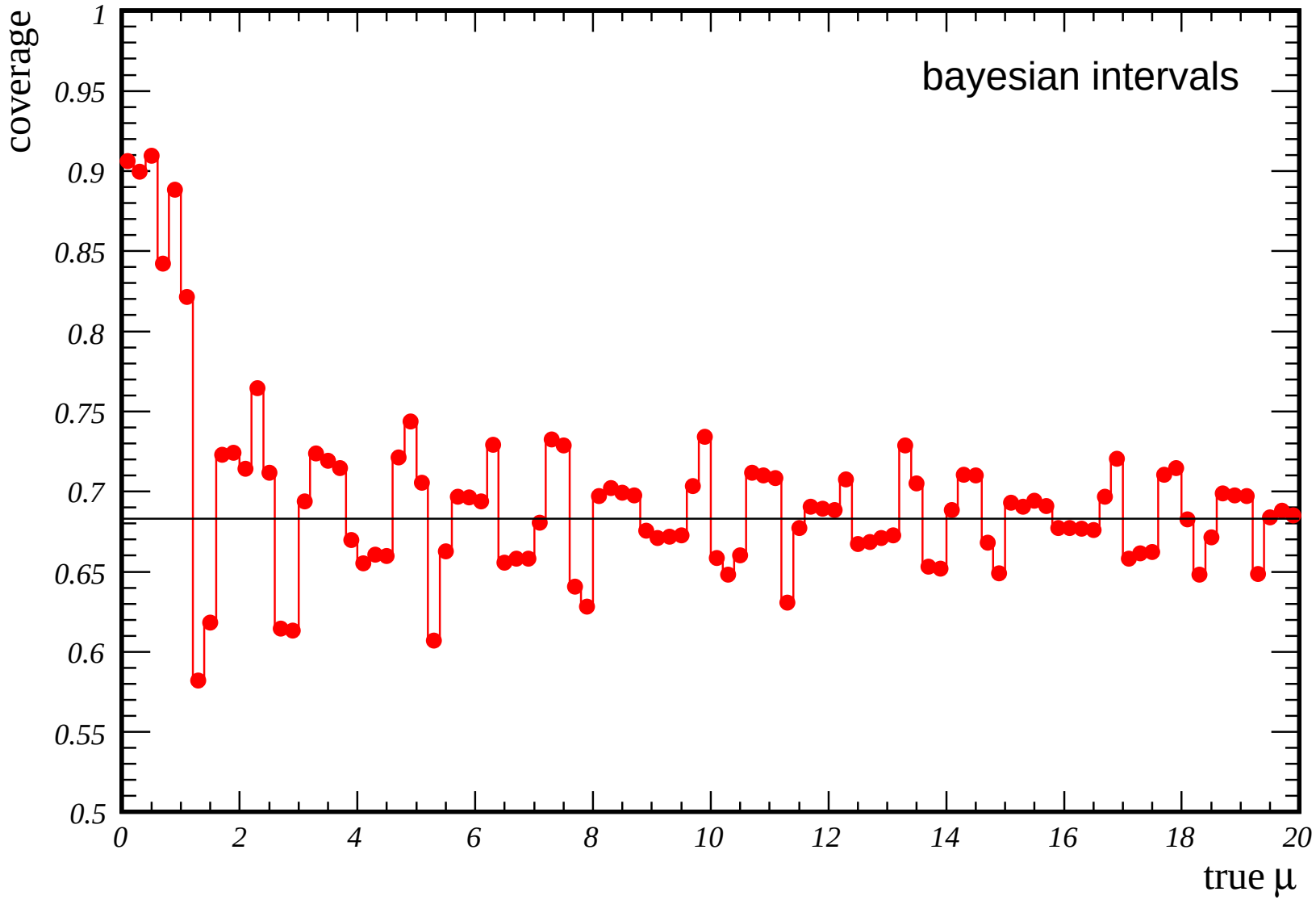
68.3% confidence level intervals



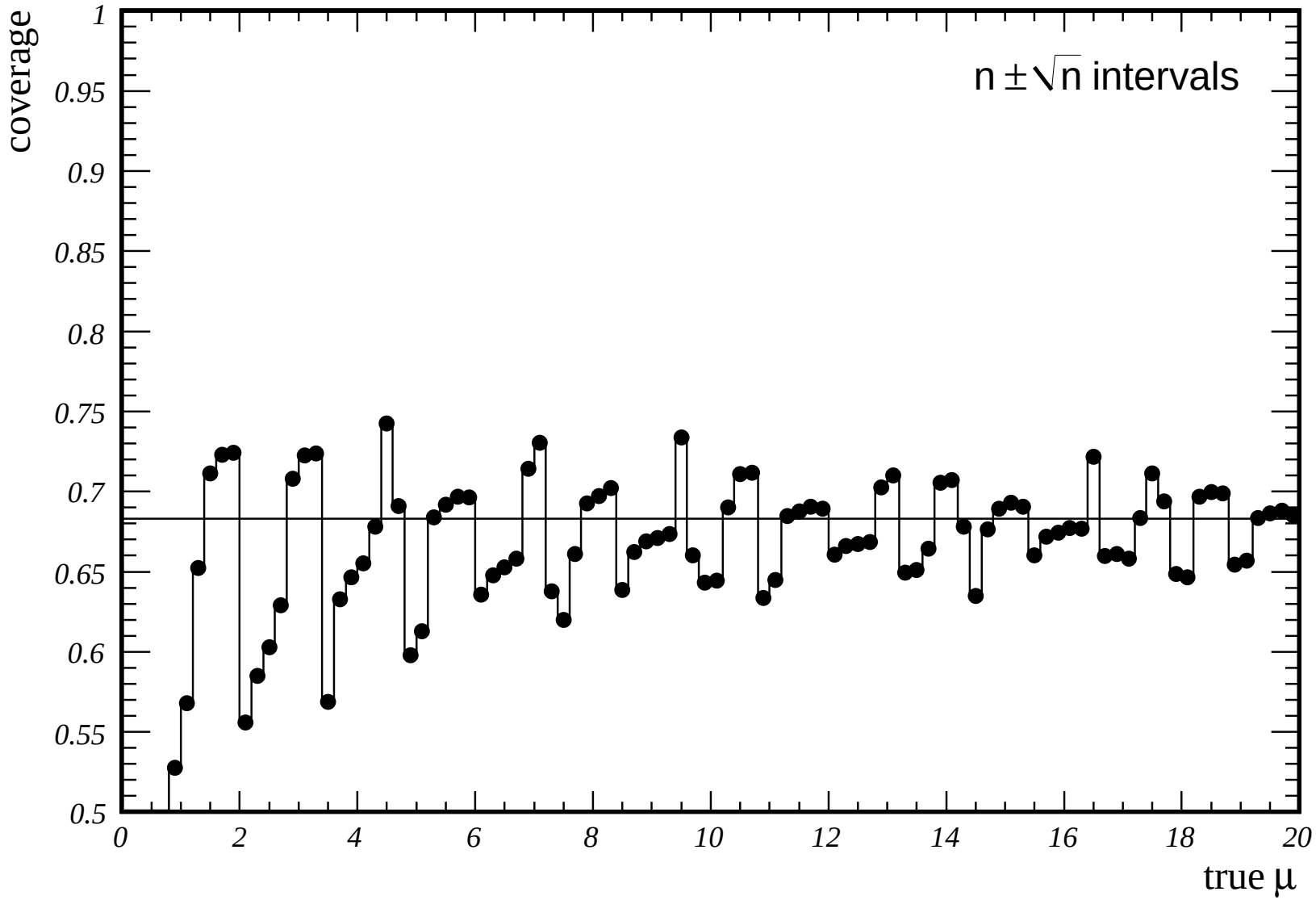
Coverage of 68.3% confidence level intervals



Coverage of 68.3% confidence level intervals



Coverage of 68.3% confidence level intervals



Concluding remarks

- bayesians makes statements about the theory
  - ➜ "The true value μ is with 90% probability inside the 90% confidence level interval"
  - ➜ the conclusion depends on the assumed prior
- frequentists makes statements about the data
  - ➜ "90% of the 90% confidence level intervals are expected to contain the true value μ"
  - ➜ these confidence level intervals have "coverage"
  - ➜ for continuous PDFs exact coverage can be obtained
  - ➜ discrete probabilities are chosen to have over-coverage
- bayesians & frequentists base CL-intervals on the likelihood function
- confidence level intervals from maximum likelihood or least squares fits based on $\Delta\chi^2$ or $\Delta \ln L$ are exact only for gaussian PDFs. In most cases they don't have coverage.
- treating confidence level intervals like variances is questionable

➜ *extract physics parameters from a set of measurements*

❖ properties which are assumed to be satisfied:

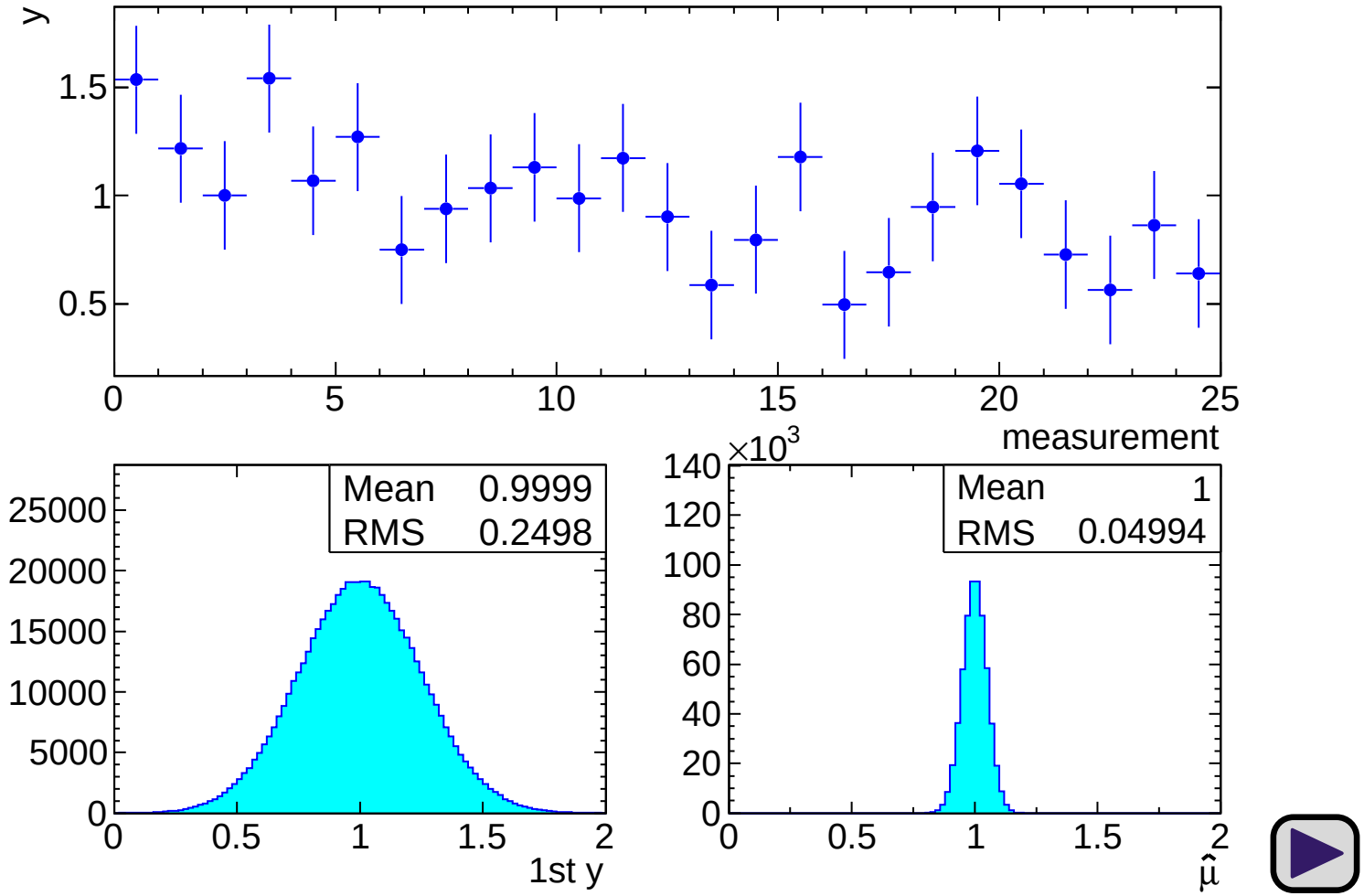
- individual measurements fluctuate with known variance
- individual measurements are unbiased

➜ *measurements of the same physical quantity*

- scenario
  - ➜ n measurements y_i with $i = 1, 2, \ldots, n$
  - ➜ all measurements fluctuate around an unknown true value μ
  - ➜ all measurements have the standard deviation σ_i
- each measurement is an estimate for μ with uncertainty σ_i
- task: combine the measurements for a better estimate of μ

➜ try the arithmetic average $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Numerical simulation



- big improvements if all variances are the same
- less/no improvement w.r.t. best measurement for different variances

➜ *modification of the arithmetic average*

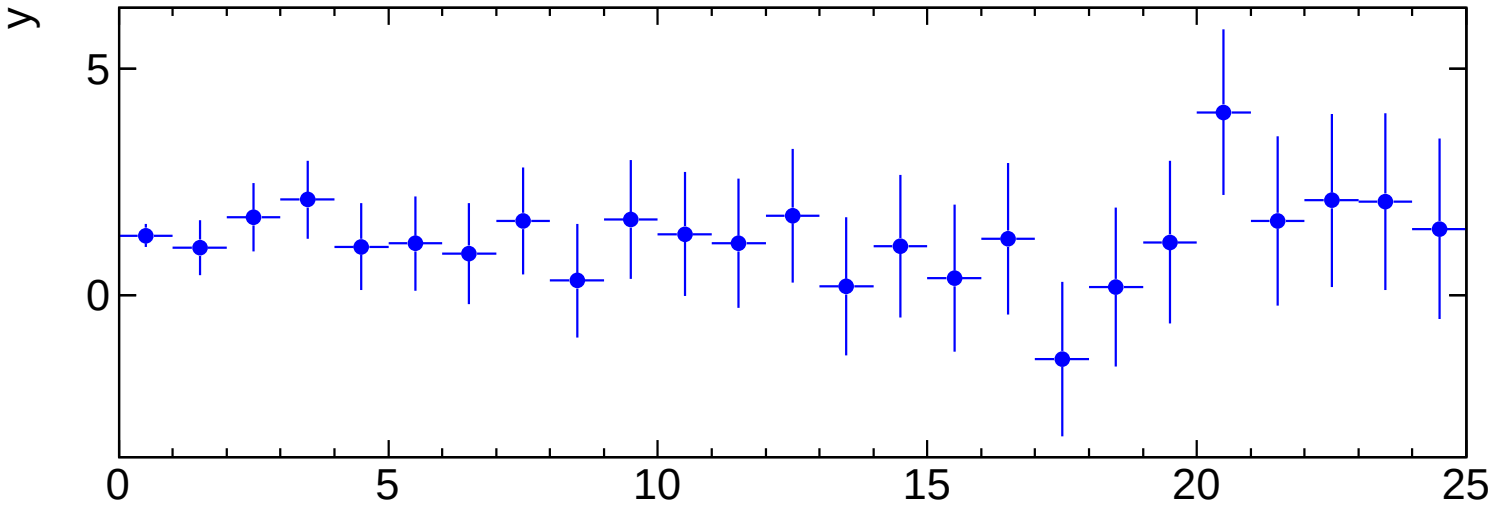
$$\hat{\mu} = \sum_{i=1}^{n} w_i y_i \qquad \text{with} \qquad \sum_{i=1}^{n} w_i = 1$$

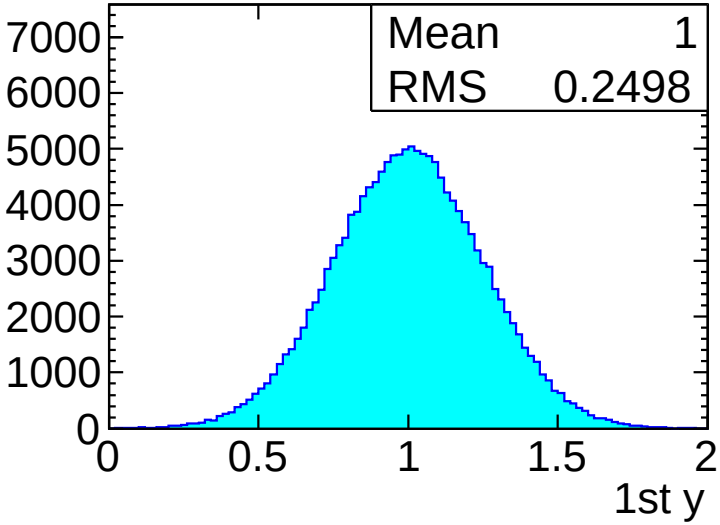
- consistent results for arbitrary weights: $\hat{\mu} = \mu$ if $y_i = \mu$
- try to find weights which minimize the variance of $\hat{\mu}$

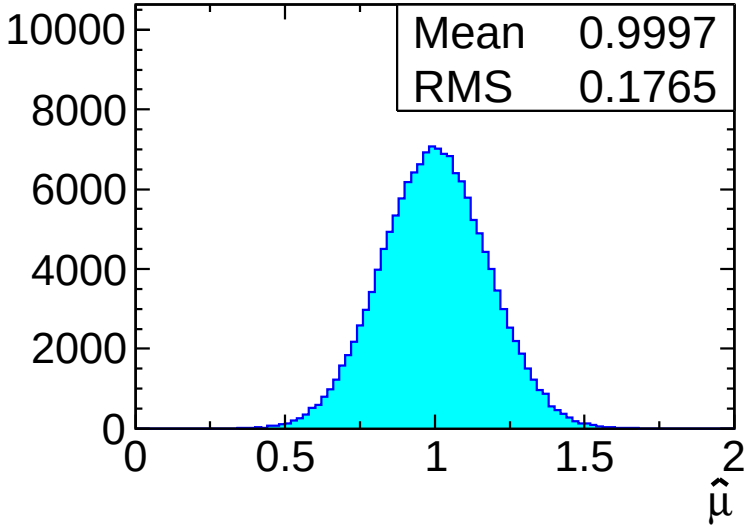
$$\sigma^2(\hat{\mu}) = \sum_{i=1}^{n} w_i^2 \sigma_i^2 \stackrel{!}{=} \min$$

- constrained minimization problem
- minimum for $w_i \propto 1/\sigma_i^2$
- recovers unweighted average if all σ_i are the same

Numerical simulation







➜ *use case: straight line fit*

Consider uncorrelated measurements $y_i, i = 1, \ldots, n$ with known variances σ_i^2, recorded for certain values x_i of a control paramater x. The expectation value of the measurements is $\langle y_i \rangle = a_0 + a_1 x_i$, where the parameters a_0 and a_1 are not known.

➜ *wanted: a method to find estimates $\hat{a}_0$ and $\hat{a}_1$ for a_0 and a_1*

❖ discussion

- control parameters x_i are known
- the measurements y_i are unbiased
- variances σ_i^2 are known
- exact shape of PDFs describing the fluctuations of the y_i is irrelevant
  - ➜ any PDF with variance σ_i^2 would do
  - ➜ different measurements can fluctuate with different PDFs

➜ *the case of two measurements*

$$\langle y_1 \rangle = a_0 + a_1\, x_1 \quad \text{and} \quad y_1 = \langle y_1 \rangle + r_1$$
$$\langle y_2 \rangle = a_0 + a_1\, x_2 \quad \text{and} \quad y_2 = \langle y_2 \rangle + r_2$$

- system of linear equations relating $\langle y_i \rangle$ and x_i
- measurements y_i have random deviation r_i from $\langle y_i \rangle$
- unbiasedness of y_i implies $\langle r_i \rangle = 0$
- estimate a_0 and a_1 by assuming $r_i = 0$, i.e. make the ansatz:

$$y_1 = \hat{a}_0 + \hat{a}_1\, x_1$$
$$y_2 = \hat{a}_0 + \hat{a}_1\, x_2$$

❖ result:

$$\hat{a}_0 = y_1 - \hat{a}_1 x_1 = \frac{x_2}{x_2 - x_1}\, y_1 - \frac{x_1}{x_2 - x_1}\, y_2$$
$$\hat{a}_1 = \frac{y_2 - y_1}{x_2 - x_1} = -\frac{1}{x_2 - x_1}\, y_1 + \frac{1}{x_2 - x_1}\, y_2$$

➜ *does the estimate make sense?*

- parameter estimates are linear combinations of the measurements
- parameter estimates are random variables
- parameter estimates fluctuate with the measurements
- check the expectation values . . .

$$\langle \hat{a}_0 \rangle = \left\langle \frac{1}{x_2 - x_1}(x_2 y_1 - x_1 y_2) \right\rangle = \frac{1}{x_2 - x_1}(x_2 \langle y_1 \rangle - x_1 \langle y_2 \rangle) = a_0$$

$$\langle \hat{a}_1 \rangle = \left\langle \frac{1}{x_2 - x_1}(-y_1 + y_2) \right\rangle = \frac{1}{x_2 - x_1}(-\langle y_1 \rangle + \langle y_2 \rangle) = a_1$$

❖ conclusion:

➜ the estimates for the unknown parameters are unbiased

➜ the parameter errors can be determined by error propagation

yes, the parameter estimates make sense!

➜ *the case of $n > 2$ measurements*

Take the lessons learnt from the case $n = 2$ and try to estimate the unknown parameters by a linear combination of the measurements.

$$\hat{a}_0 = \sum_{i=1}^{n} p_i \, y_i \quad \text{and} \quad \hat{a}_1 = \sum_{i=1}^{n} q_i \, y_i$$

- this is a convenient ansatz, not derived from any “first principles”
- it is not the only possible generalization of the case $n = 2$
- nor will it give the best possible estimates for a_0 and a_1
- but it is simple and robust, requiring only minimal input
- and turns out to be surprisingly powerful . . .

➜ determine parameters p_i and q_i . . .

Optimizing the parameter estimates

➜ *exploit the freedom of the linear ansatz to. . .*

- make sure that the estimates are unbiased
- and that the estimates are as accurate as possible

❖ condition for unbiased estimates:

$$\langle \hat{a}_0 \rangle = \sum_{i=1}^{n} p_i \langle y_i \rangle = \sum_{i=1}^{n} p_i (a_0 + a_1 x_i) = a_0 \sum_{i=1}^{n} p_i + a_1 \sum_{i=1}^{n} p_i x_i \stackrel{!}{=} a_0$$

$$\langle \hat{a}_1 \rangle = \sum_{i=1}^{n} q_i \langle y_i \rangle = \sum_{i=1}^{n} q_i (a_0 + a_1 x_i) = a_0 \sum_{i=1}^{n} q_i + a_1 \sum_{i=1}^{n} q_i x_i \stackrel{!}{=} a_1$$

one obtains 4 conditions:

$$\sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} q_i = 0 \qquad \sum_{i=1}^{n} p_i x_i = 0 \qquad \sum_{i=1}^{n} q_i x_i = 1$$

Discussion

- only 4 constraints for $2n$ parameters
- easy to satisfy both for p_i and q_i
  - ➜ start from a set of random numbers e.g. for p_i
  - ➜ subtract a constant such that the “0-constraint” is satisfied
  - ➜ scale the numbers such that the “1-constraint” is satisfied
- additional criterion needed to fix the coefficients
- require minimal variance for the parameter estimates
  - ➜ constrained minimization problem

❖ variance of parameter estimates from error propagation:

$$\sigma^2(\hat{a}_0) = \sum_{i=1}^{n} \left(\frac{\partial \hat{a}_0}{\partial y_i} \right)^2 \sigma_i^2 = \sum_{i=1}^{n} p_i^2 \, \sigma_i^2 \quad \text{and} \quad \sigma^2(\hat{a}_1) = \sum_{i=1}^{n} q_i^2 \, \sigma_i^2$$

➜ constrained minimization

Fixing the coefficients p_i

➜ *minimization using Lagrange multipliers for the constraints*

$$\sum_{i=1}^{n} p_i^2\,\sigma_i^2 + 2\alpha_0\left(1-\sum_{i=1}^{n} p_i\right) + 2\beta_0\left(-\sum_{i=1}^{n} p_i\,x_i\right) \stackrel{!}{=} \min$$

requiring zero derivatives with respect to p_i then yields:

$$2p_i\,\sigma_i^2 - 2\alpha_0 - 2\beta_0\,x_i = 0 \quad \Rightarrow \quad p_i = \frac{1}{\sigma_i^2}(\alpha_0 + \beta_0 x_i)$$

α_0 and β_0 follow from the constraint to have unbiased estimates:

$$\sum_{i=1}^{n} p_i = \alpha_0 \sum_{i=1}^{n} \frac{1}{\sigma_i^2} + \beta_0 \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} = \alpha_0\,S_1 + \beta_0\,S_x \;= 1$$

$$\sum_{i=1}^{n} p_i\,x_i = \alpha_0 \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} + \beta_0 \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} = \alpha_0\,S_x + \beta_0\,S_{xx} = 0$$

Fixing the coefficients q_i

➜ *minimization using Lagrange multipliers for the constraints*

$$\sum_{i=1}^{n} q_i^2\,\sigma_i^2 + 2\alpha_1\left(-\sum_{i=1}^{n} q_i\right) + 2\beta_1\left(1-\sum_{i=1}^{n} q_i x_i\right) \stackrel{!}{=} \min$$

requiring zero derivatives with respect to q_i then yields:

$$2\,q_i\,\sigma_i^2 - 2\alpha_1 - 2\beta_1 x_i = 0 \quad \rightarrow \quad q_i = \frac{1}{\sigma_i^2}(\alpha_1 + \beta_1 x_i)$$

α_1 and β_1 follow from the constraint to have unbiased estimates:

$$\sum_{i=1}^{n} q_i = \alpha_1 \sum_{i=1}^{n} \frac{1}{\sigma_i^2} + \beta_1 \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} = \alpha_1\,S_1 + \beta_1\,S_x \;= 0$$

$$\sum_{i=1}^{n} q_i\,x_i = \alpha_1 \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} + \beta_1 \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} = \alpha_1\,S_x + \beta_1\,S_{xx} = 1$$

Putting it all together

Solving the linear equations for the Lagrange parameters $\alpha_{\{0,1\}}$ and $\beta_{\{0,1\}}$

$$\begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 & \alpha_1 \\ \beta_0 & \beta_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and substituting the results into p_i, q_i, with $D = S_1 S_{xx} - S_x^2$, yields

$$p_i = \frac{1}{\sigma_i^2}(\alpha_0 + \beta_0 x_i) = \frac{1}{D}\left(S_{xx}\frac{1}{\sigma_i^2} - S_x\frac{x_i}{\sigma_i^2}\right)$$

$$q_i = \frac{1}{\sigma_i^2}(\alpha_1 + \beta_1 x_i) = \frac{1}{D}\left(-S_x\frac{1}{\sigma_i^2} + S_1\frac{x_i}{\sigma_i^2}\right)$$

and thus

$$\hat{a}_0 = \frac{1}{D}(S_{xx}S_y - S_x S_{xy}) \quad \text{and} \quad \hat{a}_1 = \frac{1}{D}(S_1 S_{xy} - S_x S_y)$$

where

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

➜ *linear error propagation*

$$C_{kl}(\hat{a}) = \sum_{i=1}^{n} \frac{\partial \hat{a}_k}{\partial y_i} \frac{\partial \hat{a}_l}{\partial y_i} \sigma_i^2$$

yields

$$C_{00}(\hat{a}) = \sum_{i=1}^{n} p_i^2 \sigma_i^2 = \frac{S_1}{D^2}(S_{xx} S_1 - S_x^2) = \frac{S_1}{D}$$

$$C_{11}(\hat{a}) = \sum_{i=1}^{n} q_i^2 \sigma_i^2 = \frac{S_{xx}}{D^2}(S_{xx} S_1 - S_x^2) = \frac{S_{xx}}{D}$$

$$C_{01}(\hat{a}) = \sum_{i=1}^{n} p_i q_i \sigma_i^2 = \frac{-S_x}{D^2}(S_{xx} S_1 - S_x^2) = \frac{-S_x}{D}$$

. . . the well known textbook formulae for straight line fits.

➜ *re-write the solution derived before. . .*

$$\hat{a}_0 = \frac{1}{D}(S_{xx}S_y - S_x S_{xy}) \quad \text{and} \quad \hat{a}_1 = \frac{1}{D}(S_1 S_{xy} - S_x S_y)$$

to make the structure more evident:

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \frac{1}{D}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \cdot \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} \;\rightarrow\; \begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \cdot \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

or

$$S_1\ \hat{a}_0 + S_x\ \hat{a}_1 - S_y\ = 0$$
$$S_x\ \hat{a}_0 + S_{xx}\ \hat{a}_1 - S_{xy} = 0$$

i.e. two equations which define the best fit parameters as the zero of a two-dimensional function. Now exploit the fact that it's always possible to interpret the zero of a function as a stationary point (e.g. minimum) of its primitive.

p.t.o. ➜

➜ *introducing $F(a_0, a_1)$ such that*

$$\left.\frac{\partial F}{\partial a_0}\right|_{\{a_0,a_1\}=\{\hat{a}_0,\hat{a}_1\}} = 0 \quad \text{and} \quad \left.\frac{\partial F}{\partial a_1}\right|_{\{a_0,a_1\}=\{\hat{a}_0,\hat{a}_1\}} = 0$$

it follows (from dimensional considerations)

$$\frac{\partial F}{\partial a_0} = S_1\, a_0 + S_x\, a_1 - S_y \quad \text{and} \quad \frac{\partial F}{\partial a_1} = S_x\, a_0 + S_{xx}\, a_1 - S_{xy}\ .$$

Integration of the first equation yields

$$F = \frac{1}{2} S_1\, a_0^2 + S_x\, a_0\, a_1 - a_0\, S_y + g(a_1)$$

where $g(a_1)$ does not depend on a_0. Taking the derivative with respect to a_1 and comparing with the known derivative determines $g'(a_1)$:

$$\frac{\partial F}{\partial a_1} = S_x\, a_0 + g'(a_1) = S_x\, a_0 + S_{xx}\, a_1 - S_{xy}$$

p.t.o. ➜

It follows

$$g'(a_1) = S_{xx}\,a_1 - S_{xy} \quad \text{and thus} \quad g(a_1) = \frac{1}{2}S_{xx}\,a_1^2 - S_{xy}\,a_1 + \frac{C}{2}$$

with an arbitrary constant C. Asking $F_{\min} = 0$ yields $C = \sum y_i^2/\sigma_i^2$ and

$$\begin{aligned} 2F &= S_1\,a_0^2 + S_{xx}\,a_1^2 + 2S_x\,a_0\,a_1 - 2S_y\,a_0 - 2S_{xy}\,a_1 + C \\ &= \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(a_0^2 + a_1^2 x_i^2 + 2a_0\,a_1\,x_i - 2a_0\,y_i + 2a_1\,x_i\,y_i + y_i^2) \\ &= \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1\,x_i)^2}{\sigma_i^2}\,, \end{aligned}$$

and setting $2F = \chi^2$, the cost-function becomes

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f_i(a_0, a_1))^2}{\sigma_i^2} \quad \text{with} \quad f_i(a_0, a_1) = a_0 + a_1\,x_i\,.$$

Discussion

- the best parameter estimates minimize the distance between data and model, measured in units of standard deviations
- the derivation was for uncorrelated data points y_i
- general expression, also for correlated data, using $1/\sigma_i^2 = C_{ii}^{-1}$:

$$\chi^2 = \sum_{i,j=1}^{n} (y_i - f_i(a_0, a_1)) \, (y_j - f_j(a_0, a_1)) \, C_{ij}^{-1}$$

or in matrix notation

$$\chi^2 = \vec{r}^{\,T} \; C^{-1} \; \vec{r} \qquad \text{with} \qquad \vec{r} = \vec{y} - \vec{f}(a_0, a_1)$$

➜ *Invariance under linear transformations M:*

$$\vec{r}\,' = M \; \vec{r} \quad , \quad C' = M \; C \; M^T \quad , \quad C'^{-1} = (M^T)^{-1} \; C^{-1} \; M^{-1}$$

and thus $\qquad (\chi^2)' = \chi^2$

➜ *(average) measurements are linear functions of parameters $\vec{a}$*

$$\begin{aligned}\chi^2 &= (\vec{y} - M\,\vec{a})^T\ C^{-1}\,(\vec{y} - M\,\vec{a}) \\ &= \vec{y}^{\,T} C^{-1} \vec{y} - 2\vec{a}^{\,T}\left[M^T C^{-1}\vec{y}\right] + \vec{a}^{\,T}\left[M^T C^{-1} M\right]\vec{a}\end{aligned}$$

minimization:

$$\frac{\partial \chi^2}{\partial \vec{a}} = -2\left[M^T C^{-1}\vec{y}\right] + 2\left[M^T C^{-1} M\right]\vec{a} = 0$$

result: the best fit parameters are linear functions of the measurements

$$\vec{a} = Q\ \vec{y} \qquad \text{with} \qquad Q = \left[M^T C^{-1} M\right]^{-1} M^T C^{-1}$$

with covariance matrix

$$C(a) = Q\ C\ Q^T = \left[M^T C^{-1} M\right]^{-1} = \left(\frac{1}{2}\frac{\partial \chi^2}{\partial \vec{a}^2}\right)^{-1}$$

- unbiased parameter estimates (for any constant matrix C^{-1})

$$\langle \vec{y} \rangle = M \vec{a}_{\text{true}} \quad \rightarrow \quad \langle \vec{a} \rangle = \left[M^T C^{-1} M \right]^{-1} M^T C^{-1} \langle \vec{y} \rangle = \vec{a}_{\text{true}}$$

- minimum χ^2 value

$$\begin{aligned} \chi^2_{\text{min}} &= \vec{y}^T C^{-1} \vec{y} - \vec{a}^T \left[M^T C^{-1} \vec{y} \right] \\ &= \vec{y}^T C^{-1} \vec{y} - \vec{a}^T \left[M^T C^{-1} M \right] \vec{a} \\ &= \text{Tr} \left(C_y^{-1} \vec{y} \vec{y}^T - C_a^{-1} \vec{a} \vec{a}^T \right) \end{aligned}$$

- expectation value χ^2_{min}, using $C_x = \left\langle \vec{x} \vec{x}^T \right\rangle - \langle \vec{x} \rangle \langle \vec{x} \rangle^T$

$$\begin{aligned} \left\langle \chi^2_{\text{min}} \right\rangle &= \text{Tr} \left(C_y^{-1} (C_y + \langle \vec{y} \rangle \langle \vec{y} \rangle^T) - C_a^{-1} (C_a + \langle \vec{a} \rangle \langle \vec{a} \rangle^T \right) \\ &= n_y - n_a + \text{Tr} \left(C_y^{-1} \langle \vec{y} \rangle \langle \vec{y} \rangle^T - C_a^{-1} \langle \vec{a} \rangle \langle \vec{a} \rangle^T \right) = n_y - n_a \end{aligned}$$

The last step follows from $C_a^{-1} = M^T C_y^{-1} M$ and $M \langle \vec{a} \rangle = \langle \vec{y} \rangle$.

Property list of the least squares method

- formulation via the cost function . . .
  - ➜ derived for linear models and explains the name “least squares”
  - ➜ easily generalizes to multi-dimensional and non-linear problems
- least squares are a distribution-free way for parameter estimates
  - ➜ requires only data and covariance matrix of the data
  - ➜ weight matrix C^{-1} must be fixed
  - ➜ approximately gaussian errors due to the central limit theorem
- for linear models
  - ➜ unbiased estimates of the true parameters
  - ➜ parameter estimates are linear combinations of the measurements
- when using the inverse of the covariance matrix as weight matrix
  - ➜ linear estimates with minimal variance
  - ➜ independent of the shape of the PDF of the fluctuations
  - ➜ $\left\langle \chi^2_{\mathrm{min}} \right\rangle = N_{\mathrm{data}} - N_{\mathrm{par}} \equiv N_{\mathrm{ndf}}$
    - ➢ can be used to judge goodness of fit or estimate size of variances

Numerical example

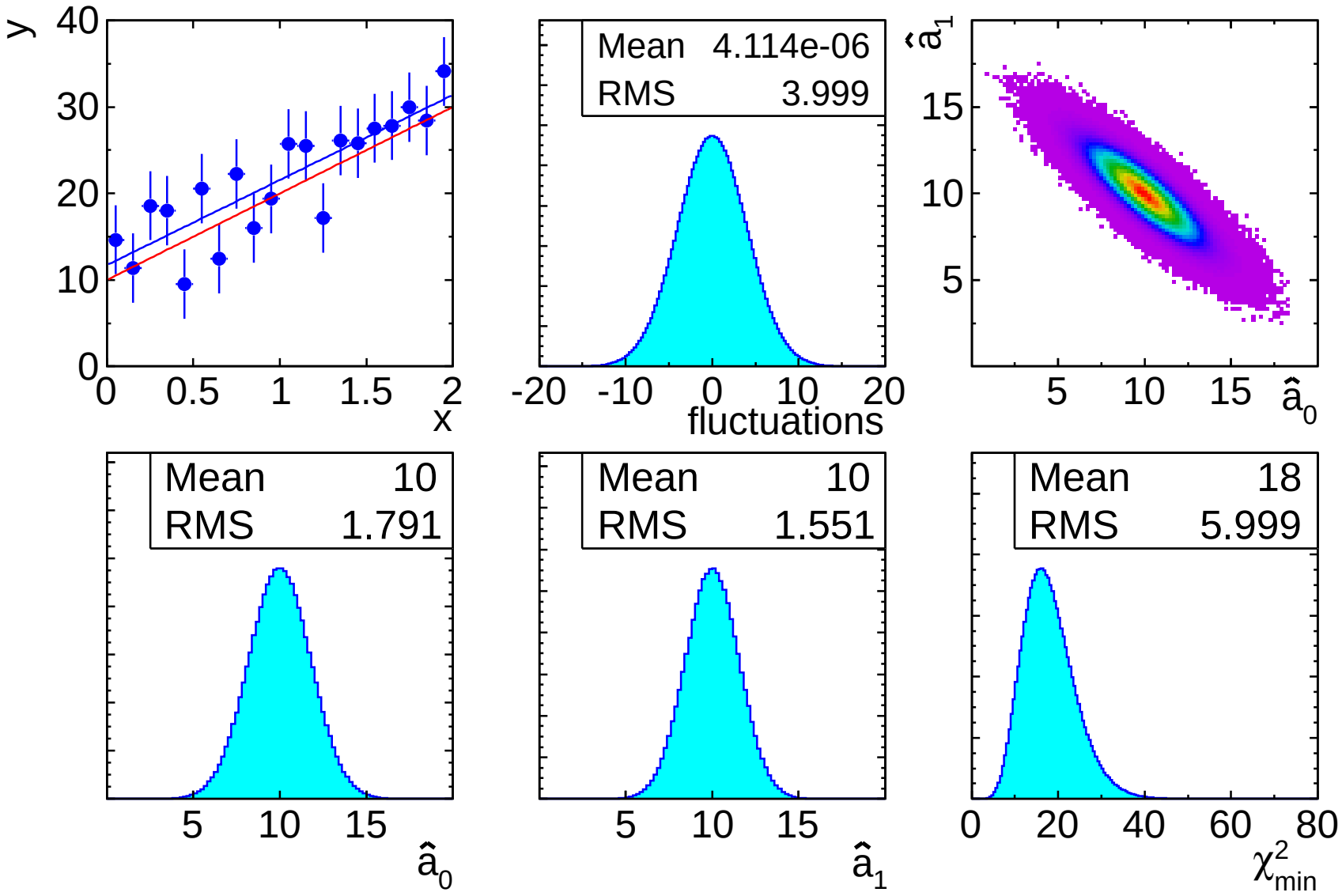
➜ *straight line fit:* $y = a_0 + a_1 x$

- expectation values of measurements $y(x)$: $\langle y \rangle = 10 + 10\,x$
- take 20 equidistant points in the range $0 < x < 2$
- measurements fluctuate with rms$= 4$ around the expectation value
  - ➜ gaussian distribution
  - ➜ exponential distribution
  - ➜ uniform distribution
- same covariance matrix and $\left\langle \chi^2_{\text{min}} \right\rangle = 18$ in all cases

$$C(a) \approx \begin{pmatrix} 3.206 & -2.406 \\ -2.406 & 2.406 \end{pmatrix} \qquad \begin{matrix} \sigma(a_0) \approx 1.7905 \\ \sigma(a_1) \approx 1.5511 \end{matrix} \qquad \rho \approx -0.8663$$

- study also poisson distributed measurements. . .
  - ➜ fit with correct standard deviations: $\sqrt{\langle y \rangle}$
  - ➜ fit with estimated standard deviations: $\sqrt{y}$

Illustration



➜ *exploring the least squares approach*

Given: measurements y_i with known variances σ_i^2, a parametric model $f_i(a)$, and positive weights $w_i > 0$. Wanted: parameter estimates $\hat{a}$.

Ansatz: $$S^2(a) = \sum_i w_i (y_i - f_i(a))^2 \stackrel{!}{=} \min$$

❖ reminder:

- the best fit $\hat{a}$ makes the model get "as close as possible" to the data
- the weights allow to (de)emphasize selected points
- a priori arbitrary weights are allowed
- for independent measurements the optimal weights are $w_i = 1/\sigma_i^2$

study an analytically solvable problem ➜

➜ *problem:*

n poisson distributed values y_i, $i = 1, \ldots, n$, such as measurements from a counting experiments, which are distributed according to the discrete probability distribution

$$p_n(\mu) = e^{-\mu} \frac{\mu^n}{n!}$$

with, for example, actual values

$$y_i = \{2, 2, 5, 2, 3, 3, 1, 3, 3, 2, 3, 2, 10, 2, 3, 1, 2, 6, 4, 3 \ldots\}$$

➜ *solution:*

- least squares fit of a constant
- study different terms for the variance in the χ^2 function

➜ *the ideal χ^2 function*

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - c)^2}{\mu} \quad \rightarrow \quad \hat{c} = \frac{1}{n} \sum_{i=1}^{n} y_i \pm \sqrt{\frac{\mu}{n}}$$

exact properties:

$$\langle \hat{c} \rangle = \mu \qquad \text{and} \qquad \frac{\langle \chi^2_{\text{min}} \rangle}{n-1} = 1$$

remarks:

- parameter estimate by arithmetic average
- ansatz questionable since μ is not known, but. . .
- $\hat{c}$ does not depend on μ, only its uncertainty and χ^2
  - ➜ determine $\hat{c}$ and use $\mu = \hat{c}$ in the χ^2 function

result: $\langle \hat{c} \rangle = \mu$ and $\frac{\langle \chi^2_{\text{min}} \rangle}{n-1} \overset{n \to \infty}{=} 1$

Variance estimate from the fit

➜ *the* `RooFit` *default*

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - c)^2}{c} \quad \rightarrow \quad \hat{c} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2} \pm \sqrt{\frac{\hat{c}}{n}}$$

asymptotic properties:

$$\langle \hat{c} \rangle \overset{n\to\infty}{=} \sqrt{\mu(\mu+1)} \qquad \text{and} \qquad \frac{\langle \chi^2_{\min} \rangle}{n-1} \overset{n\to\infty}{=} 2(\sqrt{\mu(\mu+1)} - \mu)$$

remarks:

- parameter estimate by quadratic average
- non-linear fit model (non-parabolic cost function)
- biased parameter estimate
- biased $\chi^2_{\min}$ values – p-values are of limited use

➜ *alternative* `RooFit` *setting*

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - c)^2}{y_i} \quad \rightarrow \quad \hat{c} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{y_i} \right)^{-1} \pm \sqrt{\frac{\hat{c}}{n}}$$

asymptotic properties:

$$\langle \hat{c} \rangle \overset{n\to\infty}{=} \frac{1}{\langle 1/y \rangle} \quad \text{and} \quad \frac{\langle \chi^2_{\text{min}} \rangle}{n-1} \overset{n\to\infty}{=} \frac{\mu}{1-e^{-\mu}} - \frac{1}{\langle 1/y \rangle} \,,$$

remarks:

- parameter estimate by harmonic average
- necessity to discard values $y_i = 0$
- linear model
- biased parameter estimate
- biased χ^2_{min} values – p-values are of limited use

Biased variances

➜ *avoid discarding zero bins* $z_i = y_i + 1$

$$\chi^2 = \sum_{i=1}^{n} \frac{(z_i - c)^2}{z_i} \quad \rightarrow \quad \hat{c} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{y_i + 1} \right)^{-1} \pm \sqrt{\frac{\hat{c}}{n}}$$

asymptotic properties:

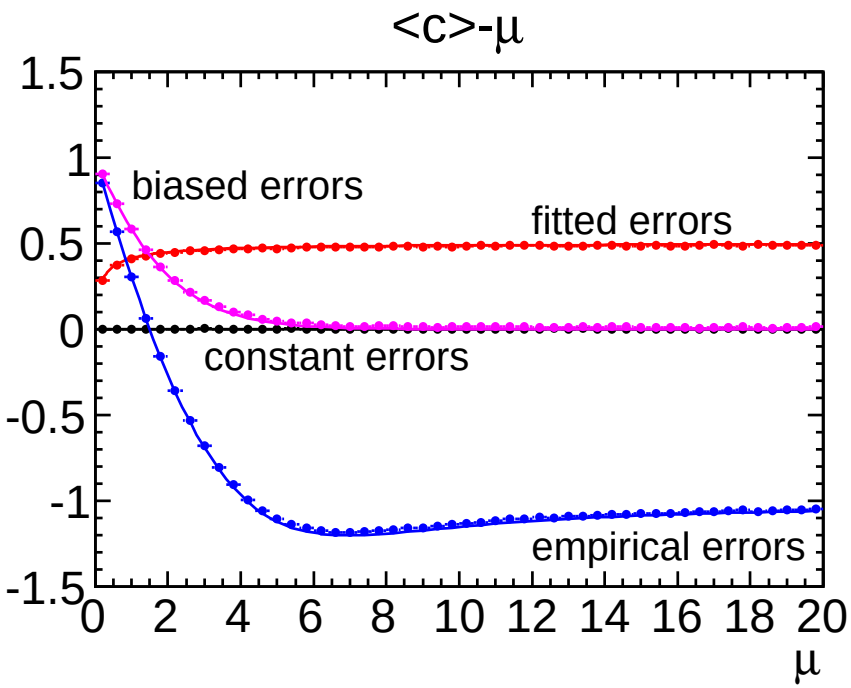
$$\langle c \rangle \overset{n\to\infty}{=} \frac{\mu}{1 - e^{-\mu}} \quad \text{and} \quad \frac{\langle \chi^2_{\min} \rangle}{n-1} \overset{n\to\infty}{=} 1 - \frac{\mu}{e^{\mu} - 1} \, .$$

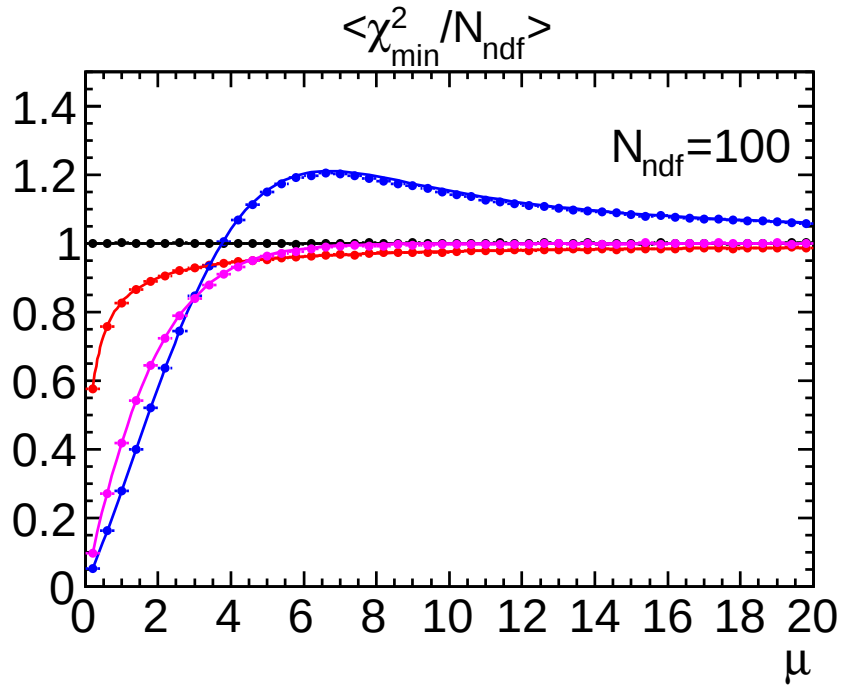
remarks:

- parameter estimate by harmonic average
- allows to include also values $y_i = 0$
- linear model
- asymptotically unbiased parameter estimate
- asymptotically unbiased $\chi^2_{\min}$ values

Numerical studies (i)

➜ *expectation values vs μ for $n = 101$*

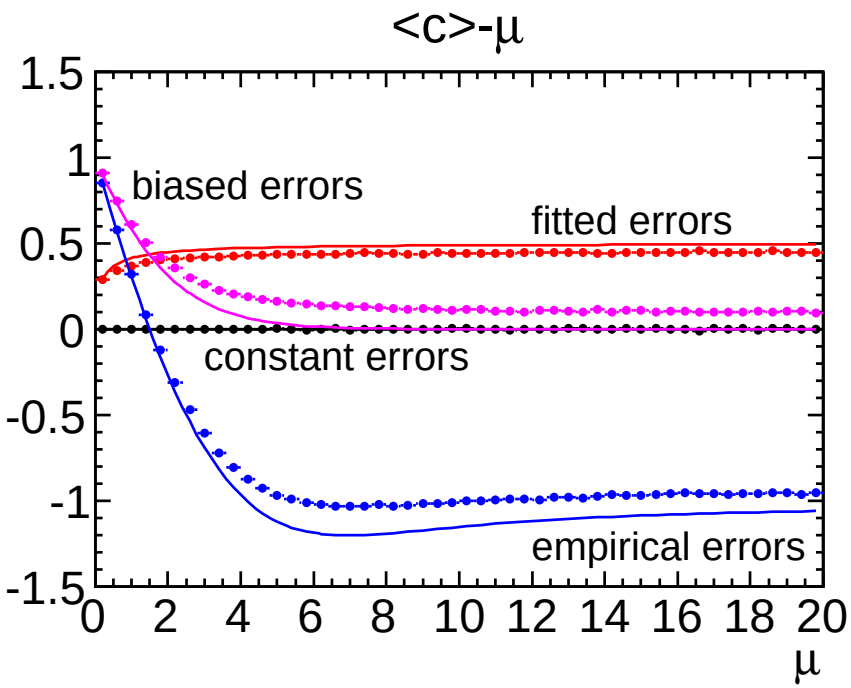


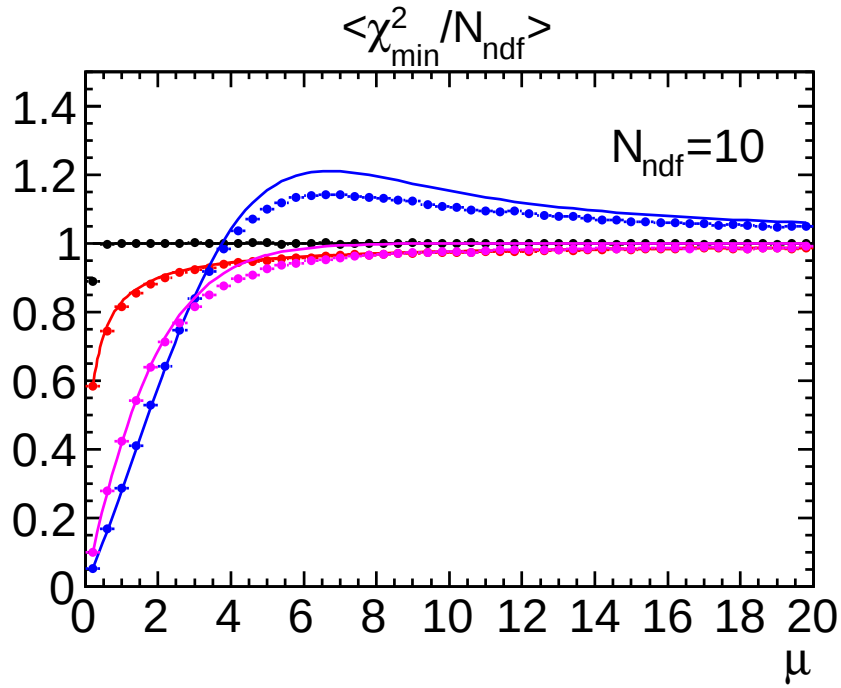


- data points: simulations for $n = 101$ data points
- curves: asymptotic expectations

Numerical studies (ii)

➜ *expectation values vs μ for $n = 11$*

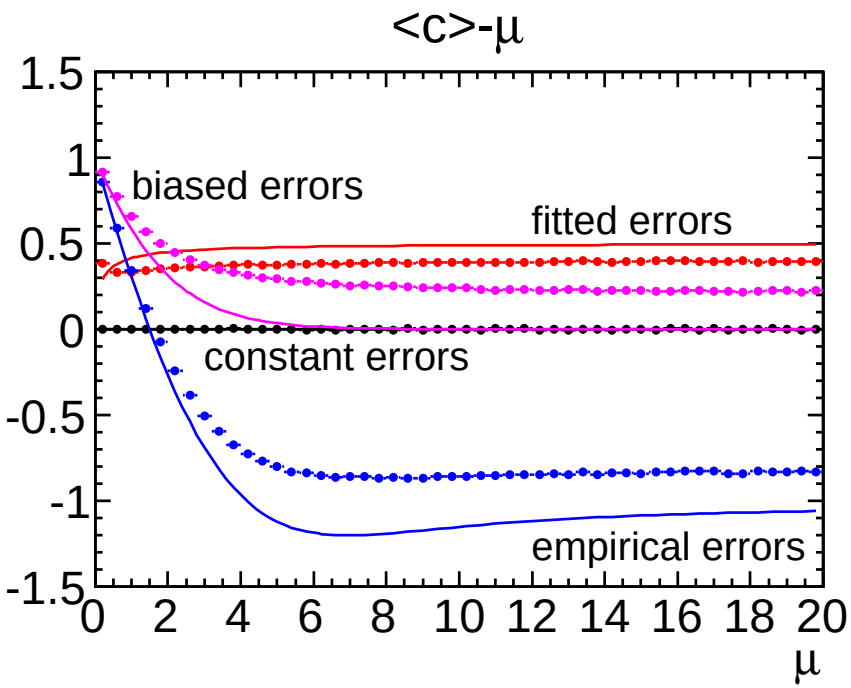


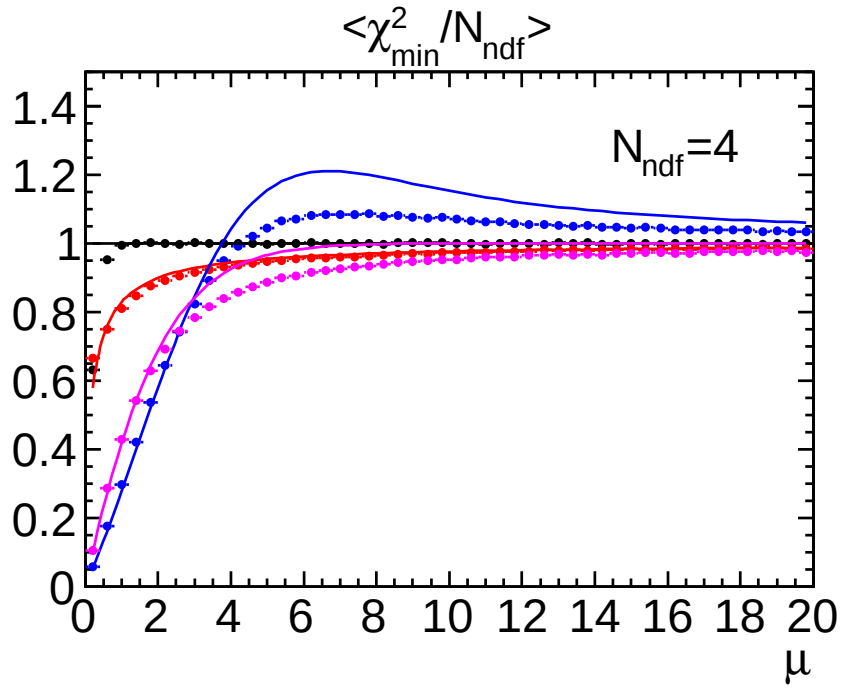


- data points: simulations for $n = 11$ data points
- curves: asymptotic expectations

Numerical studies (iii)

➜ *expectation values vs μ for $n = 5$*





- data points: simulations for $n = 5$ data points
- curves: asymptotic expectations

➜ *introductory remarks*

- common wisdom: least squares fits need. . .
  - ➜ gaussian fluctuations
  - ➜ sufficiently large event counts for poisson distributed data
- in the derivation of the method none of the above entered
  - ➜ only proper variance estimates were assumed
  - ➜ the variances are treated as constants in the χ^2 minimization
  - ➜ the variance estimates should not be correlated to the data

case study, keeping an eye on those points when doing fits ➜

Numerical example

➜ *determination of the lifetime of an unstable particle*

- lifetime distribution

$$\frac{dn}{dt} = \frac{1}{\mu} e^{-t/\mu} \qquad \text{with} \qquad \mu = 1\,\text{ns}$$

- MC study of test experiments with fixed number N of decays
  - ➜ histogram representation of the measurement
  - ➜ 100 bins for $0 < t < 10\,\text{ns}$
- optimal parameter estimate:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} t_i \qquad \text{for } \mu = 1: \qquad \hat{\mu} = 1 \pm \frac{1}{\sqrt{N}}$$

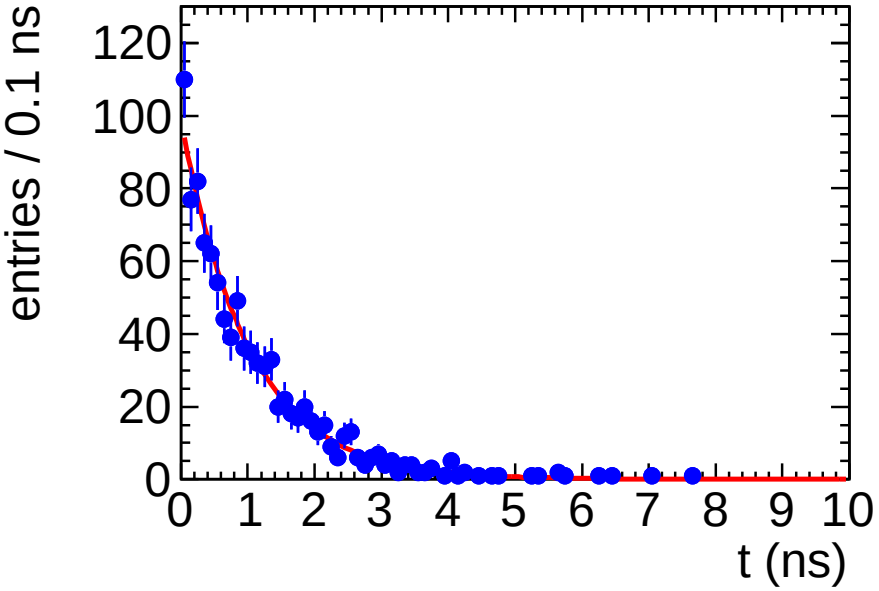
- parametric model for bin contents n_i in Least Squares fit

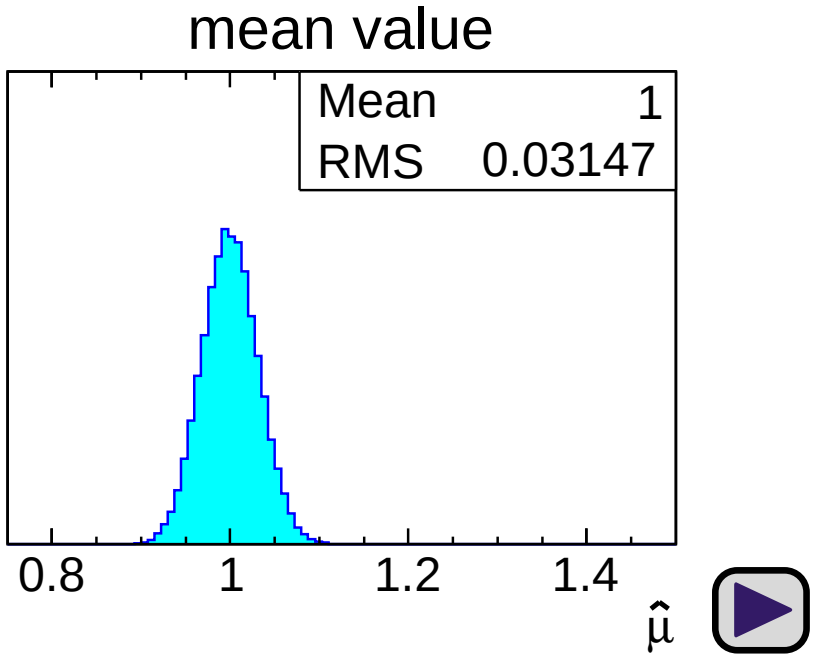
$$f_i(\mu) = N \int_{\text{bin } i} dt \; \frac{dn}{dt}$$

➜ *test different weight-assignments*

- $w_i = 1$ for all bins
  - ➜ unsophisticated but hopefully robust unweighted fit
- $w_i = 1$ for all bins with non-zero entries
  - ➜ pretend that empty bins don't have informations
- $w_i = 1/n_i$ for all bins with non-zero entries
  - ➜ use empirical variance estimates
- $w_i = 1/f_i$ for all bins
  - ➜ naive way to use the theoretical variances
- iterative fit with $w(0) = 1$ and $w_i(m) = 1/f_i(\hat{\mu}_{m-1})$ for all bins
  - ➜ proper way to use the theoretical variances
  - ➜ implements that variances must be fixed in minimization
  - ➜ weak correlation between variance estimates and data
- for comparison: simple arithmetic mean of all entries

➜ *best fit performance for $N = 1000$ events*



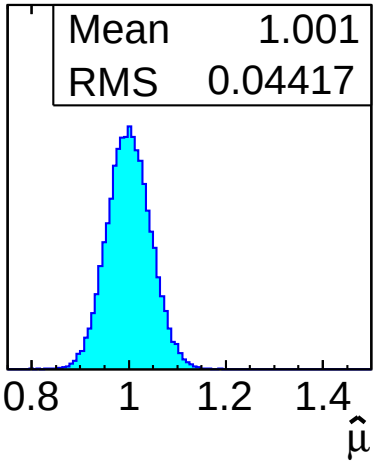


- check standard deviation and bias of fitted $\hat{\mu}$
  - ➜ as a function of available statistics
  - ➜ for the different choices of the weight function

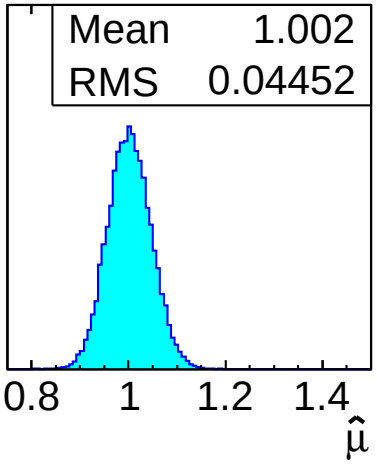
Performance comparison

➜ *parameter estimates for $N = 1000$ events*

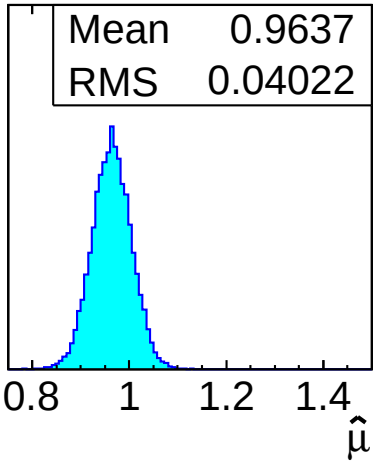
w=1, all bins



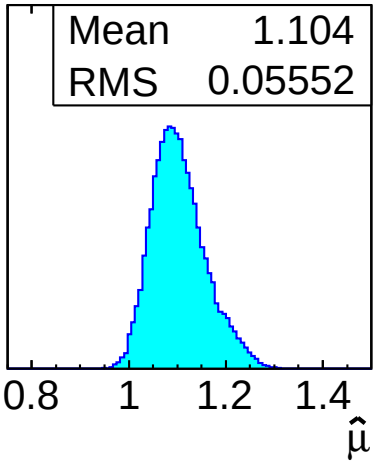
w=1, non-zero bins



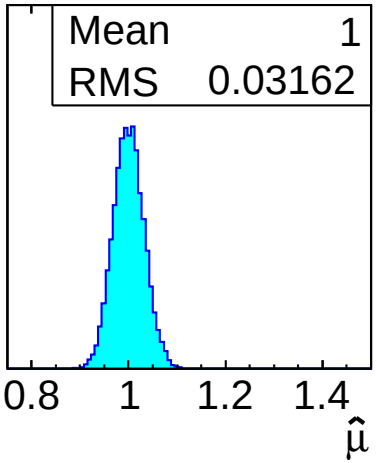
w=1/n, non-zero bins



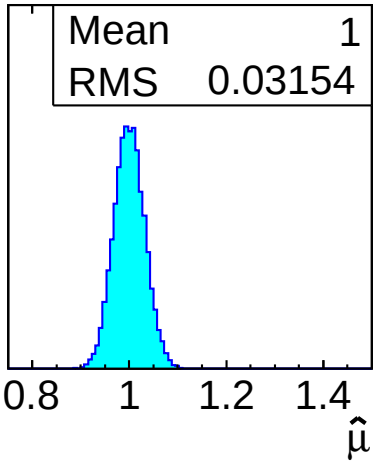
w=1/f, all bins



w=1/f_{m-1}, all bins

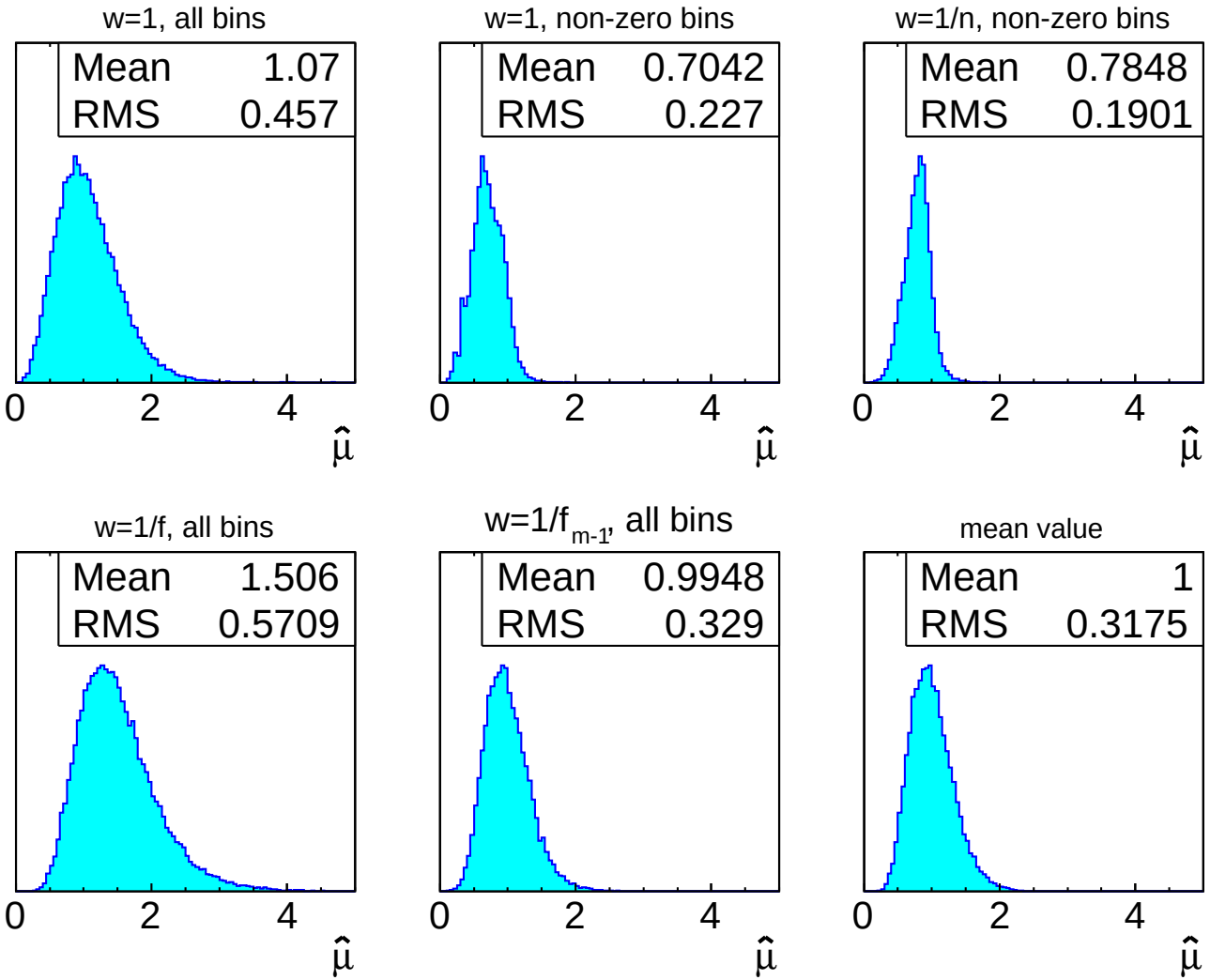


mean value



Performance comparison

➜ *parameter estimates for $N = 10$ events*



➜ *properties of different weight-assignments*

- $w_i = 1$ for all bins
  - ➜ OK, generally unbiased, but not with optimal precision
  - ➜ do not use Hessian of χ^2 function for error estimates
- $w_i = 1$ for non-zero bins
  - ➜ needless loss of information and bias at low statistics
- $w_i = 1/n_i$ for all bins with non-zero entries
  - ➜ biased – violates the least squares ansatz
- $w_i = 1/f_i$ for all bins
  - ➜ biased – violates the least squares ansatz
- iterative fit with $w(0) = 1$ and $w_i(m) = 1/f_i(\hat{\mu}_{m-1})$ for all bins
  - ➜ close to optimum (maximum likelihood fit)
  - ➜ works also at low statistics

➜ *a limiting case of the Least Squares Method*

- uncorrelated single measurements
- counting statistics
- infinitesimal bin widths - i.e. zero or one entry per bin

❖ least-squares fitting of a single parameter with a fixed number of events N:

➜ estimate the parameter a of the PDF $f(x;a)$ of the measurements

➜ iterative minimization with $\hat{a}$ the estimate from the previous step

➜ bin contents $y_i \in 0, 1$

$$\chi^2 = \sum_i \frac{(y_i - N\,p_i)^2}{N\,\hat{p}_i} \quad \text{with} \quad p_i = f(x_i;a)\Delta x \quad \text{and} \quad \hat{p}_i = f(x_i;\hat{a})\Delta x$$

expanding the numerator:

$$\chi^2 = \sum_i \frac{y_i^2}{N\hat{p}_i} - 2\sum_i \frac{y_i p_i}{\hat{p}_i} + N\sum_i \frac{p_i^2}{\hat{p}_i}$$

➜ the 1st term is arbitrary ($\propto 1/\Delta x$) and independent of a

➜ the 2nd and 3rd terms are functions of a

For infinitesimal bin widths one obtains

$$-2\sum_{\text{bins},i}\frac{y_i p_i}{\hat{p}_i} \to -2\sum_{\text{events},i}\frac{p_i}{\hat{p}_i} = -2\sum_{\text{events},i}\frac{f(x_i;a)}{f(x_i;\hat{a})}$$

and

$$N\sum_{\text{bins},i}\frac{p_i^2}{\hat{p}_i} \to N\int dx\,\frac{f^2(x;a)}{f(x;\hat{a})}$$

and minimization of χ^2 with convergence $\hat{a} \to a$ leads to:

$$\frac{\partial}{\partial a}\chi^2 = -2\sum_{\text{events},i}\frac{f'(x_i;a)}{f(x_i;\hat{a})} + N\int dx\,\frac{2\,f(x;a)\,f'(x;a)}{f(x;\hat{a})}$$

$$\stackrel{\hat{a}\to a}{=} -2\sum_{\text{events},i}\frac{f'(x_i;a)}{f(x_i;a)} + 2N\int dx\,f'(x;a)$$

$$= 2\frac{\partial}{\partial a}\left(-\sum_{\text{events},i}\ln f(x_i;a) + N\int dx\,f(x;a)\right)$$

➜ *since $f(x;a)$ is normalized when integrating over x:*

$$\frac{\partial}{\partial a}\left(\frac{1}{2}\chi^2\right) = \frac{\partial}{\partial a}\left(-\ln L(\vec{x};a)\right) \stackrel{!}{=} 0 \quad \text{with} \quad L(\vec{x};a) = \prod_{\text{events},i} f(x_i;a)$$

❖ discussion:

- the best fit parameter is obtained by maximising the likelihood of the data
- for uncorrelated measurements it is the estimate with the smallest variance
- in presence of correlations the least-squares approach with the full covariance matrix is more powerful
- going to infinitesimal bin sizes, the χ^2-minimum becomes arbitrary, i.e. the maximum of the likelihood contains no information about the quality of the fit
- maximum likelihood and least squares fits have very similar properties

$$\Delta(-\ln L) = \frac{1}{2}\Delta\chi^2$$

Extended Maximum Likelihood

➜ *ansatz to estimate also the normalisation when n events were seen:*

$$\chi^2 = \sum_i \frac{(y_i - N\,p_i)^2}{\hat{N}\,\hat{p}_i} \quad \text{with} \quad p_i = f(x_i;\,a)\Delta x \quad \text{and} \quad \hat{p}_i = f(x_i;\,\hat{a})\Delta x$$

Expanding the numerator yields:

$$\chi^2 = \sum_i \frac{y_i^2}{\hat{N}\,\hat{p}_i} - 2\frac{N}{\hat{N}}\sum_i \frac{y_i p_i}{\hat{p}_i} + \frac{N^2}{\hat{N}}\sum_i \frac{p_i^2}{\hat{p}_i}$$

- ➜ the 1st term is an arbitrary offset C
- ➜ the remaining terms depend in N and a

χ^2 function in the limit of infinitesimal bin widths:

$$\chi^2 = C - 2\frac{N}{\hat{N}}\sum_{\text{events},i}^{n} \frac{f(x_i;\,a)}{f(x_i;\,\hat{a})} + \frac{N^2}{\hat{N}}\int dx\,\frac{f^2(x;\,a)}{f(x;\,\hat{a})}$$

Derivatives w.r.t. N and a must vanish; consider $\hat{N} \to N$ and $\hat{a} \to a$.

Taking first the partial derivatives and then the limit $\hat{N} \to N$ and $\hat{a} \to a$ yields

$$\frac{\partial}{\partial N}\chi^2 = -2\frac{n}{N} + 2 = 0$$

$$\frac{\partial}{\partial a}\chi^2 = -2\sum_{\text{events},i}^{n}\frac{f'(x_i;a)}{f(x_i;a)} = 0$$

which corresponds to

$$\frac{\partial}{\partial N}(-\ln L) = \frac{\partial}{\partial a}(-\ln L) = 0$$

with

$$-\ln L = N - n\ln N - \sum_{\text{events},i}^{n}\ln f(x_i;a) = N - \sum_{\text{events},i}^{n}\ln[N\,f(x_i;a)]$$

➜ *standard and extended maximum likelihood method follow from least squares*

➜ *S.S. Wilks, March 26, 1937*

If a population with a variate x is distributed according to the probability distribution $f(x, \theta_1, \theta_2, \ldots, \theta_h)$, such that optimum estimates $\hat{\theta}_i$ of θ_i exist which are distributed in large samples according to (1), then when the hypothesis H is true that $\theta_i = \theta_{0i}, i = m+1, m+2, \ldots h$, the distribution of $-2\ln\lambda$, where λ is given by (2) is, except for terms of order $1/\sqrt{n}$, distributed like χ^2 with $h - m$ degrees of freedom.

(1) a PDF deviating from a d-dim Gaussian only by terms of order $1/\sqrt{n}$

(2) the ratio of the best fit likelihoods fitting all or only m parameters, fixing the others to the true values

$$\lambda = \frac{P(\hat{\theta}_1, \ldots, \hat{\theta}_m, \hat{\theta}_{0m+1}, \ldots, \hat{\theta}_{0h})}{P(\hat{\theta}_1, \ldots, \hat{\theta}_m, \hat{\theta}_{m+1}, \ldots, \hat{\theta}_h)}$$

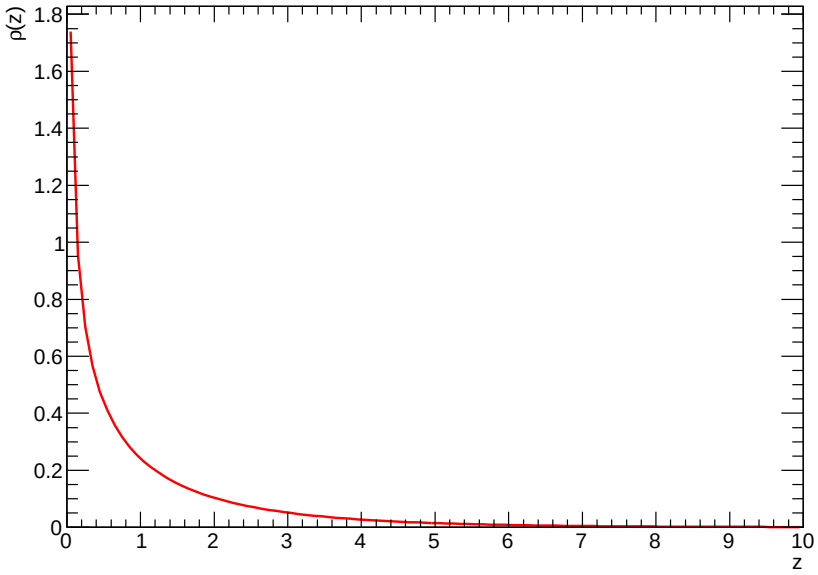
❖ likelihoods are meaningless, likelihood ratios are significant

Application of Wilks' theorem

➜ *test for the existence of a signal s component in data*

- fit with free parameter s: $F_s = -\ln L_{\text{best}}(s)$
- fit with parameter $s = 0$: $F_0 = -\ln L_{\text{best}}(s = 0)$
  - ➜ one has $F_s < F_0$ and $z = 2(F_0 - F_s) > 0$

PDF of z if $s = 0$ is true: $$\rho(z) = \frac{1}{\sqrt{2\pi z}}\, e^{-z/2}$$



➜ *p-value for observed z_{obs}*

$$p = \int_{z=z_{\text{obs}}}^{\infty} dz\ \rho(z)$$

discovery $s \neq 0$ if e.g. $p < 5.7 \cdot 10^{-7}$

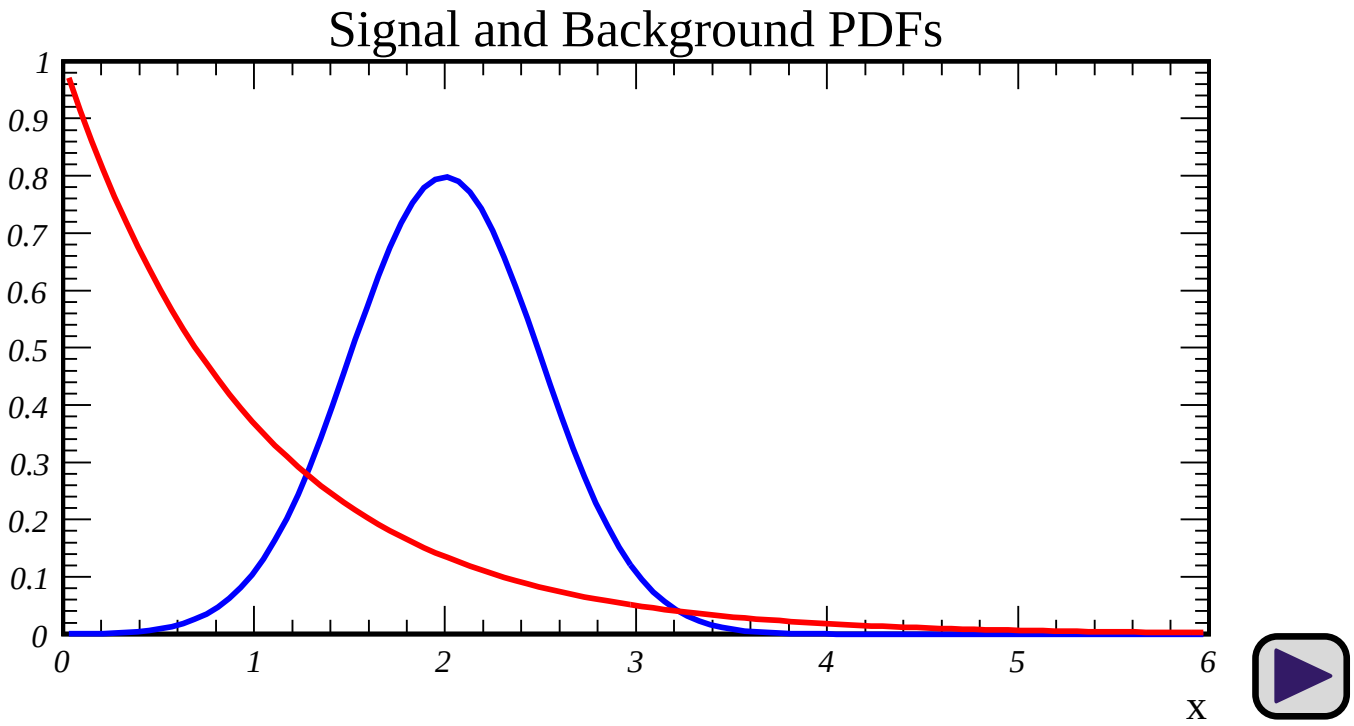
➜ *objective: decide between hypotheses*

- e.g. classification of events or candidates
  - ➜ H_0: signal
  - ➜ H_1: background
- error of 1. kind : H_0 is wrongly rejected with probability α
- error of 2. kind: H_1 is wrongly rejected with probability β

	truth	
classification	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

- in the following: PDFs of signal and background are known
- try optimal separation of both components

➜ *gaussian signal on exponential background*

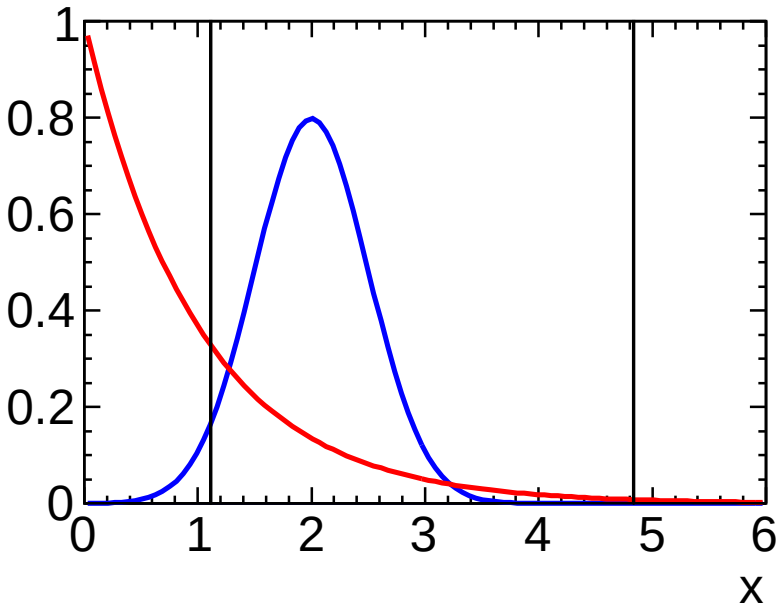


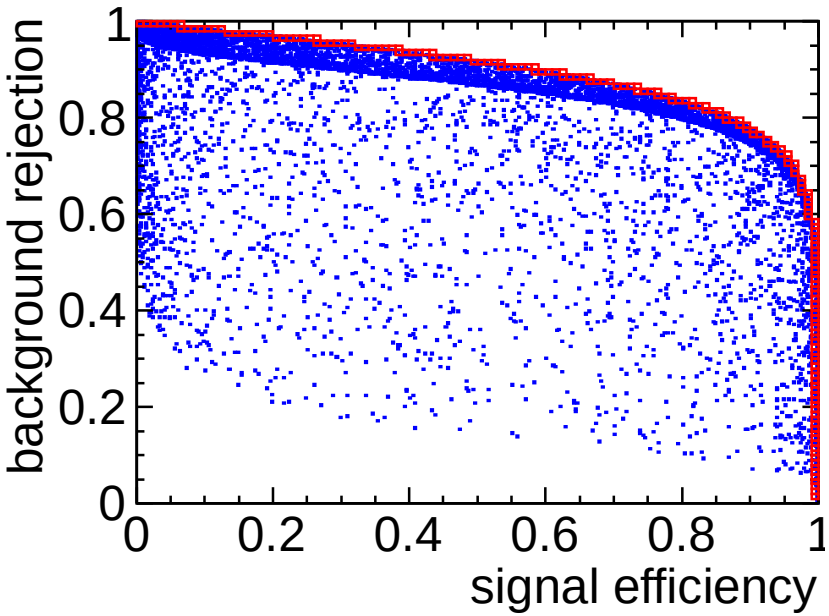
❖ study signal selection

- try different signal windows
- gauge performance by background rejection vs signal efficiency

Best selection by likelihood ratio

selection: if $\frac{f(x|H_0)}{f(x|H_1)} \leq c$ then reject x





- best "Receiver Operation Characteristic" (ROC-curve)
  - ➜ largest background rejection for fixed signal efficiency
  - ➜ smallest errors of 2nd kind for fixed errors of 1st kind
  - ➜ parameter c determines signal efficiency

➜ *definitions and conditions*

- $\vec{x}$: point in configuration space
- $f(\vec{x}|H_k)$: PDF for $\vec{x}$ in case of H_k
- critical region S: configuration space volume with probability α for H_0

$$P(\vec{x} \in S|H_0) = \int_S d^n x\, f(\vec{x}|H_0) = \alpha$$

- S_c: critical region satisfying

$$\frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} \leq c$$

➜ *conjecture:*

The critical region S_c is optimal in the sense, that it minimizes errors of the second kind (minimal probability to accept background).

➜ proof

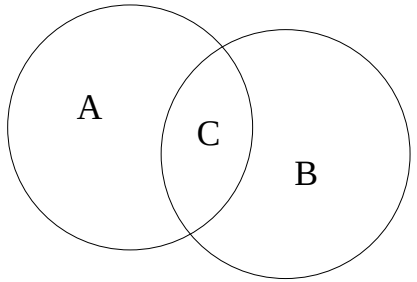
Proof of Neyman-Pearson's lemma (i)

Take two critical regions S_c and S with equal probability content for H_0

$$\int_{S_c} d^n x \; f(\vec{x}|H_0) = \int_{S} d^n x \; f(\vec{x}|H_0) = \alpha$$

In general the regions will overlap and one can write:

$$S_c = A \cup C \quad \text{and} \quad S = B \cup C$$



Thus C contributes equally to S_c and S and one has

$$\int_{A} d^n x \; f(\vec{x}|H_0) = \int_{B} d^n x \; f(\vec{x}|H_0) \; .$$

Region A is inside S_c, B is outside, i.e. by construction

$$\frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} \leq c \quad \text{if } \vec{x} \in A \qquad \text{and} \qquad \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} > c \quad \text{if} \quad \vec{x} \in B$$

It follows:

$$\int_A d^n x\, f(\vec{x}|H_0) \leq c \int_A d^n x\, f(\vec{x}|H_1)$$
$$\int_B d^n x\, f(\vec{x}|H_0) \geq c \int_B d^n x\, f(\vec{x}|H_1)$$

Compare now the H_1 (background) probabilities in S_c and S:

$$\begin{aligned} P(\vec{x} \in S_c|H_1) &= \int_A d^n x\, f(\vec{x}|H_1) + \int_C d^n x\, f(\vec{x}|H_1) \\ &\geq \frac{1}{c} \int_A d^n x\, f(\vec{x}|H_0) + \int_C d^n x\, f(\vec{x}|H_1) \\ &= \frac{1}{c} \int_B d^n x\, f(\vec{x}|H_0) + \int_C d^n x\, f(\vec{x}|H_1) \\ &\geq \int_B d^n x\, f(\vec{x}|H_1) + \int_C d^n x\, f(\vec{x}|H_1) = P(\vec{x} \in S|H_1) \end{aligned}$$

➜ *comparison of the background probability shows:*

$$P(\vec{x} \in S_c | H_1) \geq P(\vec{x} \in S | H_1) \; .$$

- critical regions are rejected for signal selections
- by construction all critical regions have the same α
  - ➜ the same signal efficiency $1 - \alpha$
- the region S_c has the largest background probability
  - ➜ largest possible rejection for given signal efficiency
  - ➜ smallest errors of 2nd kind for given errors of 1st kind
- in S_c one has

$$\frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} \leq c$$

- optimal solution of the selection problem if all PDFs are known

➜ *problem*

- PDFs are not known
- only finite samples exists to estimates the PDFs of H_0 and H_1
- multi-dimensional PDFs hard to determine ("curse of dimensionality")

➜ *general strategy*

- construct test variables or functions (classifier) in configuration space
- start with training
  - ➜ estimate PDFs L_S and L_B for signal and background
  - ➜ avoid "overtraining" (learning fluctuations in the training sample)
- performance test with independent signal and background samples

➜ *z.B. open source implementation: TMVA*

Toolkit for MultiVariate Analysis with ROOT

`arXiv:physics/0703039, CERN-OPEN-2007-007`
`http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf`

- two types of classifiers
  - ➜ optimized classifiers for predefined signal efficiency $1 - \alpha$
  - ➜ continuous probability-like classifiers t provided by TMVA
- raw ranges $t_{\min} \leq t \leq t_{\max}$, possibly peaking towards limit
  - ➜ transform to normalized classifiers to $-1 \leq t' \leq +1$

$$t' = \frac{1}{N} \ln \frac{t - t_{\min} + \delta}{t_{\max} - t + \delta} \quad \text{with} \quad \delta = \frac{t_{\max} - t_{\min}}{\exp(N) - 1}$$

  - ➜ $N \to 0$: linear rescaling to $[-1, +1]$
  - ➜ $N > 0$: remove singularities at the end points
- classifiers are not invariant under transformations of variables
  - ➜ human understanding of the problem still vital
- in most cases the theoretical optimum is not reached
- note: biased training samples lead to biased efficiency estimates
  - ➜ ongoing work to understand and control such systematics

Overview of methods

- discussed below (and available in TMVA)
  - ➜ projected 1-dim likelihood ratios
  - ➜ KNN
  - ➜ PDEFoam
  - ➜ Fisher discriminant
  - ➜ multilayer-perceptron neural networks
  - ➜ Boosted Decision Trees
- common preprocessing steps
  - ➜ decorrelation and gaussianization
  - ➜ combinations and iterations of the above
- use TMVA methods with default settings
  - ➜ no tuning of internal parameters and options
  - ➜ no preprocessing of variables
- performance classification by ROC-curves

➜ *attempt to apply the Neyman-Pearson lemma*

$$f(x_1, x_2, x_3 \ldots) \rightarrow L = f_1(x_1)\, f_2(x_2)\, f_3(x_3) \ldots$$

$$\text{with} \quad f_1(x_1) = \int dx_2\, dx_3 \ldots f(x_1, x_2, x_3, \ldots)$$

$$f_2(x_2) = \int dx_1\, dx_3 \ldots f(x_1, x_2, x_3, \ldots) \quad \text{etc.}$$

- parametrize the projected PDFs
- classifier $c(i)$ for each event i

$$c(i) = \frac{L_{\mathrm{sig}}(i)}{L_{\mathrm{sig}}(i) + L_{\mathrm{bkg}}(i)} = \frac{1}{1 + L_{\mathrm{bkg}}(i)/L_{\mathrm{sig}}(i)}$$

- projections avoid “curse of dimensionality”
- loss of performance if true PDFs do not factorize

➜ *attempt proper n-dim density estimates*

- subdivide the phase space into a given number of hyper-rectangles with (about) equal numbers of entries per cell
- search subdivision which minimizes the density variance in the cells
- assume constant density per cell
- do this separately for signal and background
- construct classifier based on likelihood ratios
- properties:
  - ➜ non-parametric description of PDFs
  - ➜ correlations are taken into account
  - ➜ very few entries per cell in high-dimensional spaces

➜ *compare a candidate event to training sample densities*

- non-parametric density estimates
- k training events (signal plus background) closest to the candidate
- classifier:

relative signal-probability $$c_{\mathrm{KNN}} = \frac{k_{\mathrm{sig}}}{k_{\mathrm{sig}} + k_{\mathrm{bkg}}} = \frac{k_{\mathrm{sig}}}{k}$$

- empirical finding: $10 < k < 100$ shows good performance
  - ➜ too large value: local density variations are not seen
  - ➜ too small value: estimates suffer from large fluctuations
- performance depends on the metric

$$R^2 = \sum_{i=1}^{n_{\mathrm{dim}}} \frac{1}{w_i^2}(x_i - y_i)^2$$

  - ➜ w_i allows adaption to spread of input variables
  - ➜ intrisically adaptive – no problem with large number of dimensions

➜ *test variable from a linear combination of the measurements x_i*

$$t(\vec{x}) = a_0 + \sum_{i=1}^{n} a_i x_i = a_0 + \vec{a}^T \vec{x}$$

❖ geometrical interpretation

- $\vec{a}$ and a_0 define a hyperplane in n dimensions
  - ➜ $\vec{a}$ is a vector normal to the plane
  - ➜ a_0 is the distance of the plane from the origin
- constant values $t(\vec{x})$ for points $\vec{x}$ on a plane parallel to the hyperplane defined by $\vec{a}$ and a_0
- adjust the orientation of $\vec{a}$ and the offset a_0 to get optimal separation between H_0 (signal) and H_1 (background)

➜ *realization:*

expectation values and covariance matrix of $\vec{x}$ for hypotheses H_k are

$$\langle \vec{x} \rangle \big|_{H_k} = \vec{\mu}_k \quad \text{and} \quad \left\langle \vec{x} \cdot \vec{x}^{\,T} \right\rangle - \langle \vec{x} \rangle \cdot \langle \vec{x} \rangle^T \Big|_{H_k} = V_k$$

for mean and variance of t under hypothesis H_k follows

$$\langle t_k \rangle = a_0 + \vec{a}^{\,T} \vec{\mu}_k \quad \text{and} \quad V_k(t) = \vec{a}^{\,T} \cdot V_k \cdot \vec{a}$$

and a measure $J(\vec{a})$ for the separation between the hypotheses is

$$J(\vec{a}) = \frac{\left(\langle t_0 \rangle - \langle t_1 \rangle\right)^2}{V_0(t) + V_1(t)} = \frac{\left(\vec{a}^{\,T}(\vec{\mu}_0 - \vec{\mu}_1)\right)^2}{\vec{a}^{\,T} \cdot (V_0 + V_1) \cdot \vec{a}}$$

or, introducing $V = V_0 + V_1$ and $\vec{\mu} = \vec{\mu_0} - \vec{\mu_1}$,

$$J(\vec{a}) = \frac{\vec{a}^{\,T} \vec{\mu} \vec{\mu}^{\,T} \vec{a}}{\vec{a}^{\,T} V \vec{a}} \stackrel{!}{=} \max \ .$$

➜ *construction of the solution*

- $J(\vec{a})$ does not depend on the normalization of $\vec{a}$ or a_0
- boundary condition $\vec{a}^T\vec{\mu} = c$ defines unique solution
- constrained maximum:

$$\frac{\partial}{\partial \vec{a}^T}\left[\frac{\vec{a}^T\vec{\mu}\vec{\mu}^T\vec{a}}{\vec{a}^T V\vec{a}} - \lambda(c - \vec{a}^T\vec{\mu})\right] = 0$$

and thus $$\frac{\vec{\mu}\vec{\mu}^T\vec{a}}{\vec{a}^T V\vec{a}} - \frac{\vec{a}^T\vec{\mu}\vec{\mu}^T\vec{a}}{(\vec{a}^T V\vec{a})^2}(V\vec{a}) + \lambda\vec{\mu} = 0$$

➜ a solution exists if $V\vec{a} \propto \vec{\mu}$, i.e.

$$\vec{a} \propto V^{-1}\vec{\mu}$$

➜ and adjustment of c and a_0 allows to have

$$\langle t_0\rangle = 1 \qquad \text{and} \qquad \langle t_1\rangle = 0$$

➜ *result:*

$$t(\vec{x}) = a_0 + c(\vec{\mu_0} - \vec{\mu_1})^T (V_0 + V_1)^{-1} \vec{x}$$

with freely choosable parameters a_0 and c.

➜ *construction of $t(\vec{x})$ requires of each hypothesis*

- n expectation values
- $n(n+1)/2$ variances and covariances
- in total $n(n+1) + 2n = n^2 + 3n$ parameters
- numerically stable determination
- very small danger of overtraining if $N \gg n^2$

➜ *consider 1-dim gaussians*

$$f(x|H_0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_0)^2/2\sigma^2} \quad \text{and} \quad f(x|H_1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_1)^2/2\sigma^2}$$

- different mean values μ_0 and μ_1 but equal standard deviations σ
- then the logarithm of the likelihood ratio

$$\ln \frac{f(x|H_0)}{f(x|H_1)} = -\frac{1}{2\sigma^2}(x^2 - 2x\mu_0 + \mu_0^2 - x^2 + 2x\mu_1 - \mu_1^2)$$

$$= \frac{\mu_1^2 - \mu_0^2}{2\sigma^2} + \frac{\mu_0 - \mu_1}{\sigma^2} x = a_0 + a_1 x$$

has the structure of a Fisher discriminant, i.e.

$$\frac{f(x|H_0)}{f(x|H_1)} \equiv r \propto e^{t(x)}$$

➜ Fisher: optimal for gaussian H_0 and H_1 with equal variance

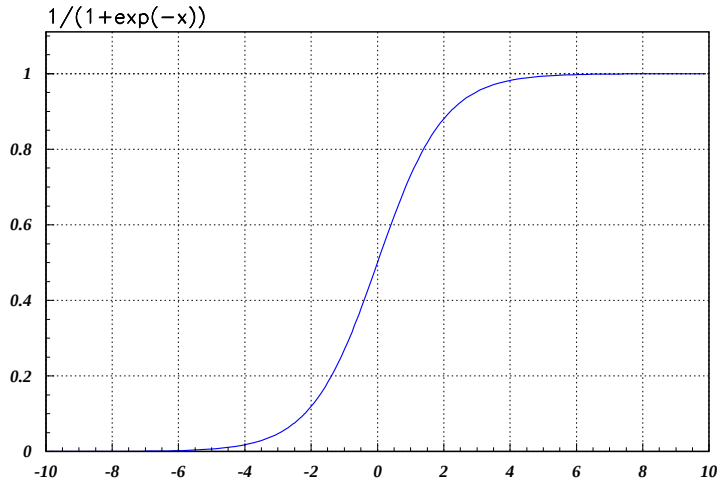
➜ *heuristic approach*

Bayes' theorem allows to formulate a relation between Fisher discriminant and bayesian signal probability. Taking equal prior probabilities $p(H_0) = p(H_1)$ for signal and background one finds:

$$P(H_0|\vec{x}) = \frac{f(\vec{x}|H_0)p(H_0)}{f(\vec{x}|H_0)p(H_0) + f(\vec{x}|H_1)p(H_1)} = \frac{1}{1+1/r}$$

For equal-width gaussians one had $r = e^t$, which leads to

$$P(H_0|\vec{x}) = \frac{1}{1+e^{-t}} \equiv s(t) \quad \in \quad [0,1]$$



➜ "logistic sigmoid function"

useful to describe decisions between hypotheses. . .

➜ *interpretation of the Fisher discriminant as "neural network"*

- Fisher discriminant as a weighted sum of the input signals

$$t(\vec{x}) = a_0 + \sum_{i=1}^{n} a_i x_i$$

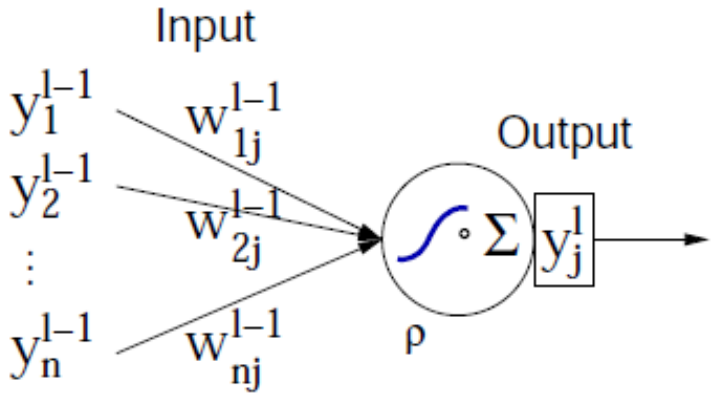
- interpretation by means of the logistic sigmoid function

$$s(t) = \frac{1}{1 + e^{-t}}$$

❖ compare:

- signal processing in nerve cells
  - ➜ several inputs x_i
  - ➜ weighted summation $\sum_i a_i x_i$
  - ➜ switching according to activation function

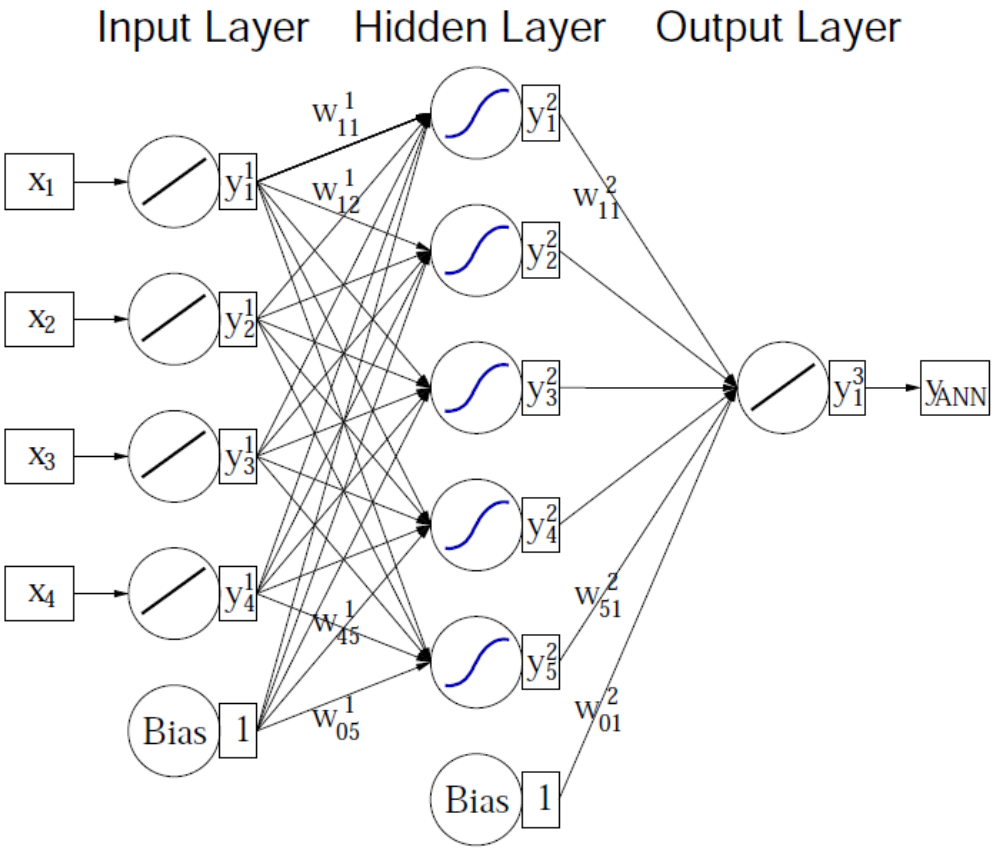
single layer perceptron



equivalent to Fisher discriminant

➜ *cascading of neurons*

- for example: double layer perceptron
- output signal is a function of an inner (hidden) layer of neurons



$$t(\vec{x}) = s\left(a_0 + \sum_{i=1}^{m} a_i h_i(\vec{x})\right)$$

with

$$h_i(\vec{x}) = s\left(w_{i0} + \sum_{k=1}^{n} w_{ik} x_k\right)$$

- n neurons in principle allow n^2 directional connections
- reduction of complexity by
  - ➜ arrangement in layers
  - ➜ restriction to feed-forward networks
- neural networks are very efficient universal approximators
  - ➜ even the most general case can be realized with a single hidden layer – but may require a very large number of neurons
  - ➜ alternatively use several hidden layers and fewer neurons
- the optimal network topology for a given application is not known
- there are many possible choices for the activation function, e.g.
  - ➜ logistic sigmoid $s(t)$, $\tanh(t)$, $\ldots$
- determination of the weight usually done numerically

- generate training samples for each hypothesis
- define decisions as a function of the output signal
- define a cost functions for the quality of the decision
- adjust the weight by minimizing the cost function

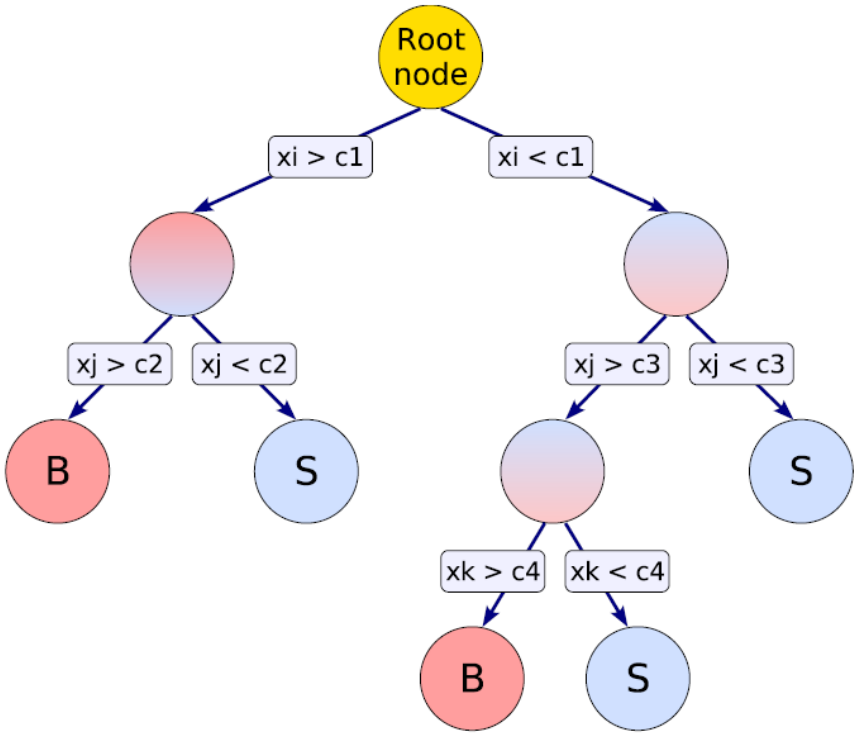
example: $$F = \sum_{\vec{x} \in H_1} t^2(\vec{x}) + \sum_{\vec{x} \in H_0} (1 - t(\vec{x}))^2$$

goal: $t(\vec{x}) = 1$ for $\vec{x} \in H_0$ (sig) and $t(\vec{x}) = 0$ für $\vec{x} \in H_1$ (bkg)

Determination of the weights is a hard non linear minimization problem with usually many local minima. It is normally sufficient to find a good minimum instead of the global one. Possible algorithm:

- ➜ take random initial values for the weights
- ➜ get the gradient of the cost function with respect to the weights
- ➜ do a (small) downhill step and iterate until a minimum is reached
- ➜ try other initial values

➜ *basic topology of a decision tree*



- sequence of binary decisions
- generalization of cut-sequence for signal selection
- each instance $\vec{x}$ is classified as signal or background

  $h(\vec{x}) = +1$ signal

  $h(\vec{x}) = -1$ background
- classification of a node as signal or background according to the majority of its instances

➜ *iterative splitting of each node*

- basic idea
  scan all variables and determine a cut which gives the best improvement in the separation of signal and background
- implementation requires a measure for separation, as e.g. the "Gini Index" S, based on the signal purity p in a node
  - ➜ separation in the mother node:
    $$S_M = p(1-p)$$
  - ➜ separation in the daughter nodes:
    $$S_T = \frac{n_1}{n_1+n_2} p_1(1-p_1) + \frac{n_2}{n_1+n_2} p_2(1-p_2)$$
- use that variable and the cut which maximizes $S_M - S_T$
- stop splitting if 100% purity or too few events in a node
- problem: small fluctuations can give radically different decision trees
- remedy: boosting

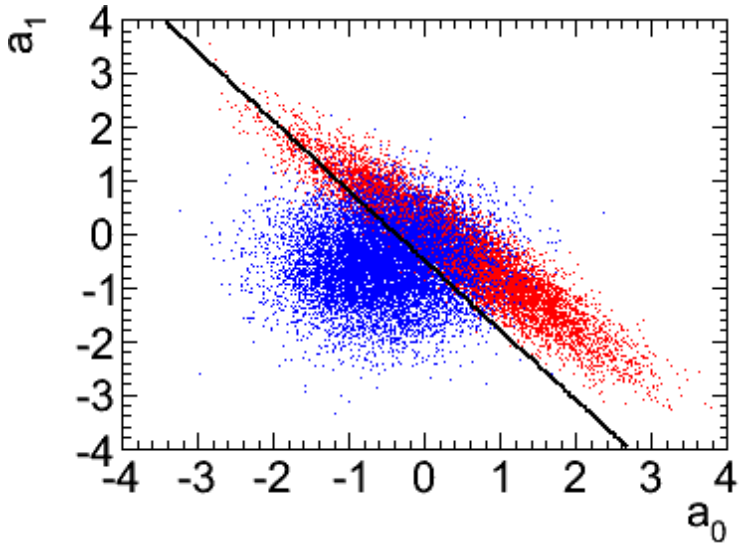
➜ *construction an average decision tree with improved performance*

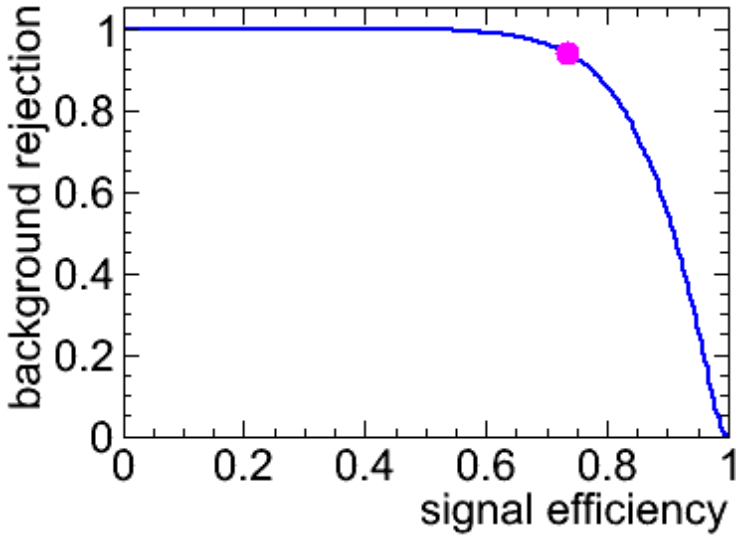
- iterative generation of decision trees by constructing new training samples based on mis-classification rate ε of the current tree
  - ➜ weight all wrongly classified instances by $(1-\varepsilon)/\varepsilon$
  - ➜ renormalize the sum of all instances to the original value
  - ➜ determine a new decision tree
- decision tree ➜ decision forest
- further improvement:
  - ➜ pruning of the trees by eliminating branches with only a negligible improvement in the separation between signal and background
- BDT-Classifier: weighted average of all classifications in the forest

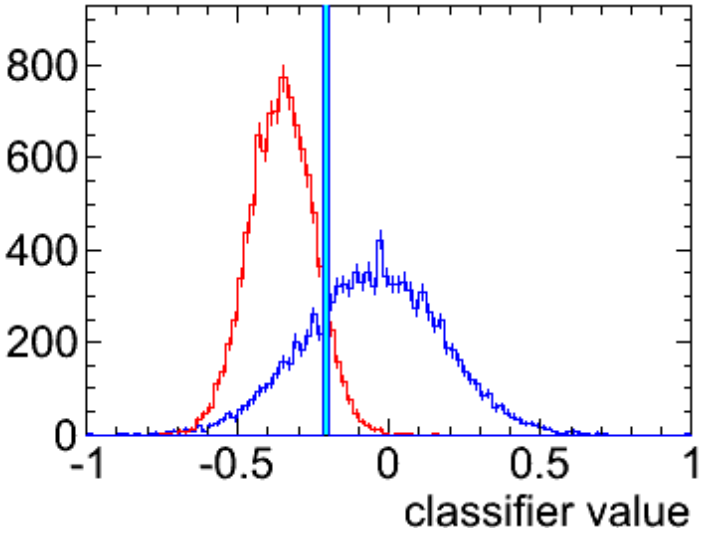
$$y(\vec{x}) = \sum_{i \in \text{forest}} \ln \frac{1-\varepsilon_i}{\varepsilon_i} \cdot h_i(\vec{x})$$

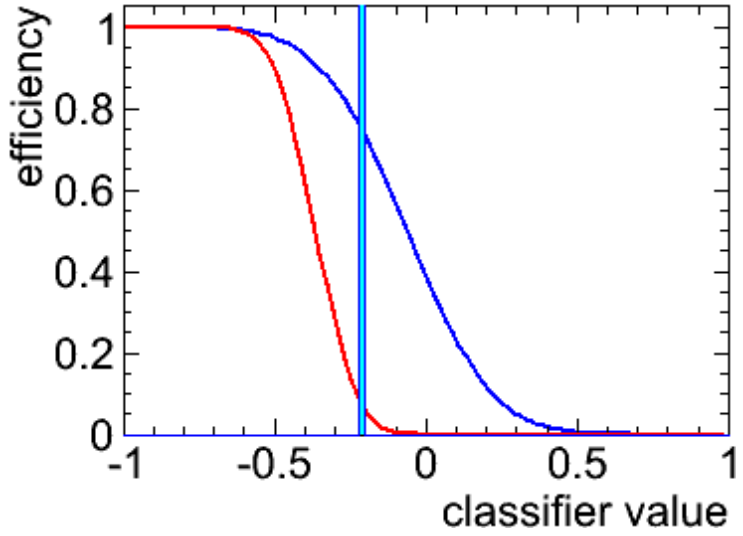
  i.e. larger weight for trees with smaller mis-identification rate
- often best: weighted mean over many weakly optimized trees

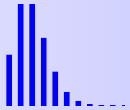
Illustration











Summary on multivariate analysis

- optimal separation of signal and background by likelihood ratios
  - ➜ in practice useful only for few dimensional problems
  - ➜ many different methods for higher dimensional problems
- projected likelihoods: ideal for uncorrelated variables
- PDEFoam and KNN: good start for n-dim likelihood ratios
- Fisher discriminant: simple and robust, optimal for gaussians
- neural networks: probably best, but hard to train
- (Boosted) Decision Trees: very good “out-of-the-box”-method
- performance of many methods depends on choice of variables
  - ➜ pre-processing can result in significant gain
    - de-correlation
    - avoid singularities ($x \to \ln x$)
    - discard insensitive variables to avoid “curse of dimensionality”
- MVA: very active field of research (data mining . . .)

➜ *separating signal and background*

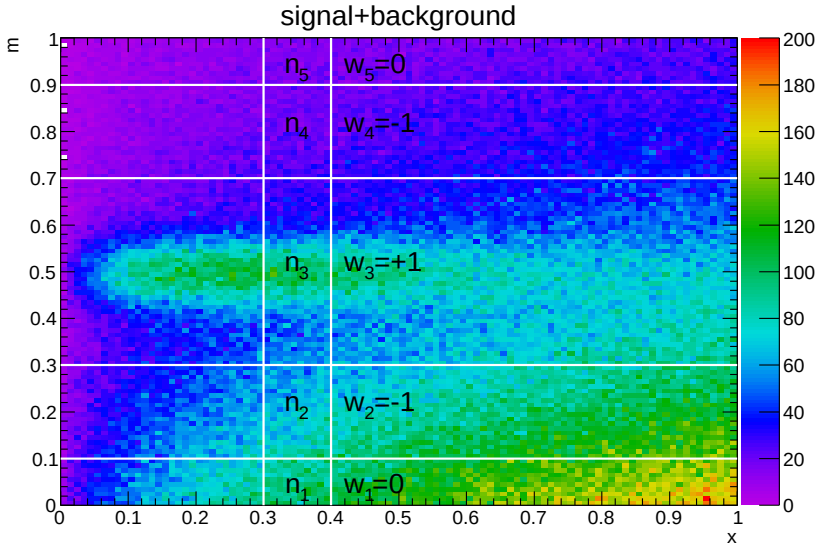
$$f(x,m) = N_s\, s(x,m) + N_b\, b(x,m)$$

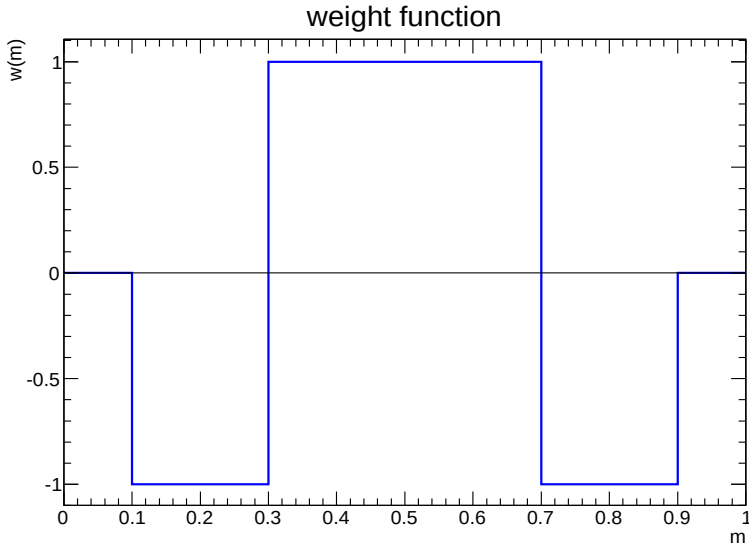
- normalized PDFs $s(x,m)$ and $b(x,m)$ for signal and background
  - ➜ signal in (dm, dx)-bin: $N_s\, s(x,m)\, dx\, dm$
  - ➜ background in (dm, dx)-bin: $N_b\, b(x,m)\, dx\, dm$
  - ➜ total entries in (dm, dx)-bin: $f(x,m)\, dx\, dm$
- normalizations N_s and N_b for signal and background
- m: "discriminant" variable to tell signal from background
  - ➜ will be treated as a scalar in the following
  - ➜ can equally well be vector
- x: "control" variable to be studied
  - ➜ will be treated as a scalar in the following
  - ➜ can equally well be vector
- try to extract the signal density as function of x

get rid of background by sideband subtraction ➜

Starting point: sideband subtraction

❖ determine the signal density $N_s\, s(x)$ for a given x





for narrow bins one has $\text{signal}(x-\Delta x/2, x+\Delta x/2) = N_s\, s(x)\, \Delta x$, and thus:

$$N_s\, s(x) = \frac{\text{signal}(x-\Delta x/2, x+\Delta x/2)}{\Delta x} = \frac{1}{\Delta x} \sum_{i, x_i \in x \pm \Delta x/2} w(m_i)$$

$$= \frac{1}{\Delta x} \int_{x-\Delta x/2}^{x+\Delta x/2} dx \int dm\, w(m)\, f(x,m) = \int dm\, w(m)\, f(x,m)$$

➜ *sideband subtraction is a special case of an integral transform*

$$N_s\, s(x) = \int dm\, w(m)\, f(x,m)$$

- the weight function $w(m)$ projects out the signal density $s(x)$
- the above equation was derived for one fixed x
- now require that the same $w(m)$ works for all x
- possible if $s(x,m)$ and $b(x,m)$ factorize as a function of x and m

$$f(x,m) = N_s\, s(x)\, s(m) + N_b\, b(x)\, b(m)$$

- assume that $s(m)$ and $b(m)$ are known

❖ Find the optimal weight function $w(m)$ for this case!

➜ *necessary condition:*

$$\int dm\, w(m)\,[N_s\, s(x)\, s(m) + N_b\, b(x)\, b(m)] = N_s\, s(x)$$

which implies

$$\int dm\, w(m)\; s(m) = 1 \qquad \text{and} \qquad \int dm\, w(m)\; b(m) = 0$$

- any $w(m)$ orthogonal to $b(m)$ can be normalized to satisfy this
- for $s(m) \propto b(m)$ signal and background cannot be separated
- select $w(m)$ with minimal variance of the total signal yield $\sum w$

$$\sum_{\text{events}} w^2 \stackrel{!}{=} \min \qquad \rightarrow \qquad \int dx\, dm\, w^2(m)\, f(x,m) \stackrel{!}{=} \min$$

➜ constrained minimization problem

➜ solved by variational calculus

➜ *constrained minimization with Lagrange parameters α and β*

$$\delta \left\{ \int dx\, dm\, w^2(m)\, [N_s\, s(x)\, s(m) + N_b\, b(x)\, b(m)] \right.$$

$$\left. + 2\alpha \left(1 - \int dm\, w(m)\, s(m) \right) - 2\beta \int dm\, w(m)\, b(m) \right\} = 0$$

- ➜ the variation is performed on $w(m)$
- ➜ the constant term 2α is irrelevant for the minimization
- ➜ integration over x gives two factors of unity ➜ single integral over m

$$\delta \left\{ \int dm\, w^2(m)\, [N_s\, s(m) + N_b\, b(m)] - 2w(m)[\alpha\, s(m) + \beta\, b(m)] \right\} = 0$$

- ➜ substituting $\delta w^2(m) = 2\, w(m) \delta w(m)$ yields

$$\int dm\, \delta w(m) \{ w(m)\, [N_s\, s(m) + N_b\, b(m)] - [\alpha\, s(m) + \beta\, b(m)] \} = 0$$

- ➜ zero integral for arbitrary $\delta w(m)$ requires $\{ \ldots \} = 0$

➜ *optimal weight function*

$$w(m) = \frac{\alpha\, s(m) + \beta\, b(m)}{N_s\, s(m) + N_b\, b(m)}$$

- $w(m)$ is a linear combination of signal purity and background purity
- numerical values are for α and β follow from the constraints

$$\int dm\, w(m)\, s(m) = 1 \qquad \text{and} \qquad \int dm\, w(m)\, b(m) = 0$$

- substituting $w(m)$ yields the system of equations:

$$\begin{pmatrix} W_{ss} & W_{sb} \\ W_{sb} & W_{bb} \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{with} \quad W_{uv} = \int dm\, \frac{u(m)\, v(m)}{N_s\, s(m) + N_b\, b(m)}$$

- with solutions

$$\alpha = \frac{W_{bb}}{W_{ss}\, W_{bb} - W_{sb}^2} \quad \text{and} \quad \beta = \frac{-W_{sb}}{W_{ss}\, W_{bb} - W_{sb}^2}$$

Numerical determination of W_{uv}

➜ *consider the binned m-distribution*

- with $i = 1, \ldots, n$ bins with widths Δm and bin centers m_i
- indices u, v referring to signal or background PDF, i.e. $u, v \in \{s, b\}$

$$W_{uv} = \int dm \, \frac{u(m)\, v(m)}{N_s\, s(m) + N_b\, b(m)} \approx \sum_i \Delta m \frac{u(m_i)\, v(m_i)}{N_s\, s(m_i) + N_b\, b(m_i)}$$

$$= \sum \Delta m \frac{u(m_i)\, v(m_i)\, \Delta m}{N_s\, s(m_i)\, \Delta m + N_b\, b(m_i)\, \Delta m} = \sum_i \frac{p_u(m_i)\, p_v(m_i)}{n(m_i)}$$

➜ with $p_{u,v}(m_i)$ the signal or background probability in bin i

$$p_s(m_i) = s(m_i)\, \Delta m \quad \text{and} \quad p_b(m_i) = b(m_i)\, \Delta m$$

➜ and $n(m_i)$ the observed number of events in the bin around m_i

Numerical example

- phase space

$$0 < x < 1 \qquad \text{and} \qquad 0 < m < 1$$

- signal $s(x, m) = s(x)\, s(m)$

$$s(x) = \frac{25}{1 - 6e^{-5}}\, x\, e^{-5x} \quad \text{and} \quad s(m) = \frac{20}{\sqrt{2\pi}}\, e^{-200(m-0.5)^2}$$

- background $b(x, m) = b(x)\, b(m)$

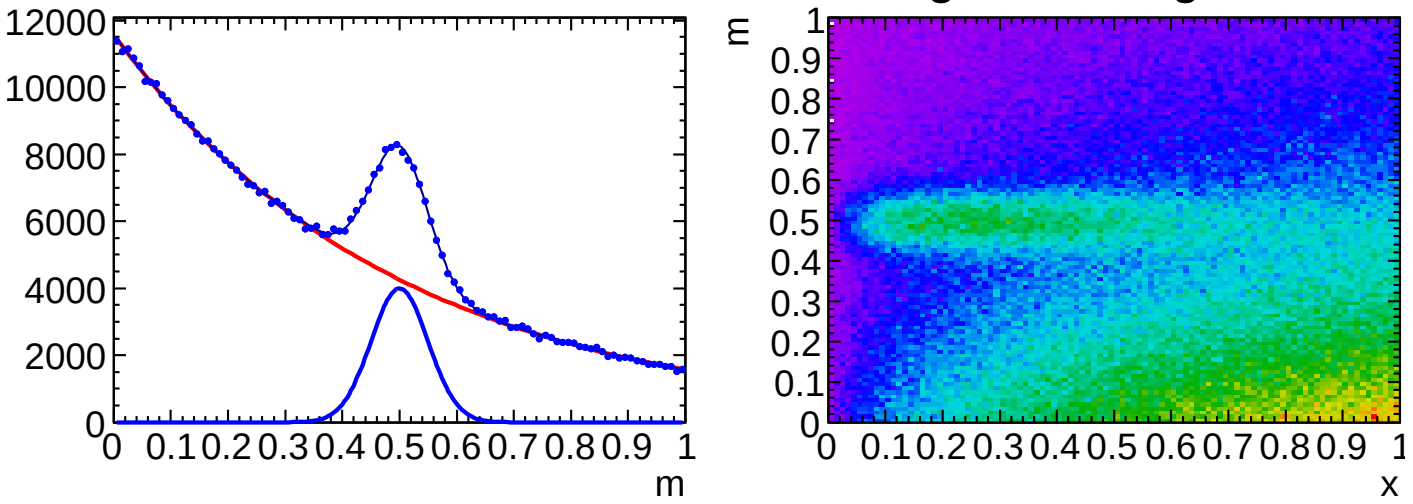
$$b(x) = 1.5\,\sqrt{x} \quad \text{and} \quad b(m) = \frac{2}{1 - e^{-2}}\, e^{-2m}$$

- event statistics

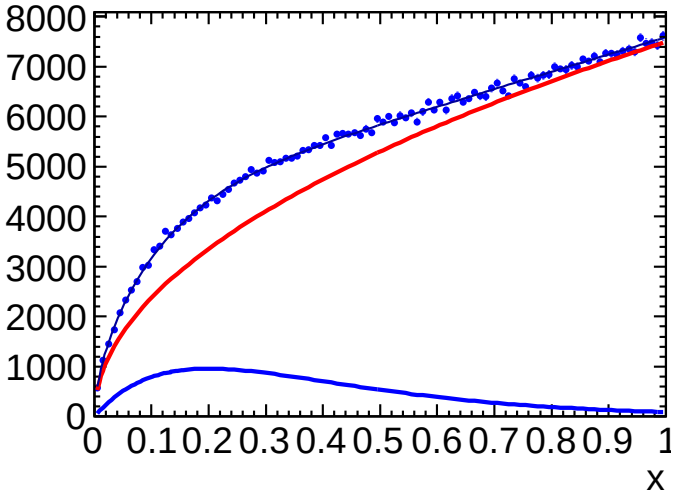
$$N_s = 50\,000 \quad \text{with} \quad N_b = 500\,000$$

➜ *sum of signal and background:* $\varepsilon(x, m) = 1$

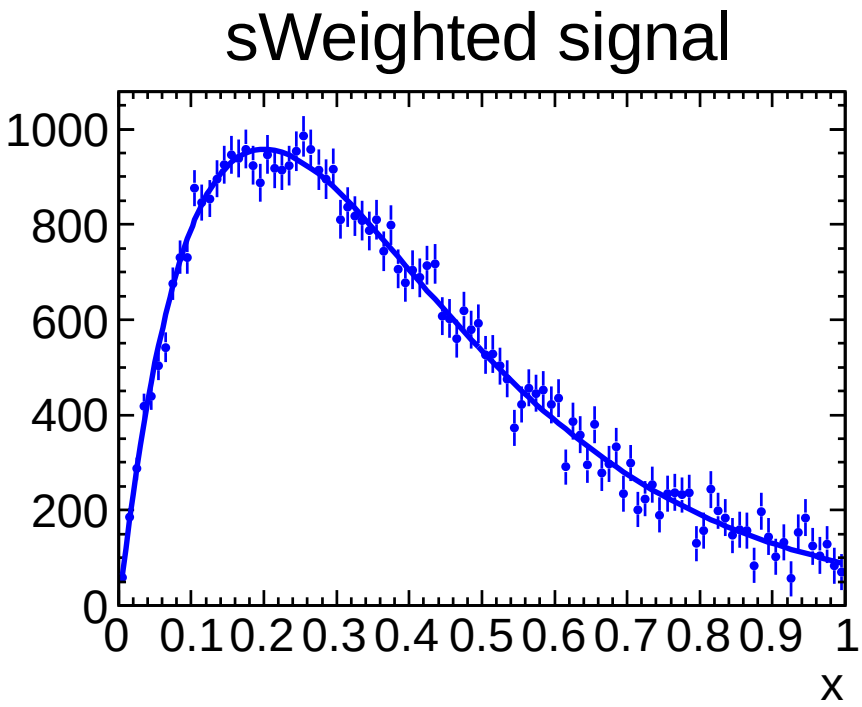
signal+background

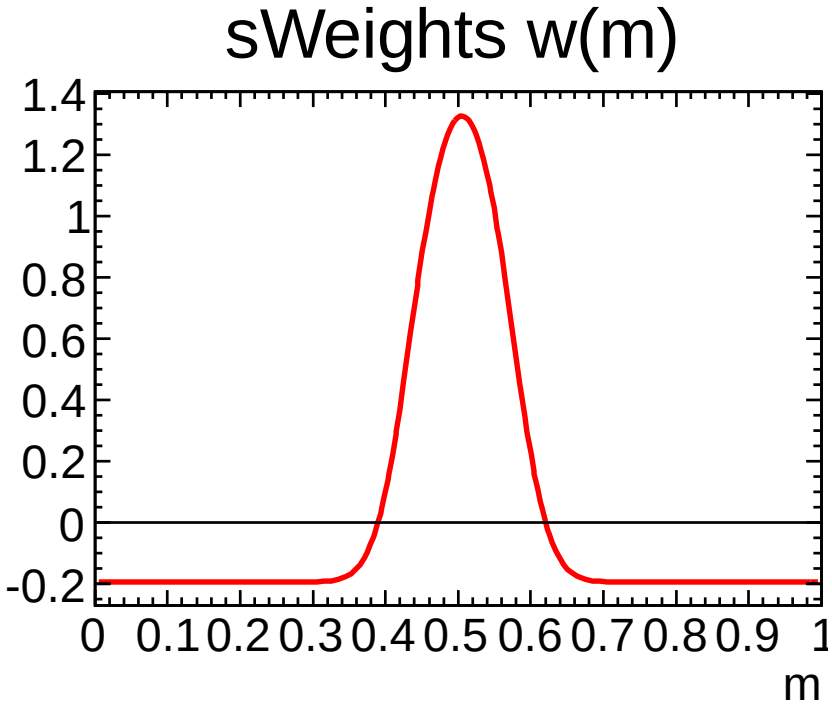


generated distributions



Extracted signal density





➜ histogram all events (x, m) with weights $w(m)$: `hx->Fill(x,w(m))`

consider the case of efficiencies $\varepsilon < 1$ ➜

Efficiencies that factorize (i)

The parameterization of the observed density $f(x,m)$ is given by

$$f(x,m) = N_s\,\varepsilon(x)\,\varepsilon(m)\,s(x)\,s(m) + N_b\,\varepsilon(x)\,\varepsilon(m)\,b(x)\,b(m)$$

which can be re-written by introducing observable quantities

$$s'(x) = \frac{\varepsilon(x)\,s(x)}{\int dx\,\varepsilon(x)\,s(x)} \quad \text{and} \quad b'(x) = \frac{\varepsilon(x)\,b(x)}{\int dx\,\varepsilon(x)\,b(x)}$$

$$s'(m) = \frac{\varepsilon(m)\,s(m)}{\int dm\,\varepsilon(m)\,s(m)} \quad \text{and} \quad b'(m) = \frac{\varepsilon(m)\,b(m)}{\int dm\,\varepsilon(m)\,b(m)}$$

$$N'_s = N_s \int dx\,\varepsilon(x)\,s(x) \int dm\,\varepsilon(m)\,s(m) \quad \text{and}$$

$$N'_b = N_b \int dx\,\varepsilon(x)\,b(x) \int dm\,\varepsilon(m)\,b(m)$$

to yield the same functional form discussed before for $\varepsilon = 1$:

$$f(x,m) = N'_s\,s'(x)\,s'(m) + N'_b\,b'(x)\,b'(m)$$

➜ $s'(x)$ can be extracted from sWeights based on the observed m-distributions

➜ *signal extraction weights $w(m)$ from the observed distributions*

$$w(m) = \frac{\alpha\, s'(m) + \beta\, b'(m)}{N'_s\, s'(m) + N'_b\, b'(m)}$$

with

$$\{\alpha, \beta\} = \frac{\{W_{bb}, -W_{sb}\}}{W_{ss}\, W_{bb} - W_{sb}^2} \quad \text{and} \quad W_{uv} = \int dm\, \frac{u'(m)\, v'(m)}{N'_s\, s'(m) + N'_b\, b'(m)}$$

Application to extract an efficiency corrected x-spectrum:

$$\int dm\, w(m)\, f(x,m) = N'_s\, s'(x) = N_s \varepsilon(x)\, s(x) \int dm\, \varepsilon(m)\, s(m)$$

Introducing a global factor F in order to get $N_s\, s(x)$:

$$F = \frac{1}{\int dm\, \varepsilon(m)\, s(m)} = \int dm \frac{1}{\varepsilon(m)} \frac{s(m)\, \varepsilon(m)}{\int dm\, \varepsilon(m)\, s(m)} = \int dm \frac{s'(m)}{\varepsilon(m)}$$

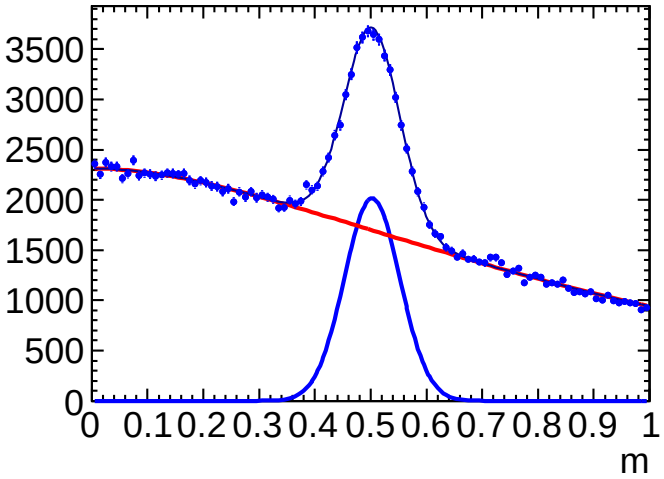
and pulling the m-independent factors $\varepsilon(x)$ and F to the LHS yields:

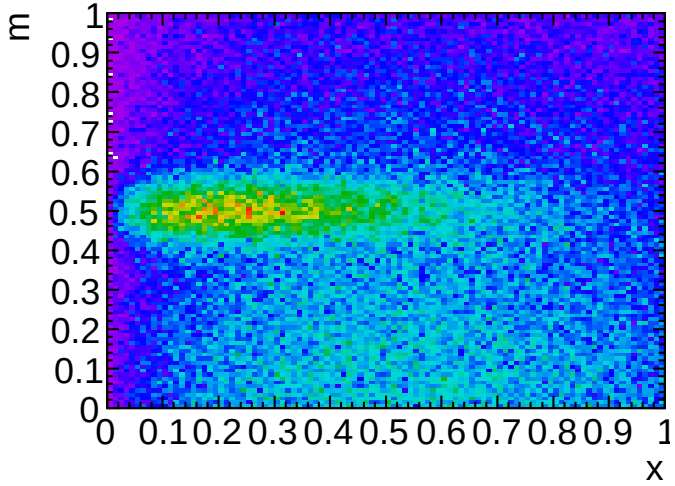
$$\int dm\; W(x,m)\, f(x,m) = N_s\, s(x) \quad \text{with} \quad W(x,m) = F\, \frac{w(m)}{\varepsilon(x)}$$

Example - factorizing efficiencies

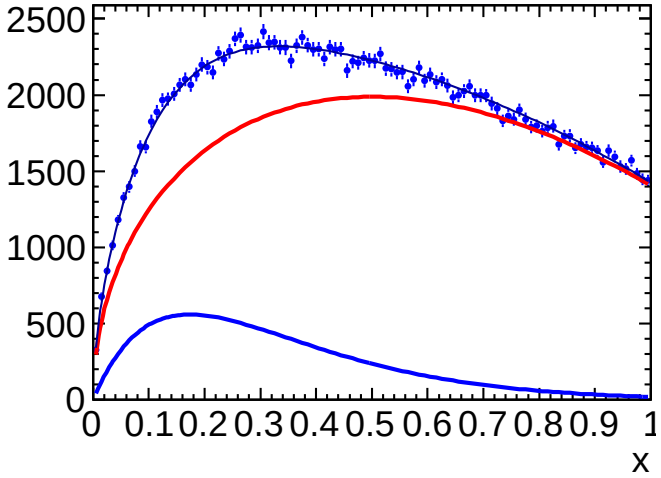
➜ *efficiency:* $\varepsilon(x, m) = ((m + 0.5)/1.5)((1.5 - x)/1.5)$

signal+background



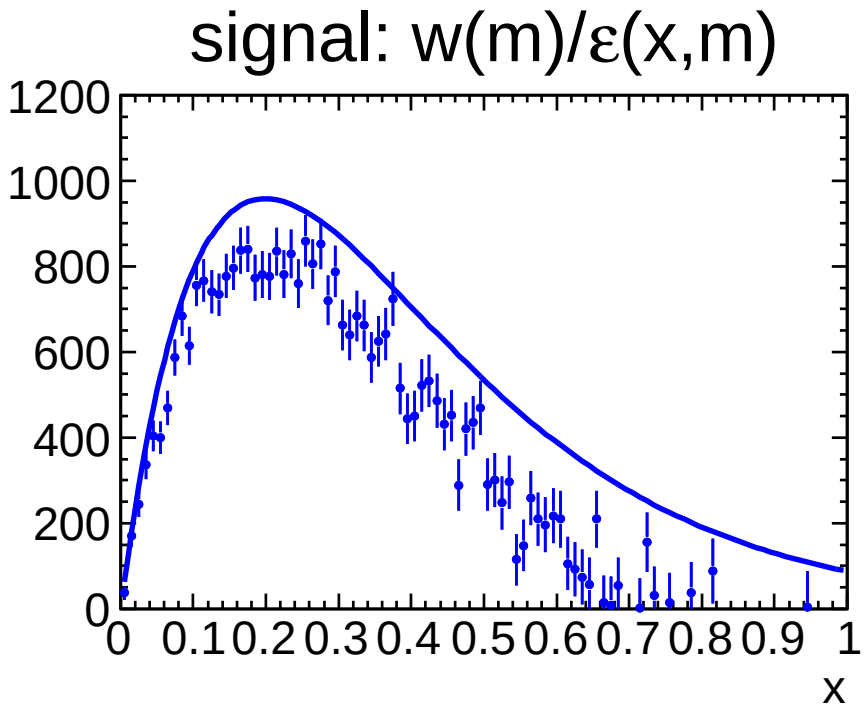


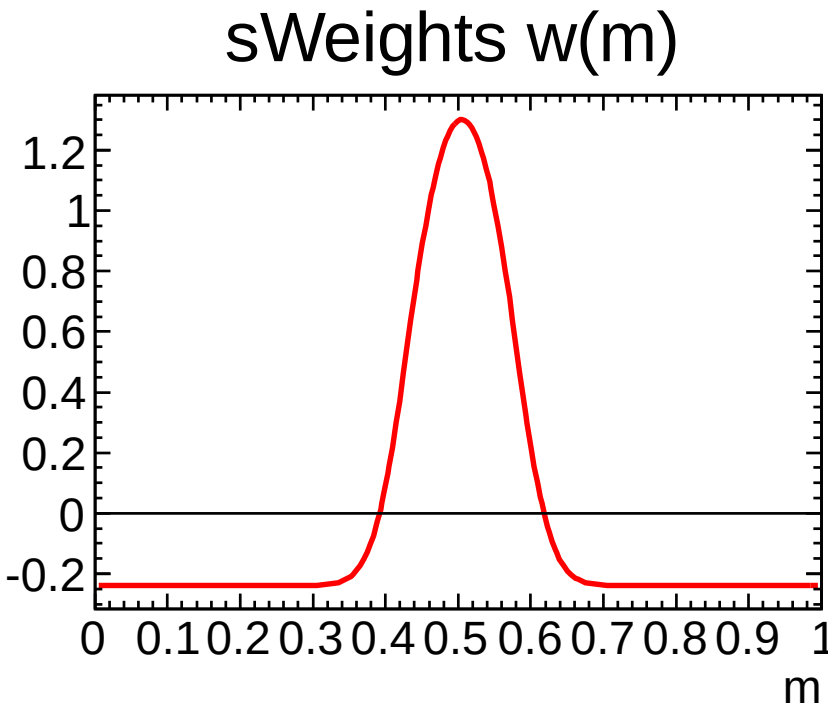
generated distributions



Naive event-by-event efficiency correction

➜ *sWeights determined for observed densities*



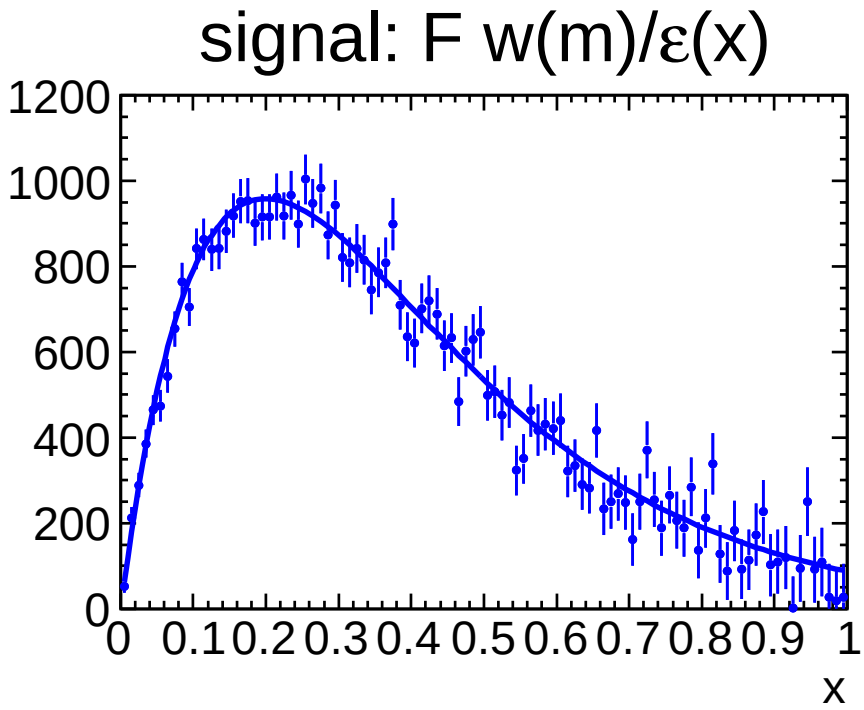


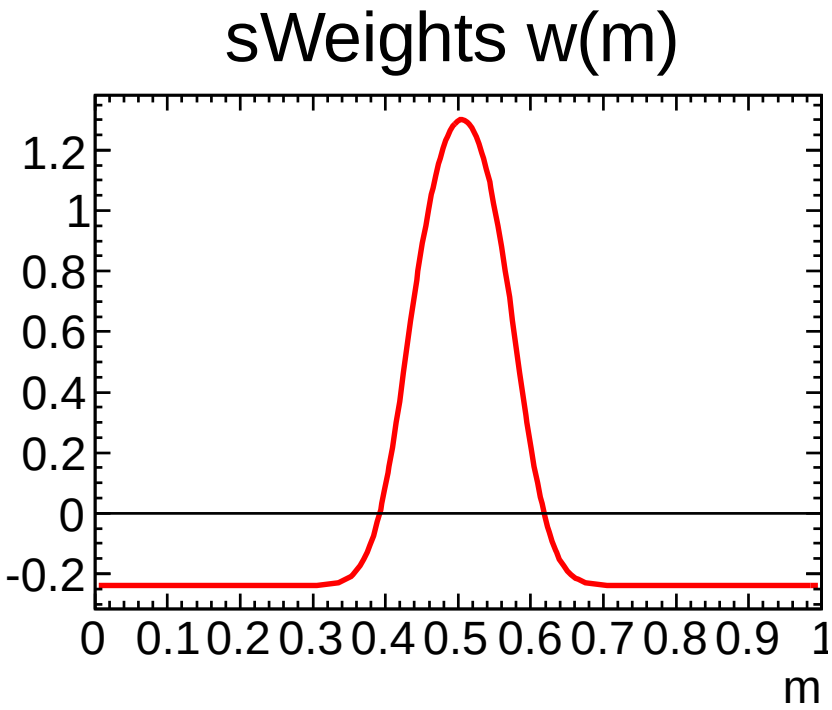
➜ events (x, m) with weights $W_0(m, x)$: `hx->Fill(x,W(m,x))` where

$$W_0(m,x) = \frac{w(m)}{\varepsilon(x,m)} = \frac{w(m)}{\varepsilon(x)\varepsilon(m)}$$

Properly extracted result

➜ *sWeights determined for observed densities*





➜ events (x, m) with weights $W(m, x)$: `hx->Fill(x,W(m,x))` where

$$W(m,x) = \frac{w(m)}{\varepsilon(x)} \cdot F \quad \text{with} \quad F = \int dm \, \frac{s'(m)}{\varepsilon(m)}$$

- sWeights based on the observed m-distribution can be used to extract $s(x)$
  - ➜ if the physics factorizes into $s(x) \cdot s(m)$ and $b(x) \cdot b(m)$
  - ➜ if the efficiency factorizes into $\varepsilon(x) \cdot \varepsilon(m)$
- correct weight: $W(x,m) = F\, w(m)/\varepsilon(x) \neq w(m)/(\varepsilon(x)\varepsilon(m))$
- per event only the efficiency $\varepsilon(x)$ is used
- $\varepsilon(m)$ enters only via the global factor F – not per event
- $\varepsilon(m)$ per event destroys normalization and orthogonality of $w(x)$

$$\int dm\, w(m)\, b'(m) = 0 \;\rightarrow\; \int dm\, \frac{w(m)}{\varepsilon(m)}\, b'(m) \neq 0$$

$$\int dm\, w(m)\, s'(m) = 1 \;\rightarrow\; \int dm\, \frac{w(m)}{\varepsilon(m)}\, s'(m) \neq 1$$

➜ incomplete background subtraction and wrong normalization!

the general case of non-factorising efficiencies ➜

➜ *parameterization of the observations*

$$f(x,m) = \varepsilon(x,m)\,[N_s\,s(x)\,s(m) + N_b\,b(x)\,b(m)]$$

- physics may be expected to factorize in x and m
- detector properties and the observed density often will not factorize

❖ new ansatz for finding a weight function:

$$\int dm\,w(m)\,\frac{f(x,m)}{\varepsilon(x,m)} = \int dm\,\left(\frac{w(m)}{\varepsilon(x,m)}\right)\,f(x,m) = N_s\,s(x)$$

with

$$\int dx\,dm\,\left(\frac{w(m)}{\varepsilon(x,m)}\right)^2 f(x,m) \stackrel{!}{=} \min \quad \text{and constraints}$$

$$\int dm\,w(m)\,s(m) = 1 \qquad \text{and} \qquad \int dm\,w(m)\,b(m) = 0$$

➜ *constrained minimization with Lagrange parameters α and β*

$$\delta \left\{ \int dx\, dm\, \frac{w^2(m)}{\varepsilon^2(x,m)} f(x,m) \right.$$
$$\left. + 2\alpha \left(1 - \int dm\, w(m)\, s(m) \right) - 2\beta \int dm\, w(m)\, b(m) \right\} = 0$$

Introducing the auxiliary function

$$q(m) = \int dx\, \frac{f(x,m)}{\varepsilon^2(x,m)} \,,$$

i.e. the per-event by $1/\varepsilon^2(x,m)$ weighted data (histogram) as a function of m (integrated over x), the defining relation for $w(m)$ becomes

$$\delta \left\{ \int dm\, w^2(m)\, q(m) - 2\alpha \int dm\, w(m)\, s(m) - 2\beta \int dm\, w(m)\, b(m) \right\} = 0$$

which yields

$$\int dm\, \delta w(m) \left\{ w(m)\, q(m) - [\alpha\, s(m) + \beta\, b(m)] \right\} = 0$$

➜ *optimal weight function to extract the efficiency corrected signal*

$$w(m) = \frac{\alpha\, s(m) + \beta\, b(m)}{q(m)} \quad \text{with} \quad q(m) = \int dx\, \frac{f(x,m)}{\varepsilon^2(x,m)}$$

with coefficients α and β from the boundary conditions

$$\{\alpha, \beta\} = \frac{\{W_{bb}, -W_{sb}\}}{W_{ss}\, W_{bb} - W_{sb}^2} \quad \text{where} \quad W_{uv} = \int dm\, \frac{u(m)\, v(m)}{q(m)}\ .$$

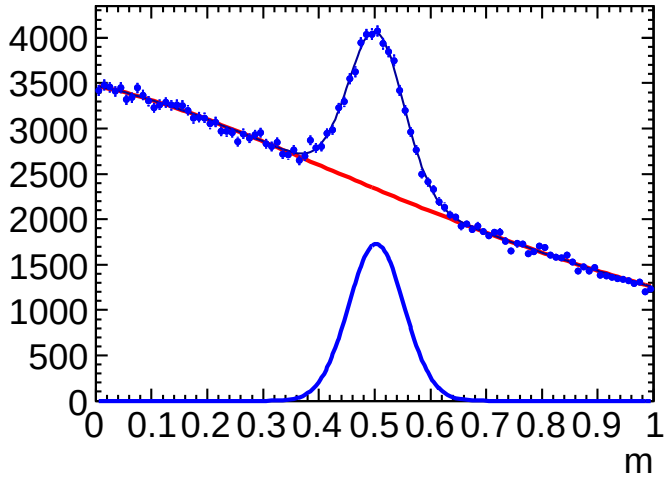
❖ discussion

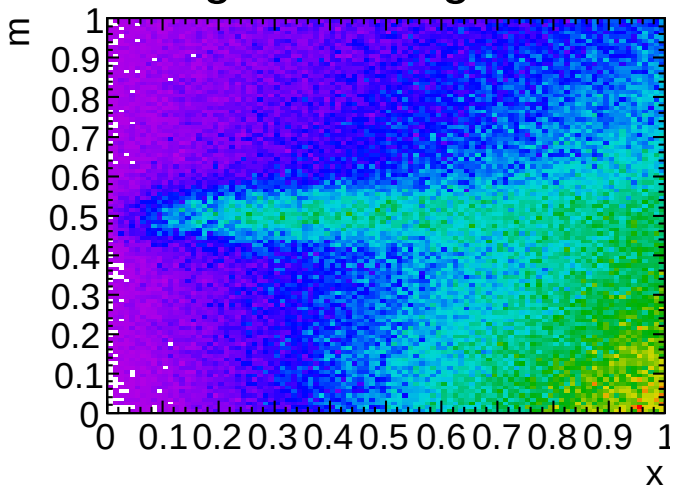
- $q(m)$ is the m-spectrum with events weighted by $1/\varepsilon^2(x,m)$,
  i.e. $q(m)\Delta m$ is the variance in each bin of the corrected m-distribution
- $s(m)$ and $b(m)$ are the efficiency corrected m-spectra
- $q(m)$, $s(m)$ and $b(m)$ are integrated over x
- the event-by-event weights to extract the signal are $w(m)/\varepsilon(x,m)$

Example - non-factorizing efficiencies

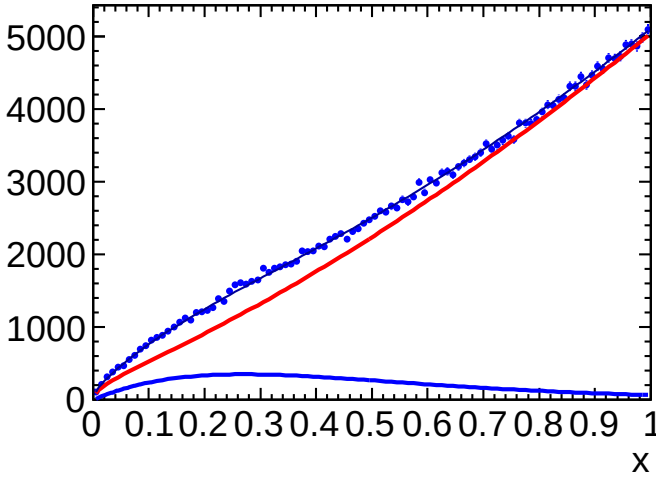
➜ *efficiency:* $\varepsilon(x, m) = (x + m)/2$

signal+background



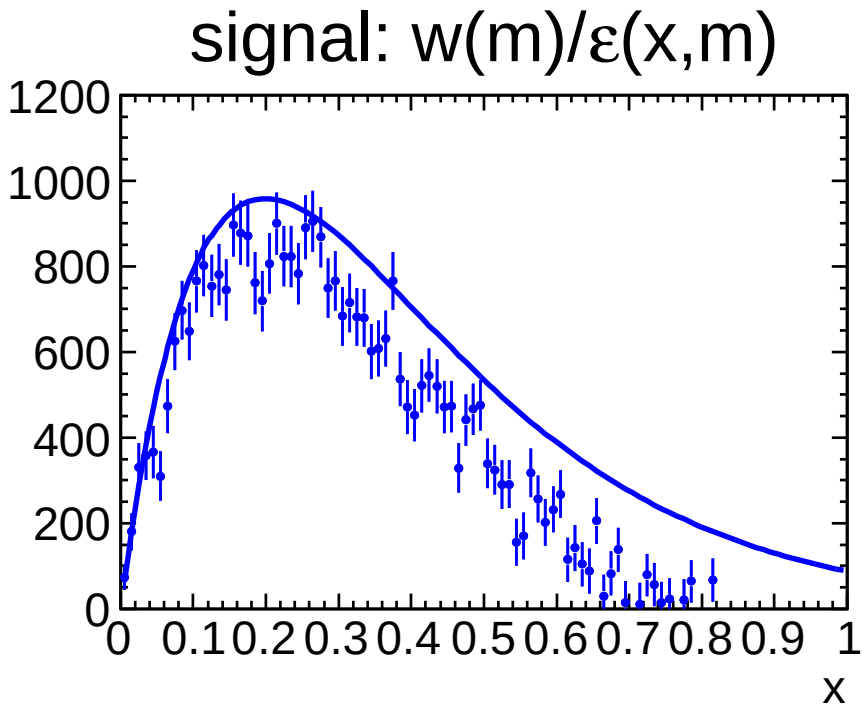


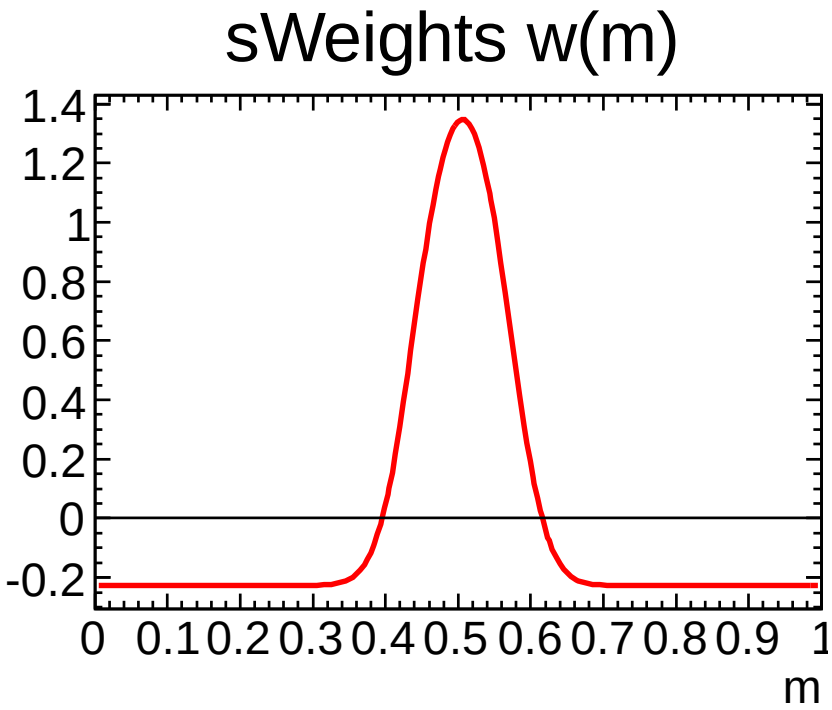
generated distributions



Naive event-by-event efficiency correction

➜ *sWeights determined for observed densities*



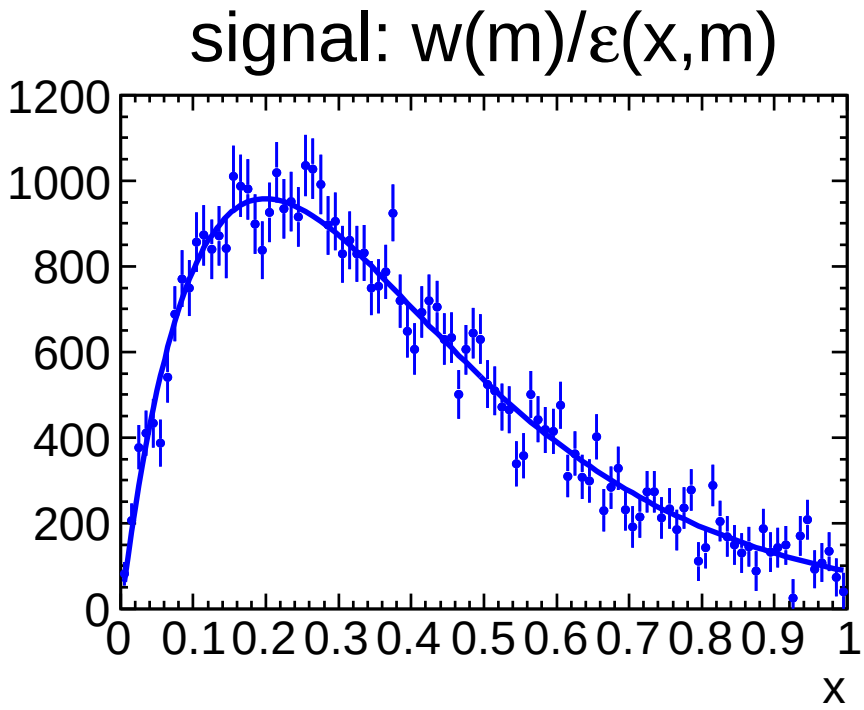


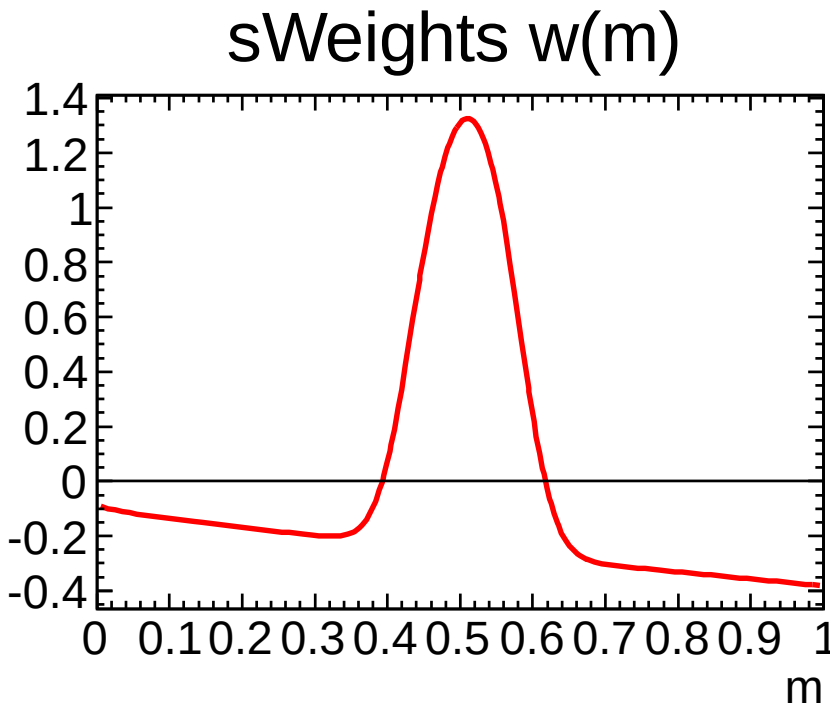
➜ events (x, m) with weights $W(m, x)$: `hx->Fill(x,W(m,x))` where

$$W(m, x) = \frac{w(x)}{\varepsilon(x, m)}$$

Properly extracted result

➜ *sWeights determined for true densities*





- determined by weighting measurements by $1/\varepsilon(x,m)$ for $s(m)$ and $b(m)$
- determined by weighting measurements by $1/\varepsilon(x,m)^2$ for $q(m)$
- event weights:

$$W(m,x) = \frac{w(m)}{\varepsilon(x,m)}$$

➜ *extract normalizations by a fit to the data*

- signal and background shapes $s(m)$ and $b(m)$ are known
- extract normalization from a least squares fit
  - ➜ assume narrow bins m and x
  - ➜ bin content and variances are n_i and σ_i^2

$$\chi^2 = \sum_i \frac{(n_i - N_s\, s(m_i)\, \Delta m - N_b\, b(m_i)\, \Delta m)^2}{\sigma_i^2} \; .$$

When events are weighted with the inverse of the efficiency one has:

$$n_i = \sum_j \Delta x \Delta m \frac{f(x_j, m_i)}{\varepsilon(x_j, m_i)} \;\rightarrow\; n_i = \Delta m \int dx \, \frac{f(x, m_i)}{\varepsilon(x, m_i)} = \Delta m \; p(m_i)$$

$$\sigma_i^2 = \sum_j \Delta x \Delta m \frac{f(x_j, m_i)}{\varepsilon^2(x_j, m_i)} \;\rightarrow\; \sigma_i^2 = \Delta m \int dx \, \frac{f(x, m_i)}{\varepsilon^2(x, m_i)} = \Delta m \; q(m_i)$$

function to be minimized:

$$\chi^2 = \sum_i \frac{(n_i - N_s\, s(m_i)\, \Delta m - N_b\, b(m_i)\, \Delta m)^2}{\sigma_i^2}$$

$$= \int dm\, \frac{(p(m) - N_s\, s(m) - N_b\, b(m))^2}{q(m)}$$

covariance matrix of normalizations N_s and N_b:

$$C_{uv}^{-1} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial N_u \partial N_v} = \int dm\, \frac{u(m)\, v(m)}{q(m)} \qquad \text{and thus}$$

$$C = \begin{pmatrix} W_{ss} & W_{sb} \\ W_{sb} & W_{bb} \end{pmatrix}^{-1} = \frac{1}{W_{ss}\, W_{bb} - W_{sb}^2} \begin{pmatrix} W_{bb} & -W_{sb} \\ -W_{sb} & W_{ss} \end{pmatrix}$$

❖ sWeights are related to the covariance matrix of the normalization fit

1. normalization of the signal distribution:

$$\sum_{\text{all events}} \frac{w(m)}{\varepsilon(x,m)} = \int dx\, dm\, \frac{w(m)}{\varepsilon(x,m)} f(x,m)$$

$$= \int dm\, w(m)\, [N_s\, s(m) + N_n\, b(m)] = N_s$$

2. variance of the normalization

$$\sum_{\text{all events}} \left(\frac{w(m)}{\varepsilon(x,m)}\right)^2 = \int dx\, dm\, \left(\frac{w(m)}{\varepsilon(x,m)}\right)^2 f(x,m)$$

$$= \int dm\, w^2(m)\, q(m) = \int dm\, w(m)\, [\alpha\, s(m) + \beta\, b(m)] = \alpha = C_{ss}$$

- normalization and variance of the signal spectrum are the same as obtained in the fit of the normalizations to the discriminant variable
- if the normalization fit had optimal precision, then sWeights are optimal to extract the signal as a function of the control variable

Summary on sWeights

- sWeights are functions orthogonal to the background density
  - ➜ signal & background are separable in a discriminant variable m
  - ➜ sWeights project out the signal component in a control variable x
  - ➜ sWeights do not quantify "signalness" ($w(m) < 0$ is allowed)
  - ➜ discriminant and control variables have to be independent
- for $\varepsilon(x,m) = \varepsilon(x) \cdot \varepsilon(m)$ sWeights for efficiency corrections can be determined from the observed densities $s'(m)$ and $b'(m)$

$$W(m,x) = F\frac{w(m)}{\varepsilon(x)} \neq \frac{w(m)}{\varepsilon(x,m)} \quad \text{with} \quad F = \int dm\, \frac{s'(m)}{\varepsilon(m)}$$

- for $\varepsilon(x,m) \neq \varepsilon(x) \cdot \varepsilon(m)$ sWeights for efficiency corrections must be determined from the corrected densities $s(m)$ and $b(m)$
  - ➜ to extract the signal use event-by-event weights $w(m)/\varepsilon(x,m)$
- everything also holds if x and m are multi-dimensional

➜ *basics about Markov chains*

- consider a system with n (discrete) states
- the system evolves in discrete time steps
- the system has no memory; subsequent steps are independent
- at time t_n the system is with probabilities $\pi_j(t_n)$ in state j
- transition probabilities $P_{kj} = p(j \to k)$ determine one time step
- conservation of probability requires

$$\sum_k P_{kj} = 1$$

- evolution of probabilities per step

$$\pi_k(t_{n+1}) = \sum_j P_{kj}\ \pi_j(t_n)$$

Chapman-Kolmogorov equation

➜ *weather forecast*

base states: $1 =$ rainy $2 =$ sunny $3 =$ cloudy

transition probabilities $p_{\mathrm{tomorrow,today}}$

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} 0.50 & 0.50 & 0.25 \\ 0.25 & 0.00 & 0.25 \\ 0.25 & 0.50 & 0.50 \end{pmatrix}$$

Columns of P are normalized. Evolution of probabilities, starting "sunny":

$$\vec{\pi}^T(t_0) = (0,1,0)$$

$$\vec{\pi}(t_1) = P\vec{\pi}(t_0) \qquad \vec{\pi}^T(t_1) = (1,0,1)/2$$

$$\vec{\pi}(t_2) = P\vec{\pi}(t_1) \qquad \vec{\pi}^T(t_2) = (3,2,3)/8$$

$$\vec{\pi}(t_3) = P\vec{\pi}(t_2) \qquad \vec{\pi}^T(t_3) = (13,6,13)/32$$

$$\vec{\pi}(t_4) = P\vec{\pi}(t_3) \qquad \vec{\pi}^T(t_4) = (51,26,51)/128$$

asymptotic $\qquad \vec{\pi}^T(t_\infty) = (0.4, 0.2, 0.4)$

➜ *convergence of the Markov chain*

Independently of the initial conditions, after a “burn-in Phase” the probabilities for the possible states approach constant values.

Subsequent steps are correlated, but the distribution of the states is described by an equilibrium distribution π_j^*.

Convergence means that detailed balance is given, i.e. if on average forward and backward steps have the same frequency

$$p(j \to k)\pi_j^* = p(k \to j)\pi_k^* \qquad \text{or} \qquad P_{kj}\pi_j^* = P_{jk}\pi_k^*$$

In this case one has for each element π^* of the probability distribution

$$(P\pi^*)_j = \sum_k P_{jk}\pi_k^* = \sum_k P_{kj}\pi_j^* = \pi_j^* \sum_k P_{kj} = \pi_j^*$$

i.e. the probabilities are constant.

➜ *basic idea*

Build transition probabilities P_{kj} which for a PDF ρ satisfy the condition of detailed balance. A Markov chain then generates samples according to ρ.

➜ *implementation*

- x, y: points in configuration space
- $P(y|x)$: probability for the transition $x \to y$
- ρ: PDF that shall be generated
- core of the algorithm: decomposition of $P(y|x)$
  - ➜ a random step $q(y|x)$ from $x \to y$, and
  - ➜ the probability $\alpha(x, y)$ to accept this step

$$P(y|x) = q(y|x) \cdot \alpha(x, y) = q(y|x) \cdot \min\left[1, \frac{q(x|y)\rho(y)}{q(y|x)\rho(x)}\right]$$

➜ show that $P(y|x)$ satisfies detailed balance for ρ

➜ *condition to be satisfied:* $P(y|x)\,\rho(x) = P(x|y)\,\rho(y)$

$$q(y|x)\rho(x)\,\min\left[1, \frac{q(x|y)\rho(y)}{q(y|x)\rho(x)}\right] = q(x|y)\rho(y)\,\min\left[1, \frac{q(y|x)\rho(x)}{q(x|y)\rho(y)}\right]$$

❖ explicitly check the different cases:

1) $q(y|x)\rho(x) = q(x|y)\rho(y)$: OK

$$q(y|x)\rho(x)\,\min\left[1, \frac{q(x|y)\rho(y)}{q(y|x)\rho(x)}\right] = q(x|y)\rho(y)\,\min\left[1, \frac{q(y|x)\rho(x)}{q(x|y)\rho(y)}\right]$$

2) $q(y|x)\rho(x) < q(x|y)\rho(y)$: OK

$$q(y|x)\rho(x)\,[1] = q(x|y)\rho(y)\,\left[\frac{q(y|x)\rho(x)}{q(x|y)\rho(y)}\right]$$

3) $q(y|x)\rho(x) > q(x|y)\rho(y)$: OK

$$q(y|x)\rho(x)\,\left[\frac{q(x|y)\rho(y)}{q(y|x)\rho(x)}\right] = q(x|y)\rho(y)\,[1]$$

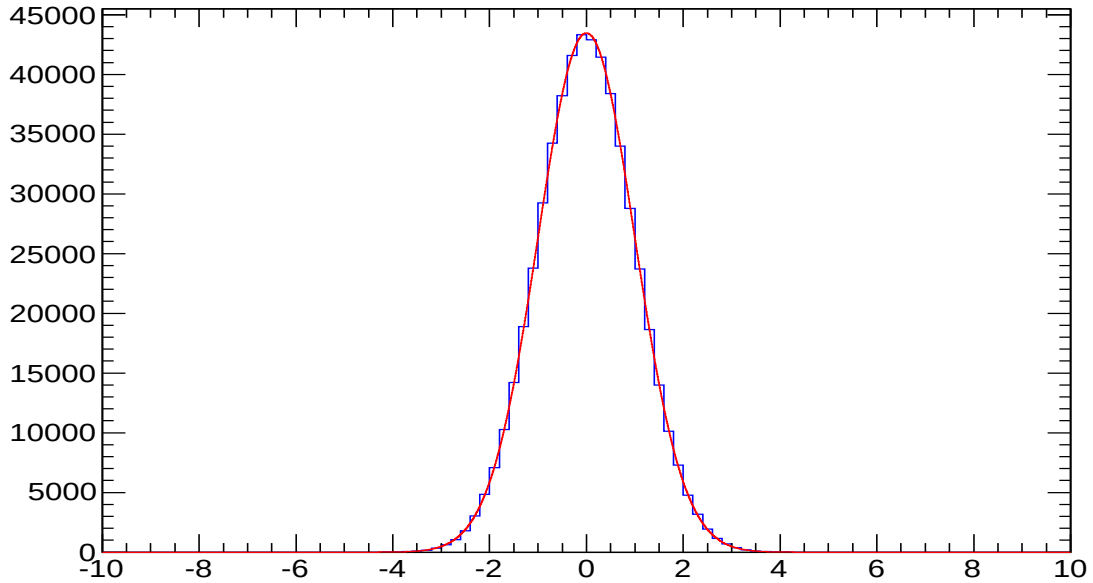
- the algorithm is independent of the dimension of $\rho(x)$
- it works for arbitrary functions $q(x, y)$
- general case: $q(y|x) \neq q(x|y)$, i.e. directional steps
- $q(y|x) = q(x|y)$: Metropolis algorithm

$$\alpha(x, y) = \min\left[1, \frac{\rho(y)}{\rho(x)}\right]$$

- normalization of ρ not needed
- choice of $q(y|x)$ affects speed of convergence
- problem: correlations between subsequent steps
  - ➜ small steps ➜ slow movement towards the next maximum
  - ➜ large steps ➜ danger to be trapped at sharp maxima
- using only every n-th value reduces correlations
- ideal jump-function depends on $\rho(x)$
- method well suited for Monte Carlo integration

➜ try it

- consider different probability densities in $-a < x < a$
  - ➜ gaussian: $\exp(-x^2/2)$
  - ➜ exponential: $\exp(-x)$ for $x > 0$
  - ➜ singular density: $1/\sqrt{x}$ for $x > 0$
  - ➜ rapidly oscillating density: $\sin^2(1/x)$
- jump function: $q(x \to y) : y = x + 0.1\ \mathrm{rndm}(-\mathrm{a}, \mathrm{a})$



➜ *integrate a gaussian over ± 3-sigma interval*

$$\langle f \rangle = \int_{-3}^{3} dx\, p(x) f(x) = \int_{-3}^{3} dx\, \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 0.9973002\ldots$$

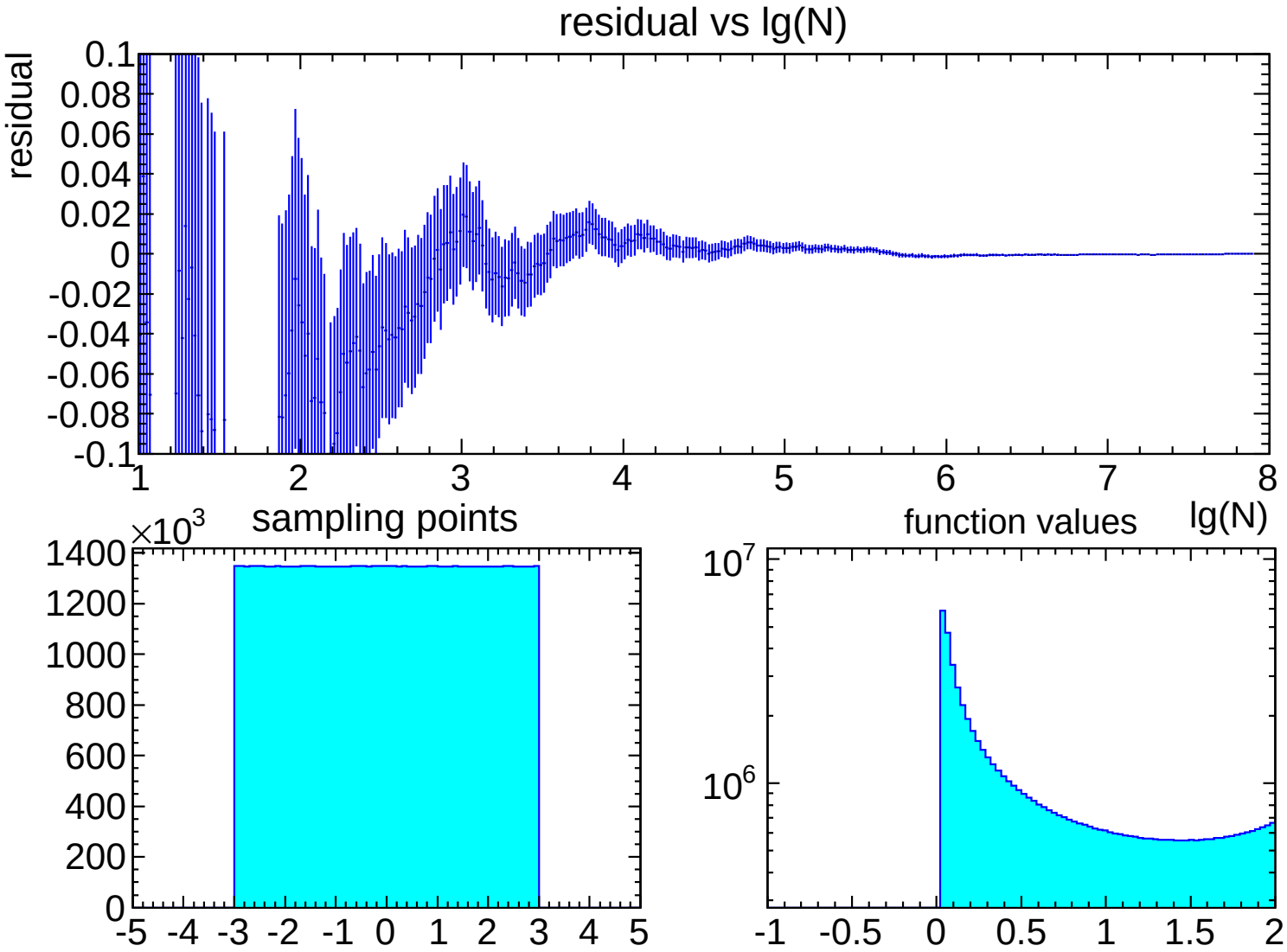
- naive Monte Carlo: uniform distribution for x

$$p(x) = \frac{1}{6} \quad \text{for} \quad x \in [-3, 3] \quad , \quad f(x) = \frac{6}{\sqrt{2\pi}} e^{-x^2/2}$$

- importance sampling “classic”: generate x according to $p(x)$

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad , \quad f(x) = \Theta(x+3)\Theta(3-x)$$

- MCMC: generate x according to $p(x)$ by a Markov chain


residual vs lg(N)
residual
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
1
2
3
4
5
6
7
8
lg(N)
×10^3
sampling points
1400
1200
1000
800
600
400
200
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
function values
10^7
10^6
-1
-0.5
0
0.5
1
1.5
2

- all tested methods converge towards the correct answer
- no reliable error estimate from MCMC
  - ➜ consequence of correlations between subsequent points
  - ➜ empirical variance underestimates the true scatter
  - ➜ similar issues with Quasi Monte Carlo
- similar convergence rates for MCMC and importance sampling
  - ➜ importance sampling requires manual tuning of random numbers
  - ➜ perfect match often difficult to get (if at all)

$$\int dx\, p(x)\, f(x) \to \int dx\, g(x)\, \frac{f(x)p(x)}{g(x)}$$

- advantages of MCMC:
  - ➜ automatically asymptotically optimal sampling
  - ➜ even in case of very strongly varying PDFs
  - ➜ ideal to determine marginal distributions

➜ *reducing the dimensionality of a PDF*

- given is a 2-dimensional PDF $f(x, y)$
- wanted are the PDFs in only one of the variables $f_1(x)$ and $f_2(y)$
  - ➜ projections of the PDFs to the coordinate axes
  - ➜ or the margin around the plot ➜ "marginal distributions"
- formally:

$$f_1(x_0) = \int dy\, f(x_0, y) \qquad \text{and} \qquad f_2(y_0) = \int dx\, f(x, y_0)$$

- special case of a much more general situation
  - ➜ given: a multidimensional space with a PDF $f(\vec{x})$
  - ➜ want: PDF on a subspace with a constraint between variables
- example from bayesian statistics
  - ➜ given a PDF of physics parameters x and "nuisance parameters" $f(x, \nu_1, \nu_2, ..)$. The PDF $f(x)$ is obtained by "marginalization", i.e. integrating over the nuisance parameters.

Marginal distributions (ii)

- a particle physics example:
  - ➜ the cross-section for the production of J/ψ mesons is $\sigma(y, p_T, \phi)$
  - ➜ determine $\sigma(p_T)$ and $\sigma(p_x)$
    - trivial marginalization for $\sigma(p_T)$
    - more complex for $\sigma(p_x)$ since $p_x = F(y, p_T, \phi)$

➜ *general case*

- given PDF: $f(\vec{x})$ ➜ wanted PDF: $g(y)$ with $y = F(\vec{x})$

  solution: $$g(y) = \int d^n x \, f(\vec{x}) \, \delta(y - F(\vec{x}))$$

- integral over PDF numerically calculable by MC/MCMC
  - ➜ sample the space according to density $f(x)$
  - ➜ histogram for each point the value $y = F(\vec{x})$.

➜ *study a toy fitting problem*

- true distribution in $0 < x < 10$:

$$f(x;p) = p_0 + p_1 \exp(-p_2 x) + p_3 \exp\left(-(x-p_4)^2 \cdot p_5\right)$$

- study binned distribution with 50 bins
- define nominal settings for the parameters (to be found by the fit)
- define ranges for the parameter values (possibly by looking at data)

	p_0	p_1	p_2	p_3	p_4	p_5
nominal values	1.0	3.0	0.4	1.0	5.0	1.0
lower bounds	0.0	0.0	0.0	0.0	0.0	0.1
upper bounds	4.0	4.0	2.0	4.0	10.	5.0

- study equivalent statistics of 10^3 and 10^4 events (values at bin center)
- gaussian fluctuations with known width proportional to $\sqrt{f(x, p_{nom})}$

Reference: classic fitting by MINUIT

➜ *frequentist approach: study the likelihood function*

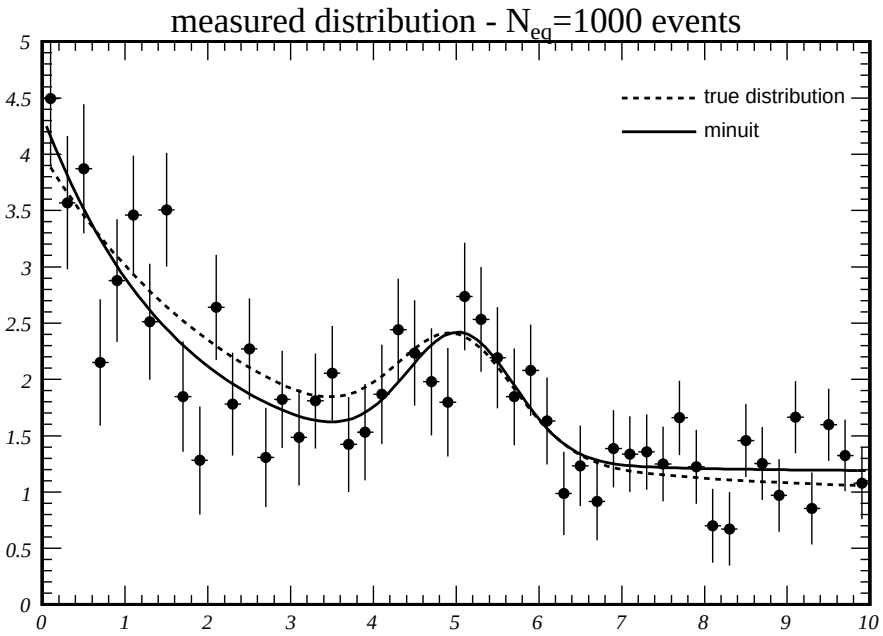
- "best" estimates for parameters minimize $-\ln L$
- two types of error estimates:
  - ➜ parabolic errors from the inverse of the second derivative of $-\ln L$
  - ➜ "MINOS" errors defined by the maximum range with $\Delta \ln L < 0.5$
- for gaussian likelihood functions the two are identical
- MINOS-errors are believed to have coverage
  - ➜ true values in $\approx 68.3\,\%$ of the cases within the error interval

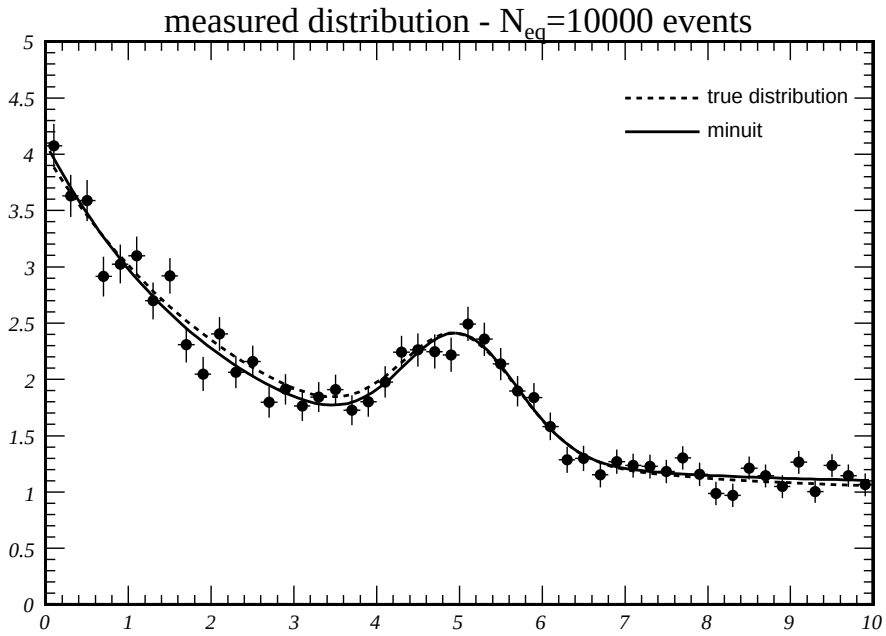
❖ numerical studies

- study sensitivity of fit results to starting values
  - ➜ 10^4 randomly chosen starting points
- study coverage when starting fit at the nominal parameter values
  - ➜ 10^4 pseudo experiments

Reference values of parameters and errors

➜ *fit results with starting point at nominal parameters*

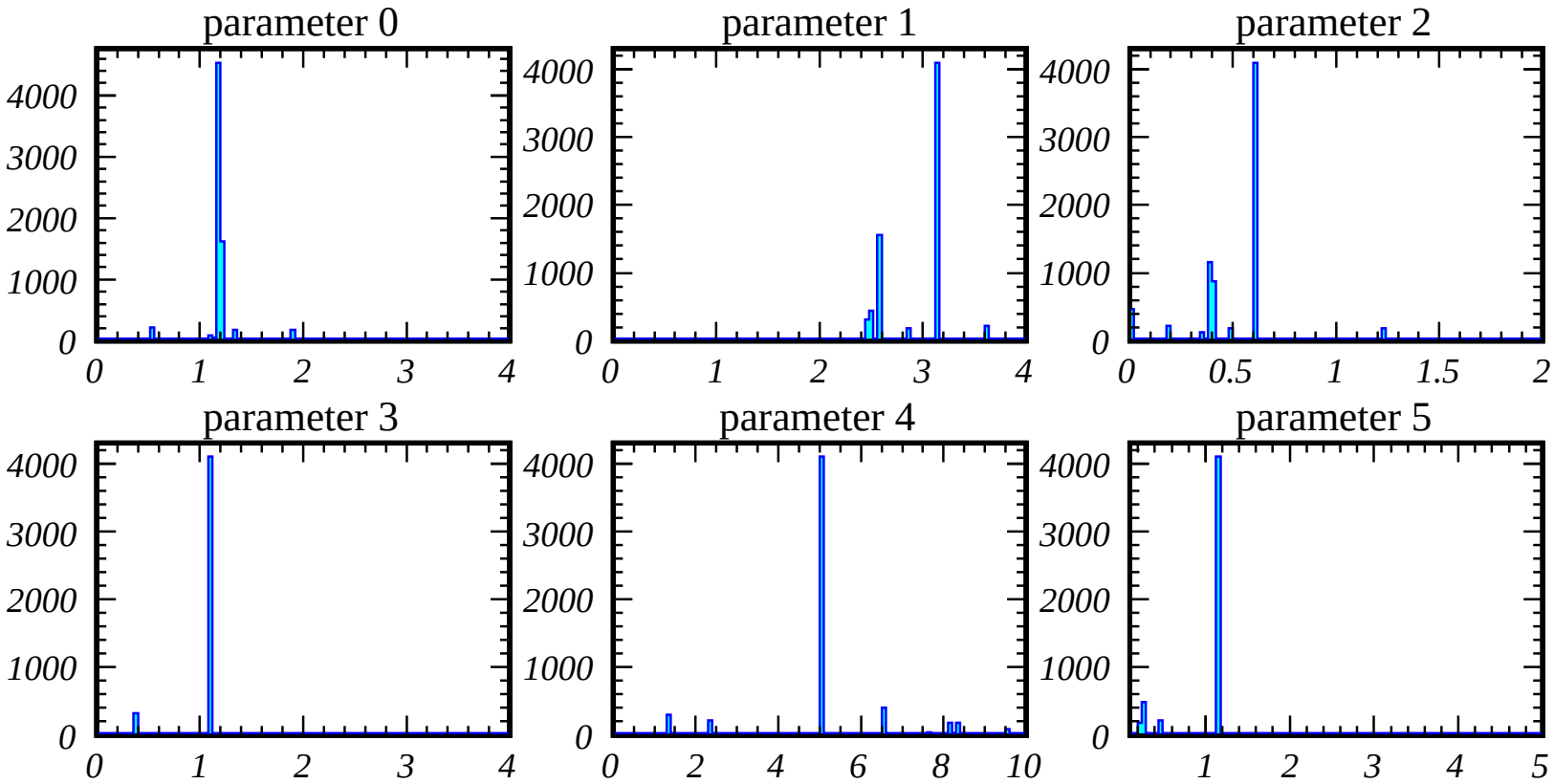




p0:	1.184 +- 0.088 [+0.086 -0.092]	1.073 +- 0.035 [+0.034 -0.036]
p1:	3.160 +- 0.421 [+0.436 -0.406]	3.025 +- 0.116 [+0.118 -0.115]
p2:	0.610 +- 0.129 [+0.138 -0.122]	0.461 +- 0.036 [+0.036 -0.035]
p3:	1.090 +- 0.237 [+0.242 -0.233]	1.034 +- 0.075 [+0.076 -0.075]
p4:	5.053 +- 0.170 [+0.167 -0.177]	5.012 +- 0.057 [+0.055 -0.056]
p5:	1.171 +- 0.546 [+0.721 -0.445]	1.033 +- 0.171 [+0.186 -0.158]

Sensitivity to inital values for $N_{eq} = 1000$

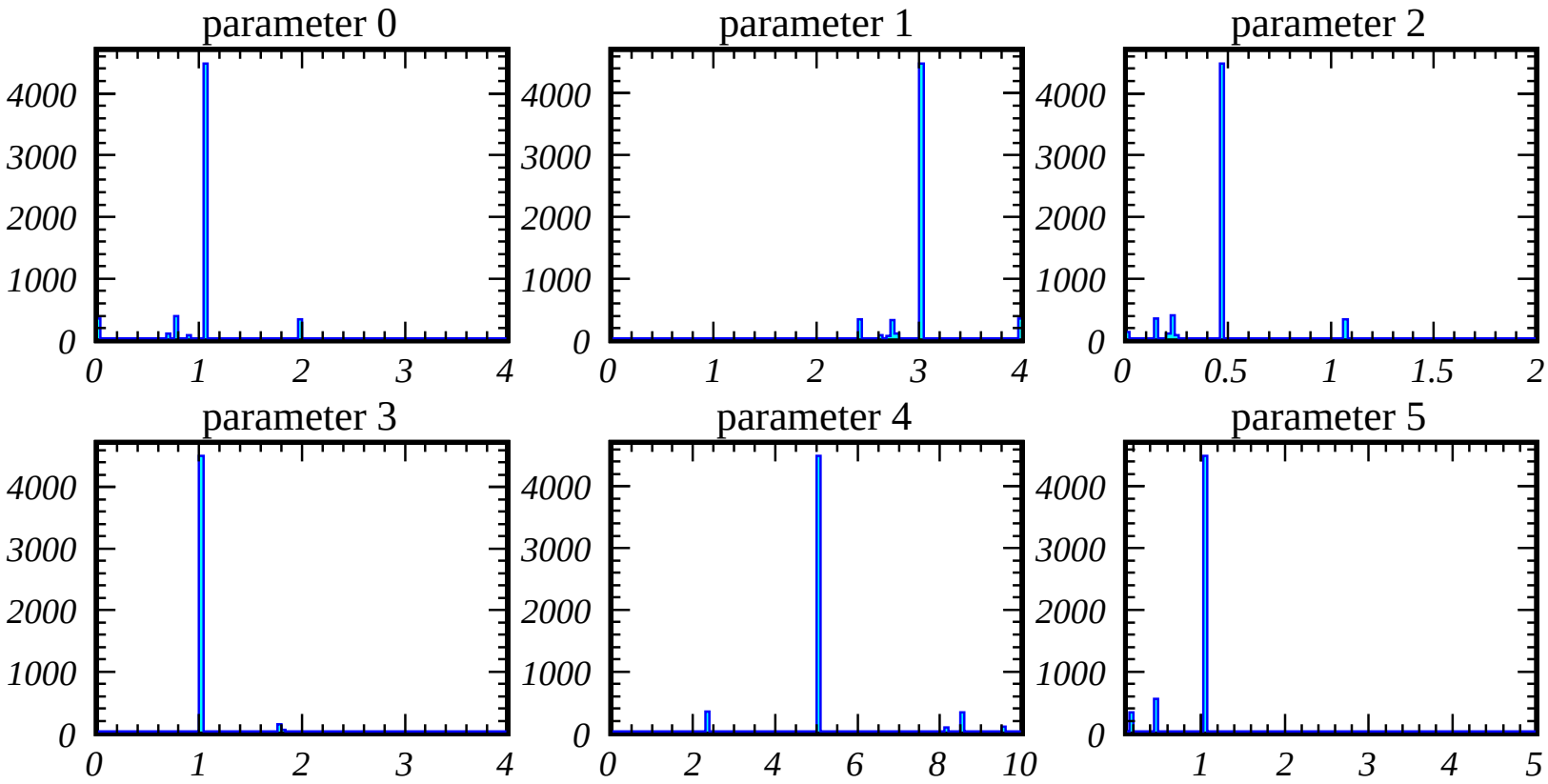
➜ *fitted parameters for random starting values*




```
10000  fits performed
 7474  fits successful
```

Sensitivity to inital values for $N_{eq} = 10000$

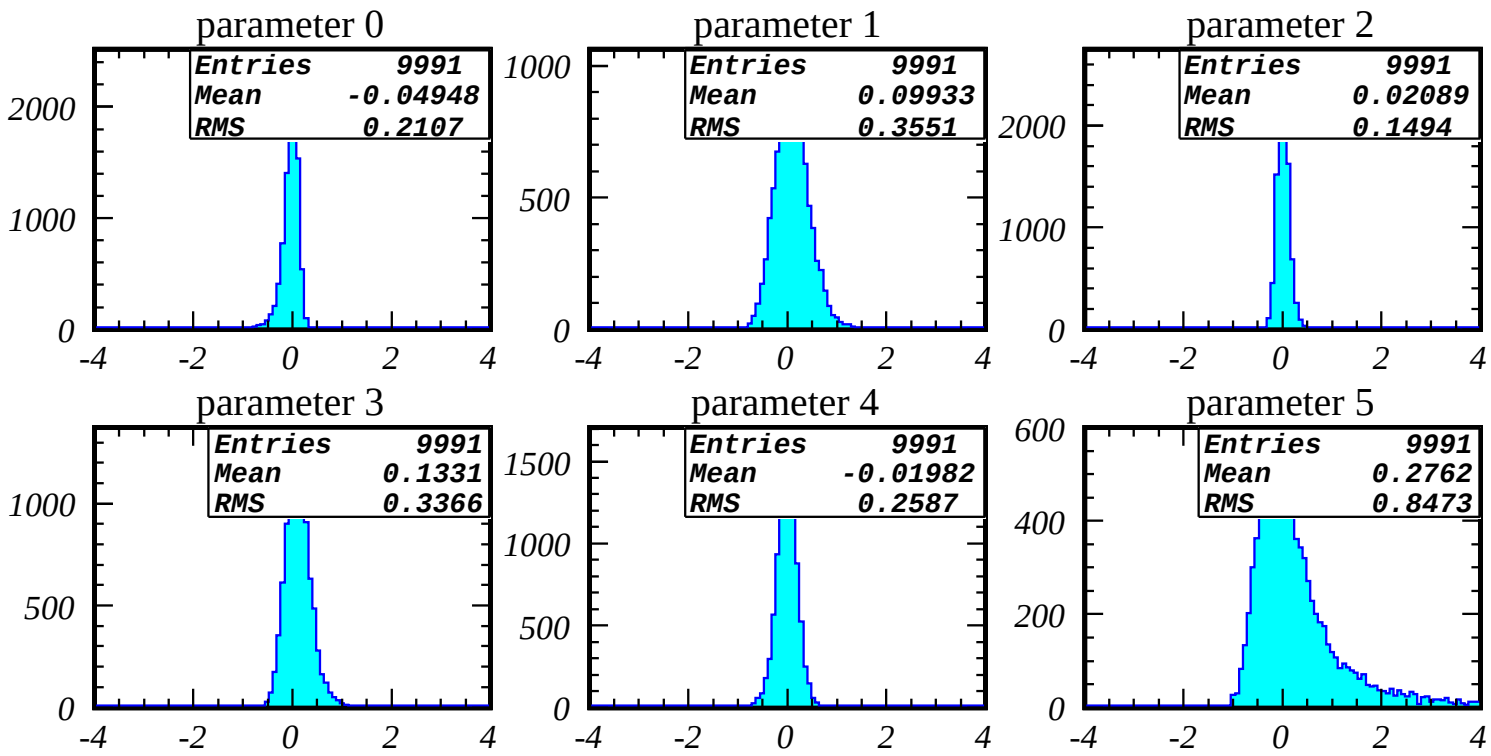
➜ *fitted parameters for random starting values*




```
10000  fits performed
 6061  fits successful
```

Coverage Check for $N_{eq} = 1000$

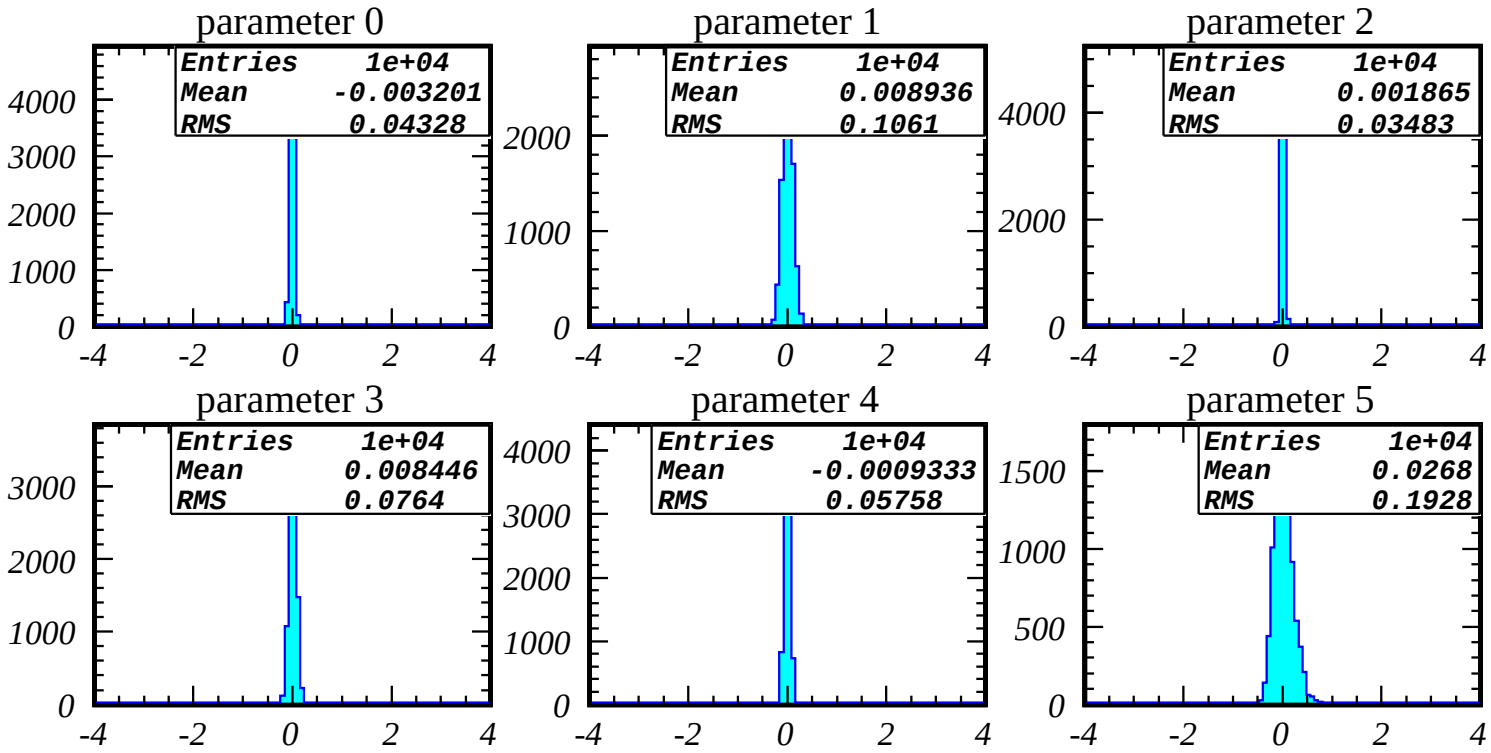
➜ *difference between fitted and true values*




```
coverage p[0] parab: 0.745 +- 0.004 minos: 0.686 +- 0.005
coverage p[1] parab: 0.698 +- 0.005 minos: 0.701 +- 0.005
coverage p[2] parab: 0.661 +- 0.005 minos: 0.652 +- 0.005
coverage p[3] parab: 0.684 +- 0.005 minos: 0.667 +- 0.005
coverage p[4] parab: 0.627 +- 0.005 minos: 0.629 +- 0.005
coverage p[5] parab: 0.661 +- 0.005 minos: 0.583 +- 0.005
```

Coverage Check for $N_{eq} = 10000$

➜ *difference between fitted and true values*




```
coverage p[0] parab: 0.695 +- 0.005 minos: 0.690 +- 0.005
coverage p[1] parab: 0.681 +- 0.005 minos: 0.682 +- 0.005
coverage p[2] parab: 0.674 +- 0.005 minos: 0.675 +- 0.005
coverage p[3] parab: 0.681 +- 0.005 minos: 0.681 +- 0.005
coverage p[4] parab: 0.686 +- 0.005 minos: 0.687 +- 0.005
coverage p[5] parab: 0.676 +- 0.005 minos: 0.671 +- 0.005
```

- minimization often fails for unlucky starting values
- problems with low statistics
  - ➜ wrong minima
  - ➜ biased parameters
  - ➜ error estimates do not have coverage
  - ➜ no significant difference between parabolic and minos errors
- generally better (more gaussian) behaviour at large statistics
  - ➜ still sensitivity to unlucky starting values
  - ➜ fewer cases of wrong minima
  - ➜ small biases
  - ➜ better coverage both for parabolic and minos errors
- note: exact coverage would require e.g. by Neyman construction
  - ➜ rarely done in more than 1 dimension

➜ *basic idea:*

- use MCMC to sample the likelihood function
- exploit information collected during the sampling process to extract information about unknown parameters
- serve simultaneously bayesian and frequentist approach
  - ➜ bayesian: calculate posteriors for uniform priors
    - best fit value from the maximum of the posterior
    - error interval from most compact 68.3% quantile
  - ➜ frequentist: keep track of minimum $-\ln L$ vs parameter value
    - best fit value from smallest $-\ln L$
    - error estimate from range with $\Delta \ln L < 0.5$
- try the approach for two different sampling schemes
  - ➜ fixed scale gaussian jumps
  - ➜ variable scale gaussian jumps

➜ *how to jump in parameter space . . .*

For each parameter:

$$p \to p + R \cdot \frac{p_{\max} - p_{\min}}{4} \cdot \texttt{rnmd.Gaus()}$$

with

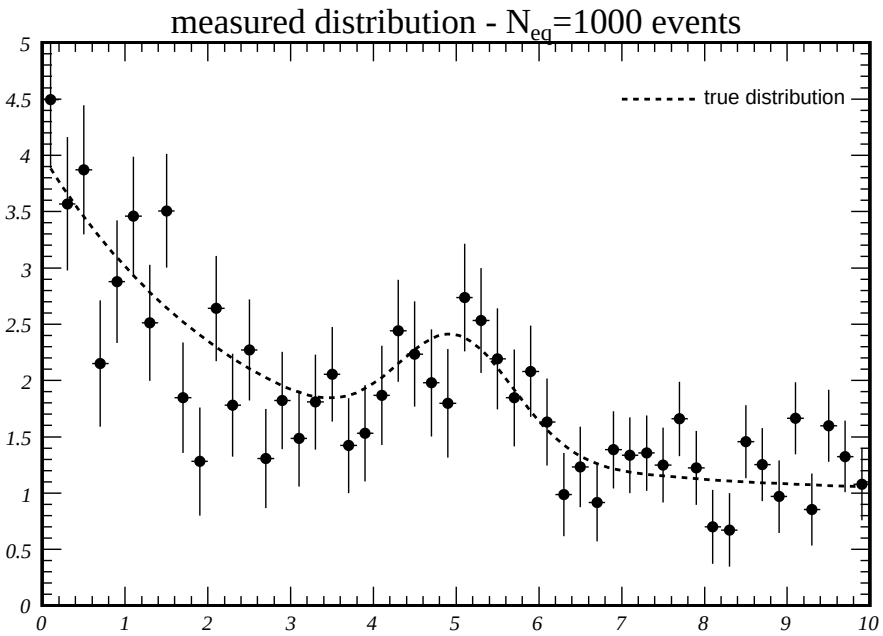
- ➜ fixed jump scale: $R = 1$
- ➜ variable jump scale: $R = (\texttt{rndm.Uniform()})^6$

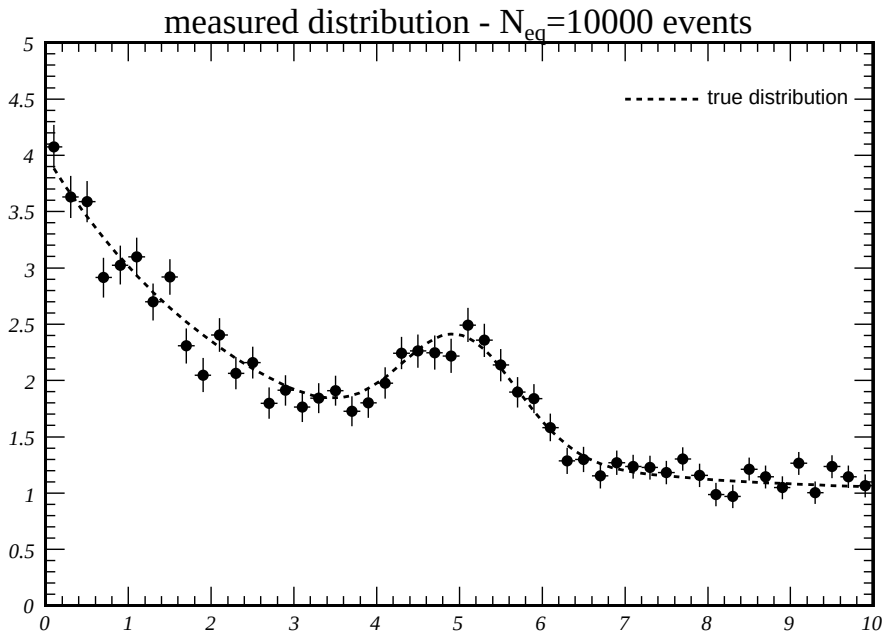
and folding back to the range $[p_{\min}, p_{\max}]$. New points are accepted with a probability given by the ratio of the new to the old likelihood. Otherwise the previous point is sampled again.

❖ properties

- gaussian jumps assure sampling with highest density around the given point
- variable scale approach has mostly small steps, i.e. emphasizes the neighbourhood of large likelihood, but features occasionally large jumps, which avoid being trapped at a local maximum

➜ *measurements to be fitted*



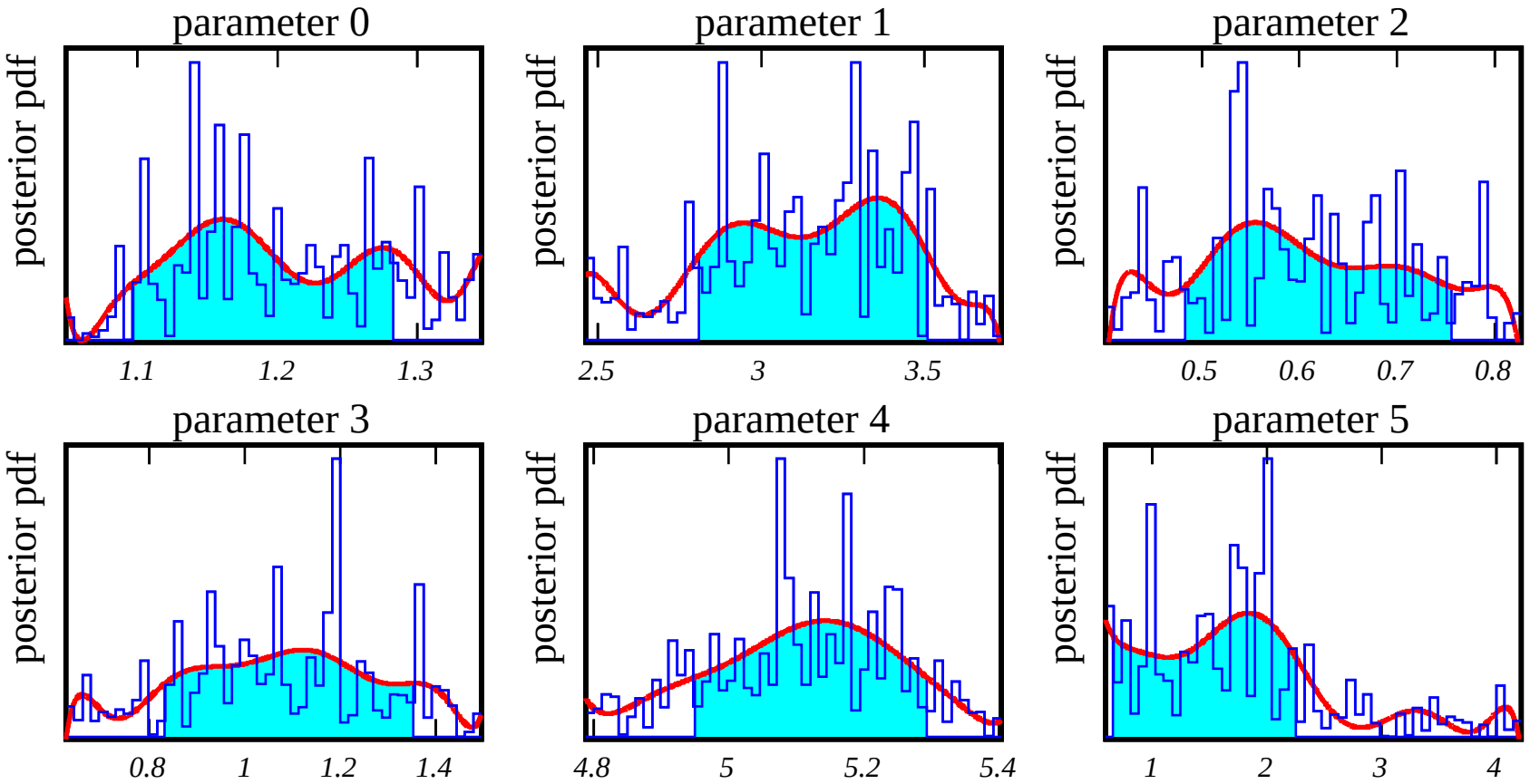


❖ try:

- 10^6 fixed scale gaussian jumps for $N_{eq} = 1000$
- 10^8 fixes scale gaussian jumps for $N_{eq} = 1000$
- 10^5 variable scale gaussian jumps for $N_{eq} = 1000$
- 10^5 variable scale gaussian jumps for $N_{eq} = 10000$

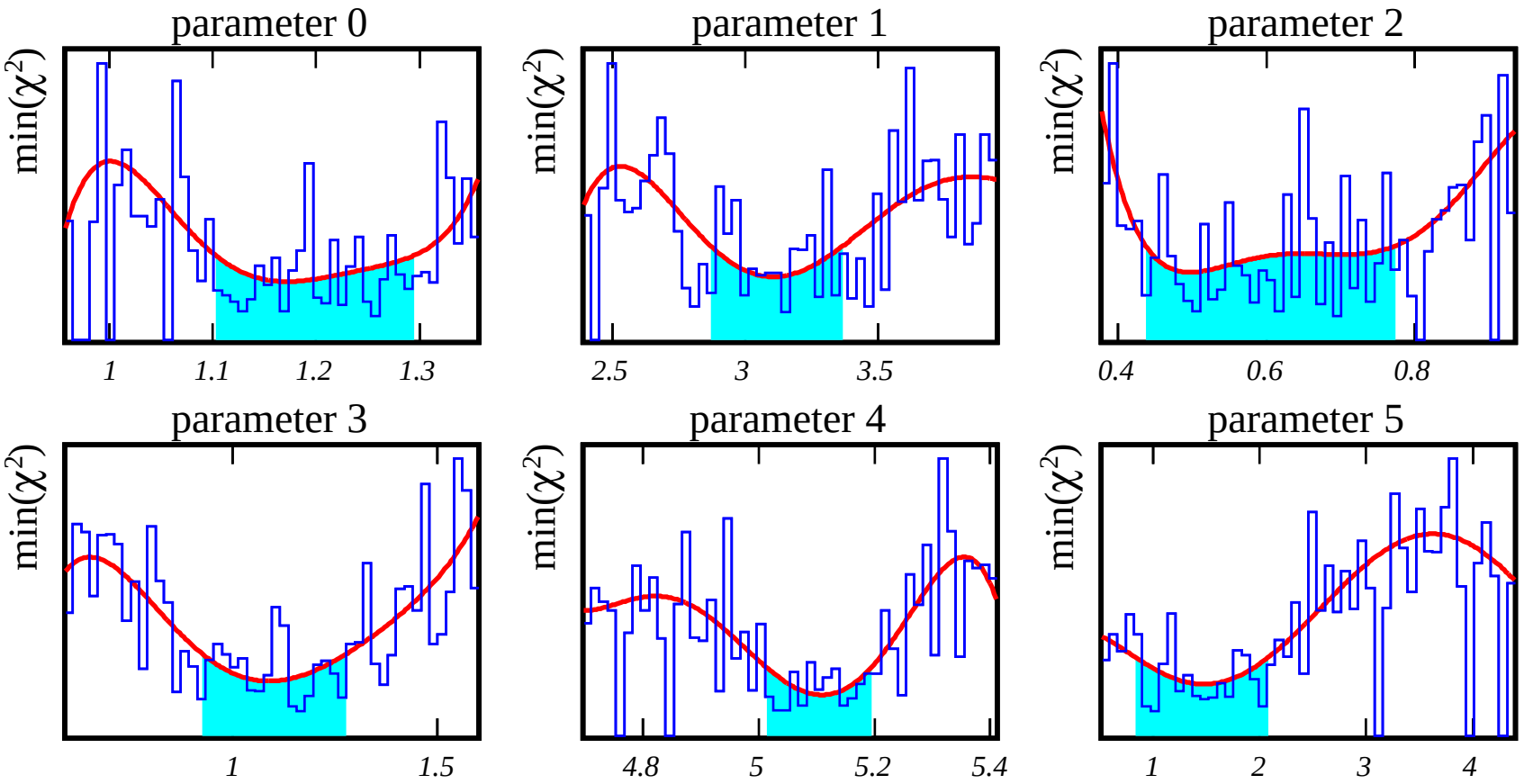
results ➜

➜ *bayesian posteriors:*



➜ not very well defined PDFs

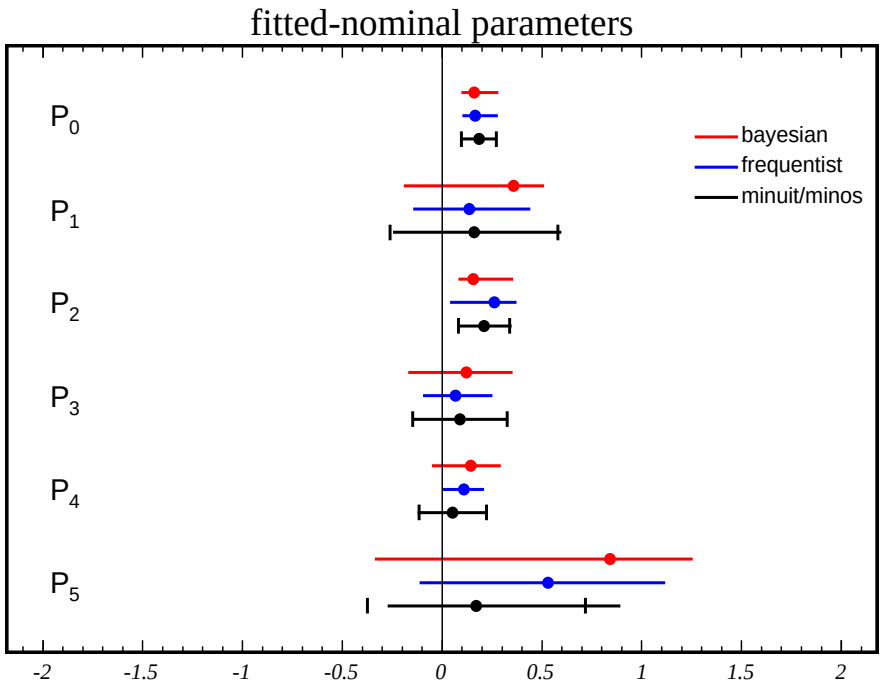
➜ *frequentist* – $\ln L$ *scans:*

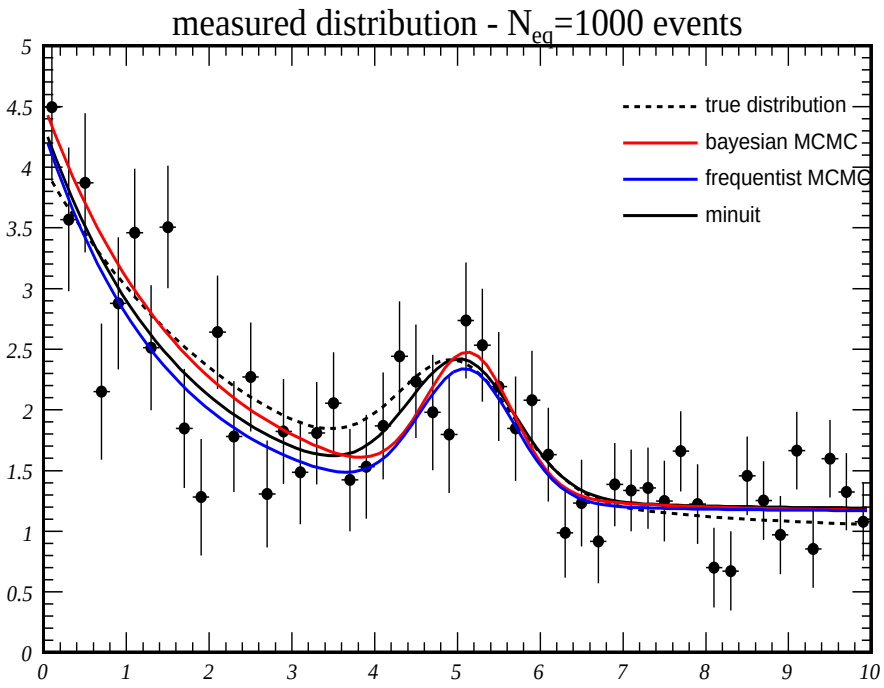


➜ not very well defined likelihood contours

$N_{eq} = 1000$: 10^6 *fixed-scale jumps*

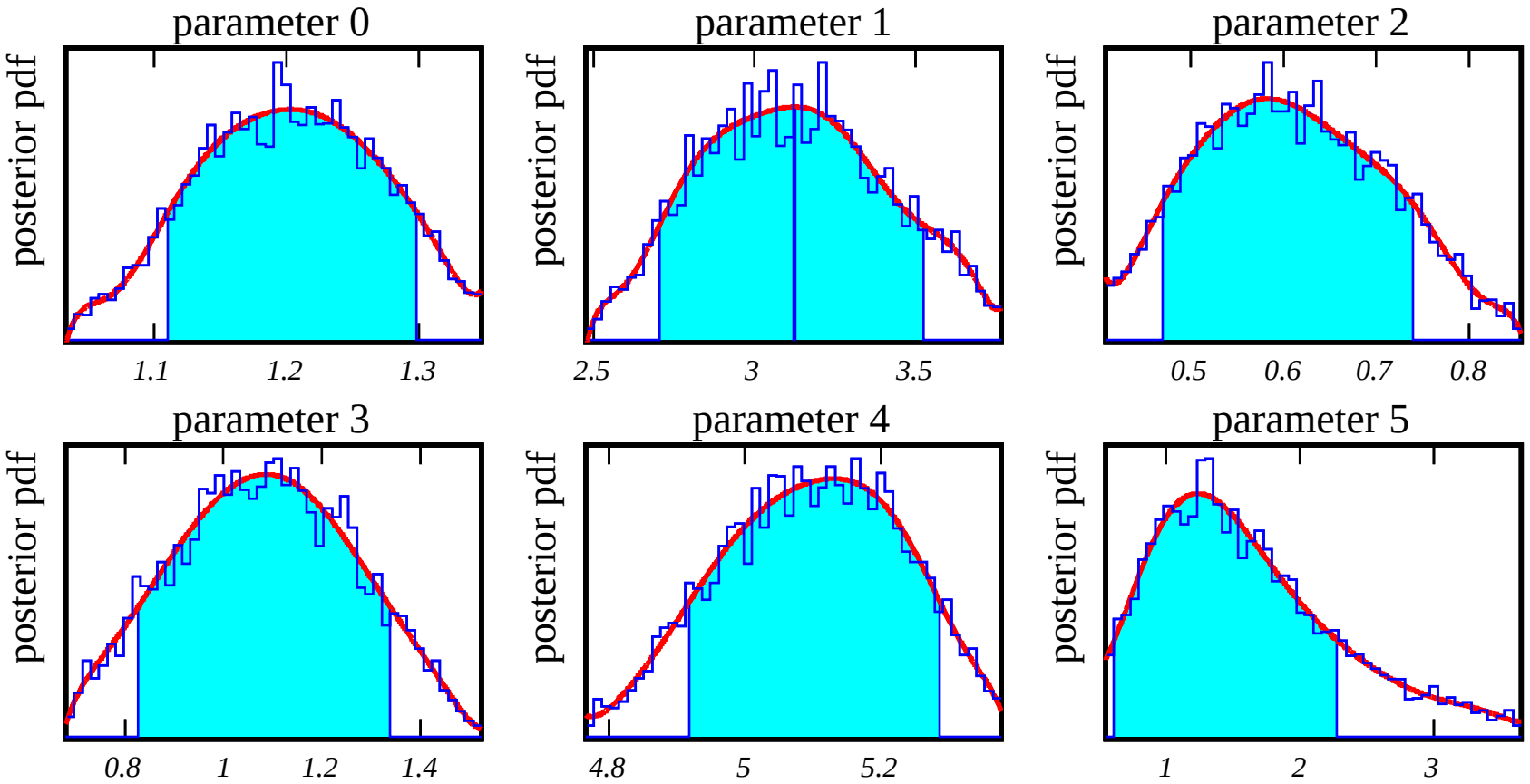
➜ *overview*





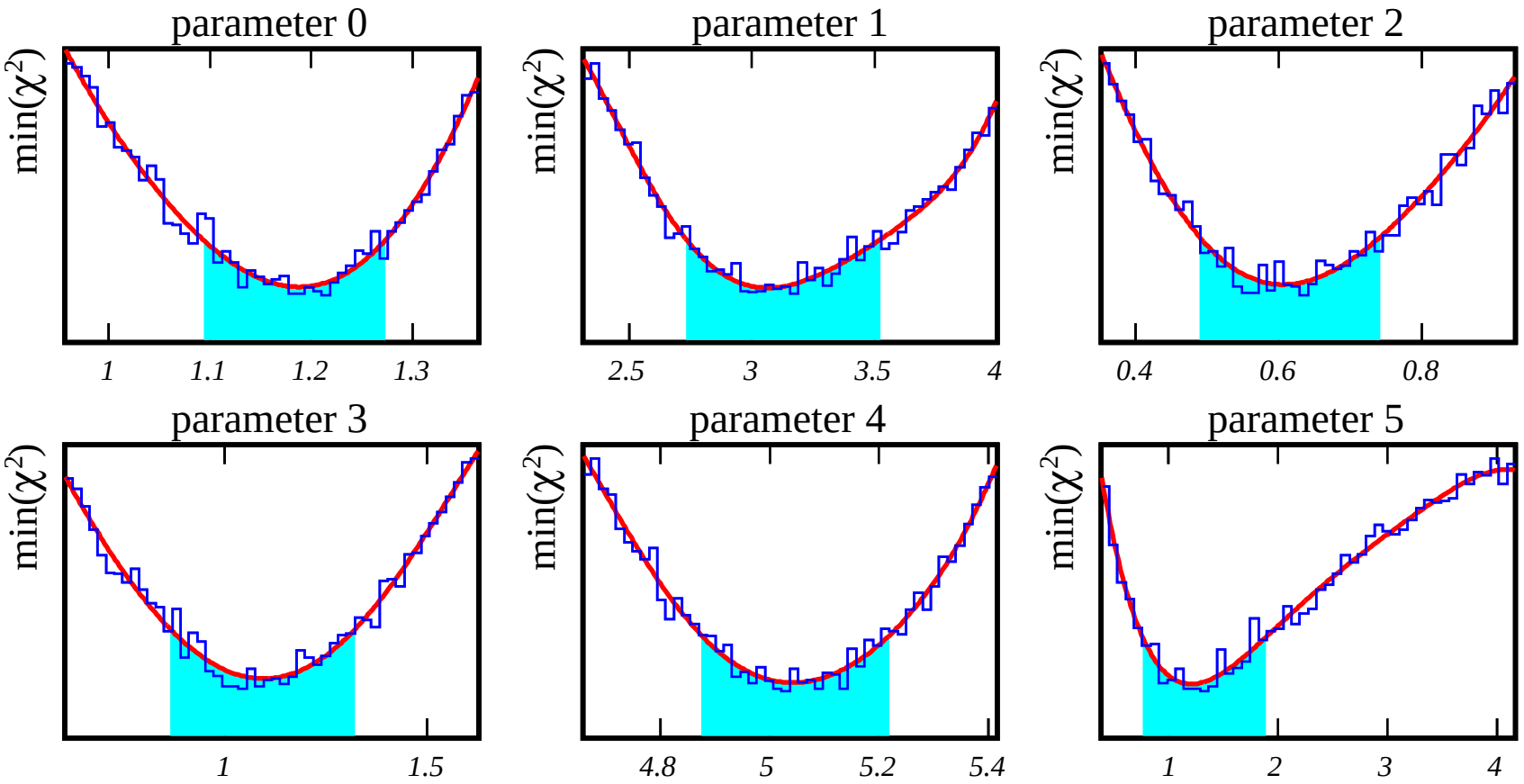
- reasonable agreement between fit results from all methods

➜ *bayesian posteriors:*



➜ reasonably very well defined PDFs

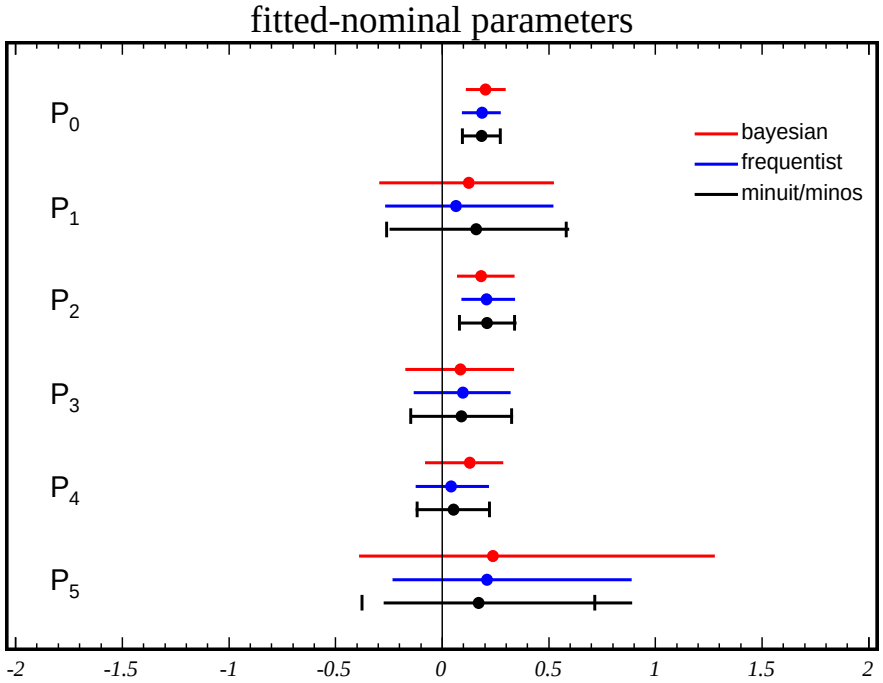
➜ *frequentist* – $\ln L$ *scans:*

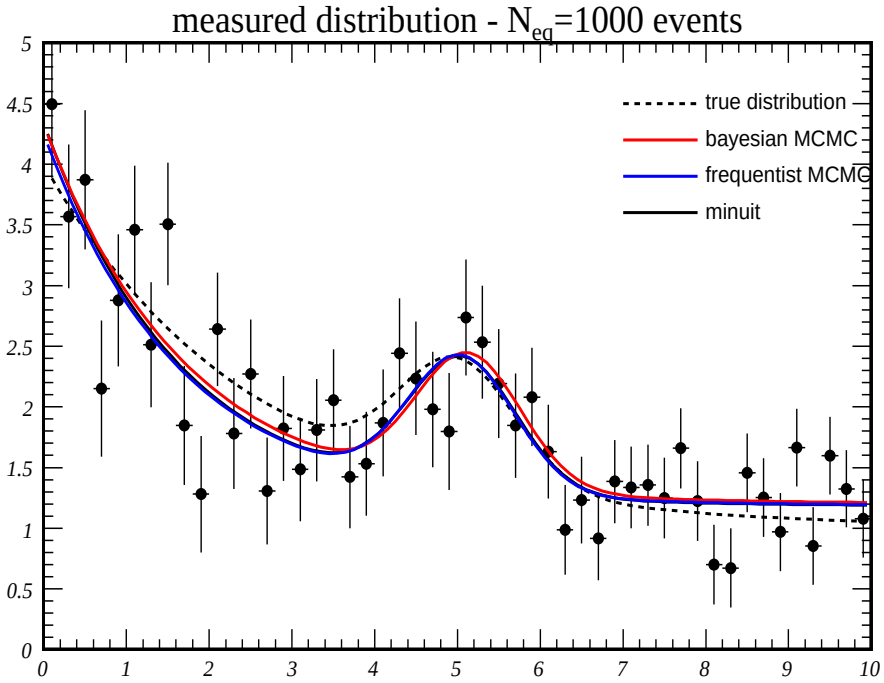


➜ reasonably well defined likelihood contours

$N_{eq} = 1000$: 10^8 *fixed-scale jumps*

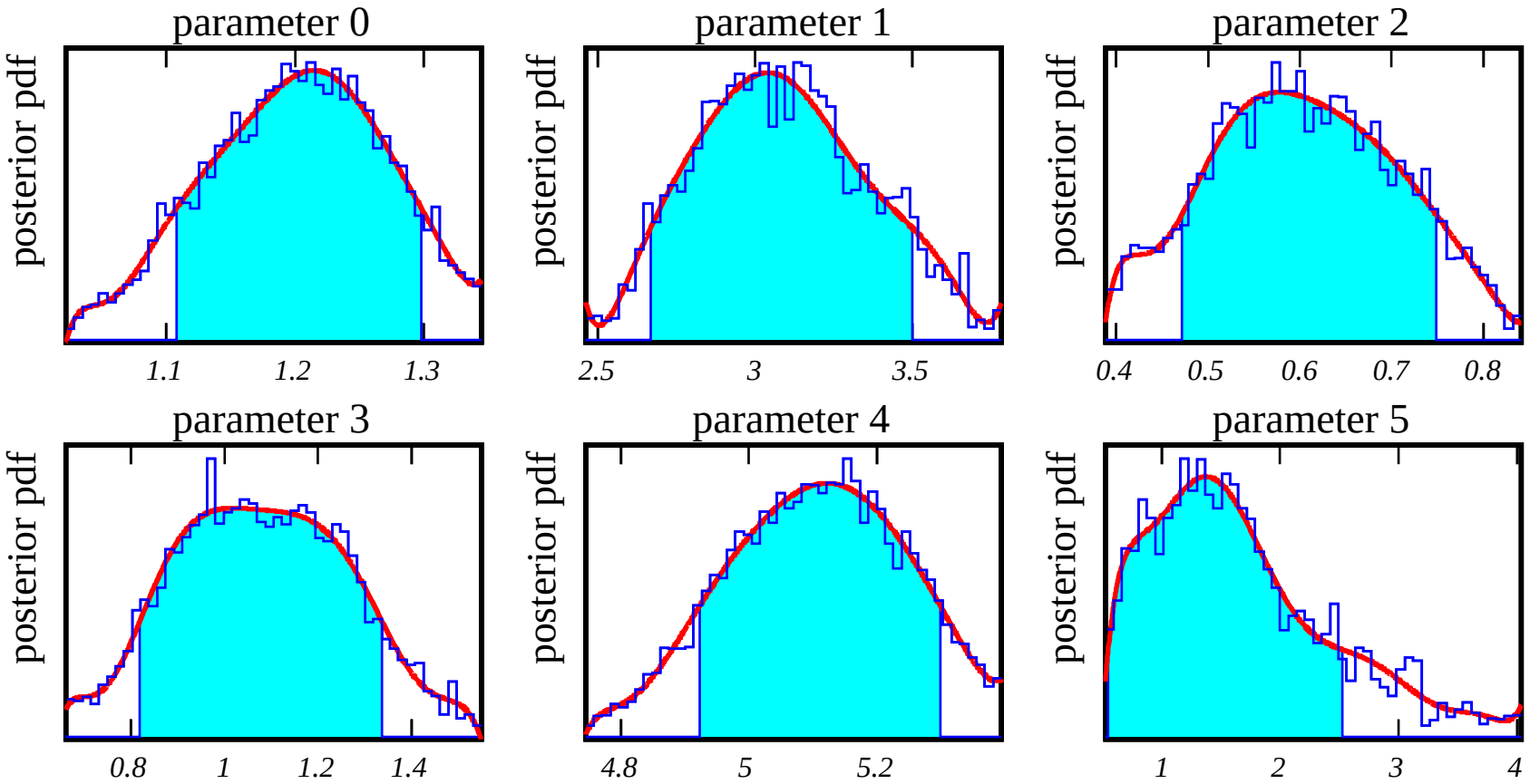
➜ *overview*





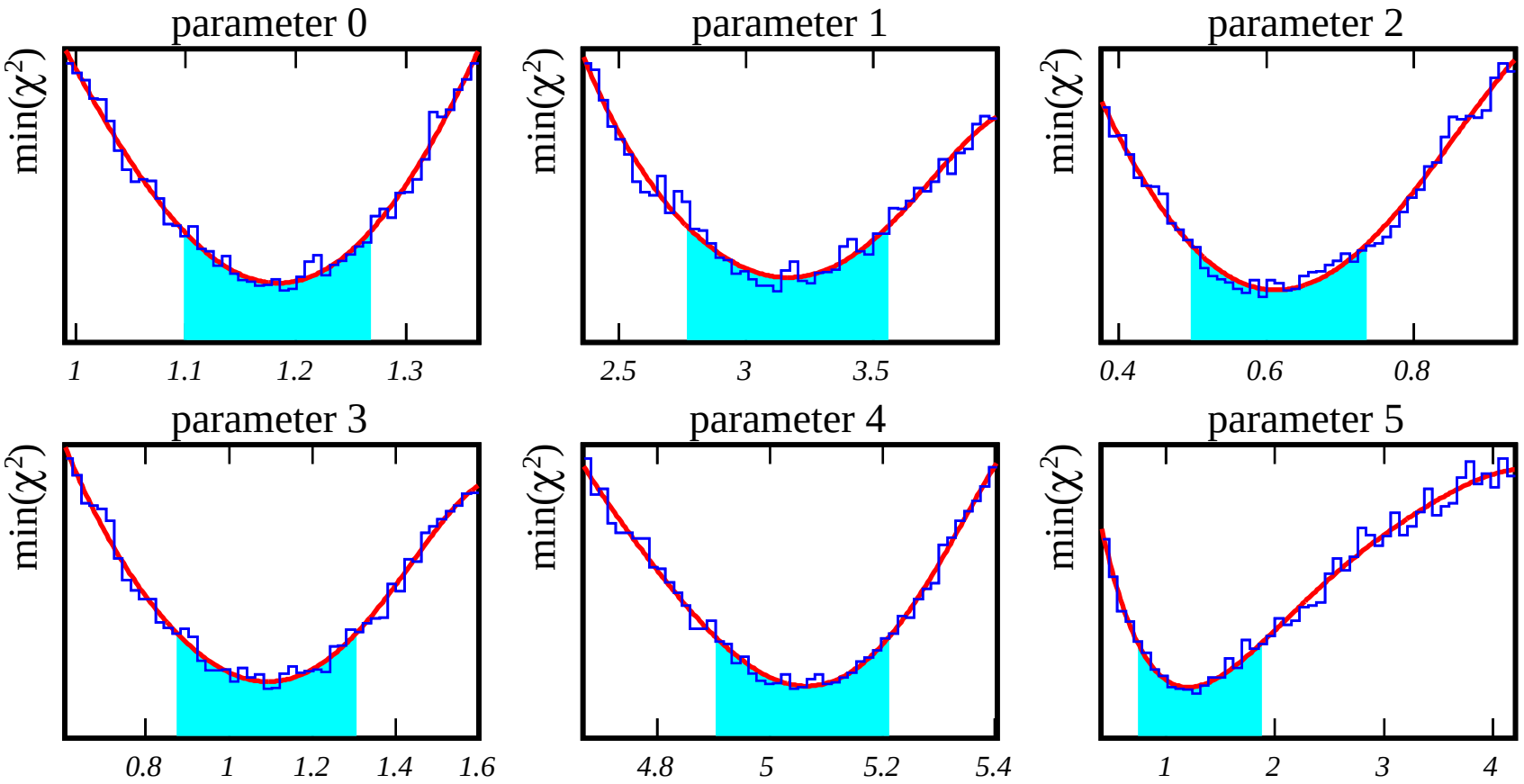
- reasonable agreement between fit results from all methods
- increased jump-statistics stabilizes results

➜ *bayesian posteriors:*



➜ very well defined PDFs

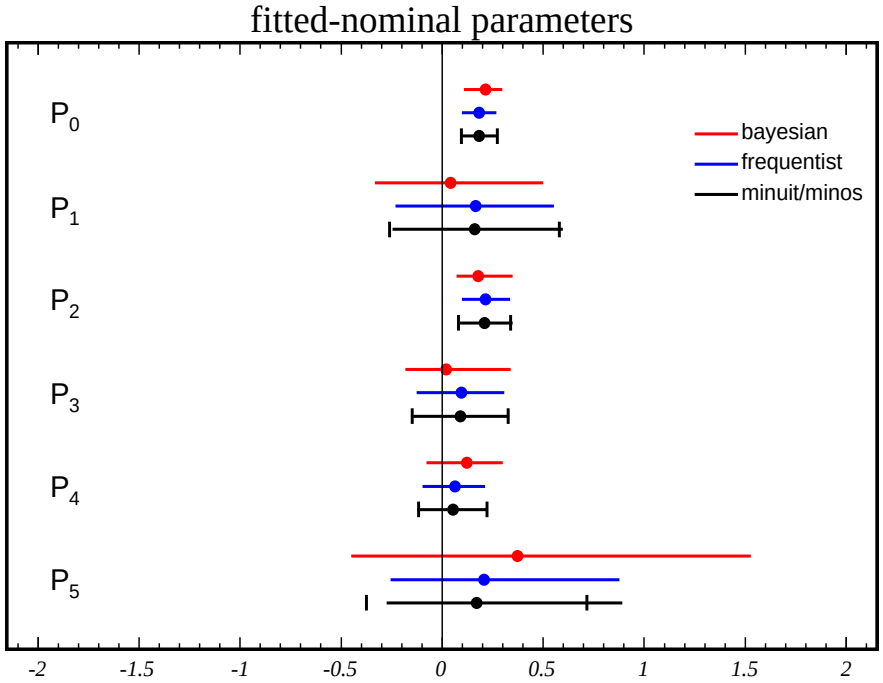
➜ *frequentist* – $\ln L$ *scans:*

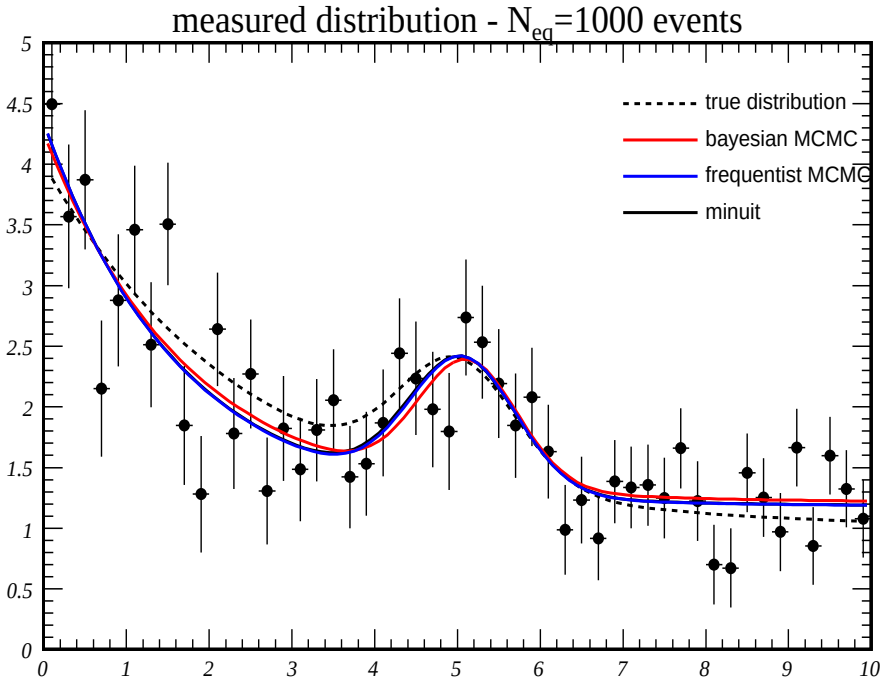


➜ well defined likelihood contours

$N_{eq} = 1000$: 10^5 variable-scale jumps

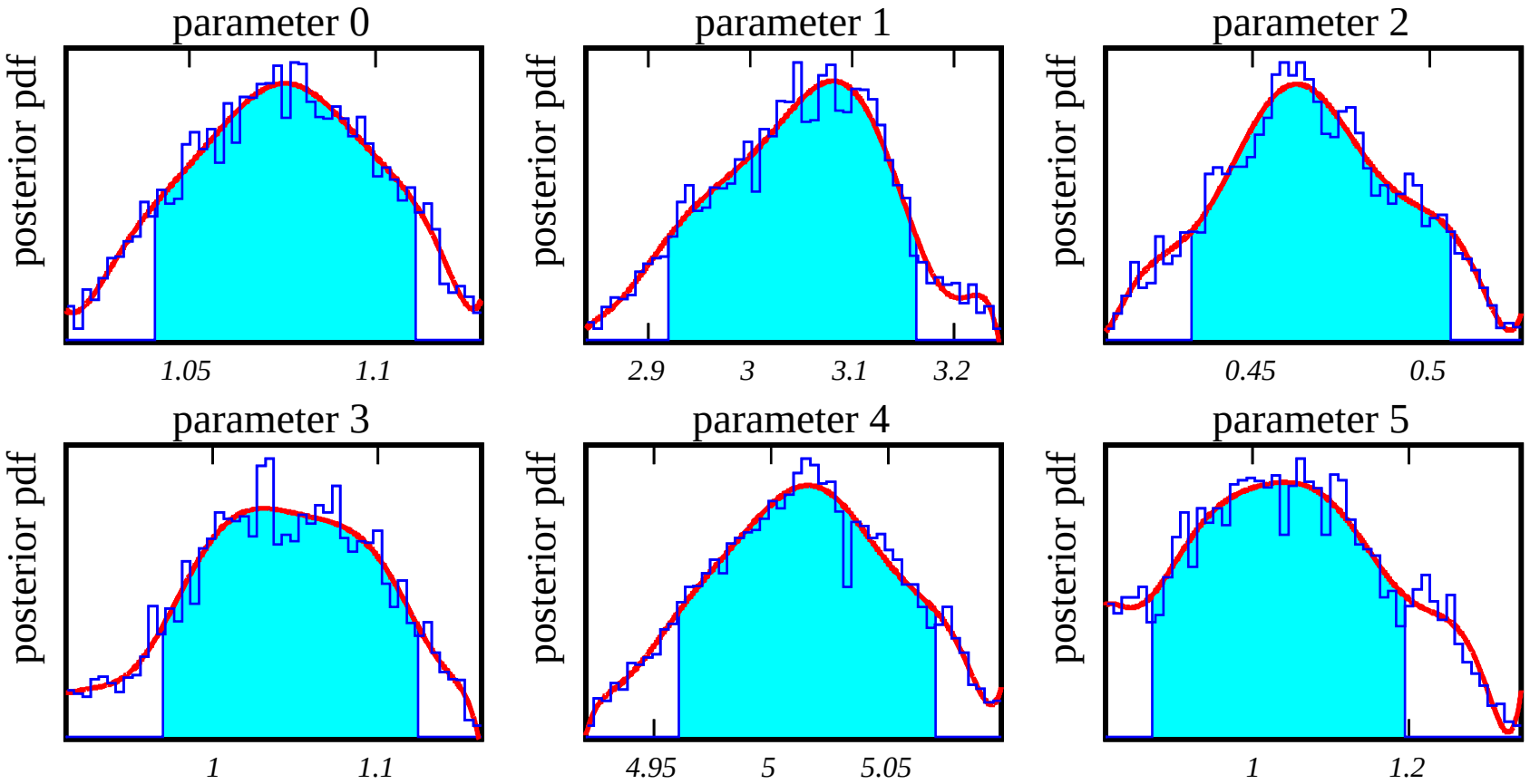
➜ *overview*





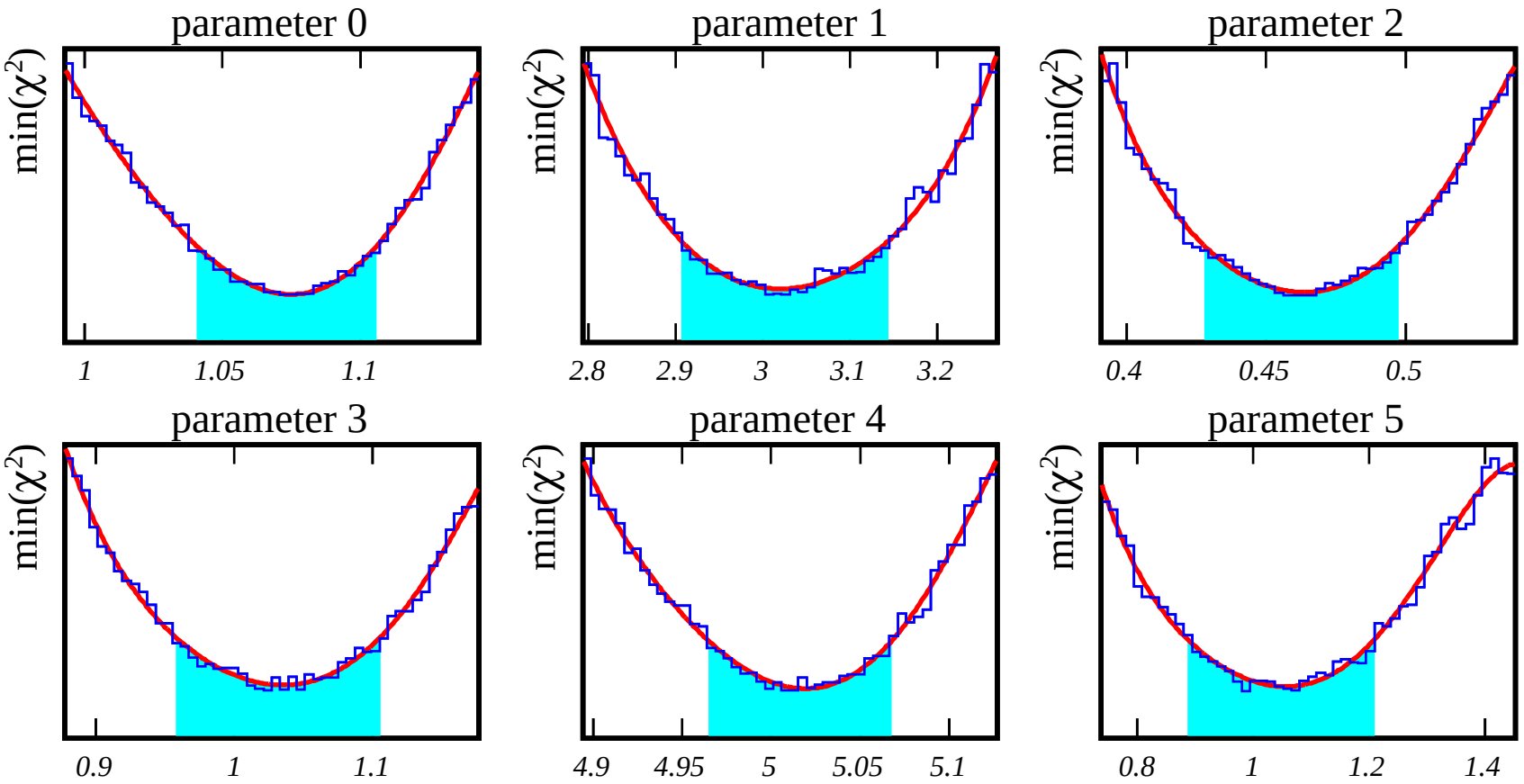
- good agreement between fit results from all methods
- variable-scale jumps do better than fixed scale

➜ *bayesian posteriors:*



➜ very well defined PDFs

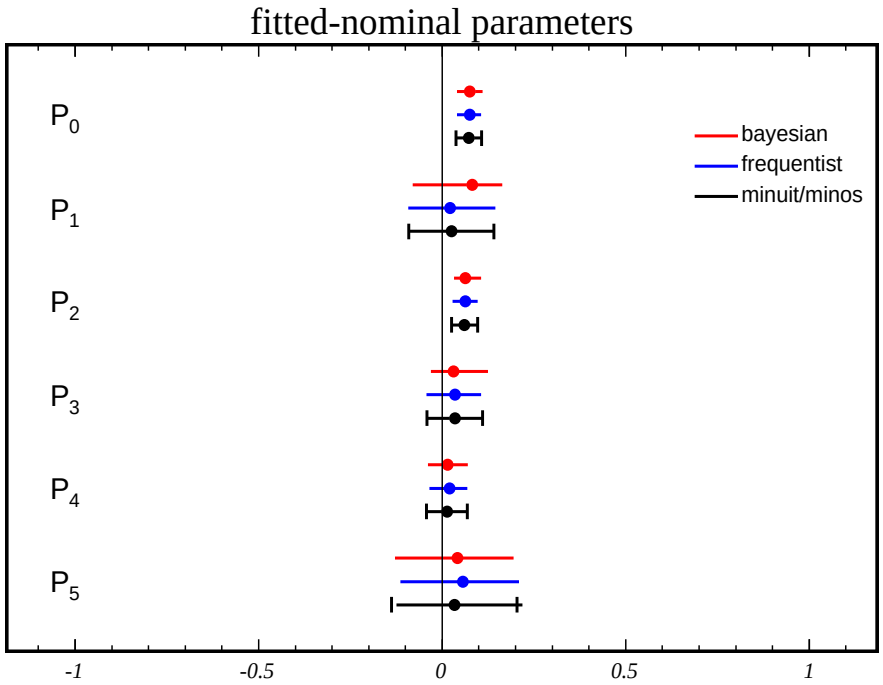
➜ *frequentist* – $\ln L$ *scans:*

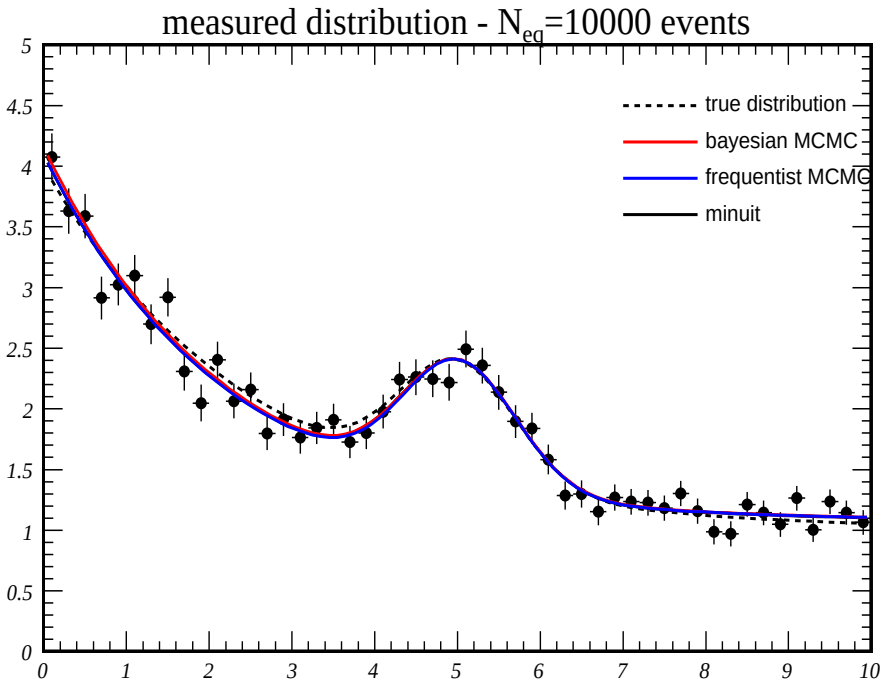


➜ well defined likelihood contours

$N_{eq} = 10000$: 10^5 variable-scale jumps

➜ *overview*





- good agreement between fit results from all methods
- variable-scale jumps do better than fixed scale

➜ *Markov Chain Monte Carlo*

- elegant method to sample arbitrary PDFs
- powerful integration tool
  - ➜ closely related to optimal importance sampling
  - ➜ complication due to burn-in phase
  - ➜ correlations between subsequent points can be a problem; could be mitigated by accepting only every n-th point
  - ➜ standard Monte Carlo errors underestimate the true uncertainties
- standard implementation: Metropolis(-Hastings) algorithm
- interesting problem: optimization of the jump-function
  - ➜ variable scale jumps look promising
- often used in connection with bayesian statistics for marginalization and integration over nuisance parameters
- sampling of likelihood function can also be used for fitting

➜ *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 $a(x)$:

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

➜ *the unfolding problem: construct an estimate for $b(y)$, given*

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

- For a parametric model $b(y;a)$ with a small number of parameters a, unfolding can be done by extracting a 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
  - ➜ truly model-independent unfolding of continuous $b(y)$ 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 $\beta_l(y)$*

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

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

- 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* $0 \leq y \leq 1$

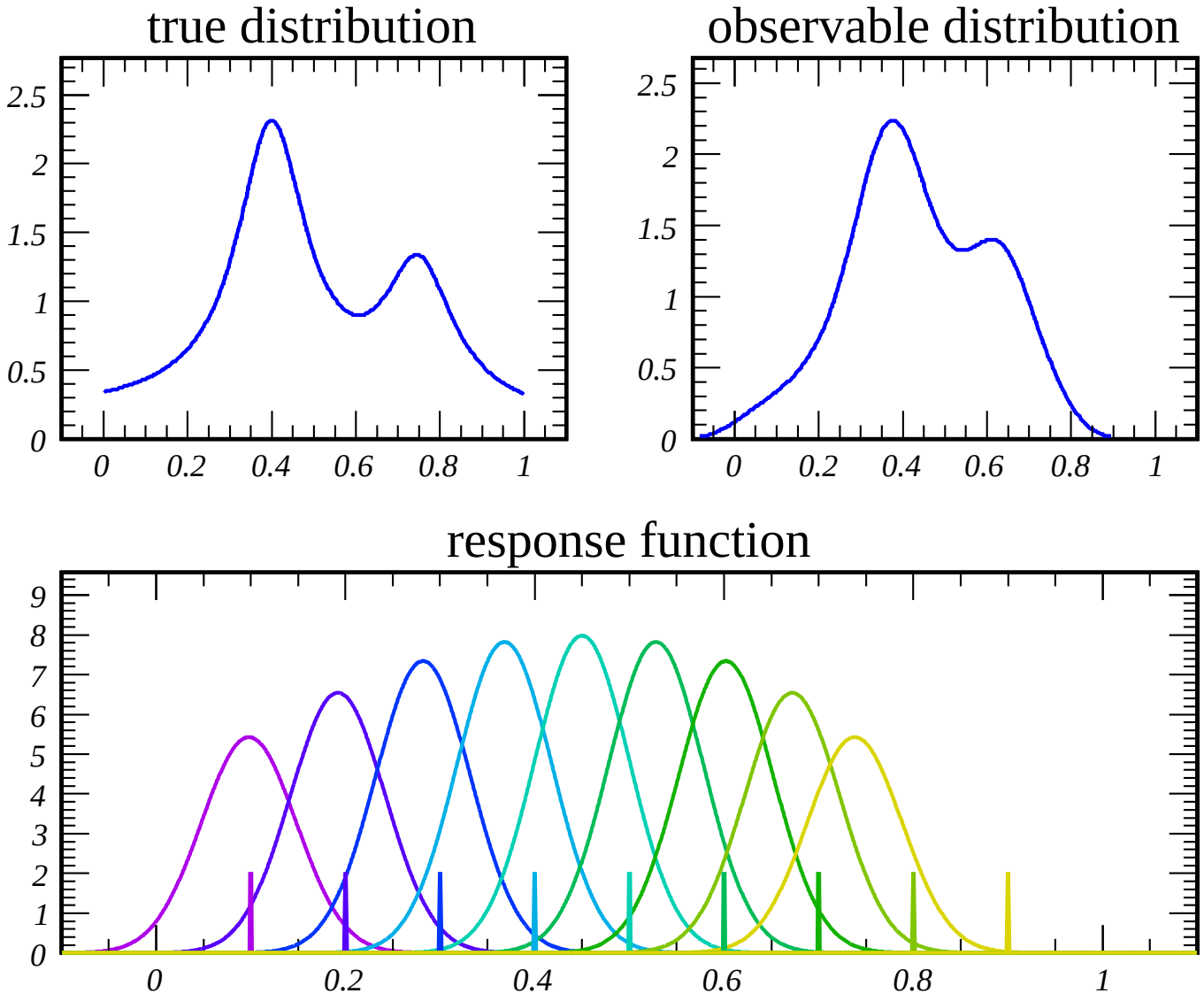
- two Breit-Wigner peaks on a smooth background

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

➜ *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 $\sigma = 0.05$)
- quadratic bias as a function of y (parameter $\beta = 0.1$)
- parabolic efficiency loss twoards phase space limits (parameter $\alpha = 0.5$)


true distribution
observable distribution
response function
0
0.5
1
1.5
2
2.5
0.2
0.4
0.6
0.8
1
3
4
5
6
7
8
9

➜ *in the following*

- take response matrix to be exact, i.e. no quantization errors
- focus on the effect of finite statistics on unfolding methods
- histogram discretisation 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 $\propto 1/\sqrt{N}$
- relation between observable and true distribution

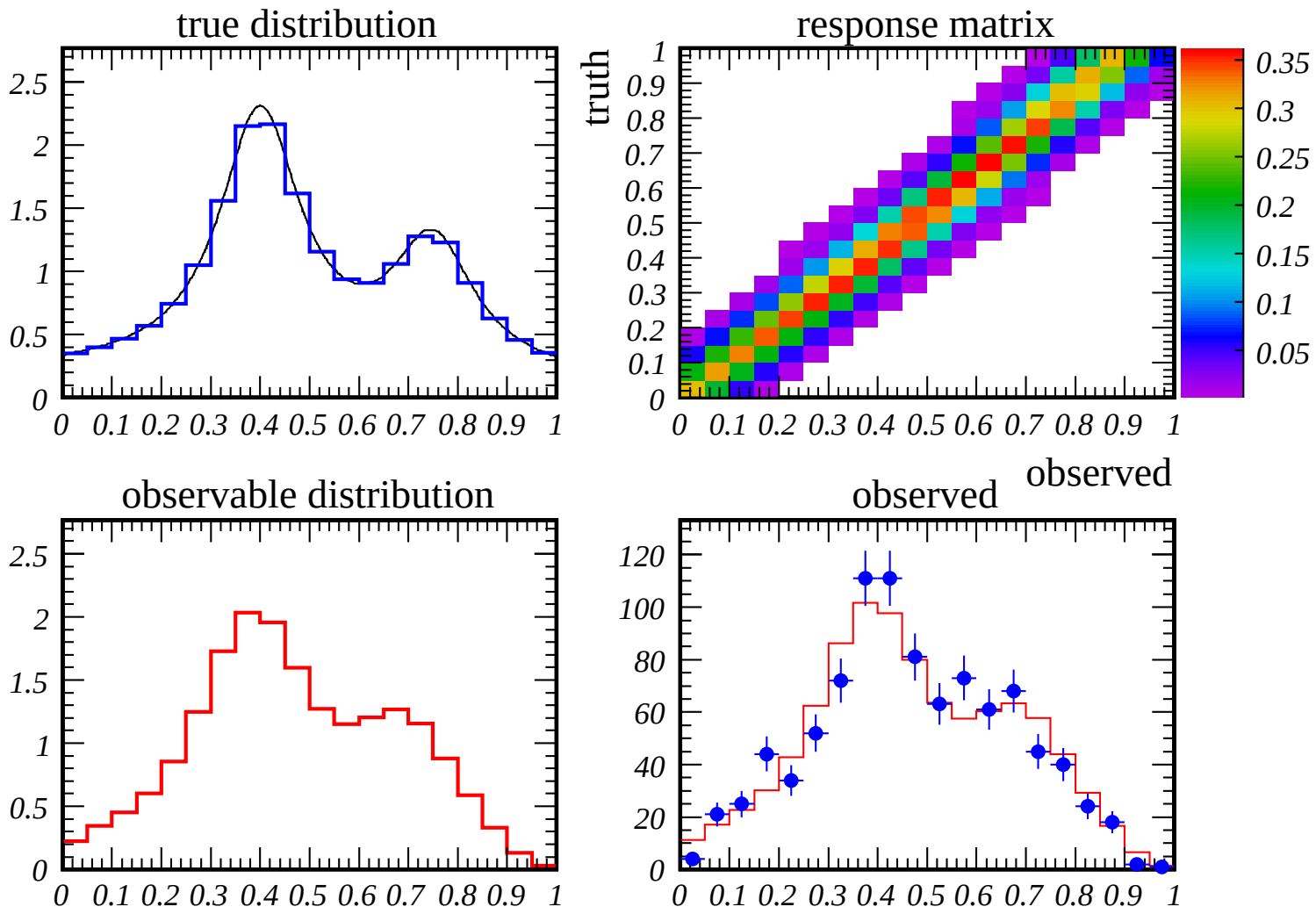
$$\langle a \rangle = G \cdot b$$

- 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. . .

Bin-by-bin correction factors

➜ *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
  - ➜ determine 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 guaranteed only for $b_k = b_k^{true}$
- in general a partial correction should be achievable

Iterated correction factors (i)

➜ *get independent of specific assumption for b_k*

Iteration starting from e.g. initial settings $b_k^{(0)} = a_k$:

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

➜ *discussion*

- correction factors work well if the true distribution is known
  - ➜ no iteration required
  - ➜ no new information from measurement
- conceptual problems
  - ➜ empty bins are corrected to zero
  - ➜ data from outside physics phase space are ignored
- iteration removes dependence on unknown distribution, but . . .
  - ➜ naive error propagation $\sigma(b_k) = c_k \cdot \sigma(a_k)$ evidently wrong
  - ➜ analytic error calculation not feasible: the iterated result is a highly non-linear function of the measurements

➜ do it properly. . .

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

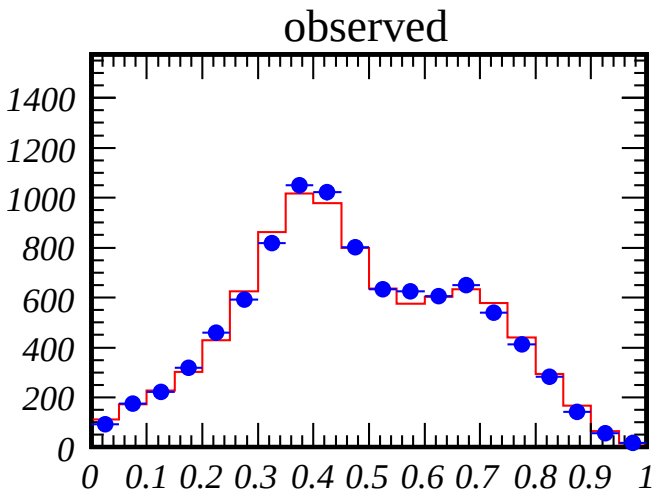
$$b_k = \frac{1}{N} \sum_{n=1}^{N} b_k^{(n)}$$

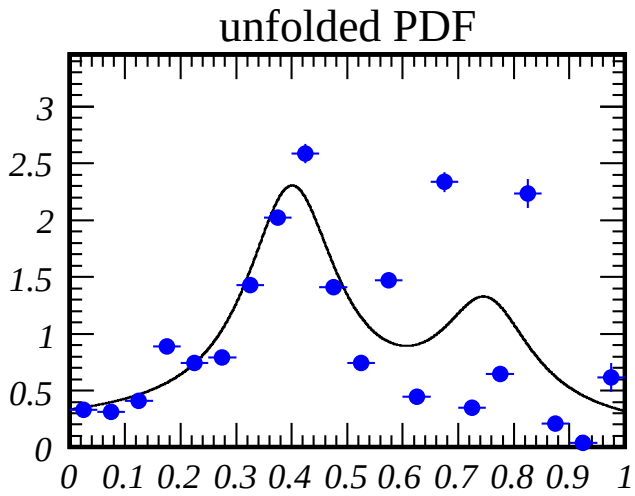
- estimate elements of the covariance matrix of the result by

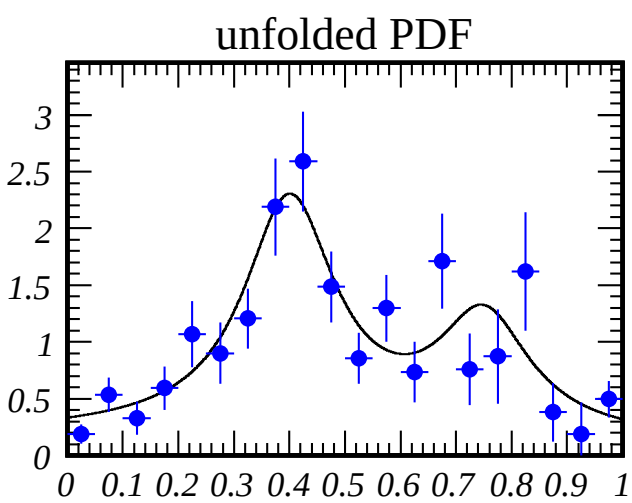
$$C_{kl}(b) = \left(\frac{1}{N} \sum_{n=1}^{N} b_k^{(n)} \; b_l^{(n)} \right) - b_k \; b_l$$

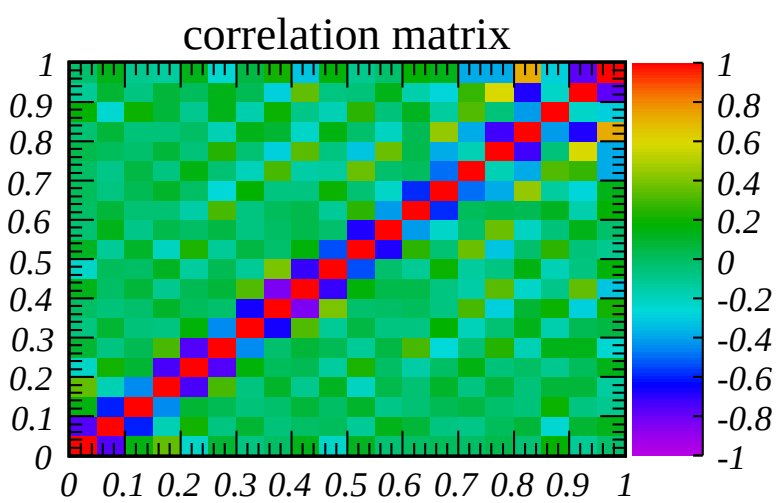
  - ➜ correlations between bins of the unfolded distribution handled properly
- numerical studies show
  - ➜ surprisingly large error in the unfolded distribution
  - ➜ strong correlations between neighboring bins
  - ➜ errors grow with the number M of iterations

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

➜ *charakteristics of bayesian unfolding*

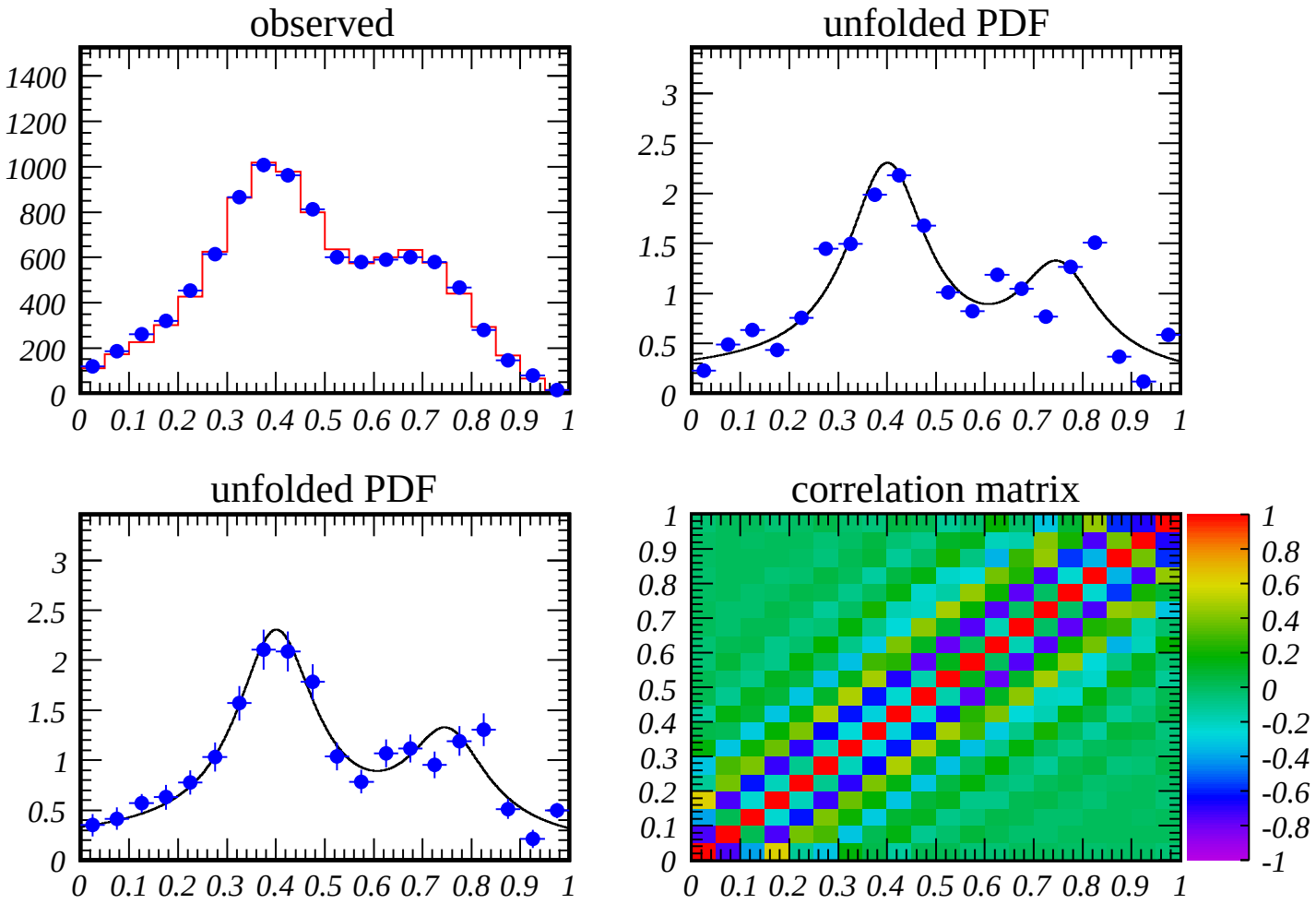
- mathematically sound aproach
- explicitly use positvity of probabilities
- can move measurements from unphysical region into allowed phase space
- no matrix inversion required
  - ➜ unfolding works also for non-square matrices G_{ik}
  - ➜ if needed the normalization B of $b_i = Bp_i$ is obtained from

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

- 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

➜ *observation*

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

➜ *conceptual 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}:

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

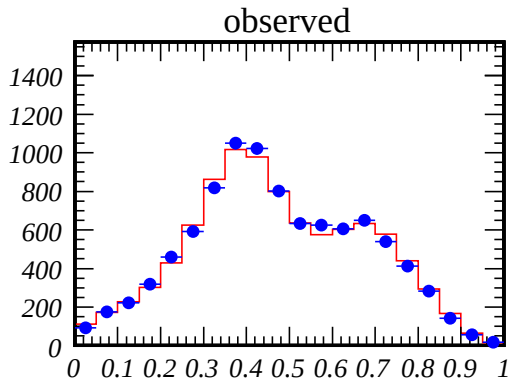
- $H = G^{-1}$ corresponds to full correction
- $H \neq G^{-1}$ implies residual distortions
  - ➜ the unfolding procedure did achieve a partial correction
  - ➜ improvement of resolution instead of full correction

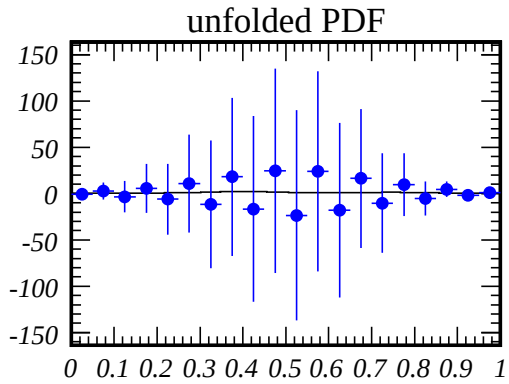
Unfolding by matrix inversion

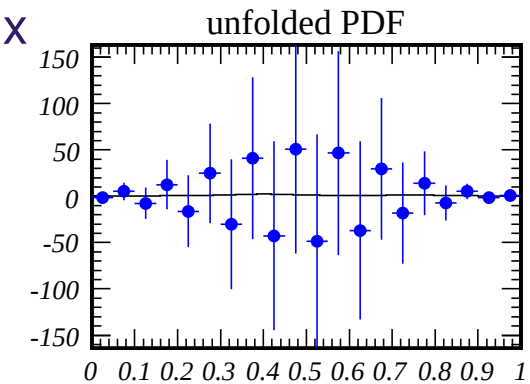
➜ *restricted to the case* $n_a = n_b$

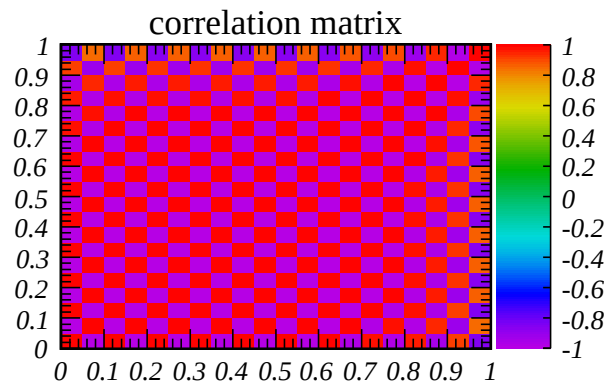
$$a = G \cdot b \quad \rightarrow \quad b = G^{-1} \cdot a \quad \text{with} \quad C(b) = G^{-1} C(a) (G^{-1})^T$$

- formally correct
- proper covariance matrix
- completely useless









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

Analysis of the response matrix

➜ *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 diag. 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

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

which diagonalize the unfolding problem

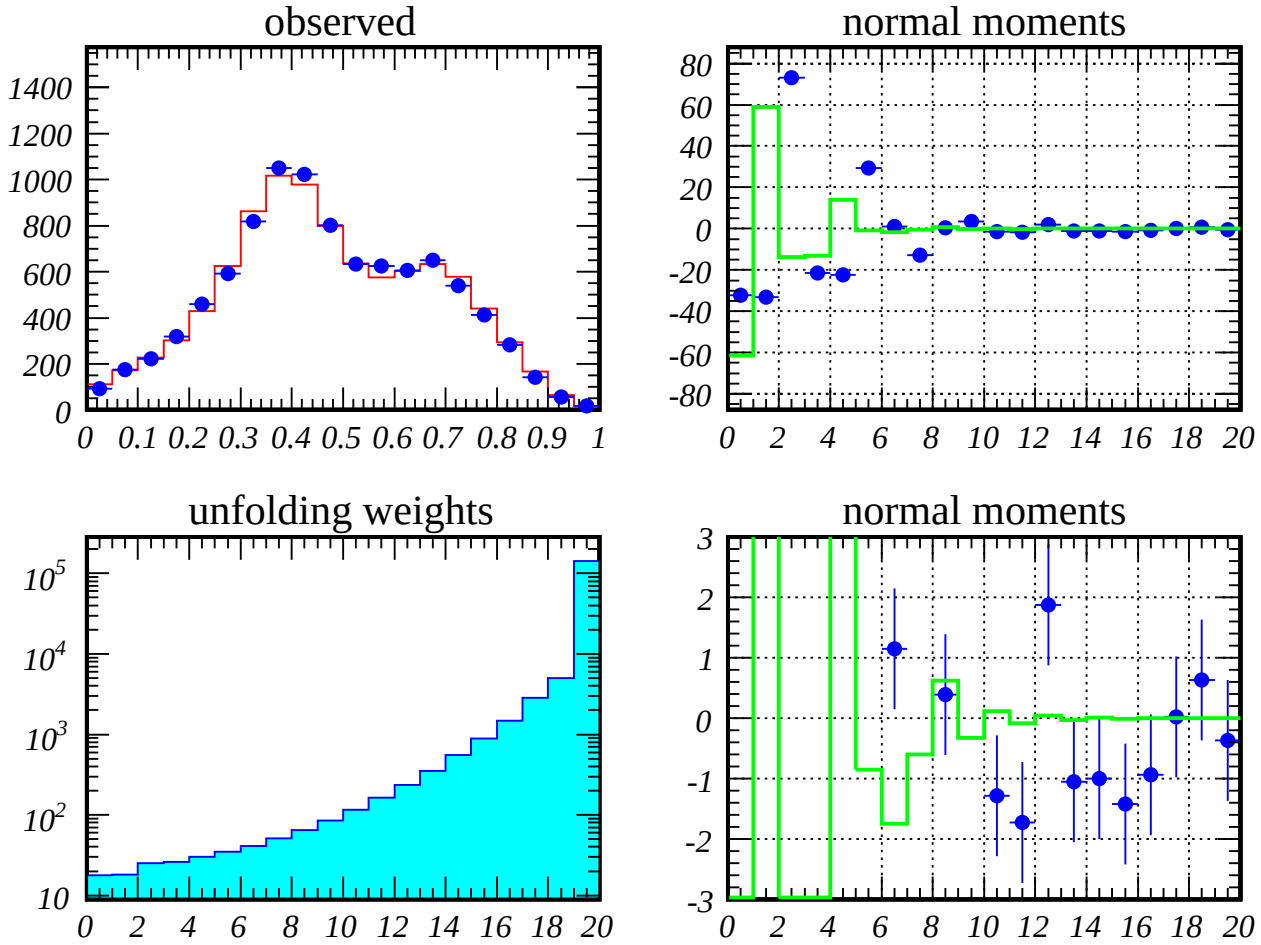
$$u = W \cdot v$$

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

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

Numerical studies

- 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.

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
- effectively there are fewer measurements with information about $b(y)$
- 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 those which are significantly different from the prior expectation
- Regularisation: for example. . .
  - ➜ take normalized moments that differ significanttly from the prior from data
  - ➜ take the others from the prior
  - ➜ construct the unfolded distribution

➜ *in general: replace missing information by assumptions*

- many possibilities to add information. . .
  - ➜ 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”

➜ *consider Fourier-analysis of a signal:*

schematically:

$$b(y) = \int d\omega \; A(\omega) \cos(\omega y)$$

include effect of finite resolution:

$$a(x) = \int dy \; \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-y)^2/2\sigma^2} \cdot b(y) = \int d\omega \; e^{-\omega^2\sigma^2} \cdot A(\omega) \cos(\omega x)$$

- high frequency components in $a(x)$ exponentially suppressed
- experimentally accessible only with very large statistics
- complete unfolding, even after discretization, usually not possible

❖ information loss is a generic, inevitable consequence of finite resolution

➜ *consider a solvable case:*

Charged particle multiplicities recorded by a particle detector:

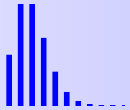
- p_n: probability to have an n-particle event
- ε: known probability to see a particle
- q_n: probability to see n particles in an event

❖ assuming that the same ε applies independently to all particles:

$$q_n = \sum_{k=n}^{\infty} p_k \binom{k}{n} \varepsilon^n (1-\varepsilon)^{k-n}$$

Given ε, the equation can be inverted, i.e. the true distribution can be inferred from the measured one. Start from the probability generating function:

$$\phi_p(z) = \sum_{n=0}^{\infty} z^n p_n \quad \Rightarrow \quad p_n = \frac{1}{n!} \left. \frac{d^n \phi_p(z)}{dz^n} \right|_{z=0}$$




➜ *probability generating function of the measurements*

$$\phi_q(z) = \sum_{n=0}^{\infty} z^n \sum_{k=n}^{\infty} p_k \binom{k}{n} \varepsilon^n (1-\varepsilon)^{k-n}$$

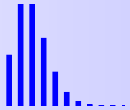
$$= \sum_{k=0}^{\infty} p_k \sum_{n=0}^{k} \binom{k}{n} (z\varepsilon)^n (1-\varepsilon)^{k-n}$$

$$= \sum_{k=0}^{\infty} p_k (1 + z\varepsilon - \varepsilon)^n = \phi_p(1 + z\varepsilon - \varepsilon)$$

The generation function of the p_n can be inferred from the measured q_n

$$\phi_p(z) = \phi_q\left(\frac{z+\varepsilon-1}{\varepsilon}\right)$$

Determine the individual p_n by differentiation




➜ *result:*

$$\left.\frac{d^n\phi_p(z)}{dz^n}\right|_{z=0} = \frac{1}{\varepsilon^n}\phi_q^{(n)}\left(\frac{\varepsilon-1}{\varepsilon}\right) \quad \text{and} \quad \phi_q^{(n)}(z)=\sum_{k=n}^{\infty}\frac{k!}{(k-n)!}q_k z^{k-n}$$

and thus

$$p_n = \frac{1}{n!}\left.\frac{d^n\phi_p(z)}{dz^n}\right|_{z=0} = \frac{1}{n!\,\varepsilon^n}\sum_{k=n}^{\infty}\frac{k!}{(k-n)!}q_k\left(\frac{\varepsilon-1}{\varepsilon}\right)^{k-n}$$

$$= \sum_{k=n}^{\infty} q_k \binom{k}{n} \frac{1}{\varepsilon^n}\left(\frac{\varepsilon-1}{\varepsilon}\right)^{k-n}$$

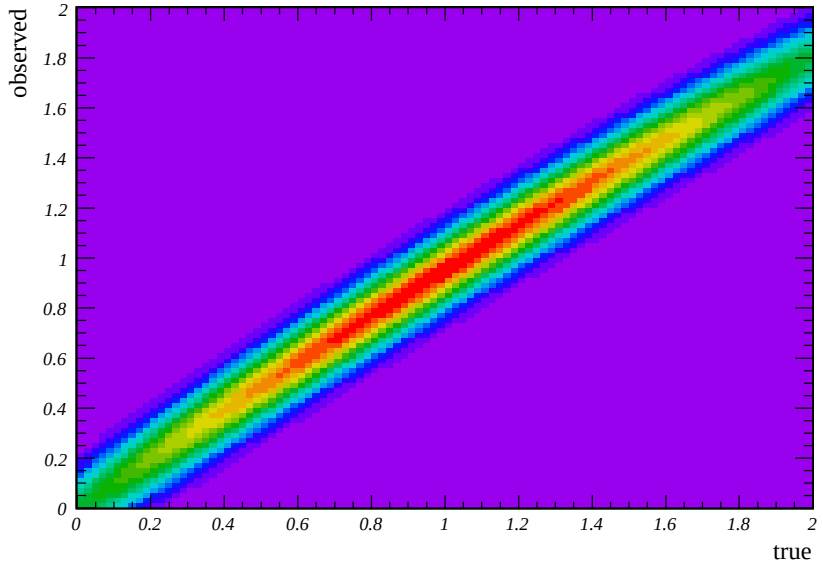
- same structure for $p \to q$ and $q \to p$
- transition $p \to q$: stable
- transition $q \to p$: powers of $(\varepsilon-1)/\varepsilon < 0$, unstable for $\varepsilon < 0.5$
- generic feature of unfolding problems!

go back to the continuous case ➜

Illustration how bad it really is . . .

➜ *consider a simple counting experiment:*

- physical phase space $x \in [0, 2]$
- measured quatities $x' \in [0, 2]$
- an imperfect detector
  - ➜ efficiency $A = 1 - 0.5\,(x - 1)^2$
  - ➜ bias $\langle x' \rangle = x - 0.05\,x^2$
  - ➜ gaussian smearing with $\sigma = 0.1$



(model by Volker Blobel)

➜ *measurement:*

- observation: $n = 8860$ entries with $x' \in [0, 2]$
- wanted: true number N of entries in $x \in [0, 2]$

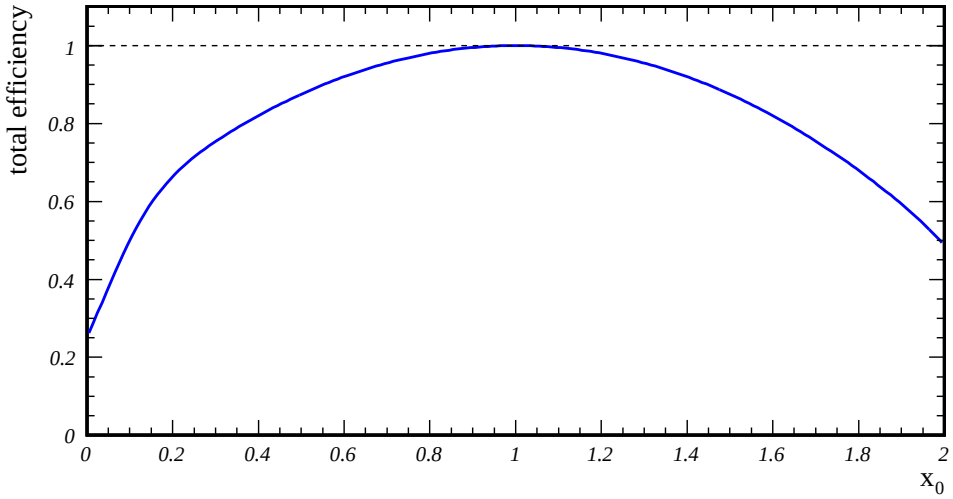
➜ study relation between n and N

- assume the true PDF to be a delta function

$$f(x) = N\delta(x - x_0)$$

➜ expected number of observed events

$$\frac{\langle n \rangle}{N} = (1 - 0.5\ (x_0 - 1)^2) \int_0^2 dx' \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x' - x_0 + 0.05\ x_0^2)^2}{2\sigma^2}\right)$$



➜ estimated value of N depends on the assumption for the true PDF

- the acceptance curve is distorted by smearing
- even a simple counting experiment is model dependent
- efficiency correction depends on the unknown true distribution
- in the example the correction factor is $1 < c < 4$

➜ *remember:*

❖ without smearing, bias and efficiency correction would be easy!

❖ with smearing a full correction becomes impossible!

➜ *in the following pursue a more modest approach:*

Use a suitable, sufficiently flexible, parametrisation of the true distribution, and try to understand what this means . . .

The parametric approach unfolding

- a: histogram of observed distribution with n_a bins
- C: covariance matrix of the measurements
- b: histogram of true distribution with n_b bins (in principle $n_b \to \infty$)
- R: response matrix describing the detector
- $g_i, i = 1, \ldots, n_k$: basis vectors, "Kernels", for expansion of b
  - ➜ a "suitable" set of independent functions
- w: vector of n_k expansion coefficients, "weights"
  - ➜ modelling of the true PDF:

$$b = \sum_{i=1}^{n_k} g_i \; w_i = K \cdot w \quad \text{with matrix} \quad K = (g_1, g_2, \ldots, g_{n_k})$$

  - ➜ relation between true and measured distribution:

$$a = R \cdot b = R \cdot K \cdot w = Q \cdot w \quad \text{with} \quad Q = R \cdot K$$

➜ *estimate the weights w from the measurements a:*

$$\chi^2 = (a - Q \cdot w)^T C^{-1} (a - Q \cdot w) \stackrel{!}{=} \min$$

which has the solution

$$w = (Q^T C^{-1} Q)^{-1} (Q^T C^{-1})\ a$$

with covariance matrix

$$C_w = (Q^T C^{-1} Q)^{-1}$$

and

$$\begin{aligned} \hat{b} &= K \cdot w \\ &= K \cdot (Q^T C^{-1} Q)^{-1} (Q^T C^{-1})\ a \\ &= K \cdot (Q^T C^{-1} Q)^{-1} (Q^T C^{-1})\ R \cdot b = R' \cdot b \end{aligned}$$

with covariance matrix

$$C_{\hat{b}} = K \cdot C_w \cdot K^T$$

and R' the posterior response matrix of the unfolding result

- for $n_a = n_b = n_k$ all matrices are square matrices and $R' = \mathbf{1}$
  - ➜ perfect unfolding for the chosen bin-widths
  - ➜ independent of the choice of kernel functions
- R' is proportional to C and C^{-1}
  - ➜ independent of the statistics in the measurements
  - ➜ depends on
    - the choice of Kernel functions, K
    - response matrix R
    - structure of the covariance matrix C (PDF of measurements)

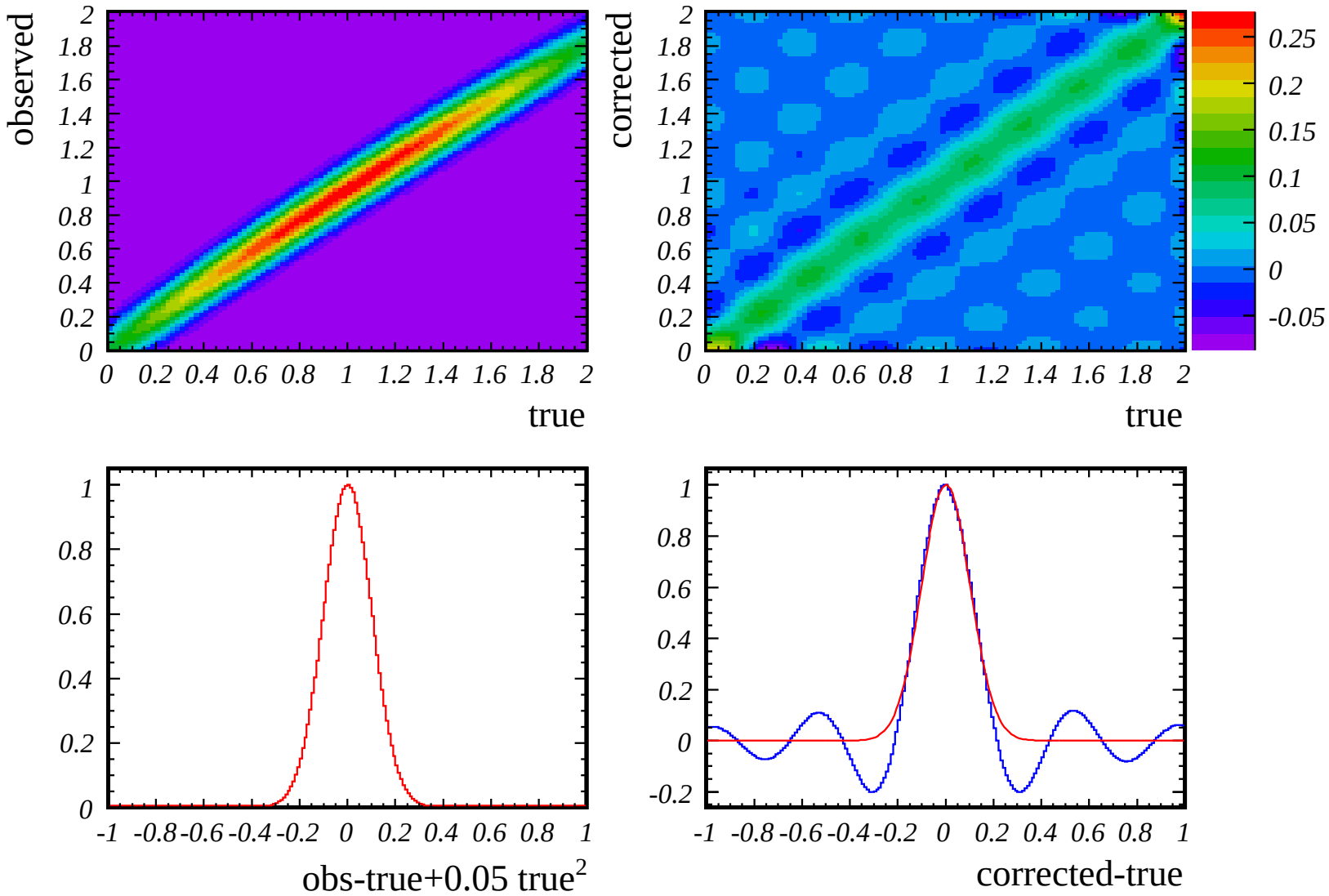
Compare R and R' for numerical examples!

➜ *study response matrices for. . .*

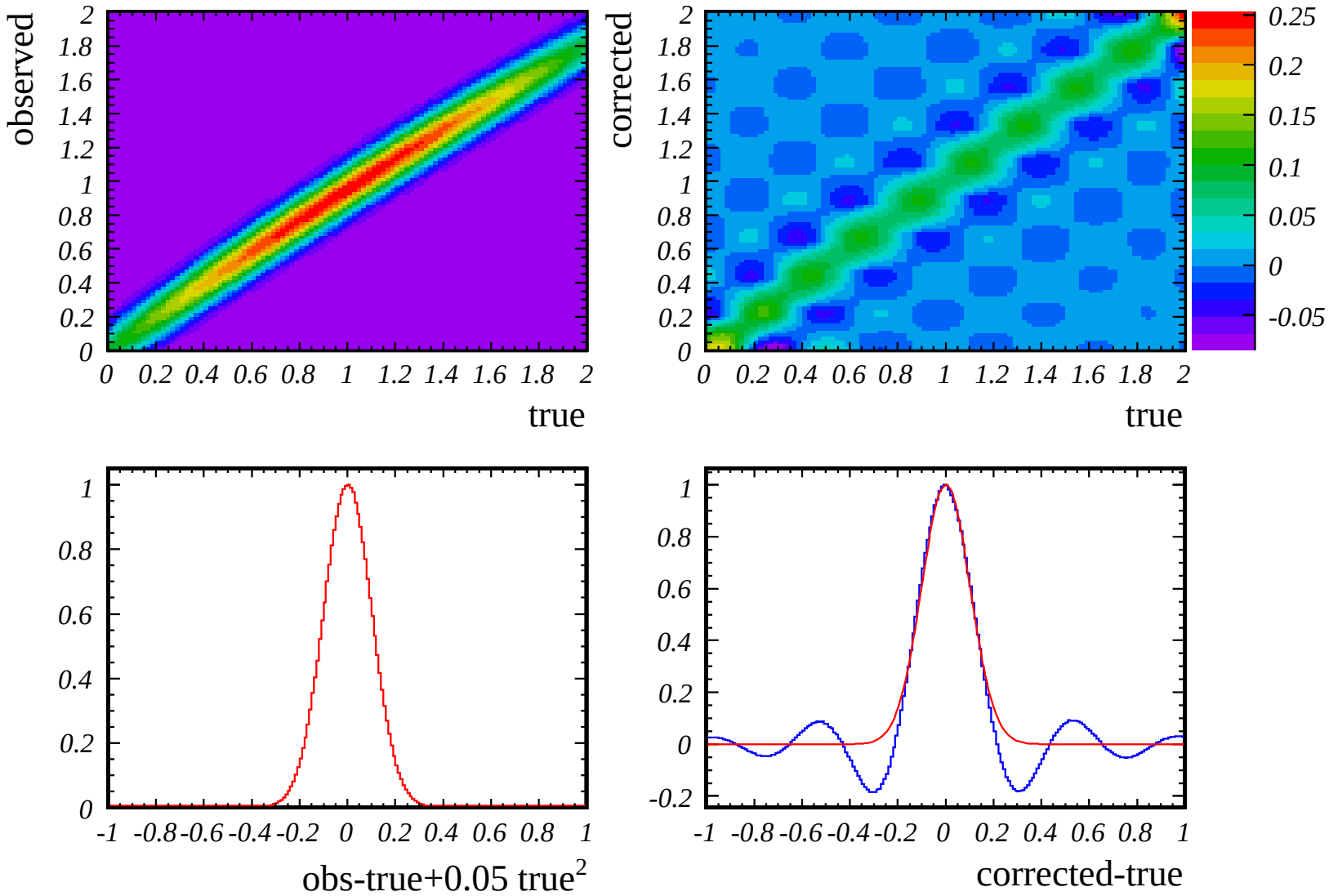
- binning $n_a = n_b = 100$, number of kernels $n_k = 10$ and $n_k = 15$
- different types of kernel functions, with $u = (x - x_i)/\lambda$, and
  - ➜ gaussian kernels: $\propto \exp(-u^2/2)$
  - ➜ Epanechnikov kernels: $\propto (1 - u^2)$ for $|u| < 1$
  - ➜ Cauchy kernels: $\propto 1/(1 + u^2)$
- equidistant kernel centers x_i with $x_1 = 0$ and $x_{n_k} = 2$
- bandwidth parameters $\lambda \in \{0.05, 0.10, 0.20\}$
- compare:
  - ➜ full response matrices R and R'
  - ➜ resolutions; average response to a delta-function as a function of $\Delta x = x - x_{\mathrm{true}}$ after unfolding (normalized to 1 at $\Delta x = 0$)

➜ images

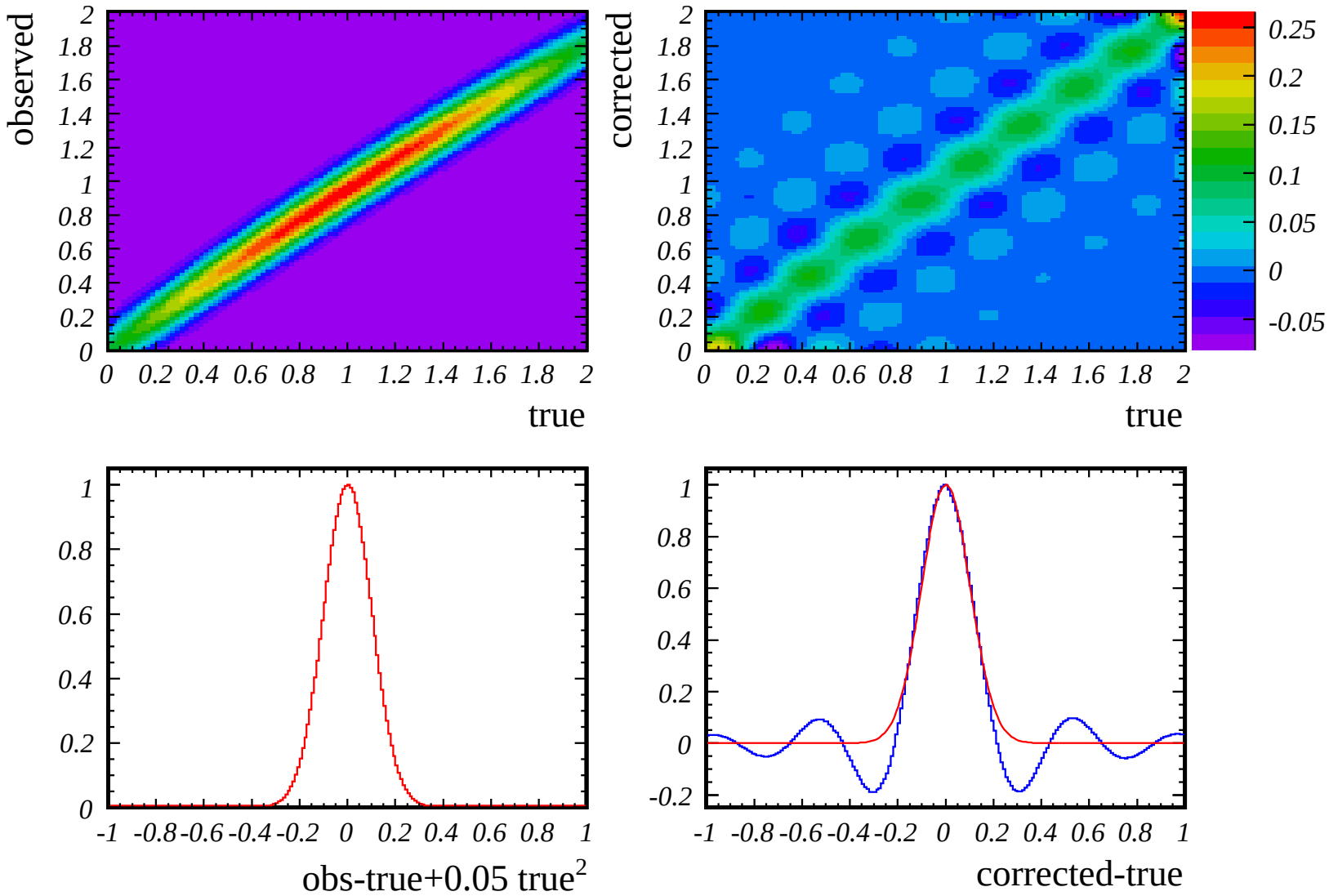
Gaussian Kernels $n_k = 10$ and $\lambda = 0.20$



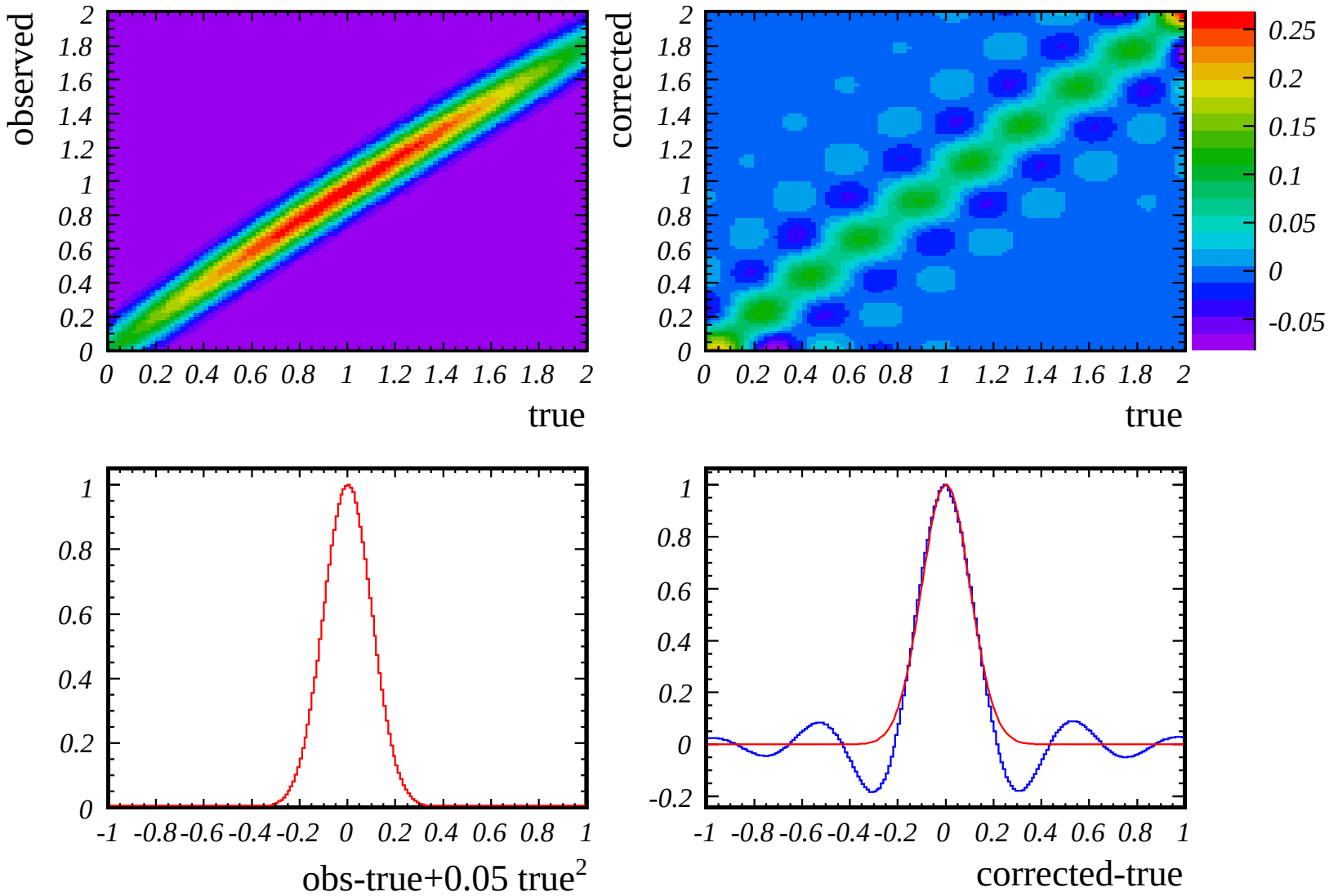
Epanechnikov Kernels $n_k = 10$ and $\lambda = 0.20$



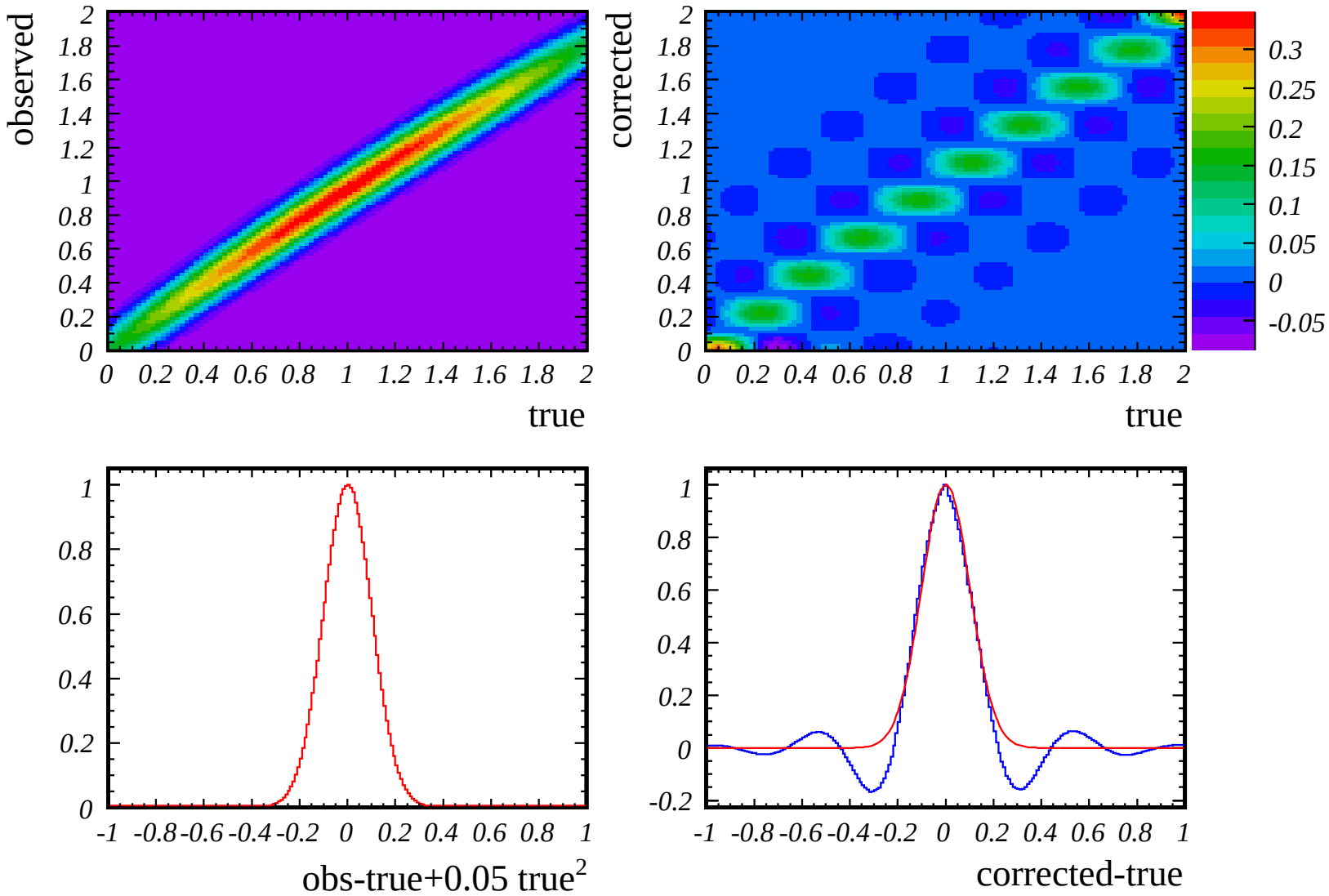
Cauchy Kernels $n_k = 10$ and $\lambda = 0.20$



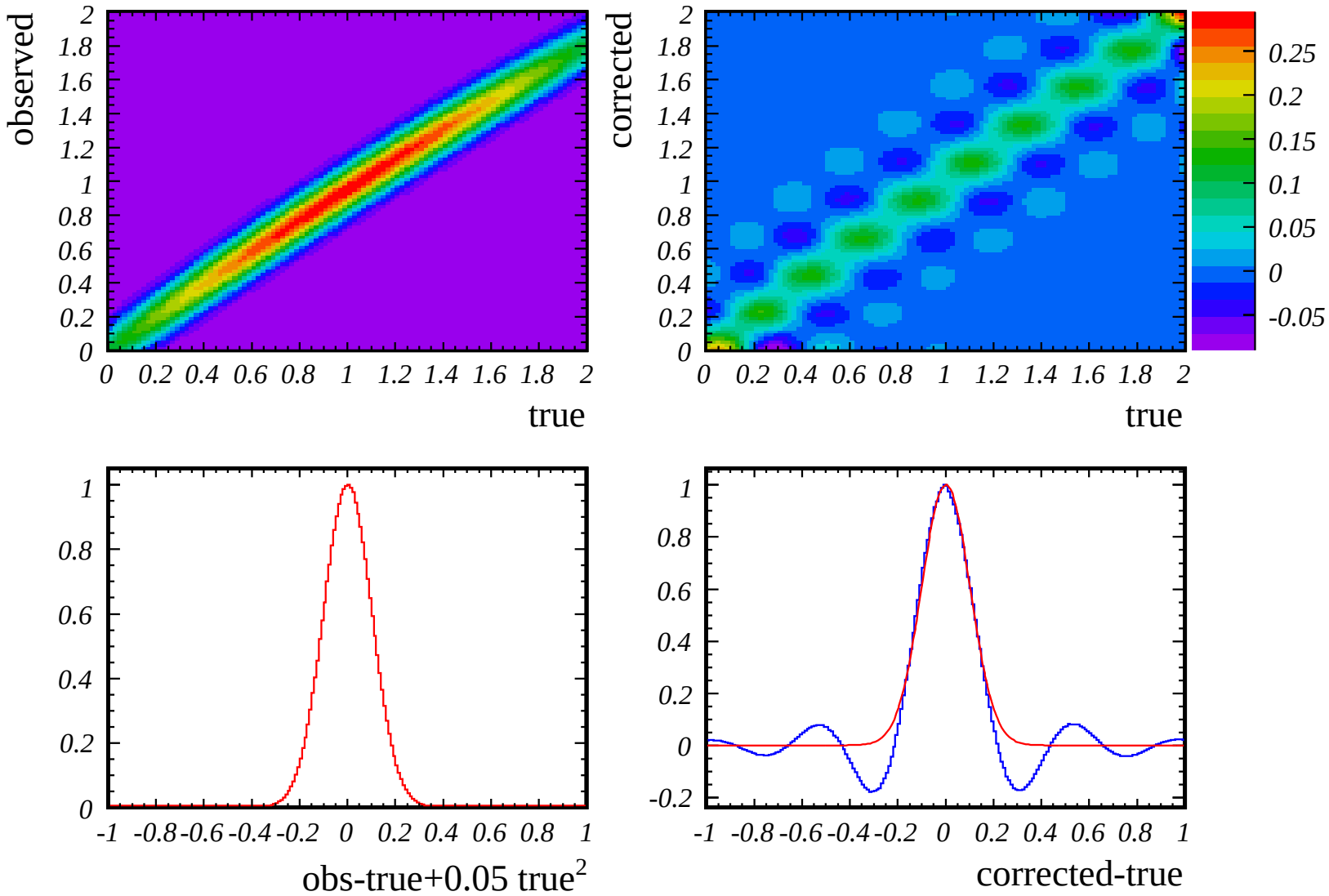
Gaussian Kernels $n_k = 10$ and $\lambda = 0.10$



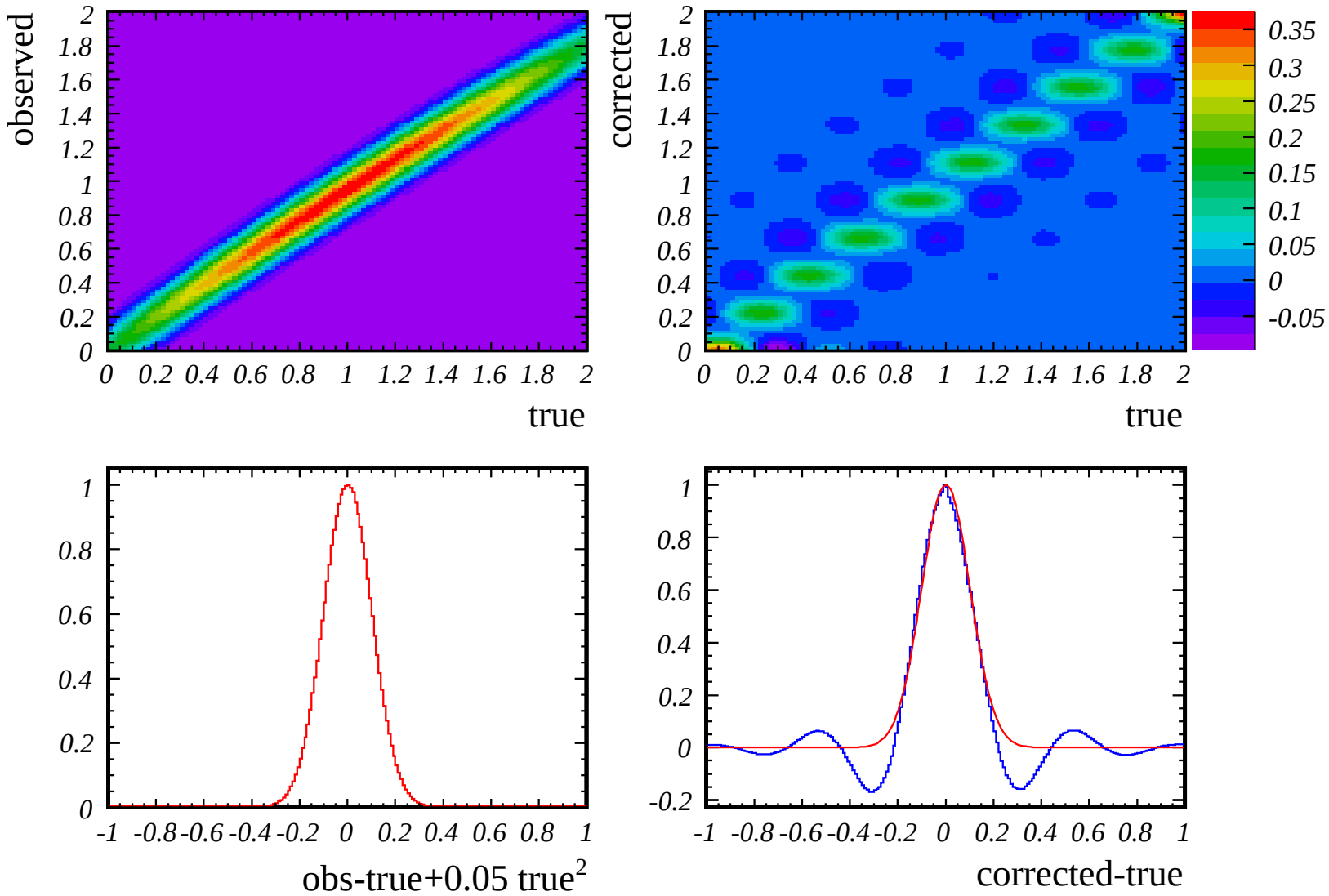
Epanechnikov Kernels $n_k = 10$ and $\lambda = 0.10$



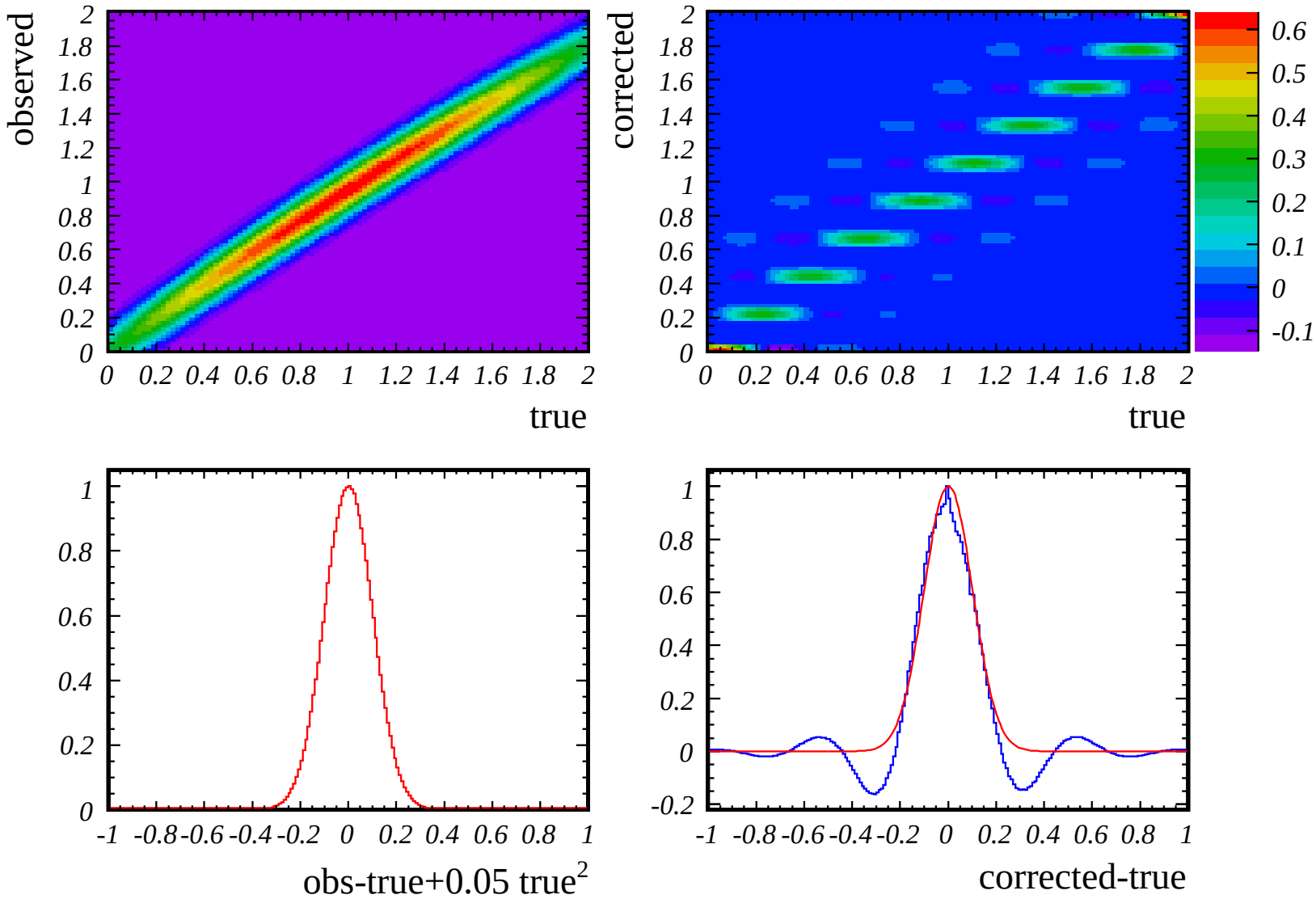
Cauchy Kernels $n_k = 10$ and $\lambda = 0.10$



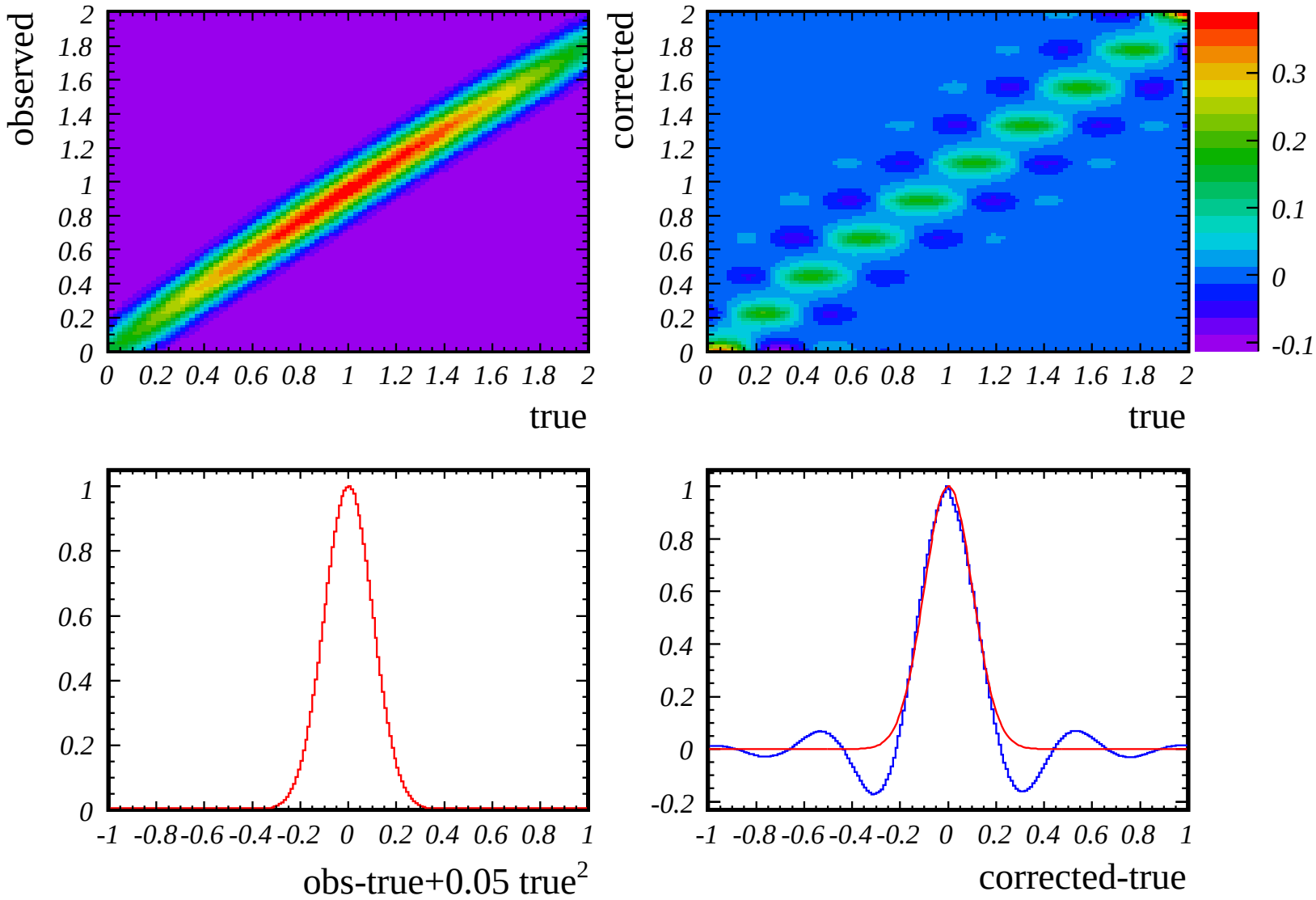
Gaussian Kernels $n_k = 10$ and $\lambda = 0.05$



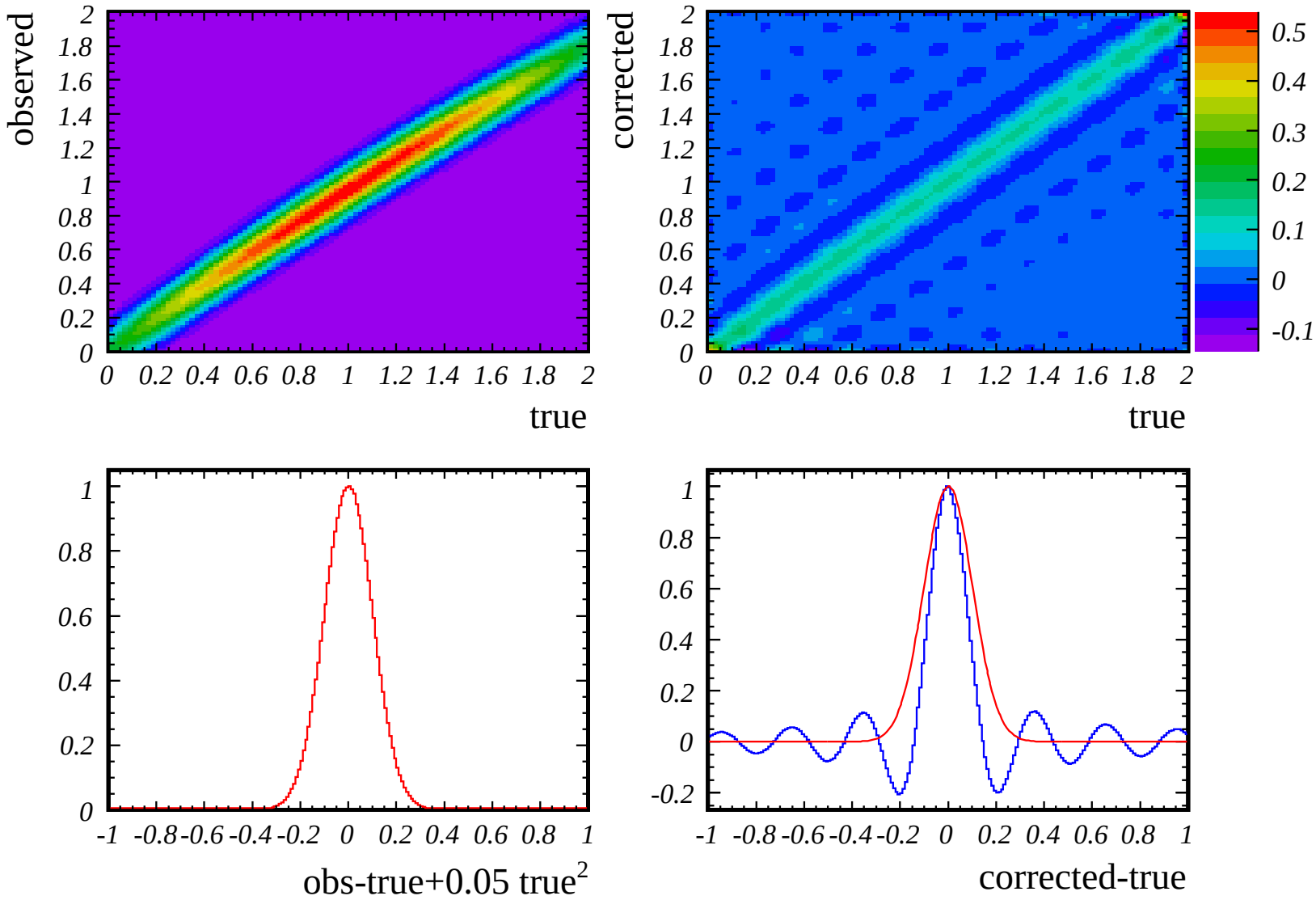
Epanechnikov Kernels $n_k = 10$ and $\lambda = 0.05$



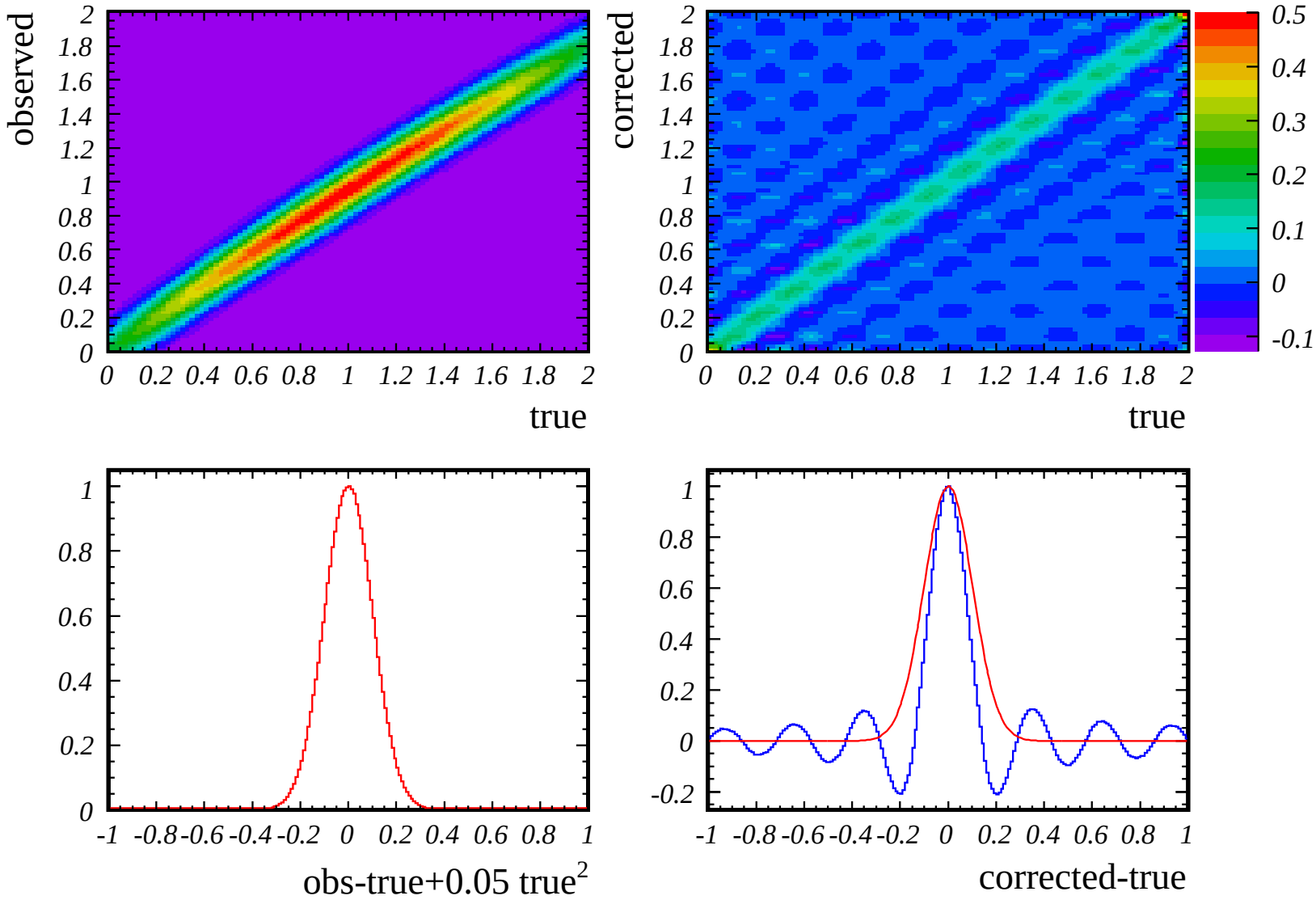
Cauchy Kernels $n_k = 10$ and $\lambda = 0.05$



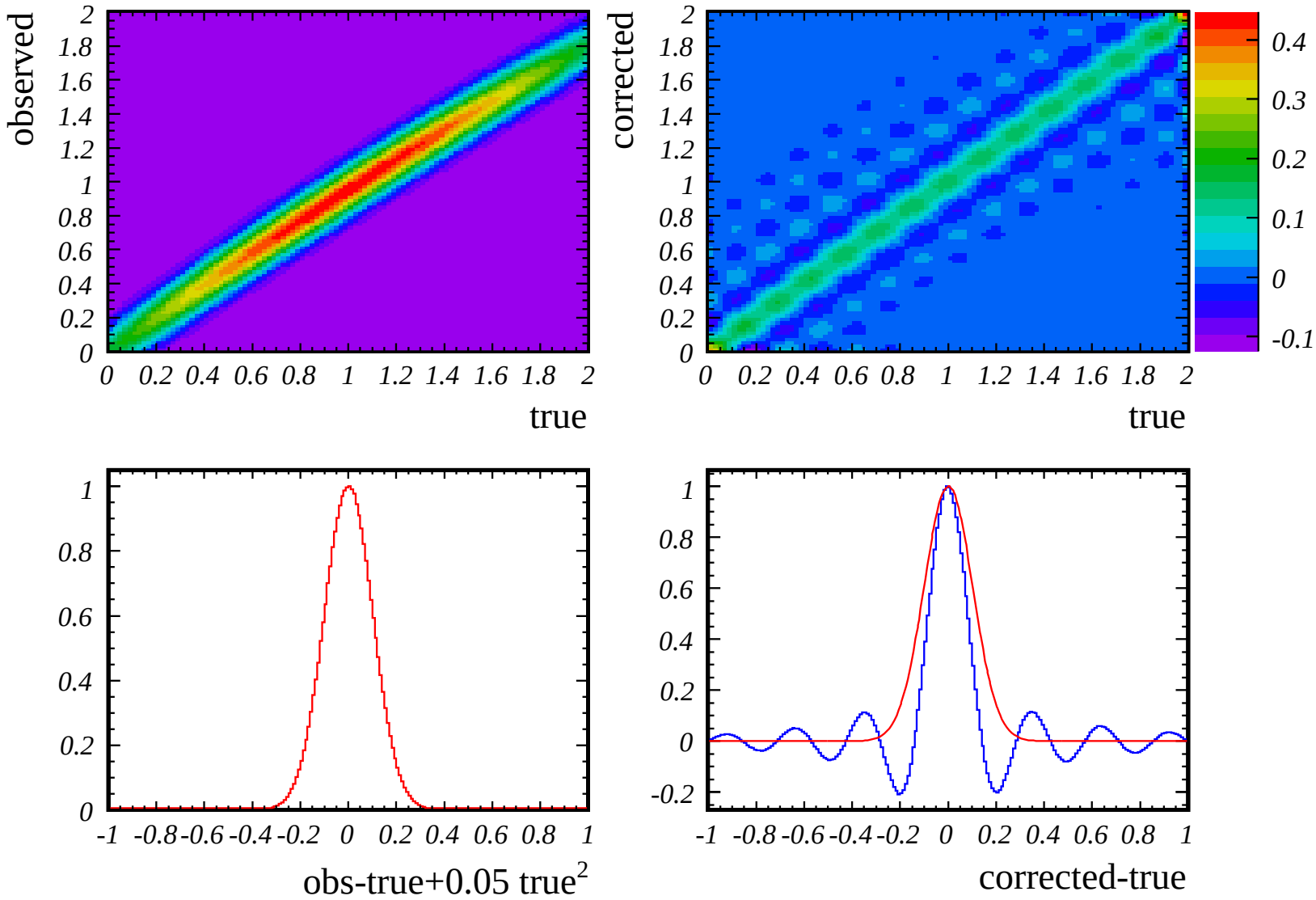
Gaussian Kernels $n_k = 15$ and $\lambda = 0.20$



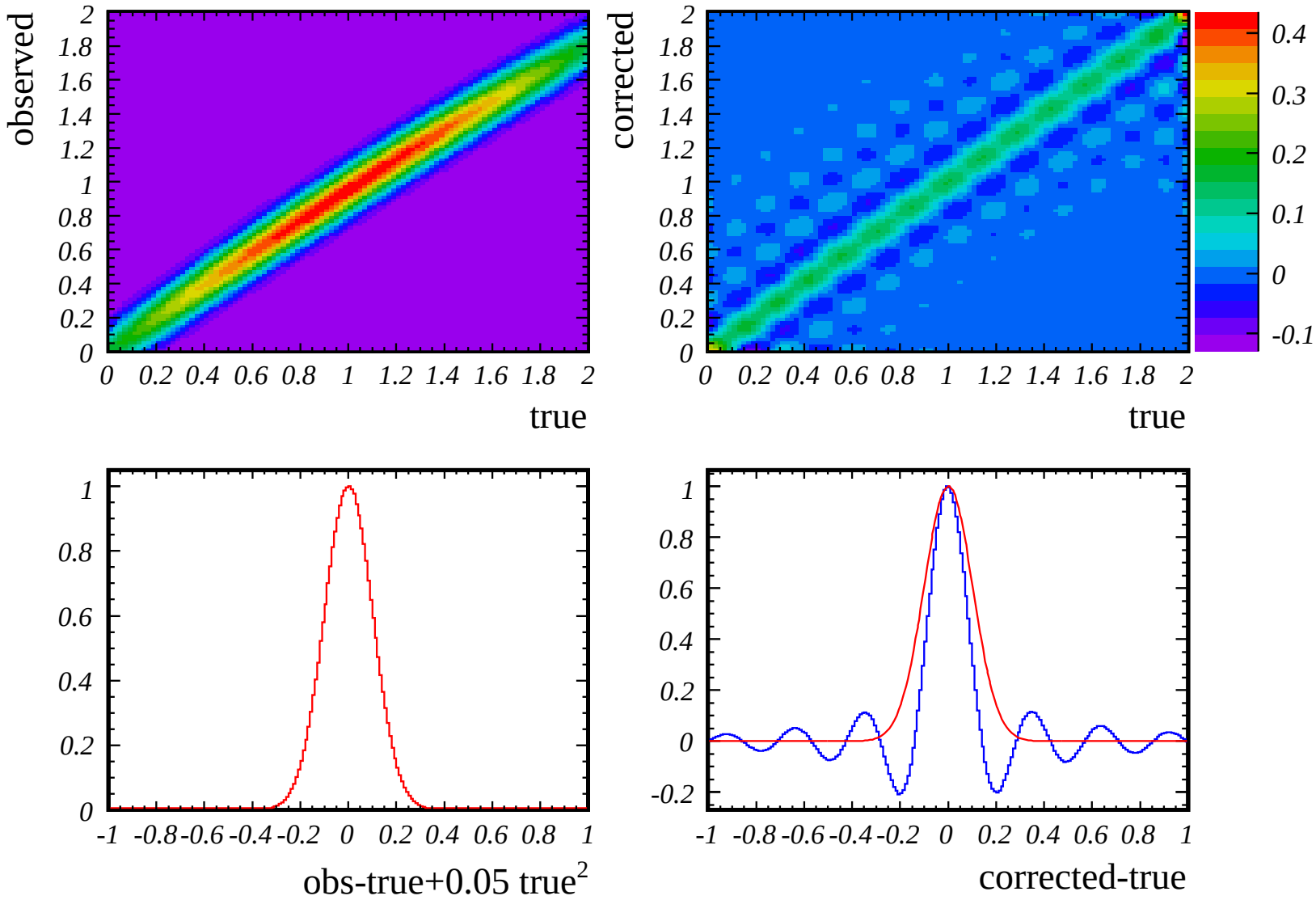
Epanechnikov Kernels $n_k = 15$ and $\lambda = 0.20$



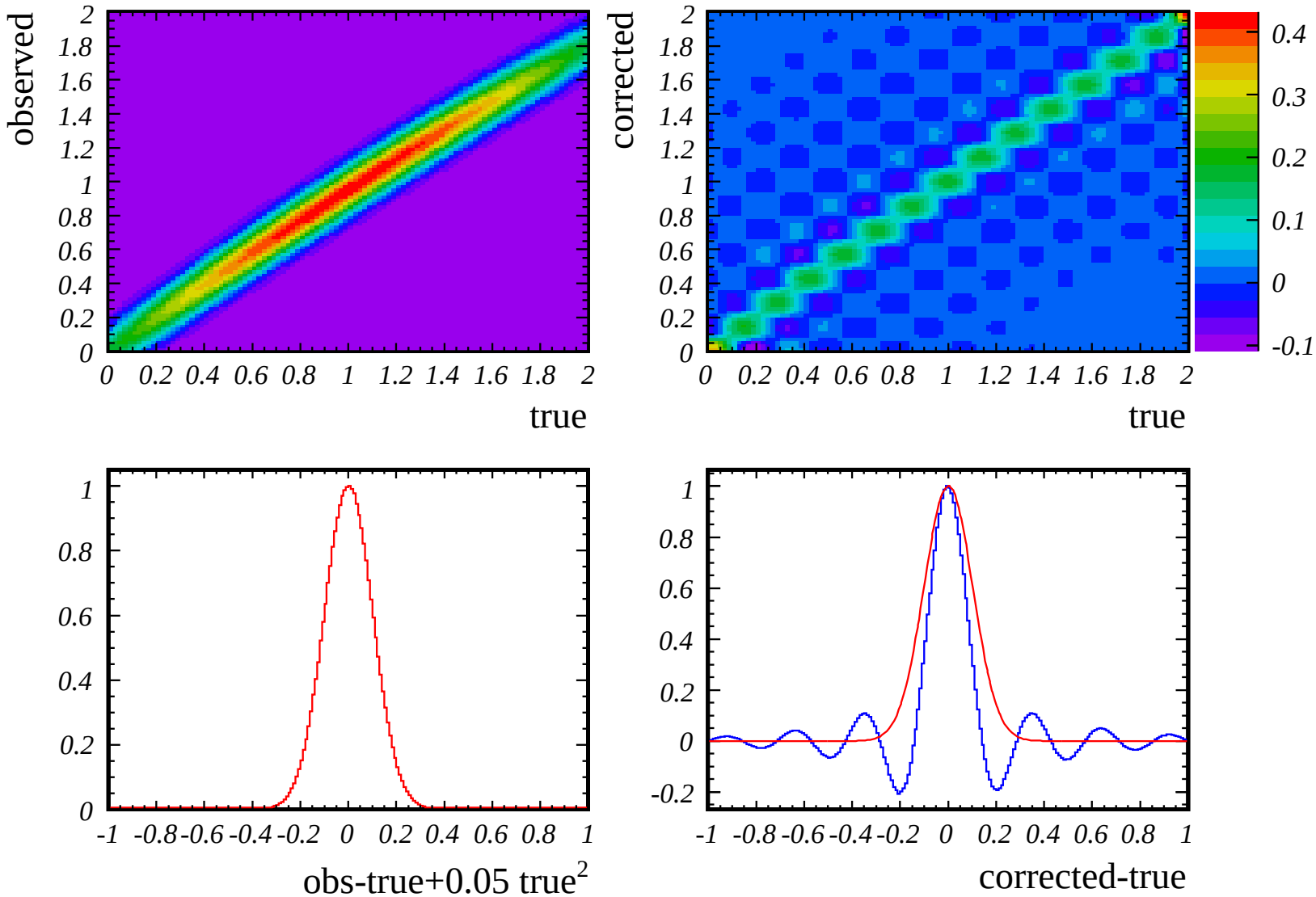
Cauchy Kernels $n_k = 15$ and $\lambda = 0.20$



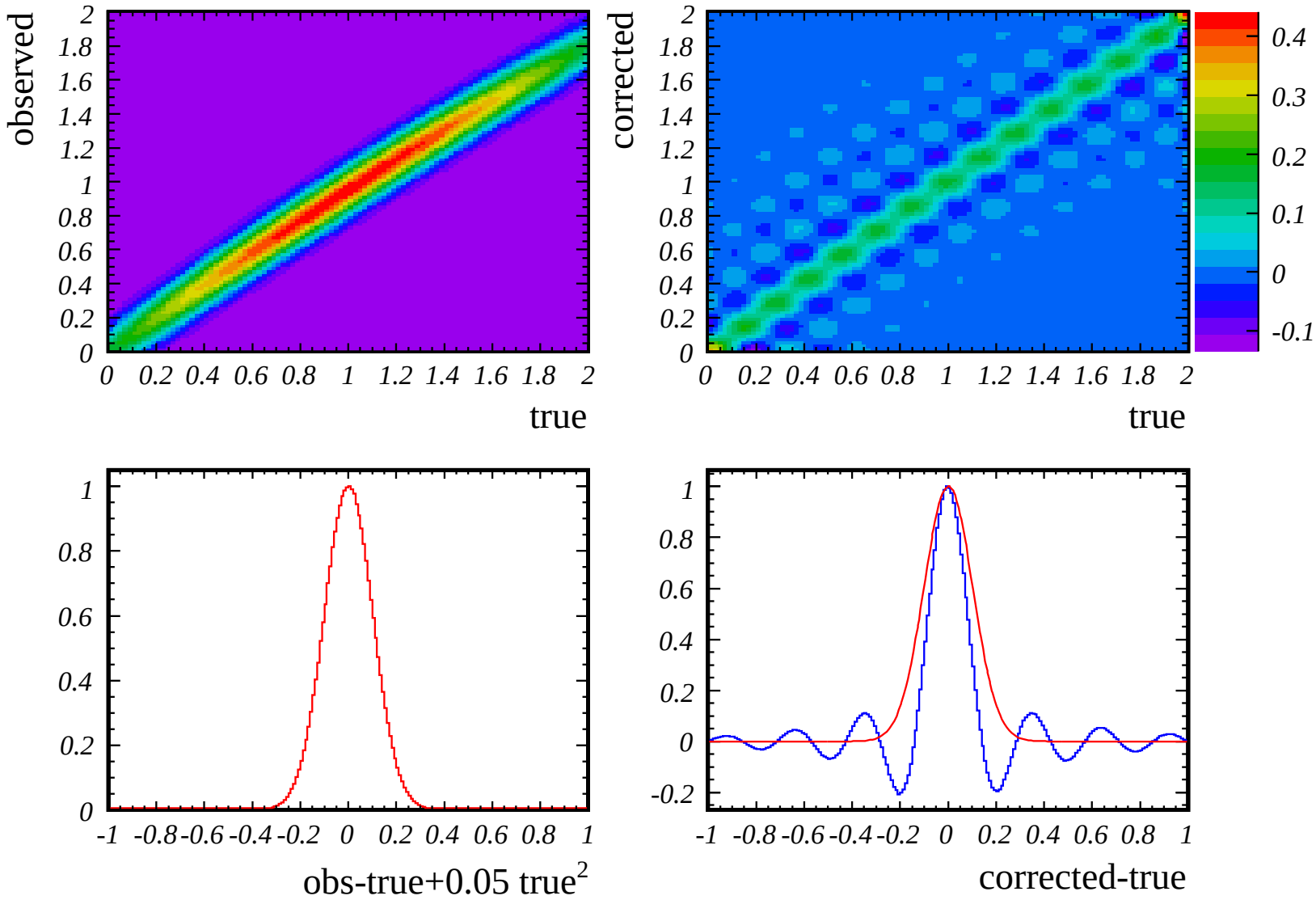
Gaussian Kernels $n_k = 15$ and $\lambda = 0.10$



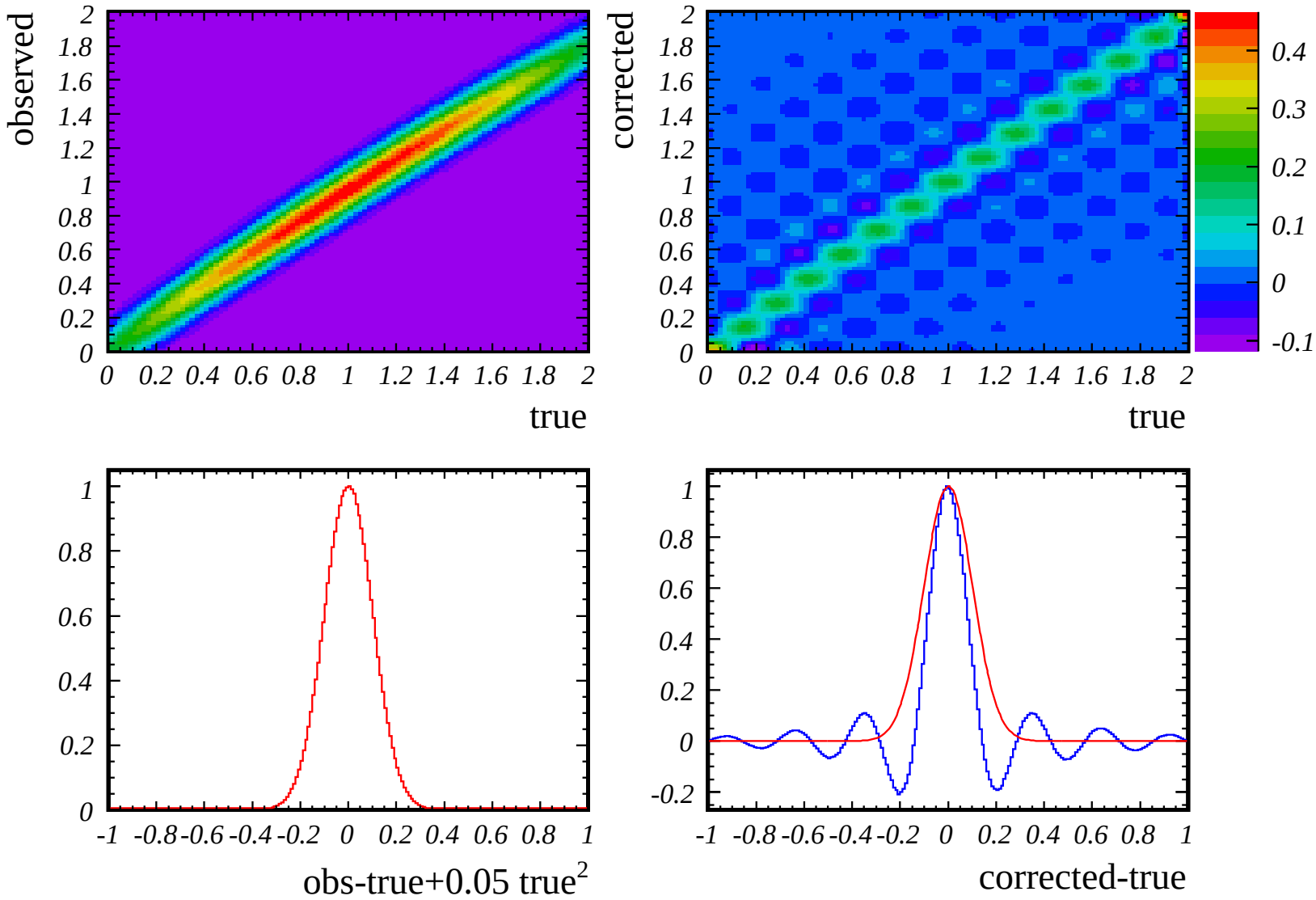
Epanechnikov Kernels $n_k = 15$ and $\lambda = 0.10$



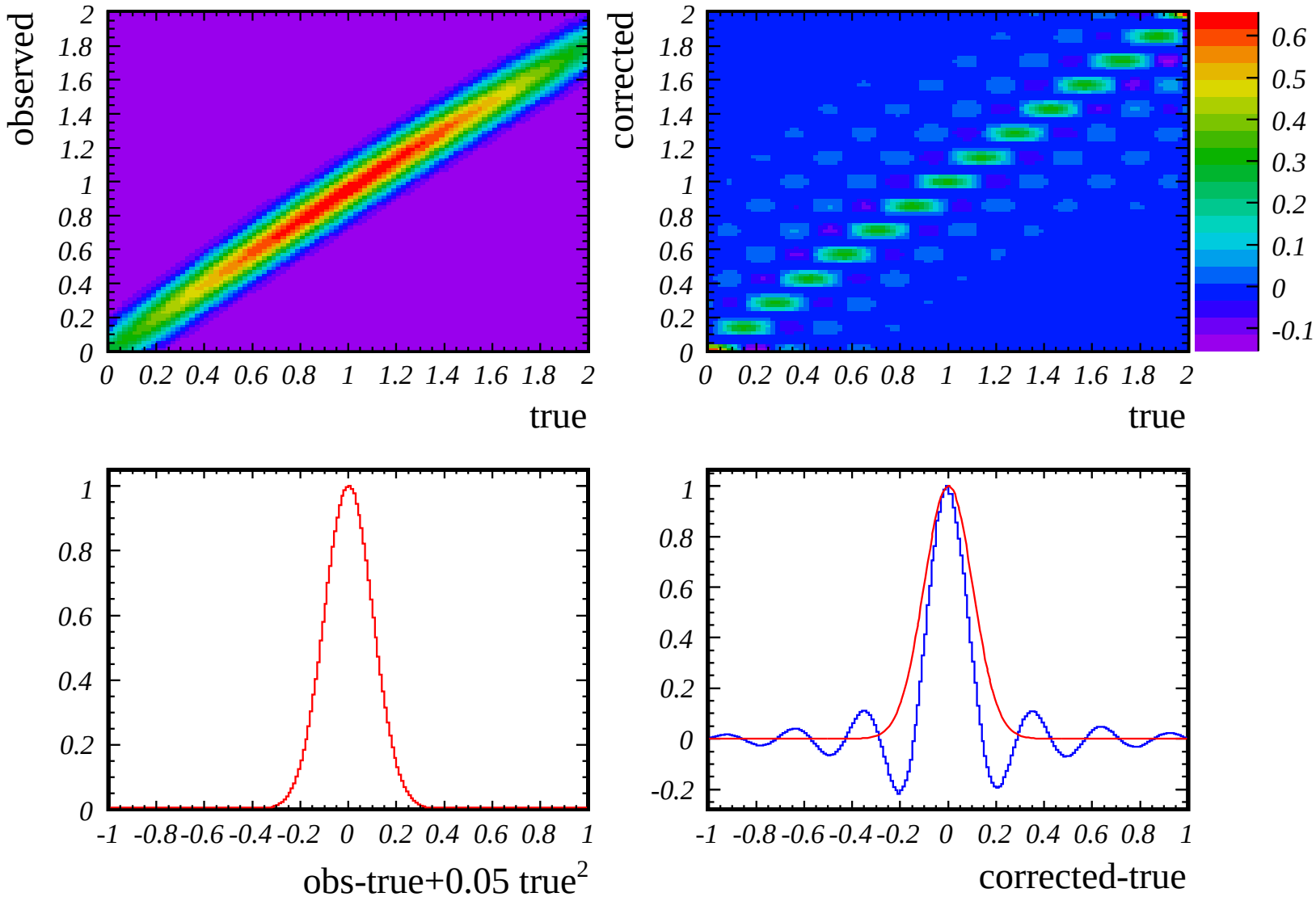
Cauchy Kernels $n_k = 15$ and $\lambda = 0.10$



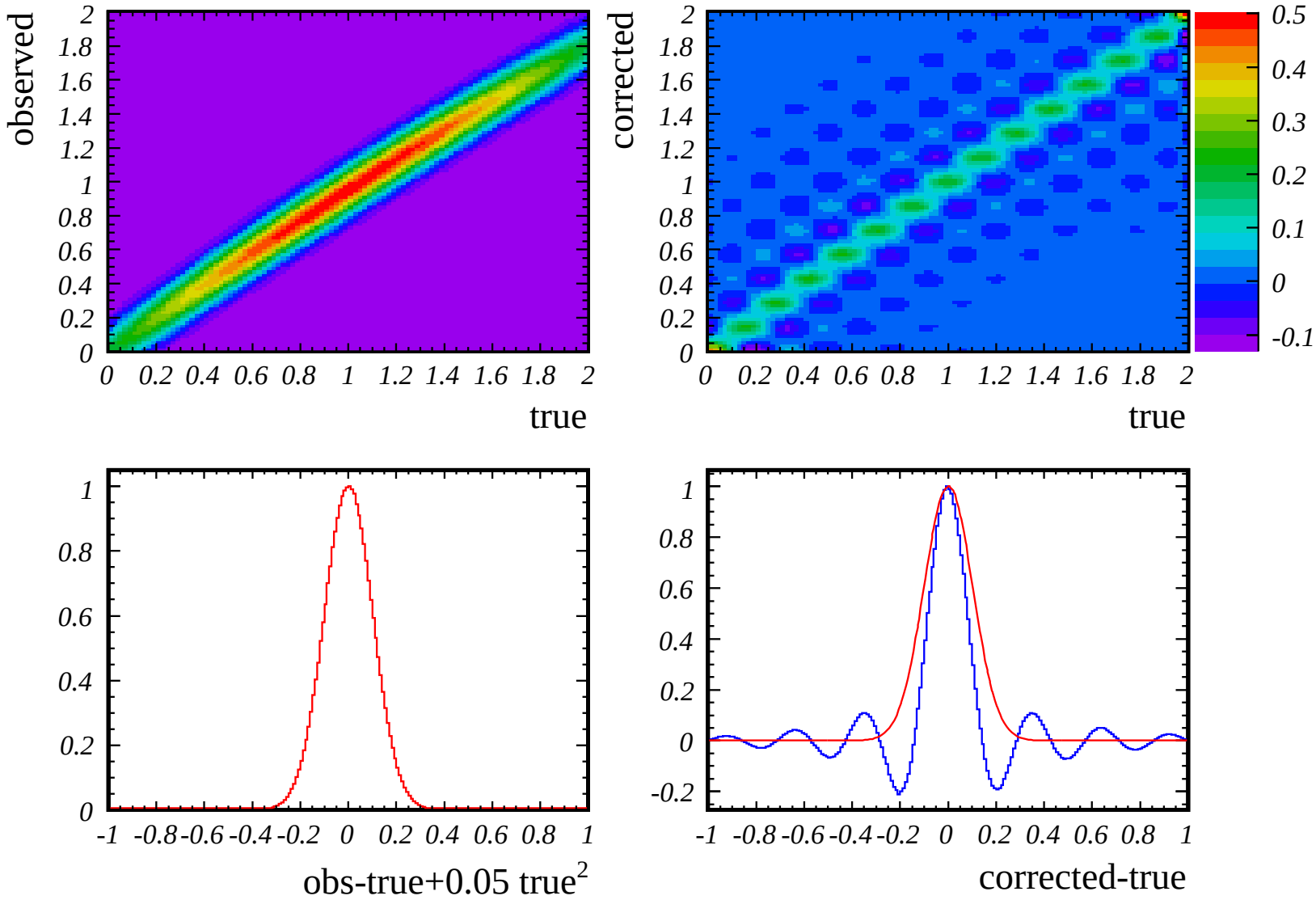
Gaussian Kernels $n_k = 15$ and $\lambda = 0.05$



Epanechnikov Kernels $n_k = 15$ and $\lambda = 0.05$



Cauchy Kernels $n_k = 15$ and $\lambda = 0.05$



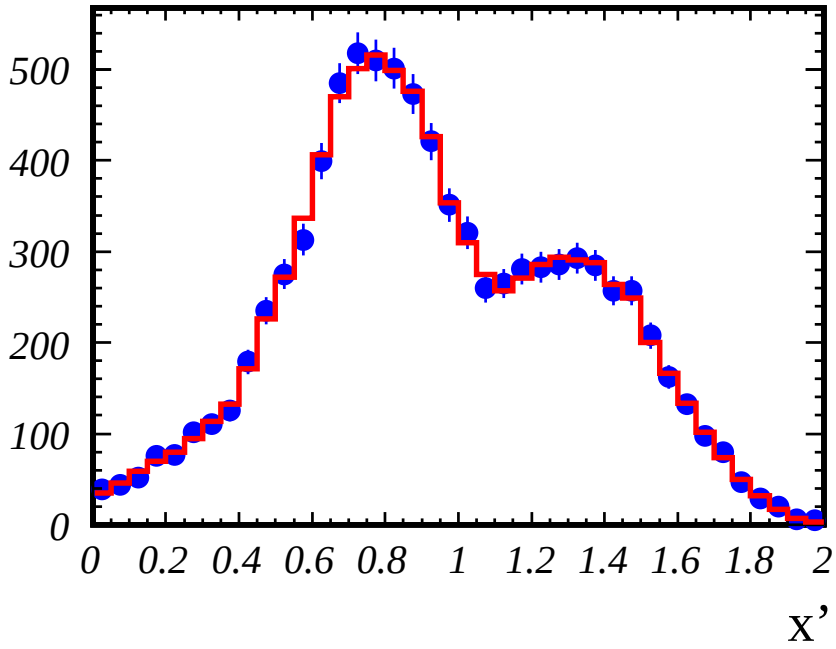
➜ *for the posterior response matrix and resolution. . .*

- the most important parameter is n_k the number of kernels
- resolution (PDF of corrected-true) shows very little sensitivity to
  - ➜ type of kernel function
  - ➜ kernel bandwidth λ
- smoothness of $\hat{b}$ improves with larger λ
- quantify resolution by e.g. FWHM of PDF(corrected-true)
- in the given example . . .
  - ➜ no bias and no efficiency losses for R'
  - ➜ $n_k = 10$ has same resolution as the R
  - ➜ $n_k = 15$ has improved resolution

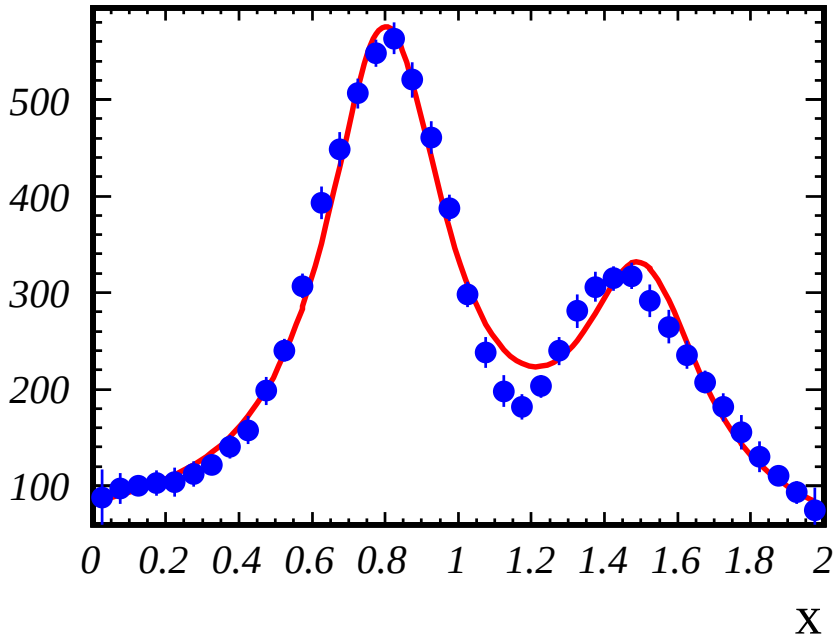
➜ do some actual unfolding

Gaussian kernels: $n_k = 10$ and $\lambda = 0.20$

observed and unfolded*smeared distribution



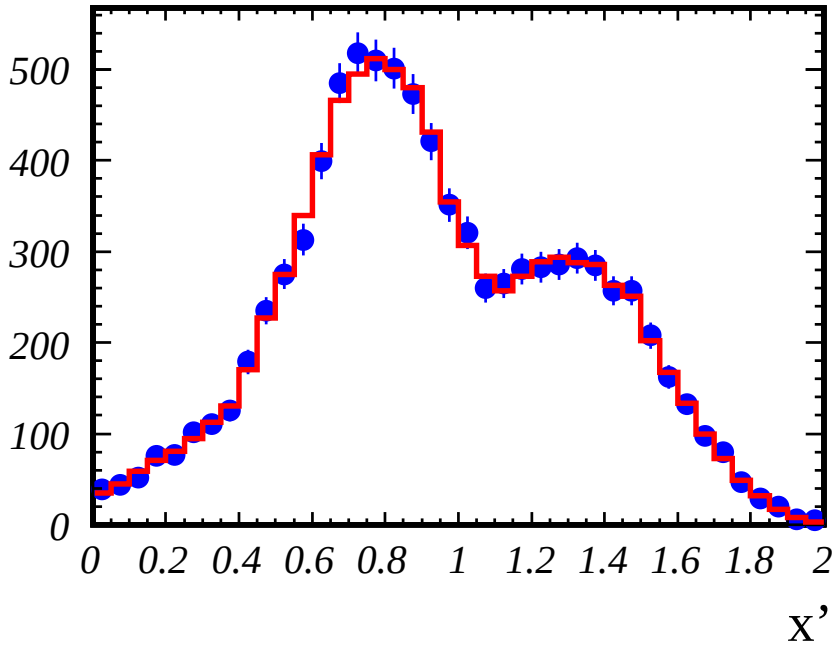
true and unfolded distribution



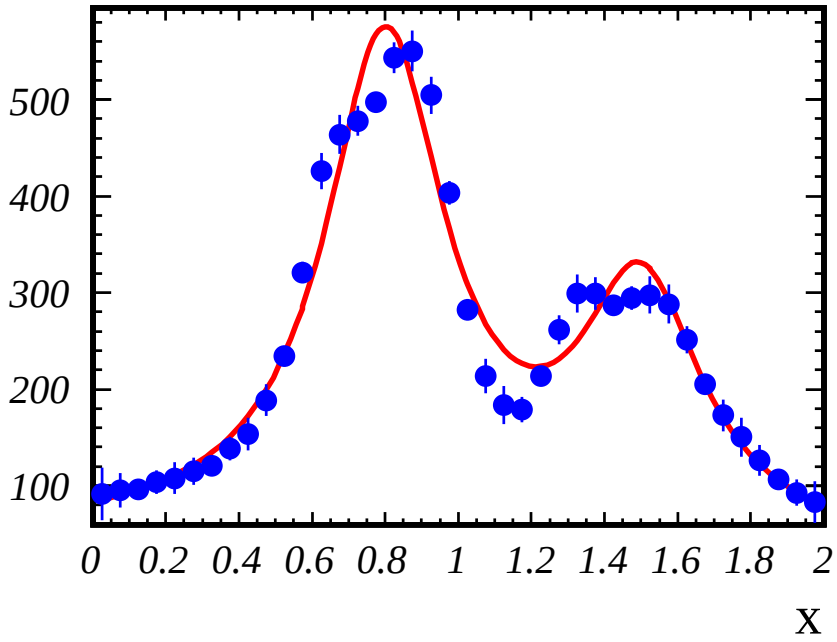
- 40 bins for true and observed distribution
- 10000 events generated for true distribution

Gaussian kernels: $n_k = 10$ and $\lambda = 0.10$

observed and unfolded*smeared distribution



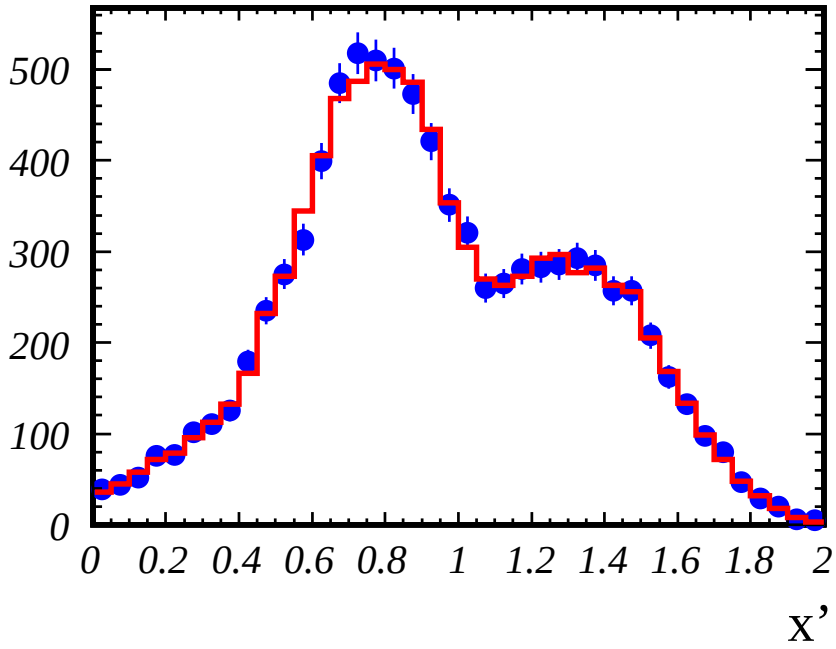
true and unfolded distribution



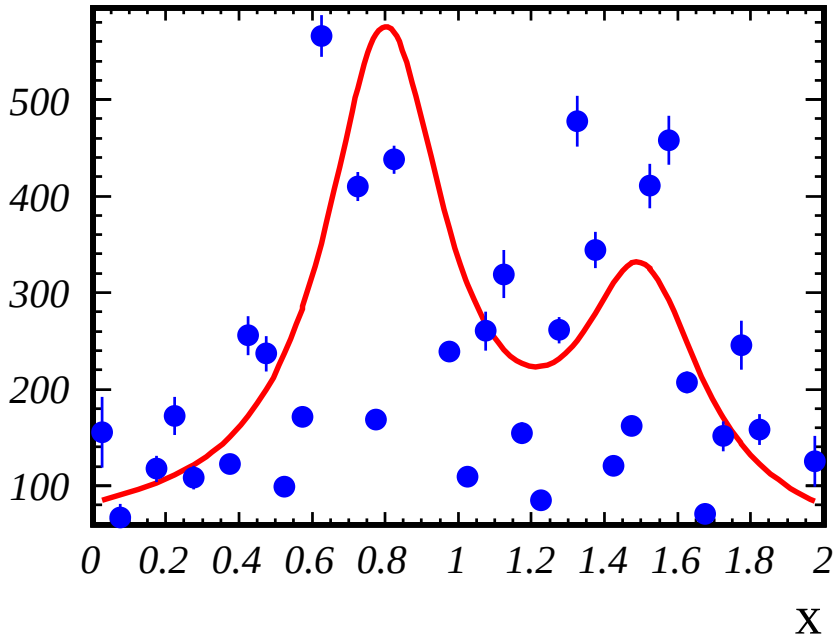
- 40 bins for true and observed distribution
- 10000 events generated for true distribution

Gaussian kernels: $n_k = 10$ and $\lambda = 0.05$

observed and unfolded*smeared distribution



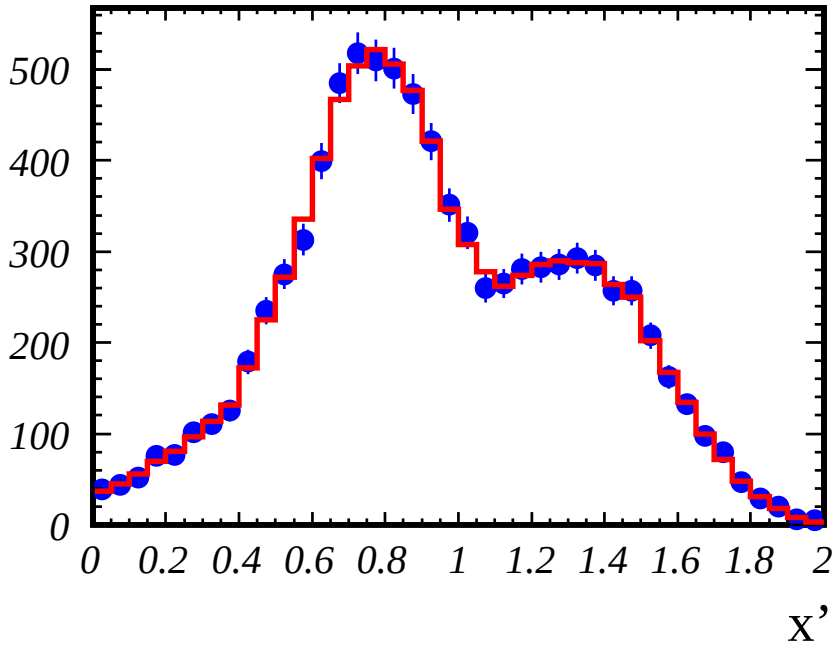
true and unfolded distribution



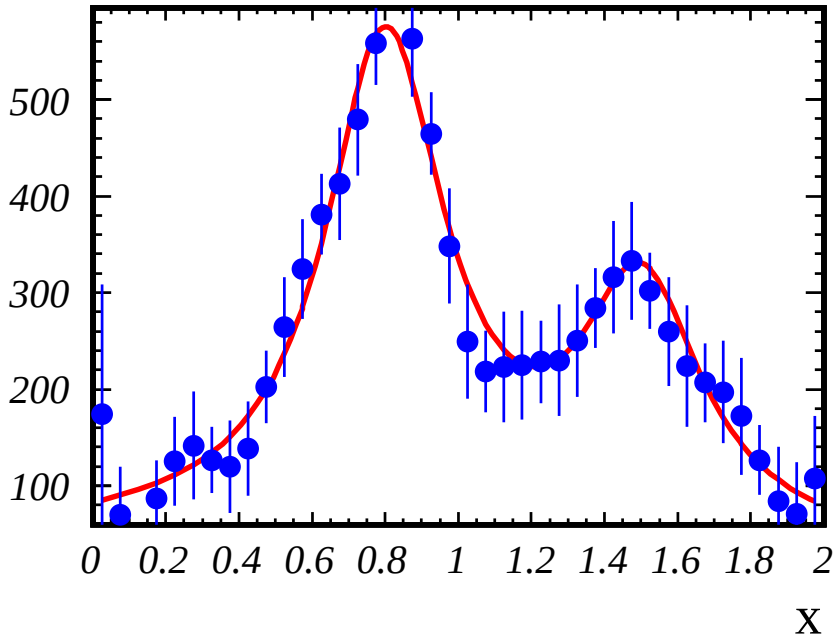
- 40 bins for true and observed distribution
- 10000 events generated for true distribution

Gaussian kernels: $n_k = 15$ and $\lambda = 0.20$

observed and unfolded*smeared distribution



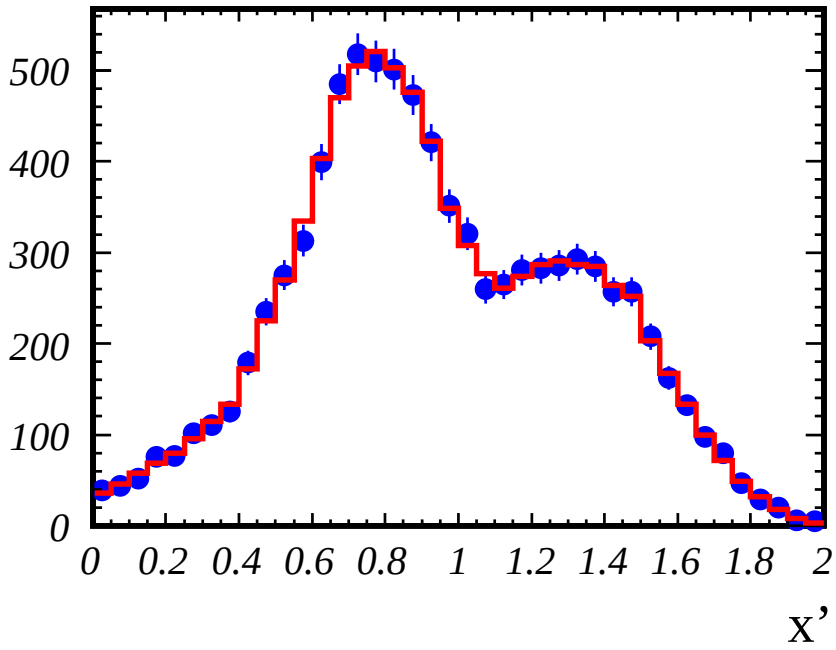
true and unfolded distribution



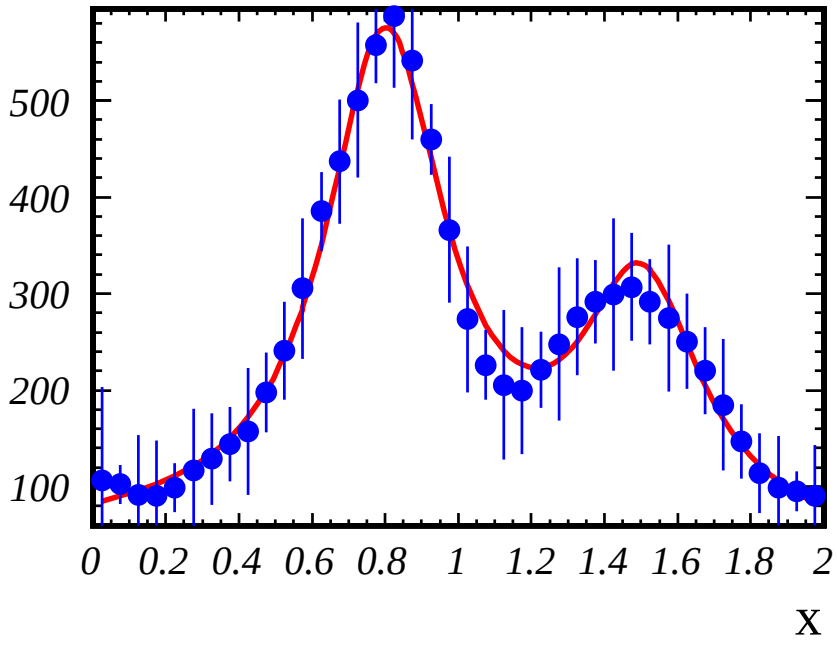
- 40 bins for true and observed distribution
- 10000 events generated for true distribution

Gaussian kernels: $n_k = 15$ and $\lambda = 0.10$

observed and unfolded*smeared distribution



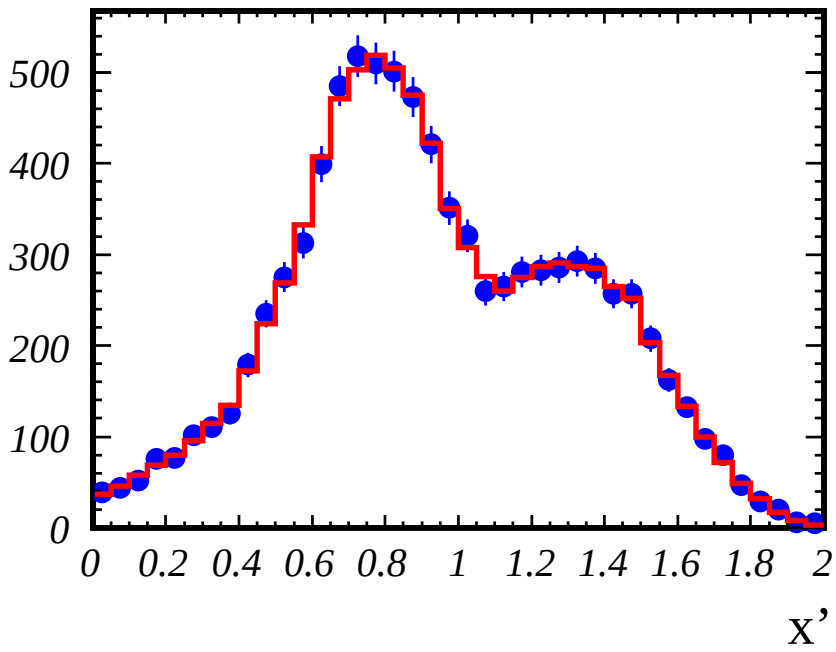
true and unfolded distribution



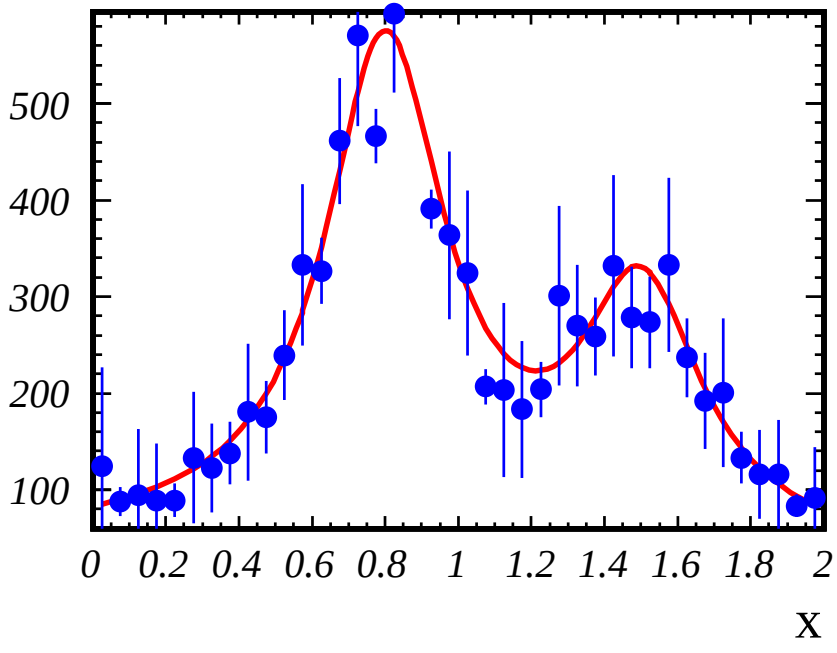
- 40 bins for true and observed distribution
- 10000 events generated for true distribution

Gaussian kernels: $n_k = 15$ and $\lambda = 0.05$

observed and unfolded*smeared distribution



true and unfolded distribution



- 40 bins for true and observed distribution
- 10000 events generated for true distribution

- all unfolding results provide good fits to the observed distribution
- improvement of resolution σ leads to larger errors
  - ➜ unsmearing is paid by loss of statistical precision
- unfolding result depends on n_k and kernel bandwidth λ
- n_k determines resolution for the unfolding result
- λ determines smoothness of the unfolding result
  - ➜ $\lambda \ll \sigma$: results look unphysical
  - ➜ $\lambda \gg \sigma$: long range correlations between errors
  - ➜ optimum bandwidth $\lambda \sim \sigma$
- conclusions:
  - ➜ unfolding corrects for bias and efficiency losses
  - ➜ resolution can/has to be chosen according to needs
  - ➜ kernel type & bandwidth determine smoothness & error matrix

❖ looking for the best kernels . . .

➜ *collecting all factors. . .*

$$R' = K\ (K^T R^T C^{-1} R\ K)^{-1} K^T R^T C^{-1} R$$

- with
  - ➜ $C[n_a, n_a]$: covariance matrix of the measured distribution a
  - ➜ $R[n_a, n_b]$: response matrix from true distribution b to a
  - ➜ $K[n_b, n_k]$: matrix of kernel functions - basis vectors to describe b
  - ➜ $R'[n_b, n_b]$: posterior response matrix after least squares fit
- simplify the expression for R' by introducing a new matrix:

$$M = R^T C^{-1} R$$

  - ➜ square matrix with dimensions $M[n_b, n_b]$ and rank n_a
- result:

$$R' = K\ \left(K^T M\ K\right)^{-1} K^T M$$

➜ *write down R' with matrix dimensions*

$$K[n_b, n_k] \left(K^T[n_k, n_b] M[n_b, n_b] K[n_b, n_k]\right)^{-1} K^T[n_k, n_b] M[n_b, n_b]$$

- the posterior response has dimensions $R'[n_b, n_b]$
- R' can only be determined if the inverse of $S = K^T M\ K$ exists
  - ➜ the dimensions are $S[n_k, n_k]$, the rank is $\min(n_k, n_a)$
  - ➜ the inverse exists for $n_k \leq n_a$
  - ➜ one can fit at most $n_k = n_a$ weights to n_a data points
- for the special case $n_k = n_b = n_a$ one finds

  $$R' = K\ \left[K^T M\ K\right]^{-1} K^T M = K\ \left[K^{-1} M^{-1} (K^T)^{-1}\right] K^T M = 1$$

  - ➜ perfect unfolding
  - ➜ usually impossible in practical applications since $n_k \ll n_b$

➜ *independent of statistics*

$$R' = K \left[K^T M K\right]^{-1} K^T M$$

➜ with N events in the measured distribution one finds

$$M = R^T C^{-1} R \propto \frac{1}{N}$$

$$(K^T M K)^{-1} \propto N$$

➜ i.e. the N-dependence cancels

➜ *invariance under regular transformations of the basis K*

$$K[n_b, n_k] \rightarrow K[n_b, n_k] \cdot V[n_k, n_k] \quad \text{where} \quad V^{-1} \quad \text{exists}$$

- proof: substitute $K\ V$ into the definition of R'

$$\begin{aligned}
R'' &= (K\ V)\ \left[(K\ V)^T M\ (K\ V)\right]^{-1} (K\ V)^T M \\
&= K\ V\ \left[V^T K^T M\ K\ V\right]^{-1} V^T K^T M \\
&= K\ V\ \left[V^T (K^T M\ K)\ V\right]^{-1} V^T K^T M \\
&= K\ V\ \left[V^{-1} \left(K^T M\ K\right)^{-1} (V^T)^{-1}\right] V^T K^T M \\
&= K\ \left[K^T M\ K\right]^{-1} K^T M = R'
\end{aligned}$$

- kernels (basis functions) which can be (approximately) transformed into each other will have (approximately) the same posterior response
- the approximation of one set of kernels through another will get better with increasing number of kernels n_k
- asymptotically, i.e. for $n_k = n_b$, the choice of kernels does not matter and the posterior response will always be a unit matrix
- convergence to the asymptotic case often is surprisingly fast
- to understand the properties of the posterior response one can select a set of kernel functions which makes problem most transparent
- a natural basis is given by the eigenvectors of $M = R^T C^{-1} R$
  - ➜ positive definite symmetric matrix
  - ➜ ortho-normal set of eigenvectors (basis functions)

Introducing the diagonal matrix E of eigenvalues, the eigenvector equation becomes (with explicit dimensions for the matrices)

$$M[n_b, n_b]\, K[n_b, n_k] = K[n_b, n_k]\, E[n_k, n_k]$$

and, using the ortho-normality of the eigenvectors

$$K^T[n_k, n_b] K[n_b, n_k] = 1[n_k, n_k]$$

as well as the symmetry of M and E, one finds for R':

$$\begin{aligned} R' &= K\ \left[K^T M\ K\right]^{-1} K^T M \\ &= K\ \left[K^T (M\ K)\right]^{-1} (M\ K)^T \\ &= K\ \left[K^T K\ E\right]^{-1} (K\ E)^T \\ &= K\ \left[E\right]^{-1} E^T K^T = K\ E^{-1} E\ K^T = K\ K^T \end{aligned}$$

Discussion

- the posterior response matrix will be symmetric
- asymptotically for $n_k = n_b$ one has $R' = 1$
- for small values n_k the eigenvectors need to be known
- if M decribes detector smearing, then it is not only symmetric but approximately of the structure $M_{ij} \approx f(i-j)$
- for $M_{ij} = f(i-j)$ the eigenvectors are harmonic functions
  - ➜ the kernel matrix K is given by the discrete Fourier transform

$$K_{i0} = \frac{1}{\sqrt{n_b}} \quad \text{and} \quad K_{ik} = \sqrt{\frac{2}{n_b}} \cos\left(\frac{k\pi}{2n_b}(2i+1)\right) \quad \text{for} \quad k > 0$$

  - ➜ posterior response matrix R' can be calculated
  - ➜ posterior resolution function $\langle x_{\mathrm{true}} - x_{\mathrm{unf}} \rangle$ can be calculated

➜ *normalized resolution curve for the leading n_k kernels:*

$$g(\phi) = \frac{1}{n_k}\frac{\sin(n_k\phi)}{\sin\phi}\cos((n_k-1)\phi) - \frac{1}{n_k}\sum_{k=1}^{n_k-1}\frac{\sin(2k\phi)}{k(\pi-2\phi)}$$

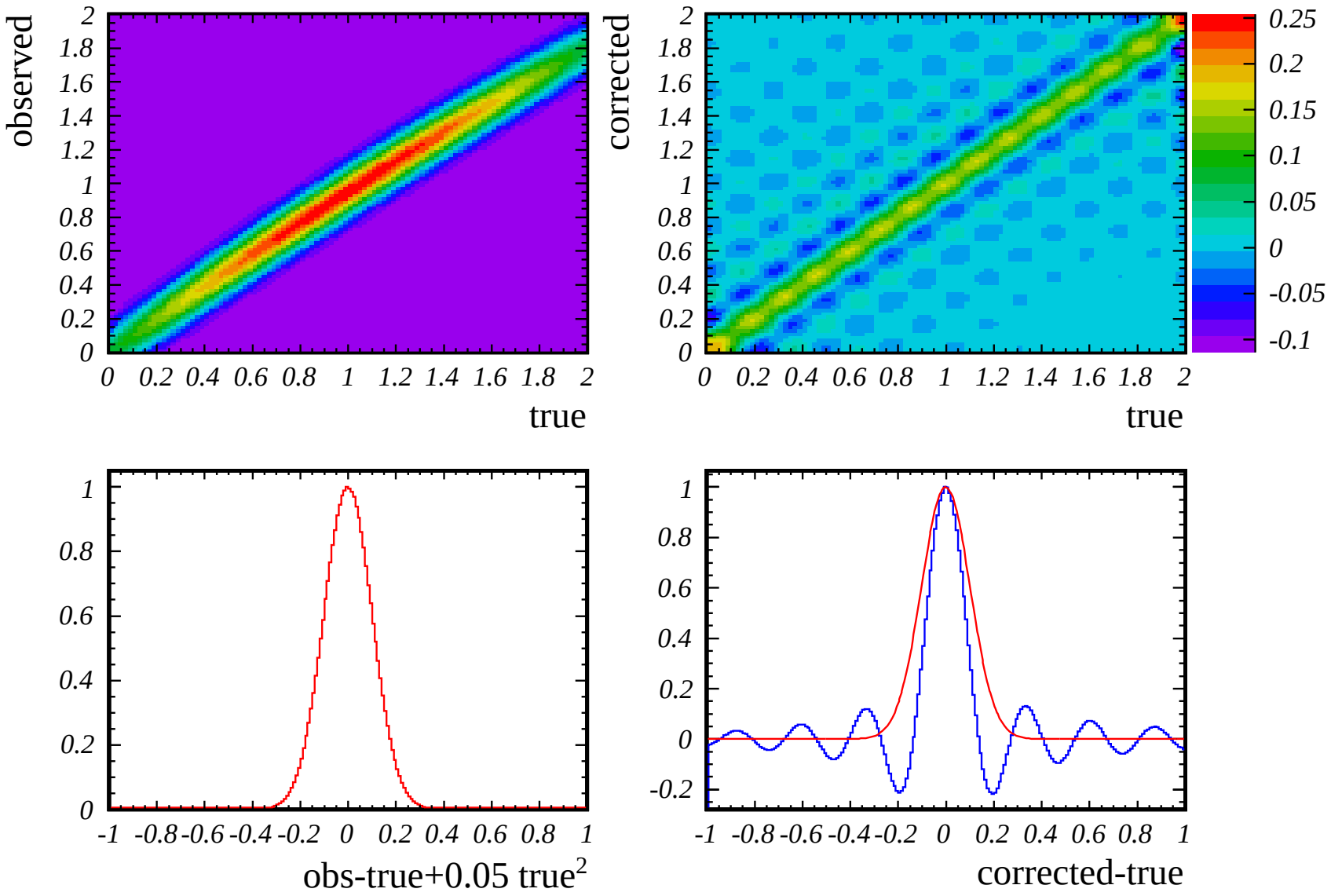
with, denoting with Δ the full x-range,

$$\phi = \left|\frac{\pi(x_{\text{true}} - x_{\text{unf}})}{2\Delta}\right|$$

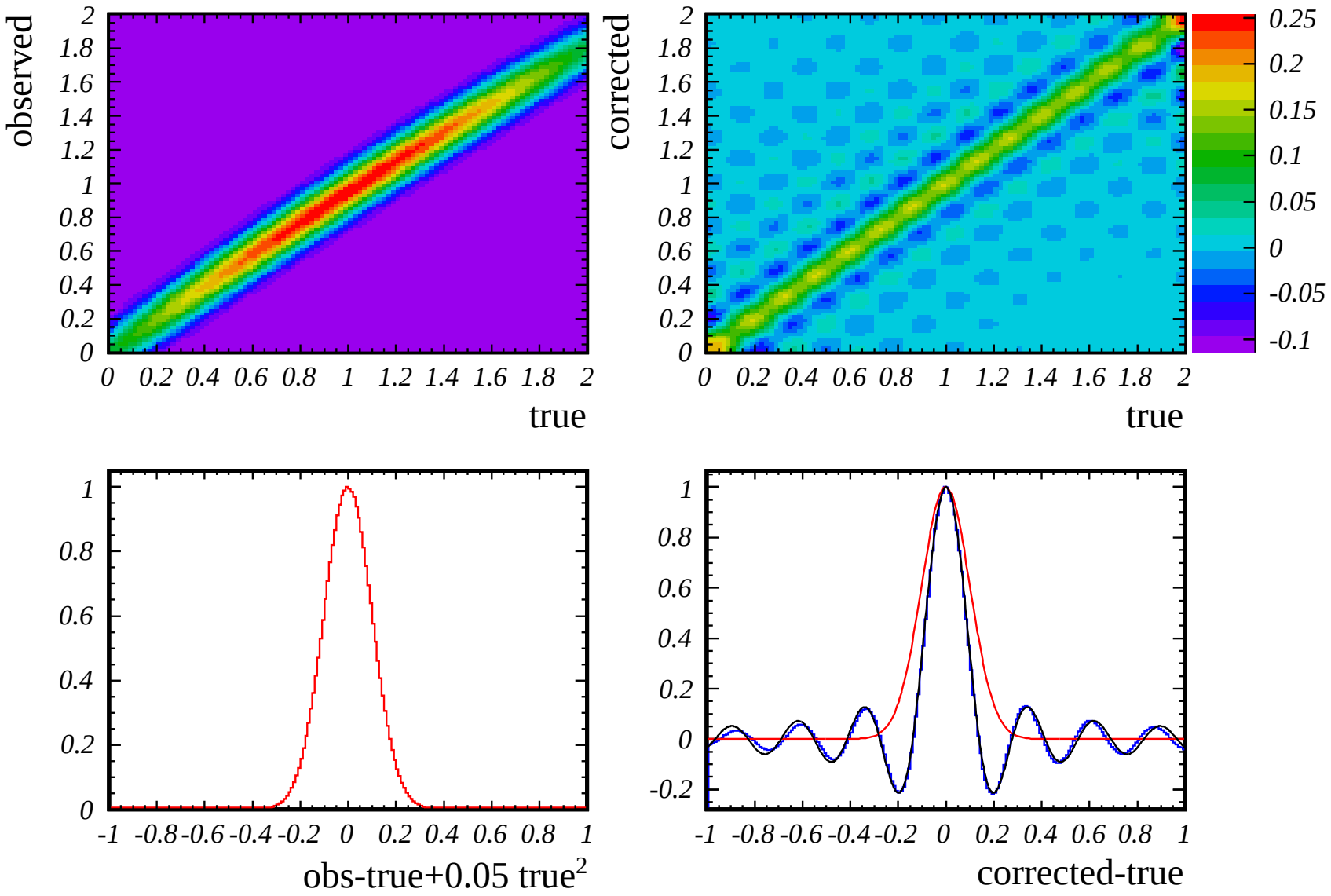
- damped oscillatory behaviour
- asymptotically independent of the choice of kernels
  - ➜ even polynomials would do; invariance under coordinate transformation implies equivalence to orthogonal polynomials, with similar properties as harmonic functions

numerical check ➜

Numerical result with 15 gaussian kernels



overlayed with analytical prediction



➜ *lessons learned*

- perfect unfolding is not possible
  - ➜ ill-conditioned problem
  - ➜ sensitive to statistical and numerical noise
  - ➜ unfolding results are prone to instabilities
  - ➜ huge literature exists on regularization schemes
- efficiency losses and biases are corrected
- resolution can be improved, but . . .
  - ➜ improvement of resolution entails loss in statistical precision
- the resolution of the unfolding result is driven by the numbers of terms used to build a representation of the truth
- a universal (asymptotic) function seems to exist for the resolution as a function of the number of terms used

➜ *basic idea:*

Representation of a function h as superposition H of harmonic functions.

For complex valued functions $h(t)$ one has:

$$H(\nu) = \int_{-\infty}^{\infty} dt\, h(t) e^{2\pi i \nu t} \quad \text{with} \quad h(t) = \int_{-\infty}^{\infty} d\nu\, H(\nu) e^{-2\pi i \nu t}$$

remark:

Using frequency ν instead of $\omega = 2\pi\nu$ yields, up to the $i \leftrightarrow -i$ symmetric expressions in the time and frequency domain.

note:

- determination of all frequencies requires infinitely fine sampling
- in practical applications:
  - ➜ finite resolution with time steps Δ
  - ➜ frequencies in the range $[-\nu_c, +\nu_c]$ can be determined
  - ➜ the maximum frequency is $\nu_c = 1/(2\Delta)$

Consider sampling times t_n and frequencies ν_+ and ν_- :

$$t_n = n \cdot \Delta \quad , \quad \nu_+ = \frac{1-c}{2\Delta} \quad \text{and} \quad \nu_- = \frac{c-1}{2\Delta} = -\nu_+$$

For $0 \leq c \leq 1$ both frequencies are in $[-\nu_c, +\nu_c]$. The weight functions of the Fourier transform on the sampling times become

$$w_n^+(c) = e^{2\pi i \nu_+ t_n} = e^{\pi i n(1-c)} = e^{+i\pi n} e^{-i\pi nc} = (-1)^n e^{-i\pi nc}$$
$$w_n^-(c) = e^{2\pi i \nu_- t_n} = e^{\pi i n(c-1)} = e^{-i\pi n} e^{+i\pi nc} = (-1)^n e^{+i\pi nc}$$

Replacing $+c$ by $-c$, i.e. going to frequencies outside $[-\nu_c, +\nu_c]$ yields

$$w_n^+(-c) = w_n^-(+c) \quad \text{and} \quad w_n^-(-c) = w_n^+(+c) \, ,$$

No new information is obtained. The same holds for $c \to c + m$, with m integer.

➜ *result:*

sampling with time steps Δ probes frequencies in $\left[-\frac{1}{2\Delta}, +\frac{1}{2\Delta}\right]$

Take N samples h_k, $k = 0, \ldots, N-1$, with uniform spacing Δ, and sampling times $t_k = k \cdot \Delta$. Assume even values N. According to the sampling theorem only frequencies in $[-\nu_c, +\nu_c]$ with $\nu_c = 1/(2\Delta)$ can contribute.

For uniform sampling determine $N+1$ uniformly spaced frequencies:

$$\nu_n = \frac{n}{N\Delta} \quad \text{with} \quad n = -\frac{N}{2}, \ldots, +\frac{N}{2}$$

The Fourier components H_n at the discrete frequencies are:

$$H(\nu_n) = \int dt\, h(t) e^{2\pi i \nu_n t} \approx \sum_{k=0}^{N-1} \Delta \cdot h_k e^{2\pi i \nu_n t_k} = \Delta \cdot \sum_{k=0}^{N-1} h_k e^{2\pi i n k/N} = \Delta \cdot H_n$$

and since

$$H_{n+N} = \sum_{k=0}^{N-1} h_k e^{2\pi i (n+N) k/N} = \sum_{k=0}^{N-1} h_k e^{2\pi i n k/N} e^{2\pi i k} = H_n$$

i.e. $H_{-N/2} = H_{N/2}$, there are N independent components for N samples h_k.

➜ *substituting $H_{-n} \equiv H_{N-n}$ to map everything to positive frequencies:*

$$H_n = \sum_{k=0}^{N-1} h_k \, e^{2\pi i k n/N} \quad \text{and} \quad h_k = \frac{1}{N} \sum_{n=0}^{N-1} H_n \, e^{-2\pi i k n/N}$$

(positive frequencies for $0 < n < N/2$, negative frequencies for $n < N/2 < N$)

❖ proof of closure:

$$\frac{1}{N} \sum_{n=0}^{N-1} H_n \, e^{-2\pi i k n/N} = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} h_m \, e^{2\pi i m n/N} \, e^{-2\pi i k n/N}$$

$$= \frac{1}{N} \sum_{m=0}^{N-1} h_m \sum_{n=0}^{N-1} e^{2\pi i n (m-k)/N} = \frac{1}{N} \sum_{m=0}^{N-1} h_m \sum_{n=0}^{N-1} e^{(2\pi i (m-k)/N) \cdot n}$$

$$= \frac{1}{N} \sum_{m=0}^{N-1} h_m N \cdot \delta_{km} = h_k$$

Note that, except for $m = k$, the sum over $\exp((2\pi i(m-k)/N)n)$ is a sum over equidistant points on a unit circle, which is zero.

➜ *efficient numerical implementation for $N = 2^m$ points (Gauss)*

$$H_n = \sum_{k=0}^{N-1} h_k \, e^{2\pi i k n/N} = \sum_{k=0}^{N-1} W^{n\cdot k} h_k \quad \text{with} \quad W = e^{2\pi i/N}$$

In general one need $O(N^2)$ operations to get all H_n. However, splitting the sums in contributions for even and odd indices shows:

$$H_n = \sum_{k=0}^{N-1} h_k \, e^{2\pi i k n/N} = \sum_{k=0}^{N/2-1} h_{2k} \, e^{2\pi i (2k) n/N} + \sum_{k=0}^{N/2-1} h_{2k+1} \, e^{2\pi i (2k+1) n/N}$$

$$= \sum_{k=0}^{N/2-1} h_{2k} \, e^{2\pi i k n/(N/2)} + \sum_{k=0}^{N/2-1} h_{2k+1} \, e^{2\pi i k n/(N/2) + 2\pi i n/N}$$

$$= \sum_{k=0}^{N/2-1} h_{2k} \, e^{2\pi i k n/(N/2)} + W^n \sum_{k=0}^{N/2-1} h_{2k+1} \, e^{2\pi i k n/(N/2)} = H_n^e + W^n \cdot H_n^o$$

➜ *for an even number of sampling points one finds:*

$$H_n = H_n^e + W^n \cdot H_n^o \quad \text{with} \quad W = e^{2\pi i/N}$$

- ➜ all H_n^e and all H_n^o can be determined with $(N/2)^2$ operations
- ➜ all H_n can be determined with $2(N/2)^2 + N$ operations
- ➜ for $N = 2^m$ the approach can be iterated $m-1$-times

$$\begin{aligned} N^2 &\to 2(N/2)^2 + N \\ &\to 2(2(N/4)^2 + N/2) + N = 4(N/4)^2 + 2N \\ &\to 4(2(N/8)^2 + N/4) + 2N = 8(N/8)^2 + 3N \quad \text{etc.} \end{aligned}$$

minimum number of operations:

$$2^m \left(\frac{N}{2^m}\right)^2 + m \cdot N = N \cdot (m+1) = N(1 + \log_2 N) = O(N \log_2 N)$$

The power spectrum

Consider a signal $h(t)$ that was sampled with spacing Δ at N points h_k.
The H_n are the signal amplitudes at frequency $\nu = n/(N \cdot \Delta)$ Hz.
The "power spectrum" is the square of the absolute values of the amplitudes.

$$\sum_{k=0}^{N-1} |h_k|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |H_k|^2$$

For a single harmonic $A_k \sin(2\pi\, k\, t)$, one finds

$$\sum_{k=0}^{N-1} |h_k|^2 \approx A_k^2 \frac{N}{2} = \frac{1}{N} \left(|H_k|^2 + |H_{N-k}|^2 \right) \ ,$$

and thus

$$A_k^2 = \frac{2}{N^2} \left(|H_k|^2 + |H_{N-k}|^2 \right) \ .$$

For real-valued functions one has $H_{N-n} = H_n^*$, i.e. $|H_n| = |H_{N-n}|$, and thus:

$$A_k^2 = \frac{4}{N^2} |H_k|^2 \quad \text{with} \quad k < \frac{N}{2}$$

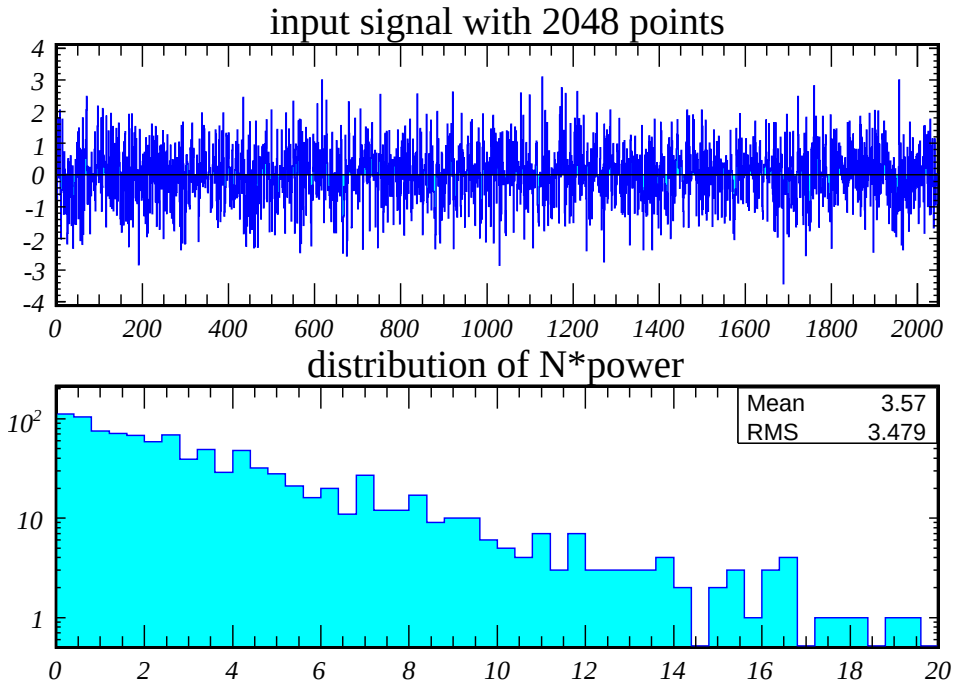
➜ *problem:*

How many samples N allow to find a frequency with power A^2 in a noisy signal?

➜ *solution:*

Study $N \cdot P$, with P the power, for purely gaussian noise with $\sigma = 1$.
Empirically finding: exponential distribution with a universal expectation value $\langle N \cdot P \rangle \approx 4$.

N	$\langle N \cdot P \rangle$
64	3.58
128	4.03
256	3.96
512	3.89
1024	3.59
2048	3.57



➜ *estimate the false-detection probability*

The probability to see a noise-signal with $P > A^2$ shall be less than p. The probability that this holds for a known frequency component is

$$p = \int_{A^2}^{\infty} dP \; \frac{1}{\langle P \rangle} \exp(-P/\langle P \rangle) = \int_{A^2/\langle P \rangle}^{\infty} dx \exp(-x) = \exp\left(-\frac{A^2}{\langle P \rangle}\right)$$

With $\langle P \rangle = 4\sigma^2/N$ this leads to:

$$p = \exp\left(-\frac{NA^2}{4\sigma^2}\right) \quad \text{or} \quad \frac{A}{\sigma} = \sqrt{-\frac{4}{N} \ln p}$$

The signal/noise should at least have this value to detect the frequency with N samples and error-probability p. The sensitivity scales with $\sqrt{N}$.

Example: $N = 1024$ and $p = 0.001$ ➜ $A/\sigma \approx 0.164$

➜ *estimate the false-detection probability*

If the frequency is not known, the look-elsewhere effect needs to be considered, i.e. none of the $N/2$ tested frequencies should have a noise power $P > A^2$. With p_1 the probability to see a single spurious frequency, the probability p to accept at least one spurious signals is

$$p = 1 - [1 - p_1]^{N/2}$$

and thus, using the previous result, $p_1 = \exp(-N\,A^2/4\sigma^2)$, one finds

$$\frac{A}{\sigma} = \sqrt{-\frac{4}{N}\ln(1-[1-p]^{2/N})} \approx \sqrt{-\frac{4}{N}\ln\frac{2p}{N}}$$

The sensitivity scales with $\sqrt{N/\ln N}$.

Example: $N = 1024$ and $p = 0.001$ ➜ $A/\sigma \approx 0.227$

❖ a larger signal/noise is required to counter the look-elsewhere effect

➜ *problem:*

A Signal $h(t)$, is sampled at in general not equidistant times t_j, $j = 1, \ldots, N$.
What is the power of a frequency ν in this signal?

- t_j can have arbitrary spacing
- smallest spacing defines the limiting frequency
- ν is a continuous variable
- for equidistant t_j a scan over ν corresponds to a Fourier transform

❖ solution: (normalized) Lomb periodogram (as a function of $\omega = 2\pi\nu$)

$$P_N(\omega) = \frac{1}{2\sigma^2}\left\{\frac{[\sum_{j=1}^{N} d_j \cos(\omega(t_j-\tau))]^2}{\sum_{j=1}^{N}\cos^2(\omega(t_j-\tau))} + \frac{[\sum_{j=1}^{N} d_j \sin(\omega(t_j-\tau))]^2}{\sum_{j=1}^{N}\sin^2(\omega(t_j-\tau))}\right\}$$

mit

$$d_j = h_j - \frac{1}{N}\sum_{i=1}^{N} h_i \;,\; \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} d_i^2 \;,\; \tan(2\omega\tau) = \frac{\sum_{j=1}^{N}\sin(2\omega t_j)}{\sum_{j=1}^{N}\cos(2\omega t_j)}$$

Properties of $P_N(\omega)$

Consider a simple periodic signal

$$h_j = A\cos(\omega t_j - \phi) \quad \text{with} \quad j = 1 \ldots N \ .$$

A lengthy elementary calculation shows:

$$P_N(\omega) = \frac{N}{2}$$

- P_N is 1/2-times the number of points containing $\omega = 2\pi\nu$
- τ is such that the result is independent of ϕ
- P_N is "per measurement", Fourier transformation "per time interval"

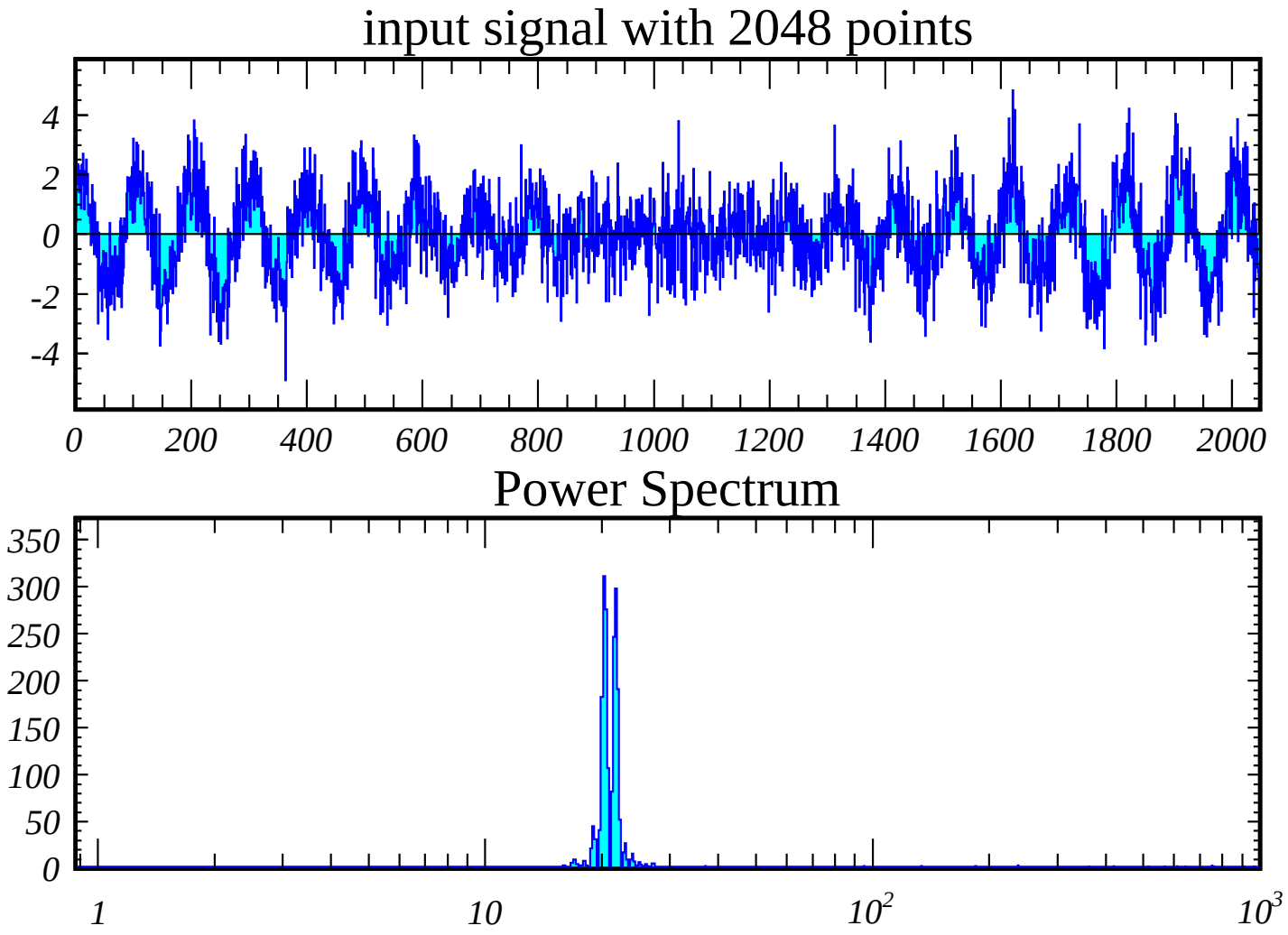
Power spectrum from the amplitude periodogram:

$$A^2(\omega) = \left(\frac{\sum_j d_j \cos(\omega(t_j - \tau))}{\sum_j \cos^2(\omega(t_j - \tau))} \right)^2 + \left(\frac{\sum_j d_j \sin(\omega(t_j - \tau))}{\sum_j \sin^2(\omega(t_j - \tau))} \right)^2$$

Response to noise, needed for signal significance, of the normalized P_N:

$$\frac{dn}{dP_N} = \exp(-P_N) \quad \text{i.e.} \quad \langle P_N \rangle = 1$$

➜ *Lomb-periodogram of two close-by frequencies plus noise*



➜ *two approaches to frequency domain*

- Fourier transformation: determine frequencies by sampling per time interval
  - ➜ yields amplitudes and phases
  - ➜ requires equidistant sampling
  - ➜ has numerically very efficient implementations
  - ➜ can be inverted
- Lomb periodogram: determine frequency component per measurement
  - ➜ better frequency resolution for given number of sampling points
  - ➜ applicable for arbitrary sampling times
  - ➜ yields only amplitude information
  - ➜ ideal to prove the existence of a frequency component
  - ➜ normalized periodogram: simple interpretation of significance