Unfolding: Regularization and error assignment

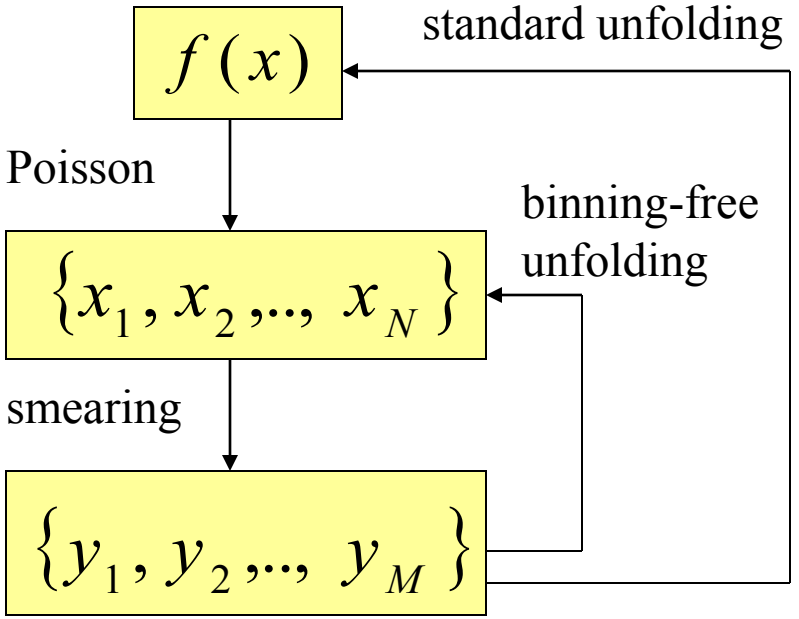
1. Introduction: some requirements
   - present data without regularization
   - fix the regularization strength
2. Example and three unfolding methods
3. Presentation and error assignment to the graph
4. Summary

The problem

unknown true distribution

true data sample

observed data sample



We got y, we want to know f(x)

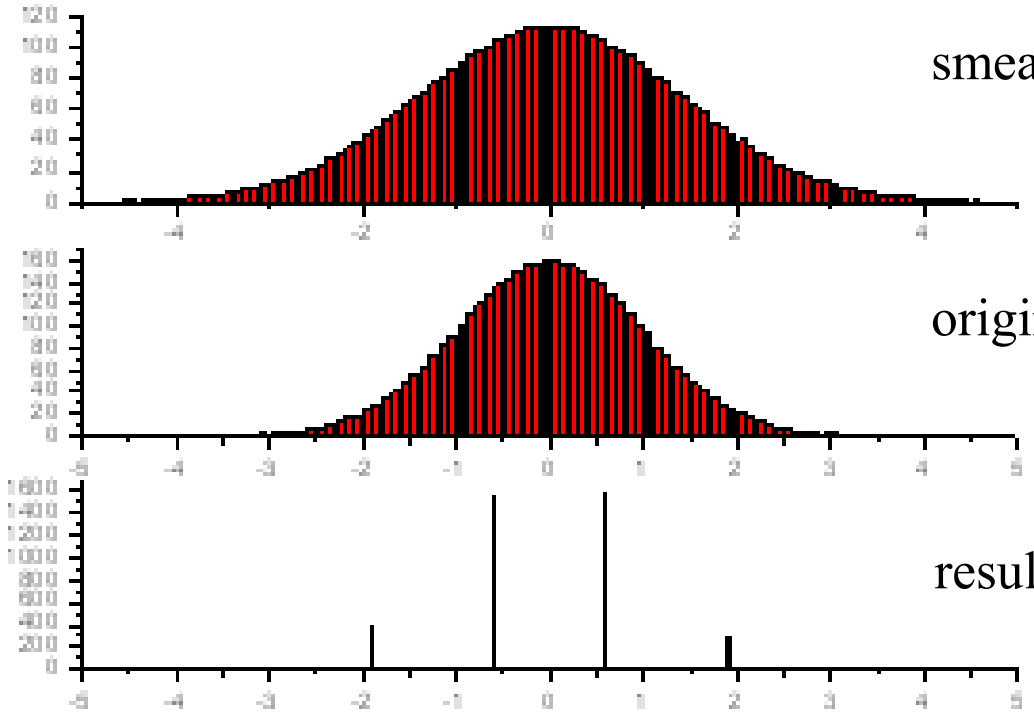
We parametrize the true distribution as a histogram (there are other attractive possibilities, i.e. splines)

As has been stressed, unfolding belongs to the class of illposed problems. Howeves when you parametrize the true distribution as a histogram with given bin widths, the problem becomes defined.

We then can determine the parameters of the bins and the corresponding error matrix. This is our mesurement which summarizes the information contained in the data. The errors have to allow for all distributions that are compatible with the observed data.

Even in the limit of infinite statistics unfolding does not necessarily lead to the correct solution.

Example: N(0,1) folded with N(0,1), avoiding statistical fluctuations.



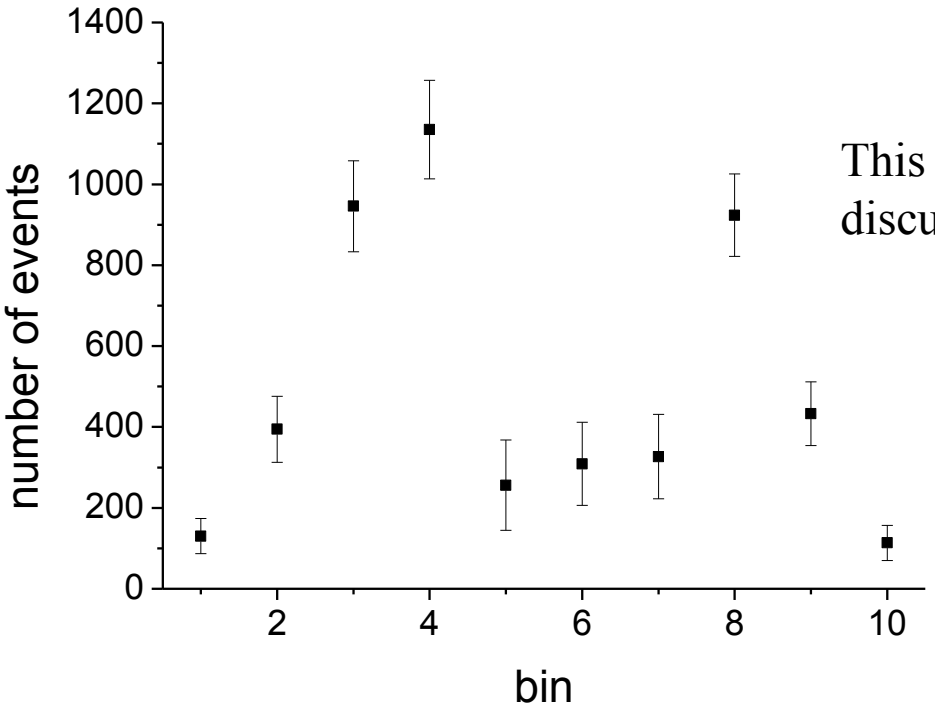
smeared distribution

original distribution

result of unfolding (MLE)

Increasing the statistics would add further delta function peaks

Unfolding without regularization, the errors are strongly correlated.
The error matrix has to be provided.



This is from an example to be discussed in more detail below

Many unfolding issues outside physics are directed towards the point estimate (unblurring pictures) and the interval estimation is of minor importance. In particle physics it is the other way round.

Essential requirements

1. All measurements have to be presented in such a way that they can be combined with other measurements. They have to be associated with error limits that permit to check the validity of theoretical predictions. This applies also to data that are distorted by measurement errors. → Publish data without regularization.

2. In addition we also need a graphical presentation. → Regularization

Three alternative proposals for requirement 1:

- Unfold result without regularization and publish with full error matrix.
- Determine eigenfunctions of the LS matrix. Publish these functions together with the uncorrelated fitted weights and corresponding errors.
- Publish observed data with their uncertainties together with the smearing matrix.

The first approach works only with reasonably large statistics (normally distributed errors of the unfolded bin entries.)
The third method leaves all the work to the user of the data.

Problem 2, regularization:

There is no general rule, how to smooth the data, but the result has to be compatible with the data. (Clearly, some methods make more sense than others.) The choice of the regularization method is subjective (in priciple Bayesian).

Smoothing is always related to a loss of information. Thus unfolded distributions with regularization are not well suited to discriminate between theories (Cousins‘ requirement cannot be fulfilled except in special situations.)

Regularization introduces constraints and for that reason reduces the nominal values of the errors. There is nothing strange about obtaining error < SQRT (number of events).

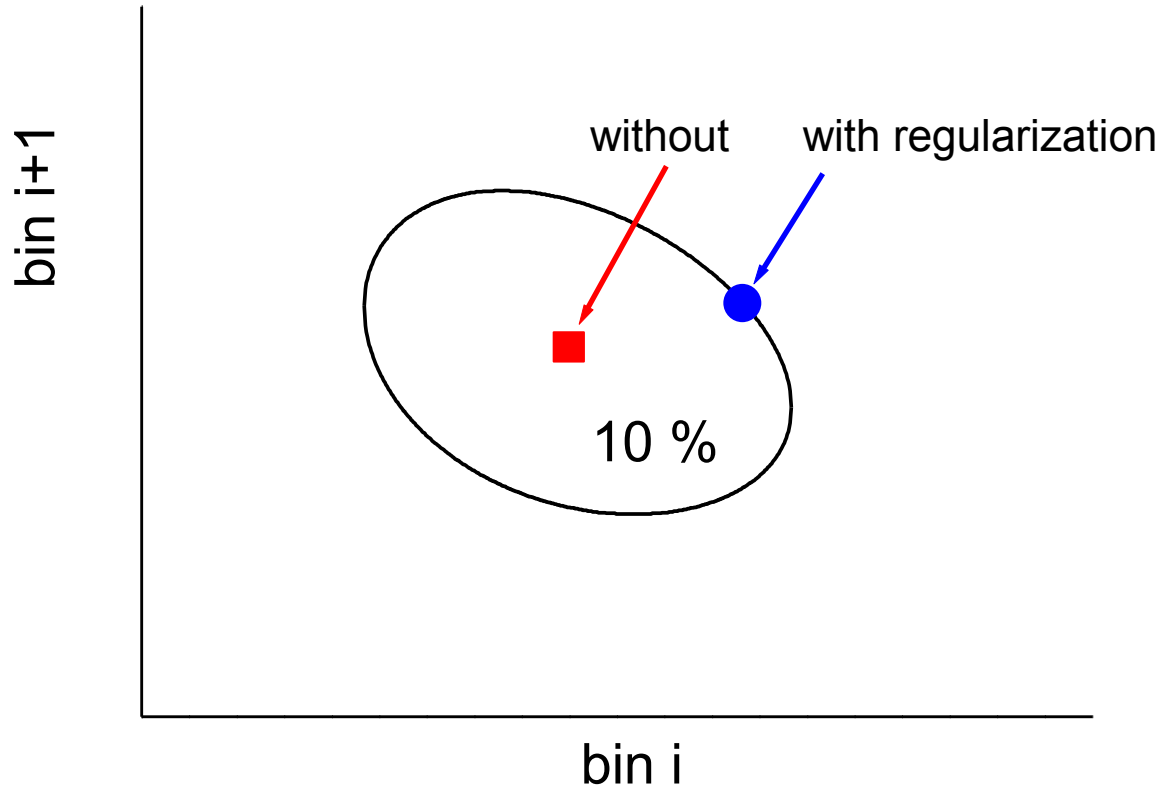
How to fix the regularization strength?

Proposal:

The regularization strength should be fixed by a commonly accepted p-value based on a goodness-of-fit statistic (normally χ^2)

The smoothed result has to be compatible with the observed data and this means to the fit without regularization. The fit without regularization is the measurement.

schematic illustration with 2 true bins, p-value 90 %.



$$\Delta\chi^2 = \chi^2_{reg} - \chi^2_{stat}$$

χ^2_{stat} = χ^2 without regularization

$$p = \int_{\Delta\chi^2}^{\infty} u_N(x)\,dx$$

p = p-value,

$u_N = \chi^2$ distribution for N degrees of freedom,

N = number of true bins

For an error ellipse corresponding to a 1-p confidence interval centered at the parameters computed without regularization the parameters with regularization is located at the surface.

Important: do not use the nacked χ^2_{reg} value of the observed distribution to compute the p value! It is not relevant. We are not checking the goodness-of-fit but want the regualized solution to be compatible with the standard fit.

Proposal: require p > 90 % (The reg. solution is close to the solution without regularization.)

Some common unfolding and regularization methods

Notation:

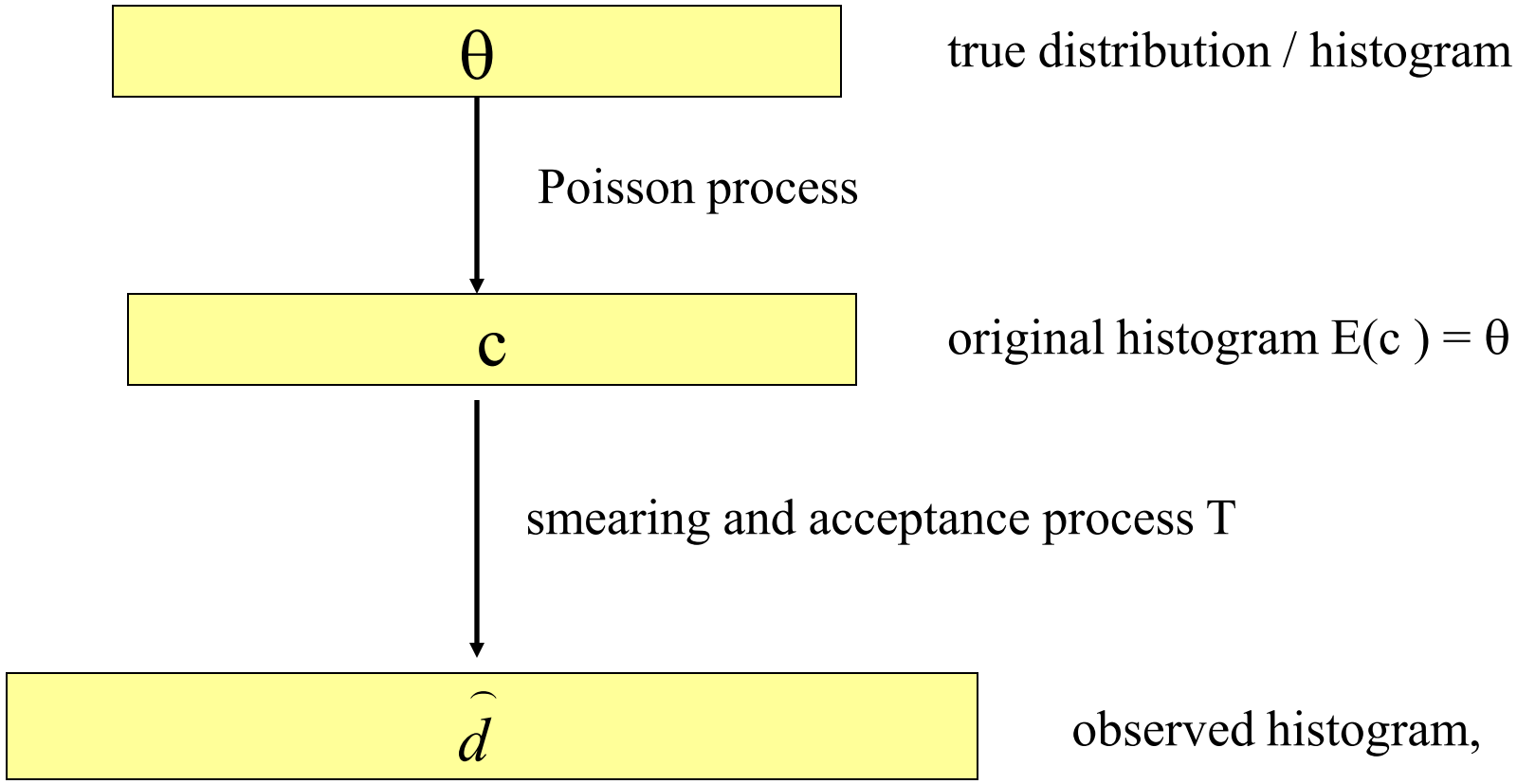
true distribution: underlying theoretical distribution, described by vector θ

original or undistorted histogram: ideal histogram before smearing, and acceptance losses, described by vector c

observed or data histogram: includes experimental effects, described by vector $\hat{d}$ (estimate of the expected value d)

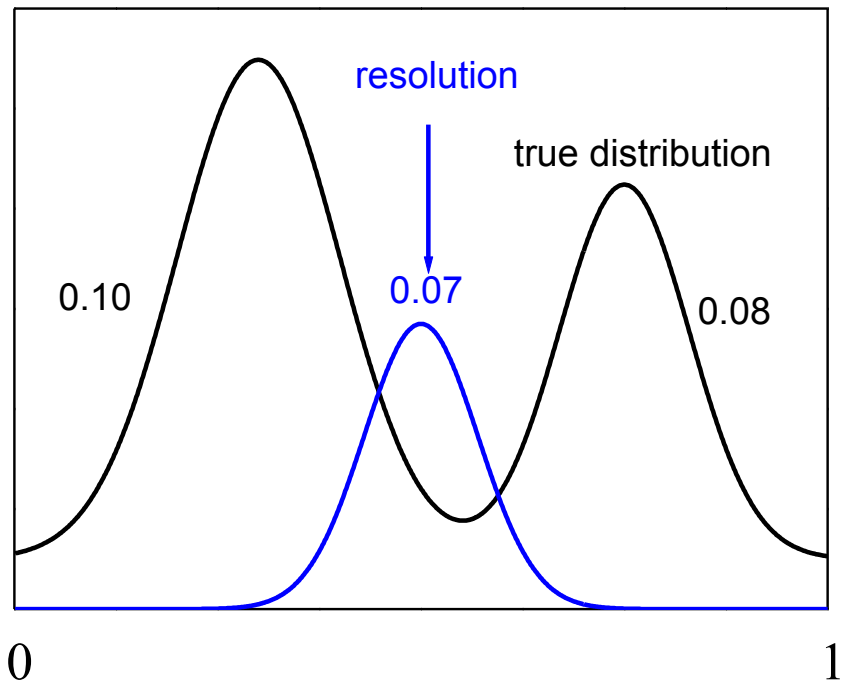
Smearing matrix T: includes acceptance losses

$$d = T\,\theta$$


θ
true distribution / histogram
Poisson process
c
original histogram E(c) = θ
smearing and acceptance process T
$\hat{d}$
observed histogram,

$d = T\theta$

We use the following example:



superposition of :

2500 events N(0.3;0.1)
1500 events N(0.75;0.08)
1000 uniform events

= 5000 events

Gaussian resolution N(0;0.07)

observed distribution: 40 bins
true distribution: 20 or 10 bins

In this example a rather bad resolution was chosen to amplify the problems.

1. χ^2 fit of the true distribution, eliminate insignificant contributions (SVD)

(LSF), minimize χ^2, Gaussian approximation

$$\chi^2 = \sum_k \frac{\left(\hat{d}_k - \sum_j T_{kj}\theta_j \right)^2}{\sum_j T_{kj}\theta_j}$$

linearized approximation (used in matrix LSF, SVD)

$$\chi^2 = \sum_k \frac{\left(\hat{d}_k - \sum_j T_{kj}\theta_j \right)^2}{\hat{d}_k}$$

In matrix notation (see for example Blobel's note):

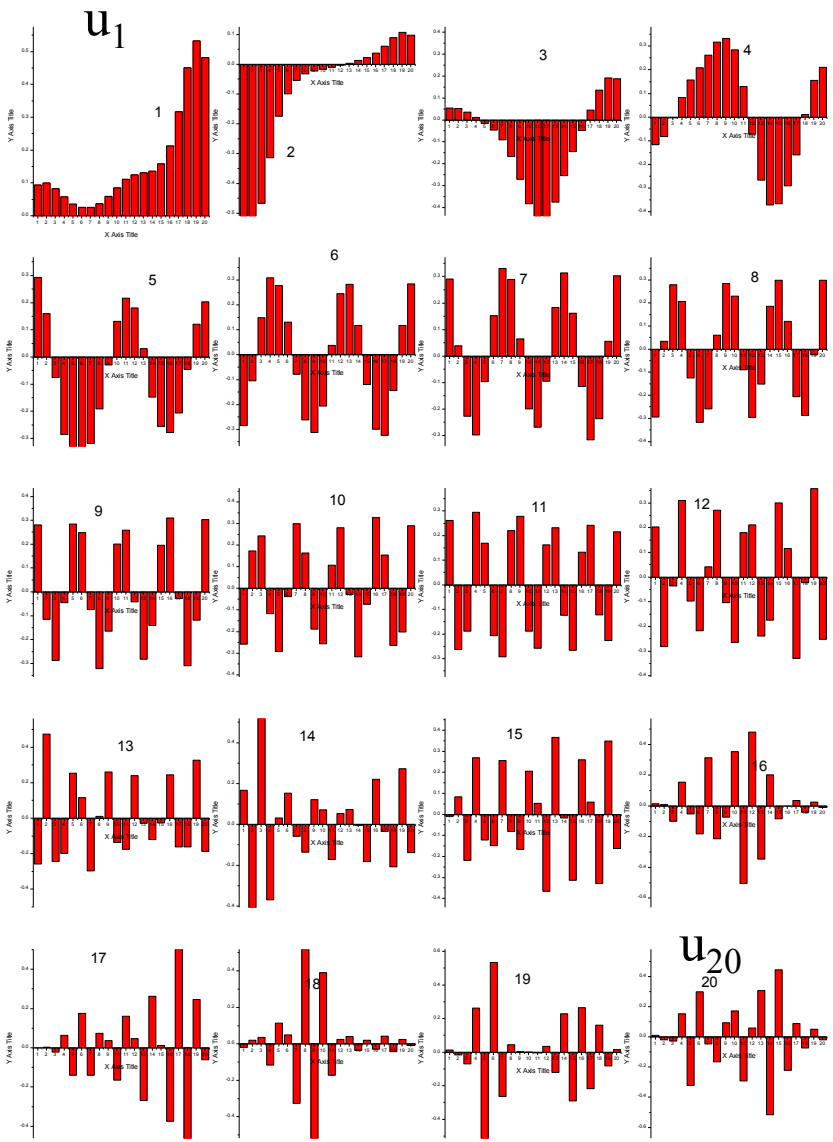
The least square fit has to include the error matrix E of the data vector d. (weight matrix $V=E^{-1}$). T is rectangular, we transform it to a quadratic matrix C:

$$T\theta = d \quad \rightarrow \quad C\theta = b$$

$$T \Rightarrow C = (T^T V)T \qquad \hat{d} \Rightarrow \hat{b} = (T^T V)\hat{d}$$

$$\hat{\theta} = C^{-1}\hat{b}$$

The eigenvectors u_i of C are uncorrelated, $\theta = \Sigma c_i u_i$

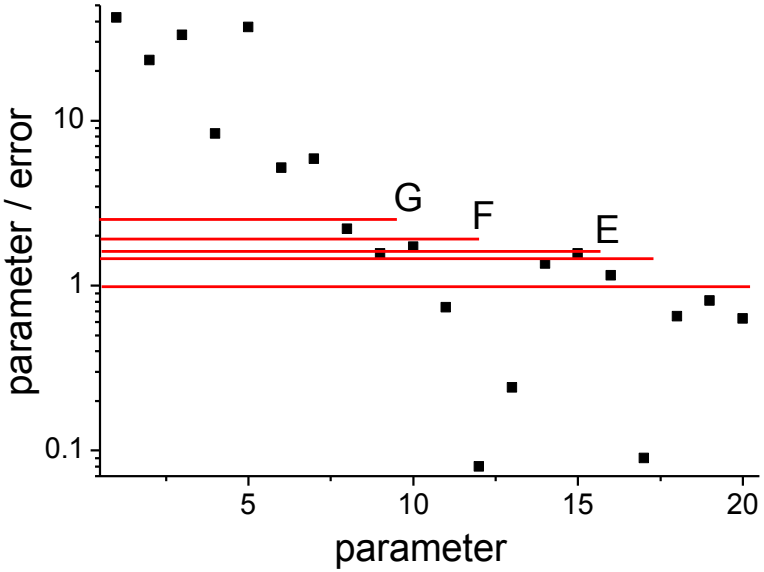


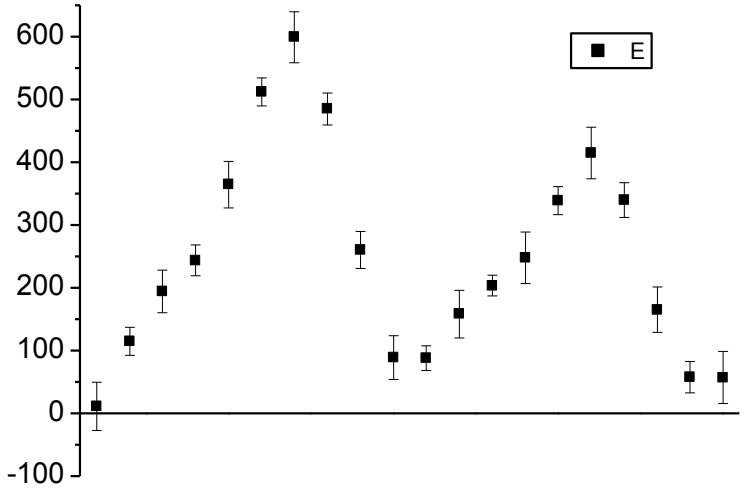
eigenvectors u_i

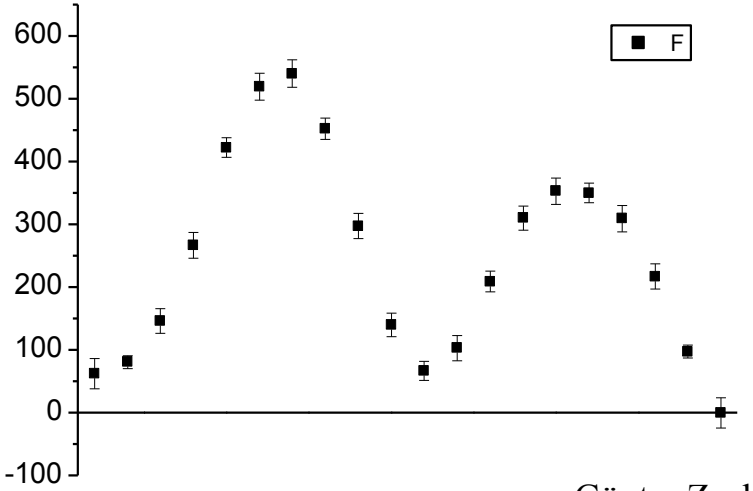
The true distribution is a superposition of these functions, ordered with decreasing eigenvalues. The coefficients (parameters) c_i are fitted. Cut on significance = |parameter/error| = $|c_i/\Delta c_i|$

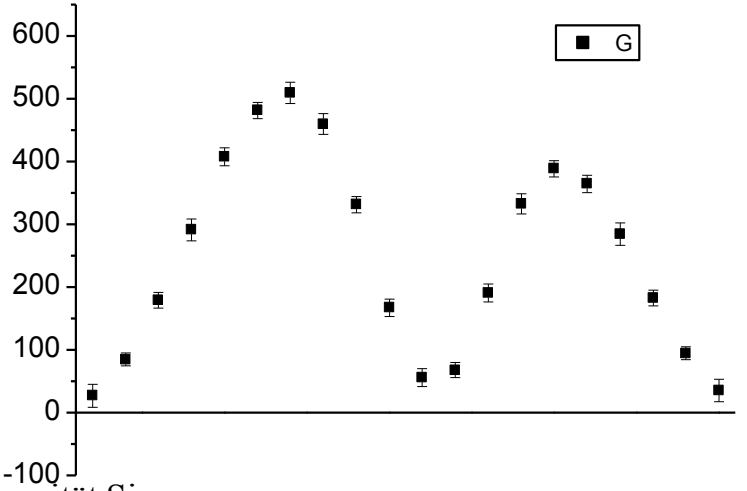
$$\theta = \sum c_i u_i$$

Unfolding result for different cuts in significance
Where should we cut?







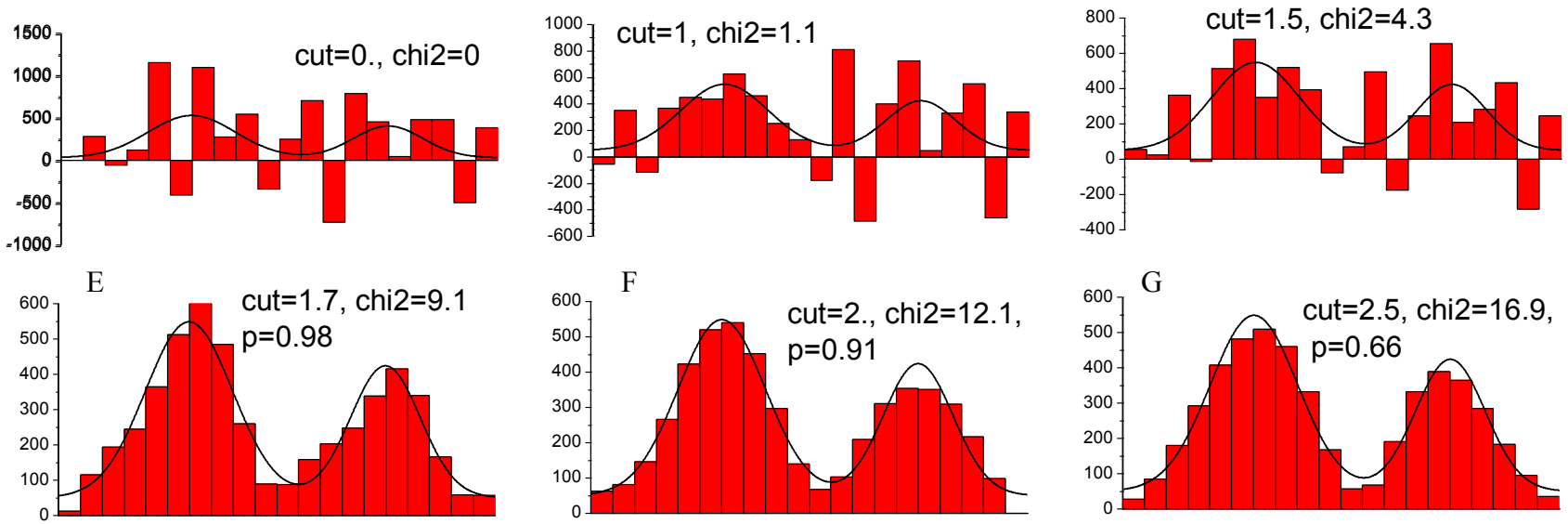


Remarks:

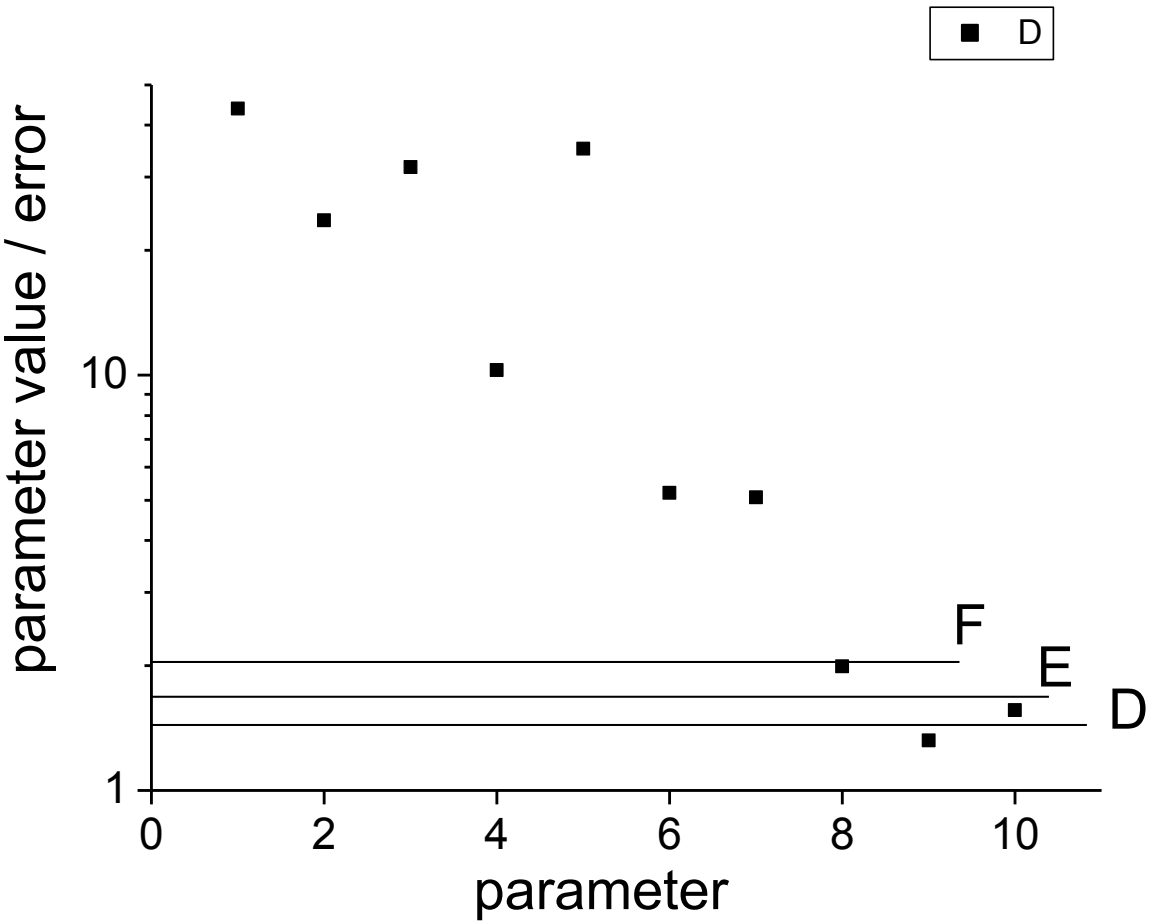
• small eigenvalues are not always correlated with insignificant parameters

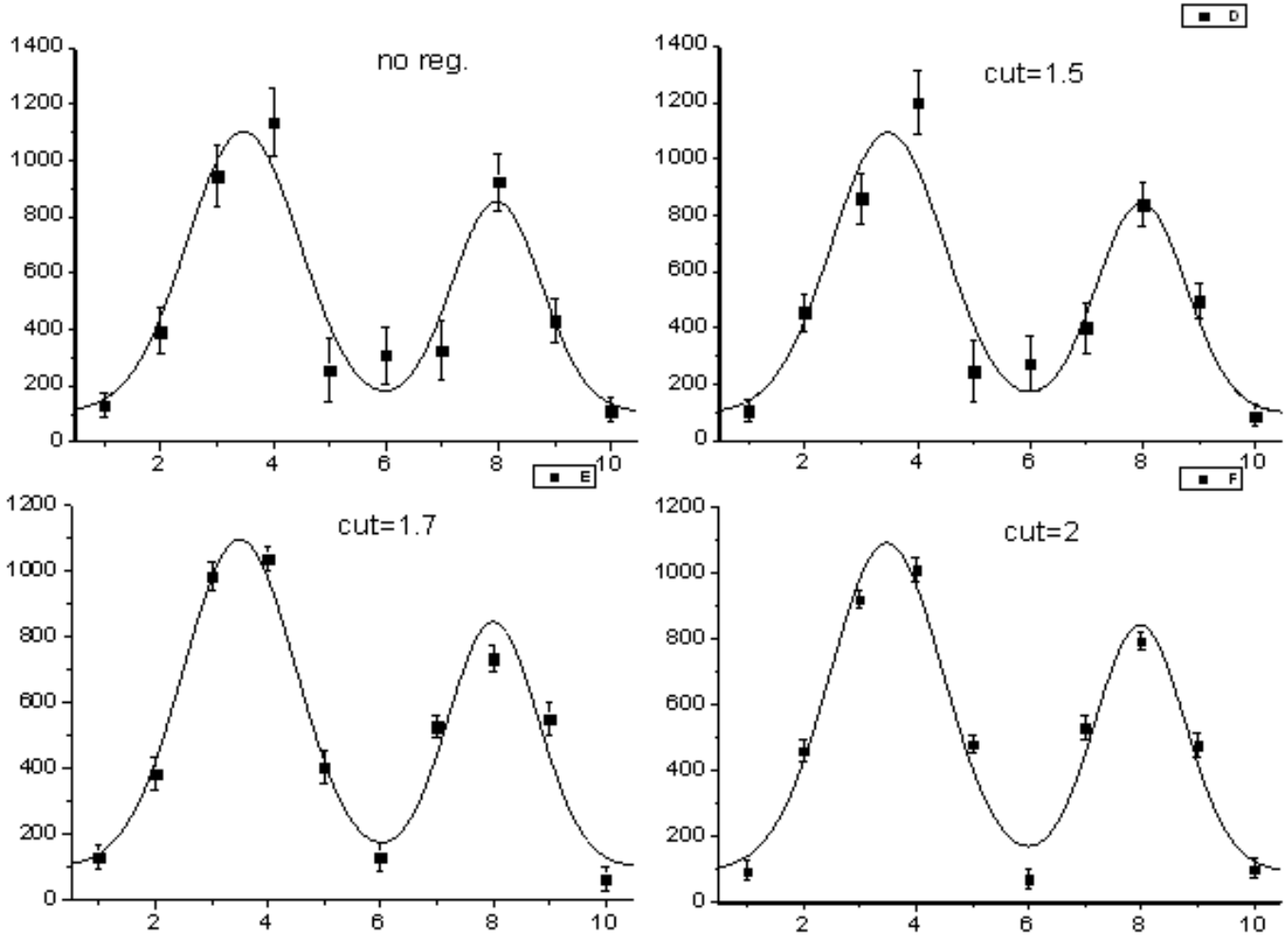
• cutting at 1 st. dev. produces unacceptable results, better cut at p=0.9

• the errors decrease with increased regularization strength. In graphs F and G some errors are smaller than SQRT(number of events), symmetric errors are not adequate.



same thing for 10 bins





With 10 bins, regularization is not necessary.

2. Poisson likelihood fit with penalty regularization

for a single bin i with expectation d_i:

$$\ln L_i = \hat{d}_i \ln d_i - d_i = \hat{d}_i \ln \sum_j T_{ij}\theta_j - \sum_j T_{ij}\theta_j$$

for the histogram with penalty term R:

$$\ln L = \sum_k \left[\hat{d}_k \ln \sum_j T_{kj}\theta_j - \sum_j T_{kj}\theta_j \right] - R$$

Frequently used penalty term

$$R = r \sum_{i=2}^{N-1} (2\theta_i - \theta_{i-1} - \theta_{i+1})^2$$

Standard: curvature regularization → prefers a linear distribution.

$$R_{regu} = r(2\vartheta_i - \vartheta_{i-1} - \vartheta_{i+1})^2$$

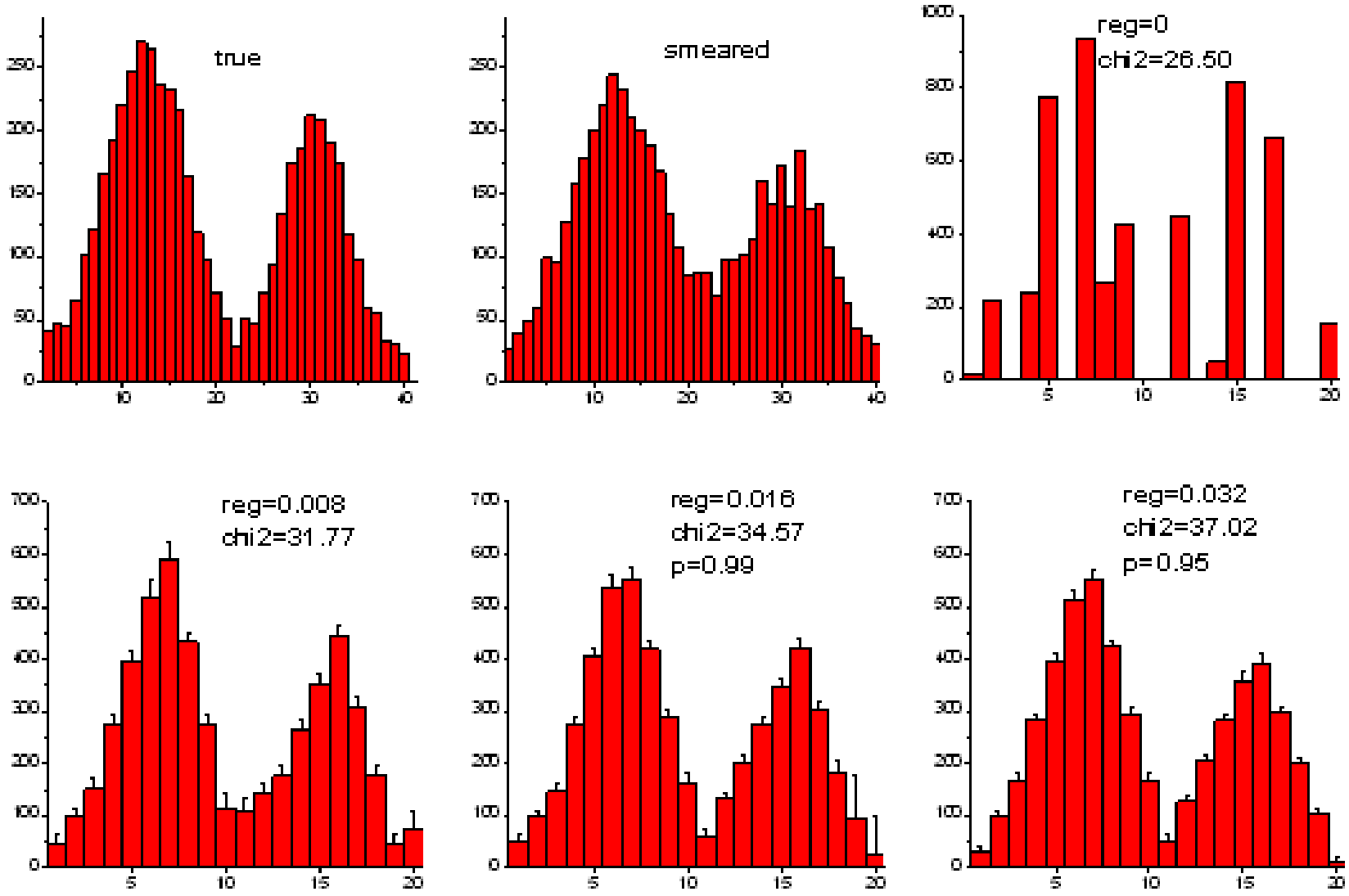
r: regularisation strength

You can use a different penalty term and apply for instance the regularization to the deviation of the fitted from the expected shape of the distribution (iterate).

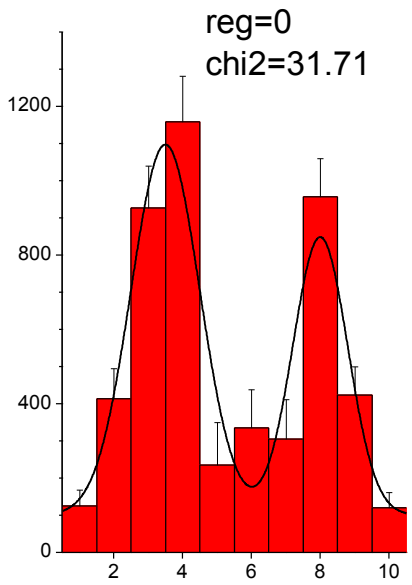
For a nearly exponential distribution:

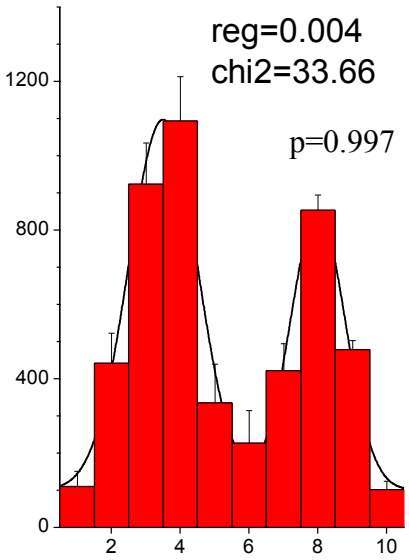
$$R_{regu} = r(2\ln\vartheta_i - \ln\vartheta_{i-1} - \ln\vartheta_{i+1})$$

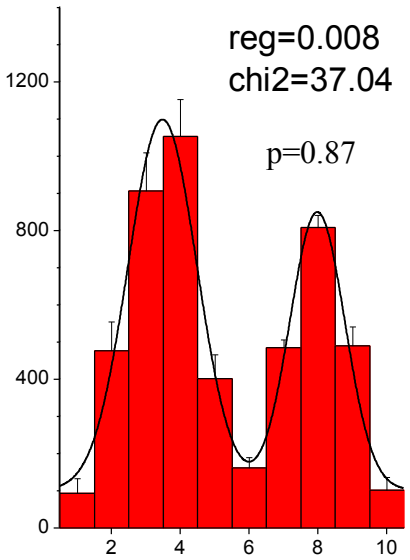
We could also normalize the expressions to the expected uncertainty.

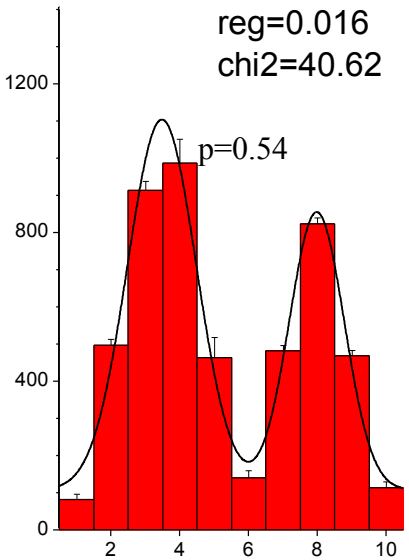

true
smeared
reg=0
chi2=26.50
reg=0.008
chi2=31.77
reg=0.016
chi2=34.57
p=0.99
reg=0.032
chi2=37.02
p=0.95

For 10 true bins









c) Iterative LSF

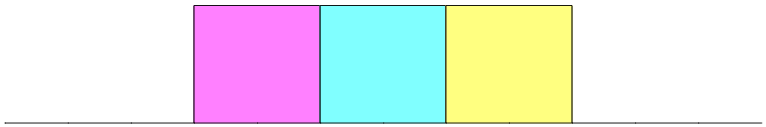
(Introduced to HEP by Mültai and Schorr 1986, see refs.)

Inverts the relation $d = T\theta$ iteratively (iterative matrix inversion)
This is equivalent to the LSF with diagonal and equal errors for the data vector.

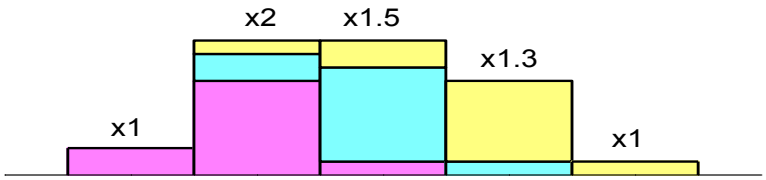
(However, in principle, the the error matrix could be implemented without major difficulty)

How does it work?

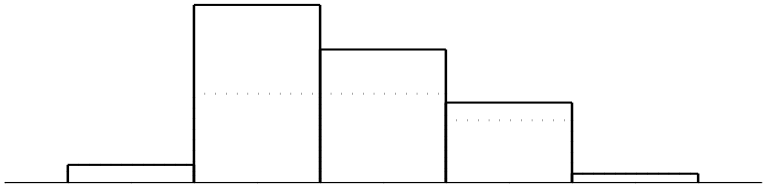
a) starting true distribution



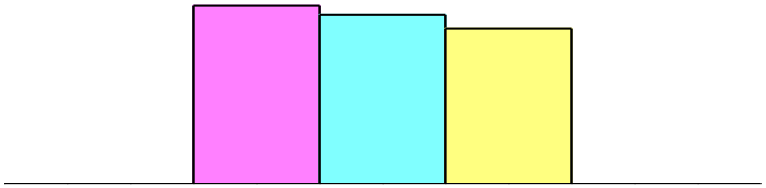
b) corresponding folded distribution.



c) observed data distribution



d) 1. Iteration



Multiply each contribution by a factor to get agreement with the observed distribution and put it back into the true distribution.

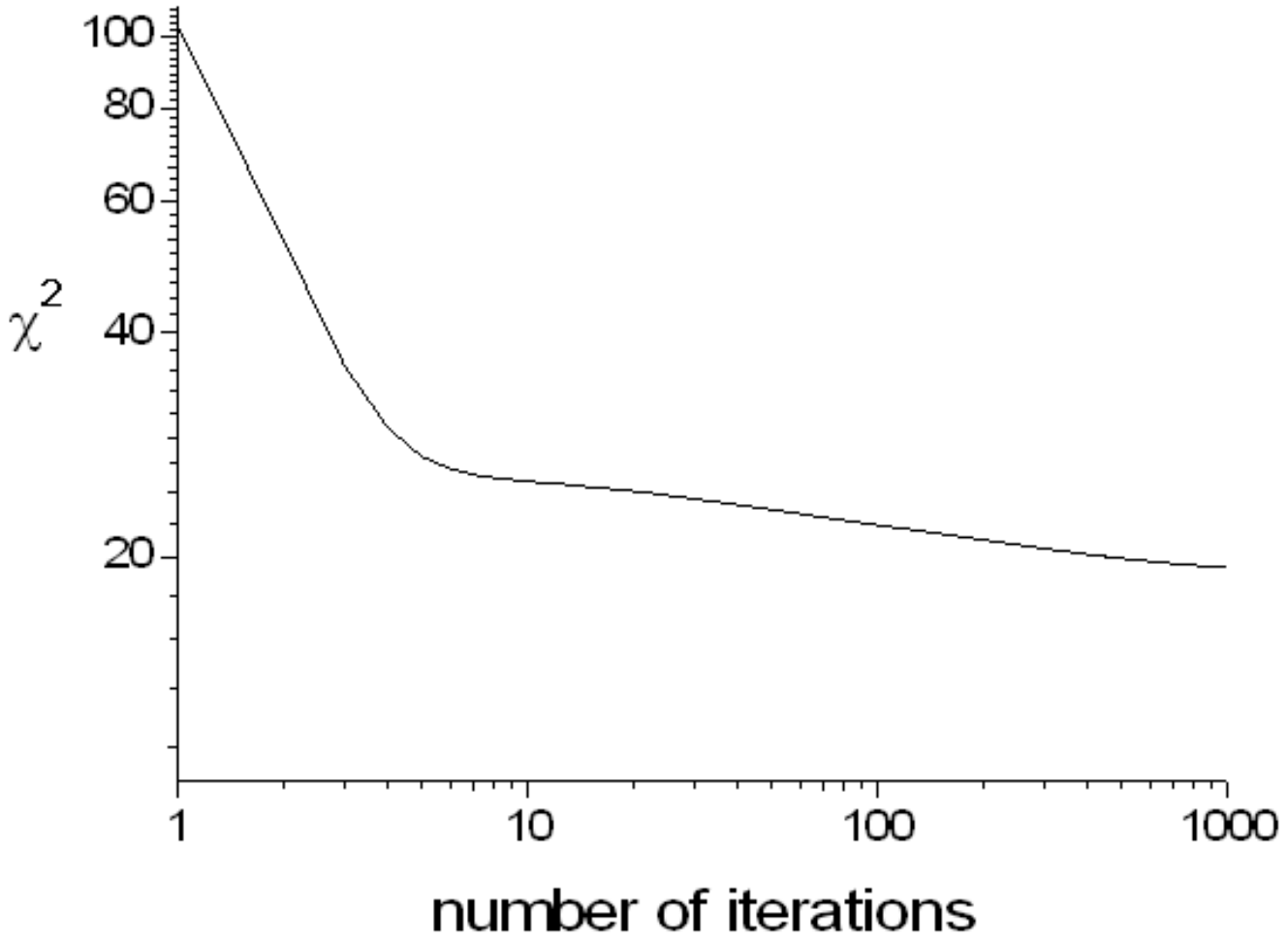
Folding step:

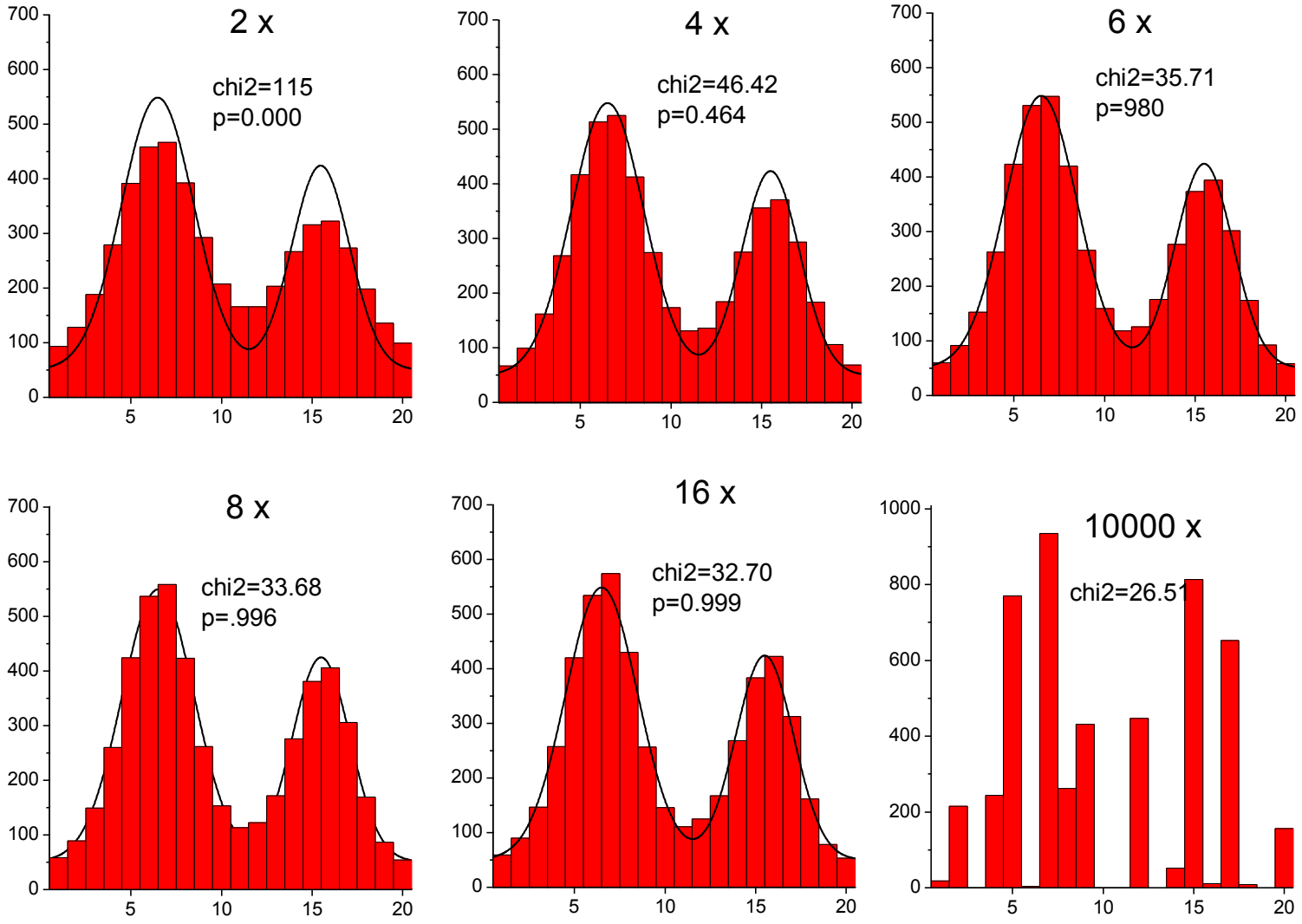
$$d_j^{(k)} = \sum_i T_{ji}\,\theta_i^{(k)}$$

Unfolding step:

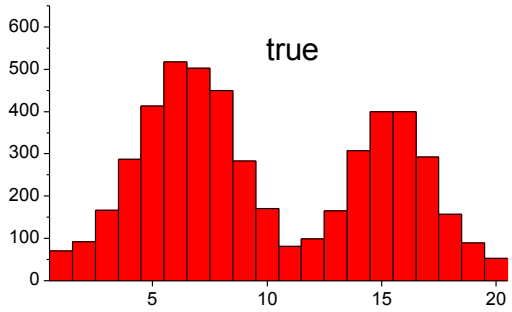
$$\theta_i^{(k+1)} = \sum_j T_{ji}\,\theta_i^{(k)}\,\frac{\hat{d}_j}{d_j^{(k)}}\,\frac{1}{\varepsilon_i}$$

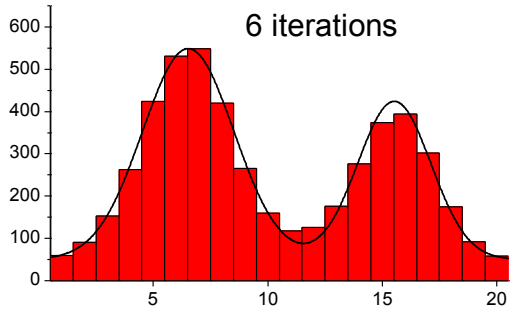
ε_i: efficiency of bin i

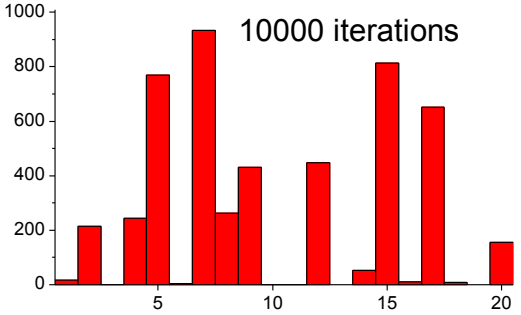

100
80
60
40
20
χ^2
1
10
100
1000
number of iterations

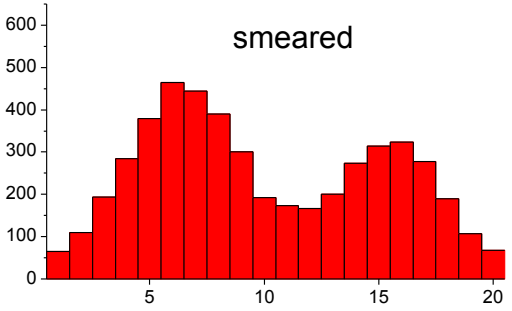

2 x
chi2=115
p=0.000
4 x
chi2=46.42
p=0.464
6 x
chi2=35.71
p=980
8 x
chi2=33.68
p=.996
16 x
chi2=32.70
p=0.999
10000 x
chi2=26.51

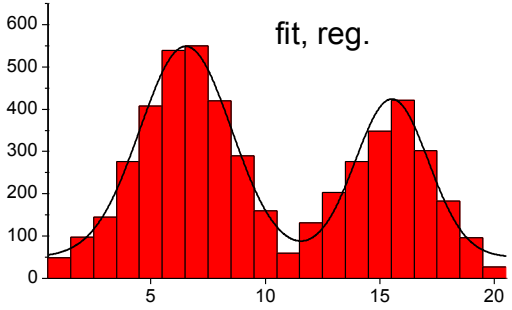
Comparison of Poisson likelihood with curvature regularization to iterative unfolding. The results in the limit of no regularization are nearly identical.

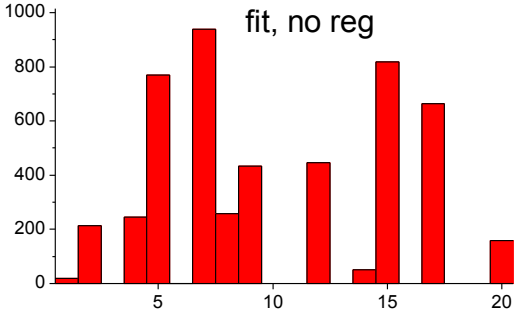




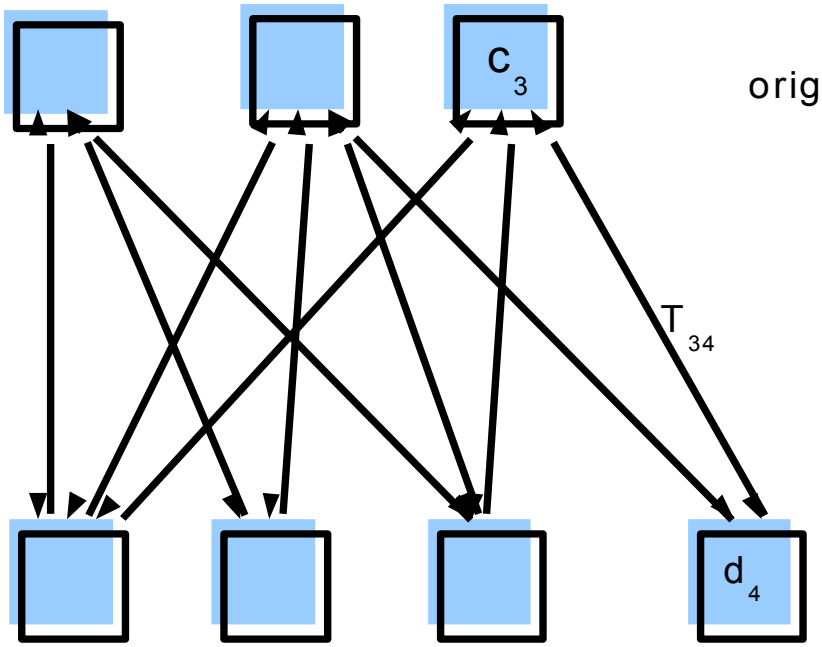








D'Agostini's Bayesian unfolding (arXiv:1010.0632v1)



original histogram (cause bins)

observed histogram (effect bins)

c → **d** (multinomial distribution, probabilities given by T)

d → **c** (using Bayes theorem)

D'Agostini reconstructs the original histogram **c** and not the true distribution. This complicates the problem (multinomial distribution). Being unable to write down a Bayesian formula relating the full histogram c and the observation d he applies a "trick":

- use Bayes' theorem to relate observed d-bins (effect bins) to individual c-bins (cause bins) with a uniform prior for the c-histogram.
- use a polynom fit (or equivalent) to smooth the c-distribution (details do not matter)
- use updated c-distribution as prior for next iteration.

By chance this leads to the same iterative process as that introduced by Múltai et al. described above (except for the intermediate smoothing)

Due to the smoothing step, the procedure converges to a smooth distribution. Regularization is inherent in the method.

The error calculation then includes a Poisson error for the d-bins in combination with the multinomial error related to the smearing matrix.

Further ingredients: (The new vesion is consistently Bayesian)

• Prior densities are used for the parameters of the multinomial and the Poisson distributions.

• The uncertainties of the smearing matrix are taken into account by sampling.

My personal opinion:

• There is no need to use the multinomial distribution. Starting from the true distribution, it is possible to introduce a prior in the standard way, but this doesn‘t produce a smooth result. The smoothing is due to the “trick“.

• The intermediate smoothing and the related convergeance are somewhat arbitrary (no p-value criterion).

• Introducing prior densities for the secondary parameters is unlikeliy to improve the result.

• I am not convinced that the error calculation is correct. (Giulio is checking this).

• I cannot see an essential advantage in D‘Agostini‘s program compared to the simple iterative method as introduced by Múltai et al. with the aditional χ^2 stopping criterion.

Comparison of the three unfolding approaches:

All three methods produce sensible results

Suppression of "insignificant" contributions in the LS approach:

- mathematically attractive, provides the possibility to document the result without regularization
- requires high statistics (normal error distribution), problematic in cases where the true distribution has bins with low number of events → negative entries.
- the cut of insignificant contributions is not straight forward.

Poisson likelihood fit with penalty term

- works also with low statistics
- very transparent
- easy to fix the regularization strength
- a curvature penalty works well, but not obvious why to choose. The penalty term can be adjusted to specific problem
- smoothing even when the data are not smeared! (but this is ok)

Iterative unfolding a la Múltai

- technically very simple.
- does not require large statistics
- offers a sensible way to implement the smoothing prejudices (but a uniform starting distribution always seems to work)
- it is not clear why it works so well!

Some further topics

Binning:

The choice of the bin with b of the true distribution depends on

1. the number of events N
2. The structure of the distribution s (1/s is typical spatial frequency)
3. The resolution σ

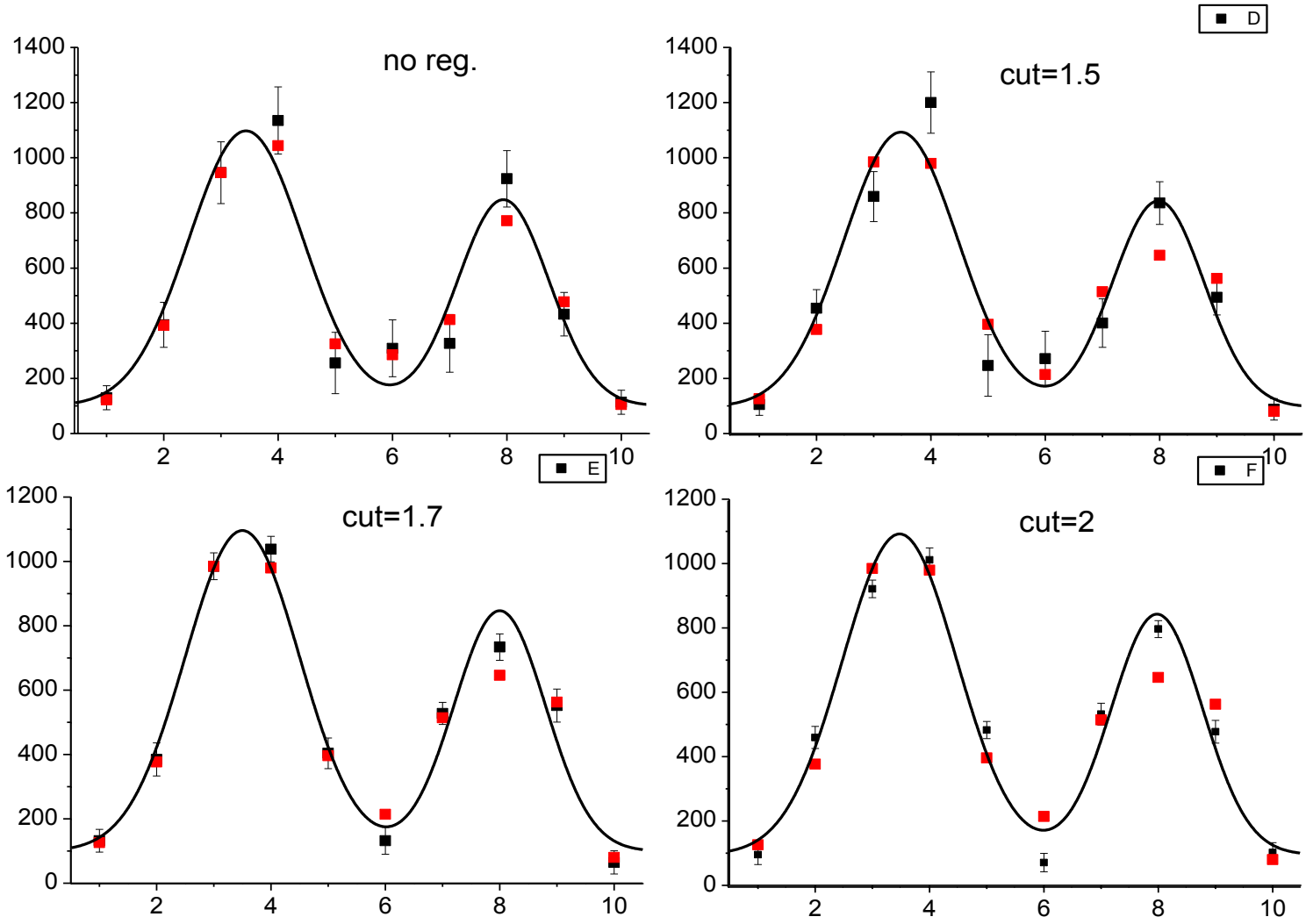
We want narrow bins to reconstruct narrow structures of the true distribution

We want wide bins to have little correlation between neighboring bins. (Wide bins suppress high frequency oscillations)

Try to fulfil: $\sigma < b < s$ (but sometimes $\sigma > s$)

For large N we can somewhat relax this requirement, but even with infinite statistics unfolding remains a problematic procedure.

Effect of the shape of the Monte Carlo distribution, flat (black) vs. exact (red)



Günter Zech, Universität Siegen
CERN, January 2011

Statistical uncertainty in the smearing matrix and background contributions.

These are purely technical problems. For the statistical uncertainty in the smearing matrix see Blobel‘s report.

A simple but computationally more involved method to determine the effects of background and smearing errors is bootstrapping:

Take the Monte Carlo event sample (N events) used to determine the smearing matrix. Contruct a bootstrap samply by drawing N events from the sample with replacement. (Some events occur more than once.) . The spread of M unfolding results from M bootstrap smearing matrices indicates the uncertainty. (This is also used by D‘Agostini)

Background can normally be treated in a simpler way:

Either add one bin in the true distribution (background bin) and a further column in the smearing matrix which gives the probability to find background in a certain observed bin. (iterative method)

or subtract the backgroung in the observed distribution and correct the error matrix of the observed histogram.

or subtract the background and estimate the introduced uncertainty by bootstrap from a background sample.

Whatever you understand, you can simulate, and what you can simulate, you can correct for, using bootstrap if you don‘t find a simpler way.

Graphical representation of the result

To indicate the shape of the true distribution, we need unfolding with regularization. Standard unfolding methods provide reasonable estimates, however there is a problem with the error estimates. The errors derived from the constained fit or computed by error propagation are arbitrary and misleading and therefore useless.

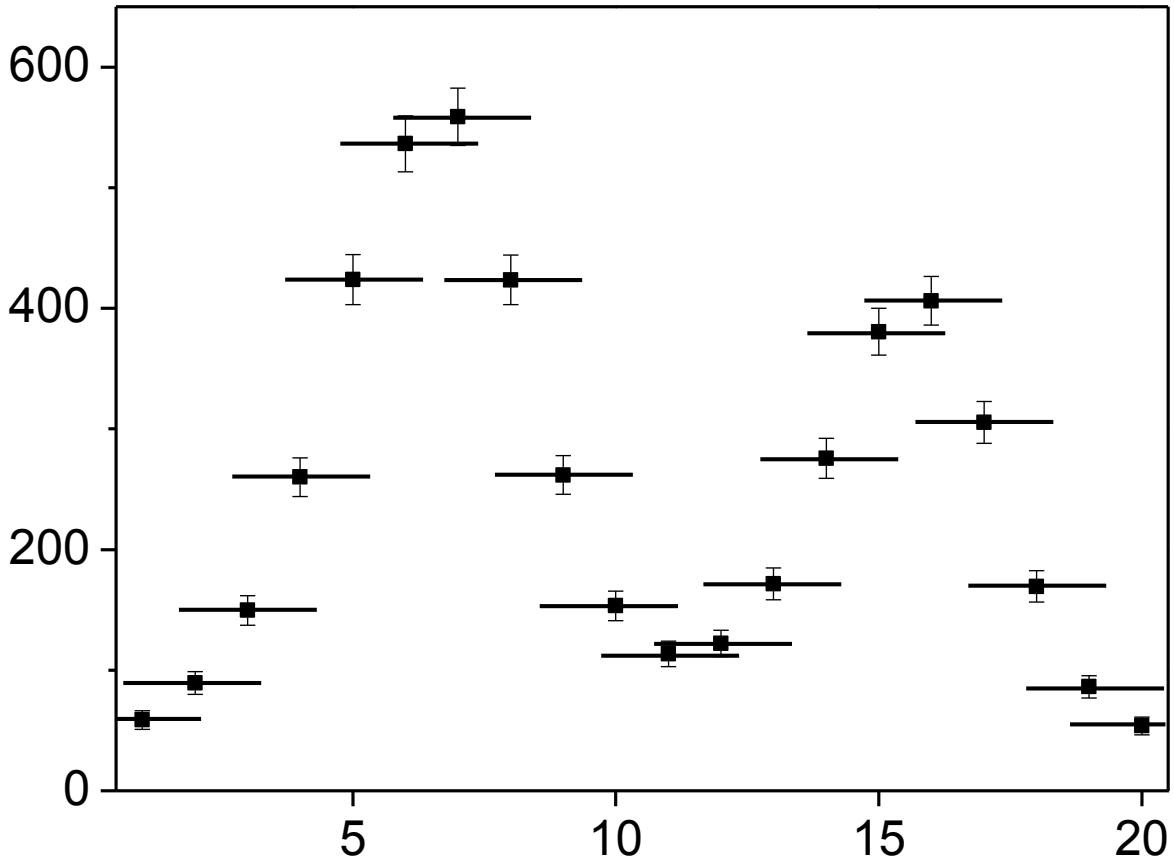
There is no perfect way to indicate the precision of the unfolded result but a better way to illustrate the quality of the point estimate is the following:

- Use as a vertical error bar the statistical error from the effective number of entries in the bin, ignoring correlations.
- Indicate by a horzontal bar the resolution of the experiment at the location of the bin.

This procedure permits to estimate qualitatively whether a certain prediction is compatible with the data and is independent from personal prejudices (regularization method and strength.)

Vertical error: SQRT (effective number of events) This is similar to the nominal errors uf regularized unfolding in the case where the bins are not correlated.

Horizontal bar: resolution.



Summary

The measurement should be published in the standard way: point estimate + error matrix (or equivalent, no regularization) → possibility to combine data and compare with theory.

For a graphical representation we have to apply regularization. (This introduces constraints, subjective input and implies a loss of information.)

The regularization strength should be fixed by a common p-value criterion. This alows to compare different approaches.

The errors produced by a fit with regularization exclude distributions with narrow structures which however are compatible with the data.

Proposal: Choose errors which are independent from the applied regularization. (effective number of entries + resolution)

My preferred unfolding methods are: Poisson likelihood fit + penalty and iterative unfolding with χ^2 stopping criterion.

References, mainly on iterative unfolding

V. Blobel, preliminary report (2020).

Y. Vardi and D. Lee, *From image deblurring to optimal investments: Maximum likelihood solution for positive linear inverse problems (with discussion)*, J. R. Ststist. Soc. B55, 569 (1993).

L. A. Shepp and Y.Vardi, *Maximum likelihood reconstruction for emission tomography*, IEEE trans. Med. Imaging MI-1 (1982) 113.

Y. Vardi, L. A. Shepp and L. Kaufmann,*A statistical model for positron emission tomography*, J. Am. Stat.Assoc. (1985) 8.

A. Kondor, *Method of converging weights - an iterativeprocedure for solving Fredholm's integral equations of the first kind*, Nucl. Instr. and Meth. 216 (1983) 177.

H. N. Mülthei and B. Schorr, *On an iterative method for the unfolding of spectra*, Nucl. Instr. and Meth. A257 (1987) 371.

G. D'Agostini, *A multidimensional unfolding method based on Bayes' theorem*, Nucl. Instr. and Meth. A 362 (1995) 487.

G. Bohm, G. Zech, *Introduction to Statistics and Data Analysis for Physicists*, Verlag Deutsches Elektronen-Synchrotron (2010), http://www-library.desy.de/elbook.html(contains also unbinned unfolding).