Binning-free Unfolding

Günter Zech, Desy May 2010

Observed event → true location

- motivation
- basic idea: likelihood, analytic resolution function
- how to find the minimum
- regularization
- results
- include Monte Carlo resolution
- further steps

Motivation

There are situations where binned unfolding suffers from serious difficulties:

• low statistics (for example 40 events)

• events located on unknown curves or points (astronomy)

• multi-dimensional distributions (structure functions)
(imagine 1000 events, 3 dimensions, 5 bins each
→ 125 bins and in average only 8 events per bin

Advantages:

- apply cuts after unfolding
- define histogram parameters after unfolding
- define histogram variables after unfolding
  (unfold p_x, p_y, plot E)
- consistent histograms of projections

Basic idea

As in parameter fitting, apply single event likelihood

$$\ln L(x_1,..,x_N) = \sum_{i=1}^{N} \ln \sum_{j=1}^{N} f(x_i'|x_j)$$

Notation:

- analytic resolution function $f(x',x_i)$
- True location of point i: x_i (free parameters in the fit. 10000 events, 2 dimensions → 20000 parameters
- Observed location x_i'

(For simplicity written in 1 dimension, but all variables could be vectors)

Minimum search

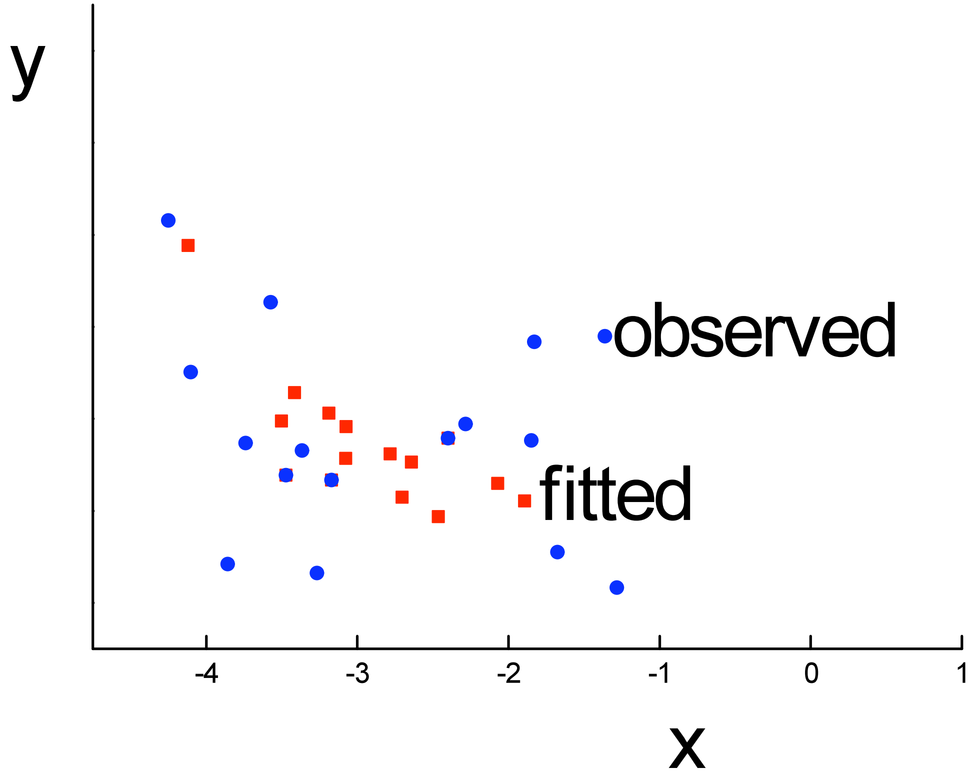
We have a huge number of parameters, but:

- it is easy to select good starting values
- the minima of –lnL are rather shallow
- no dangerous local minima

Minimum search by random migration:

- select randomly a true point move by random step according to a uniform distribution
- accept if the likelihood increases
- repeat until result is satisfying

At the beginning, the true points (red) were sitting on top of the observed data points (blue). They move in such a way that the likelihood increases.



Regularization

Two possibilities, either

1. Stop migration process, or
2. Curvature regularization by probability density estimation using side bands

$$R = r\frac{(2n_C - n_L - n_R)^2}{n_C + n_L + n_R}$$

$$\ln L = \ln L_{stat} - R$$

Correspondingly in higher dimensions

r: regularization constant

n_c: number of events in central region

n_L, n_R: number of events in left and right hand side bands

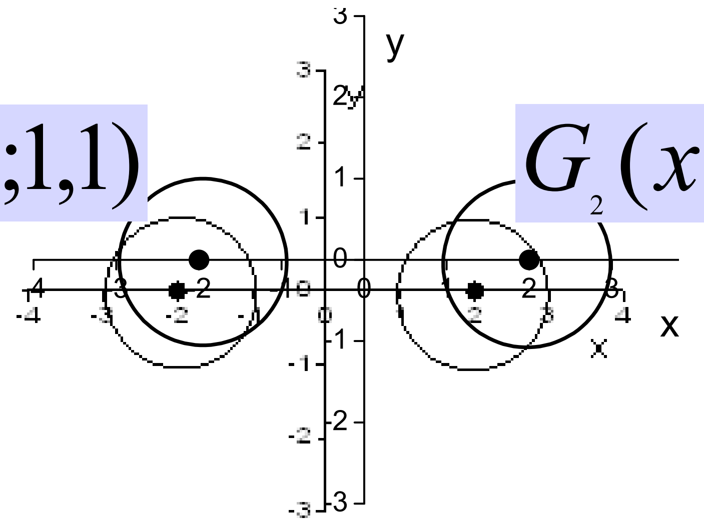
Some Results

Generate 2 Gaussians at x = -2, y = 0 and at x = +2, y = 0 and widths $s_x = 1$, $s_y = 1$ for both

fraction left: 0.6, fraction right: 0.4
1000 events

Gaussian smearing with width s = 1

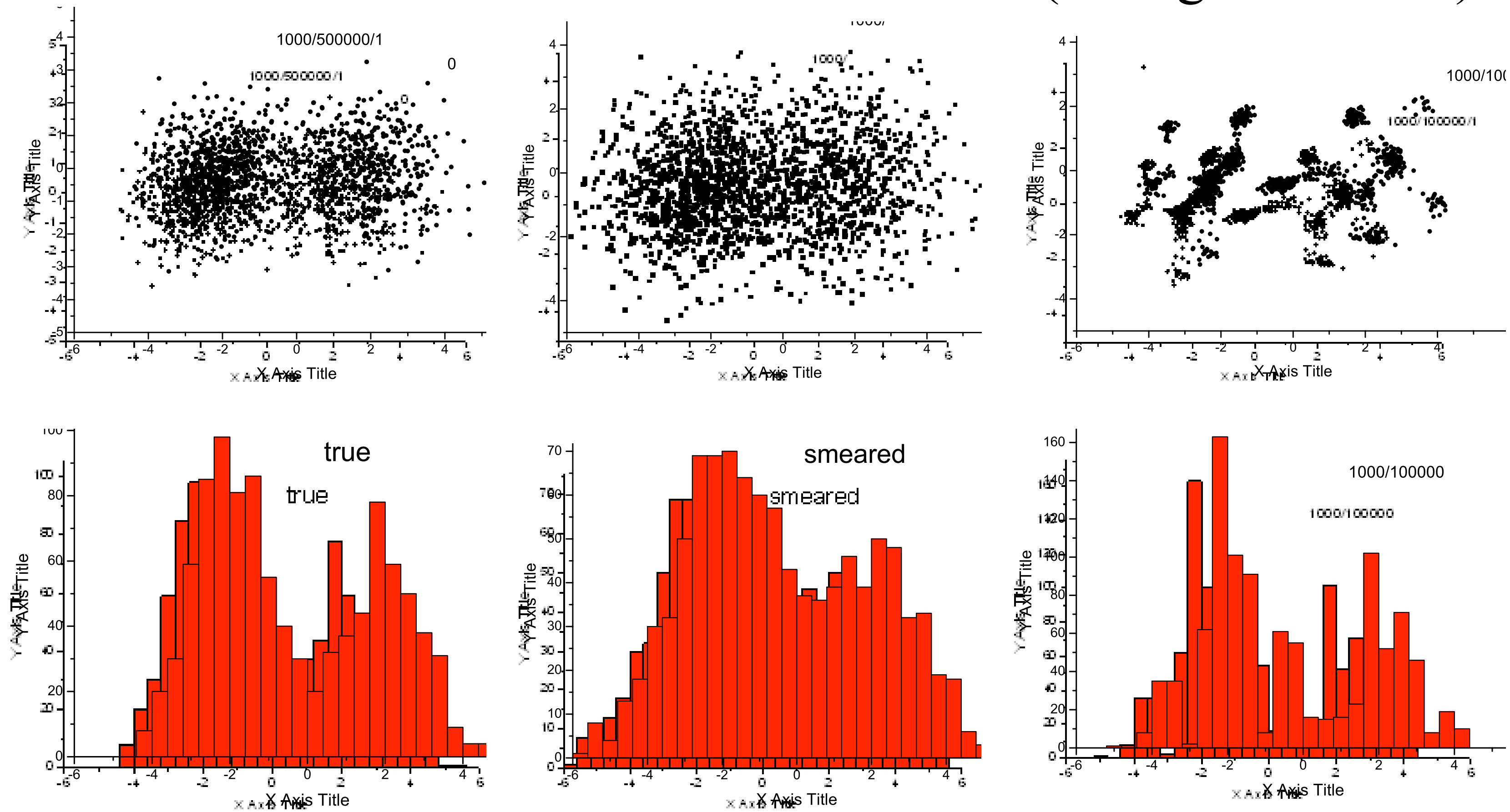
$$G_1(x, y \mid -2,0;1,1)$$

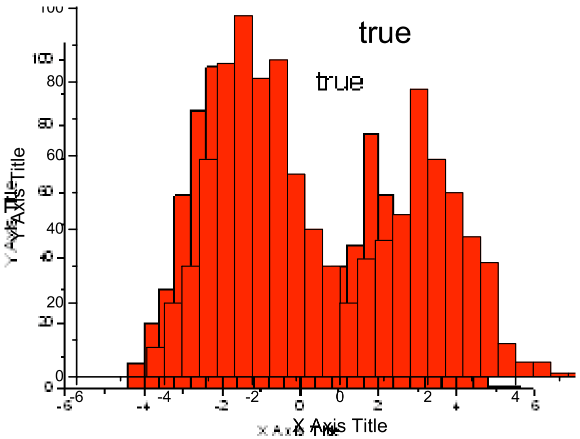


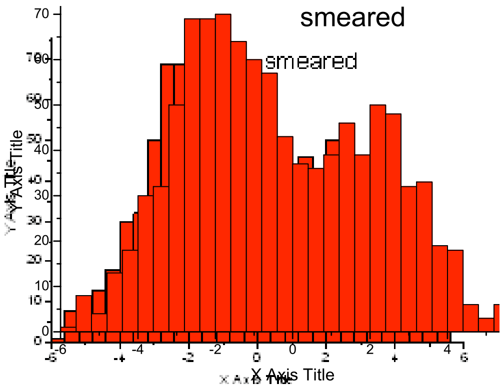
$$G_2(x, y \mid +2,0;1,1)$$

$$f(x', y' \mid x, y) = G(x', y' \mid x, y;1,1)$$

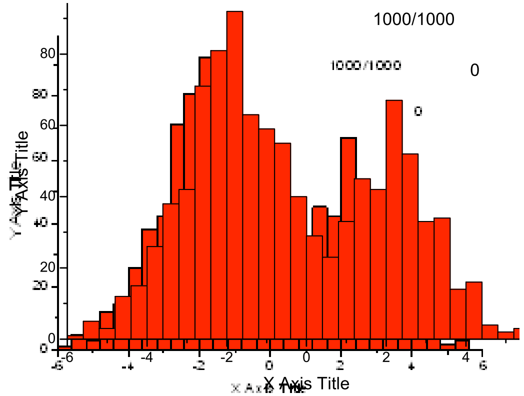
true | smeared | unfolded (no regularization)



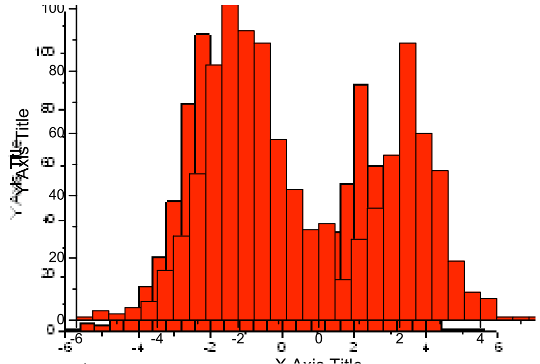




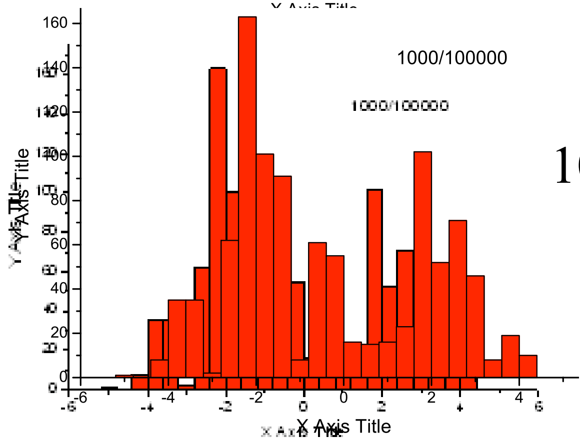
unfolded



1000 moves



2000 moves

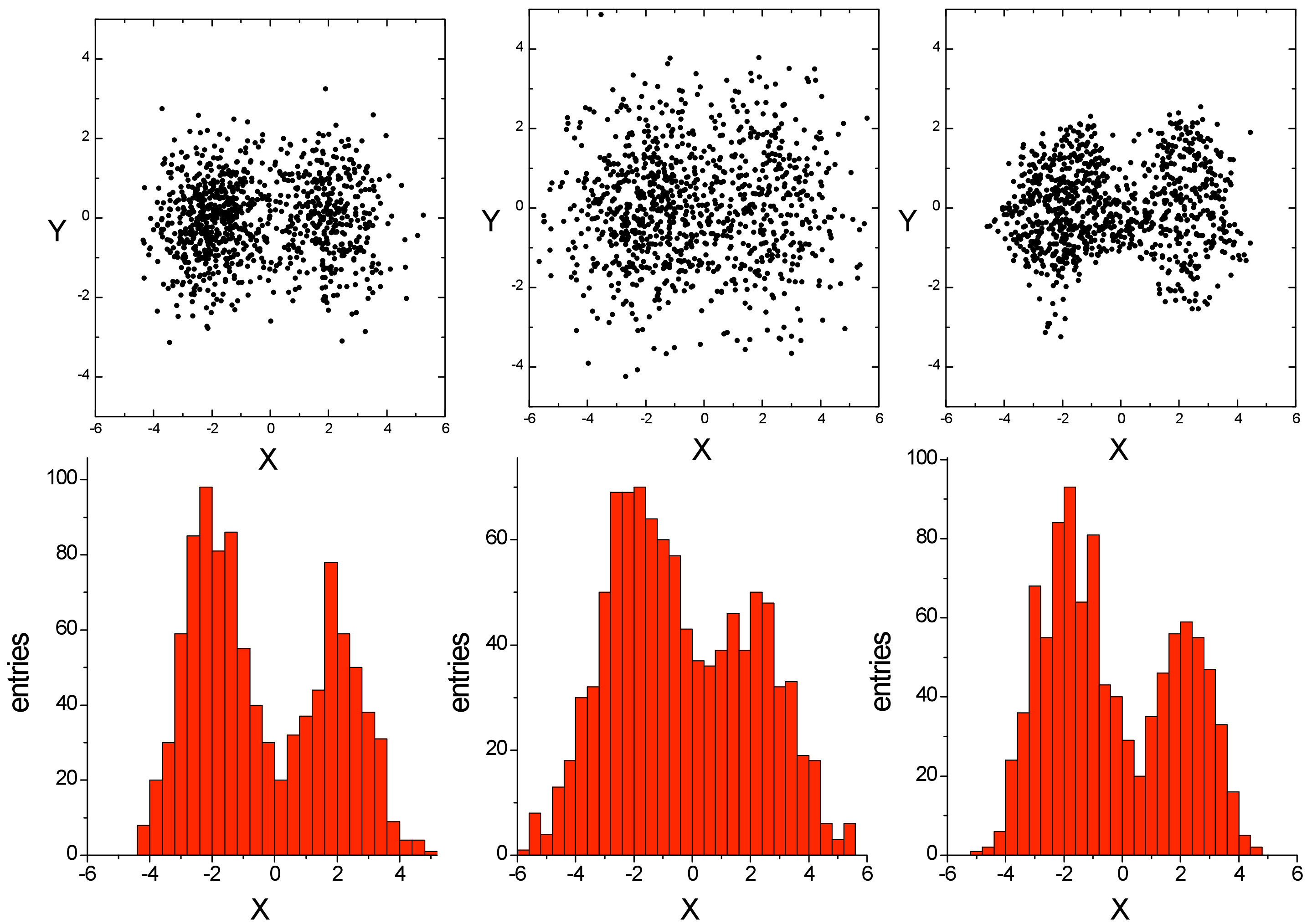


100000 moves

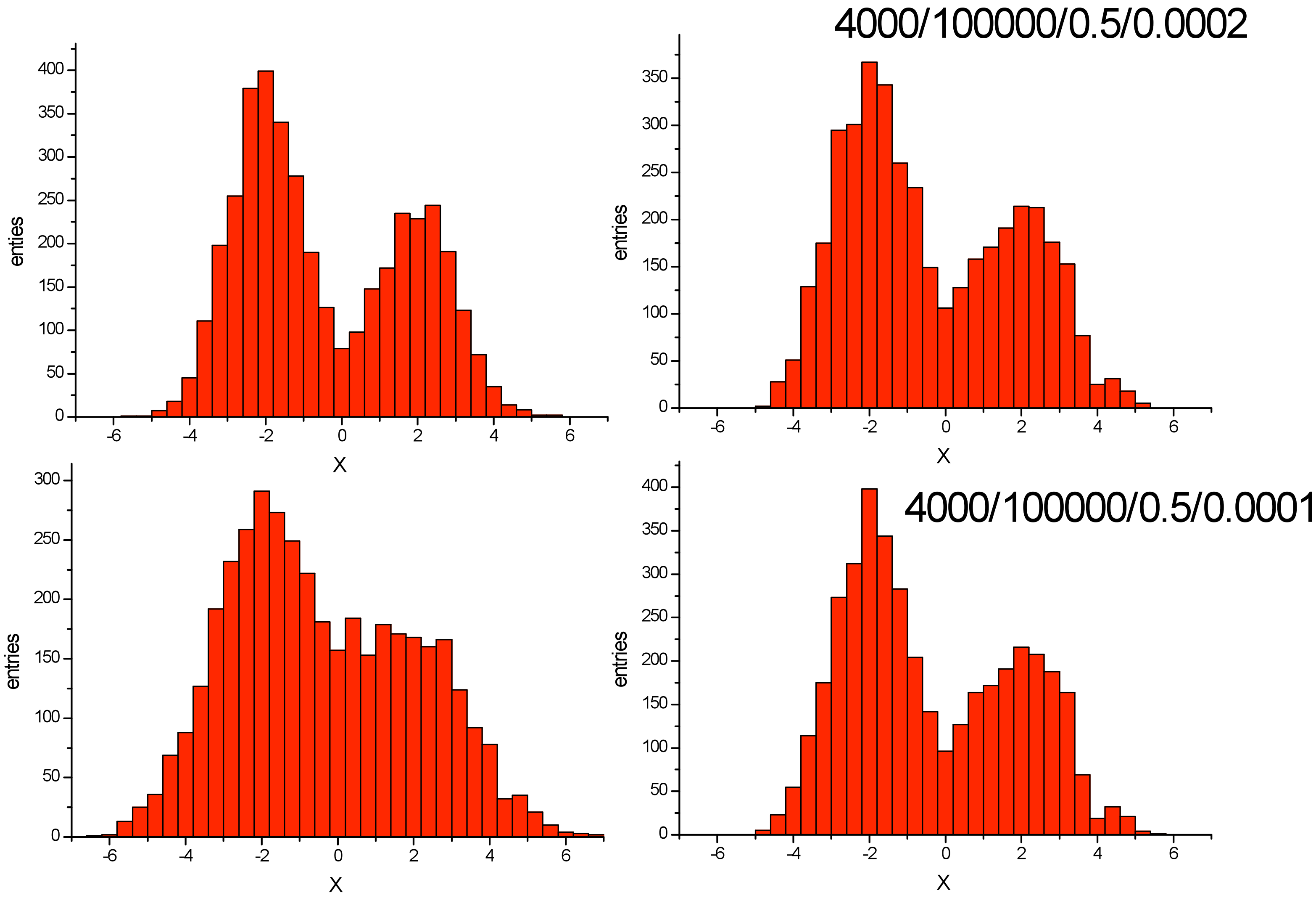
Regularization by limiting the number of moves

Side band regularization

1000/100000/0.0002/1

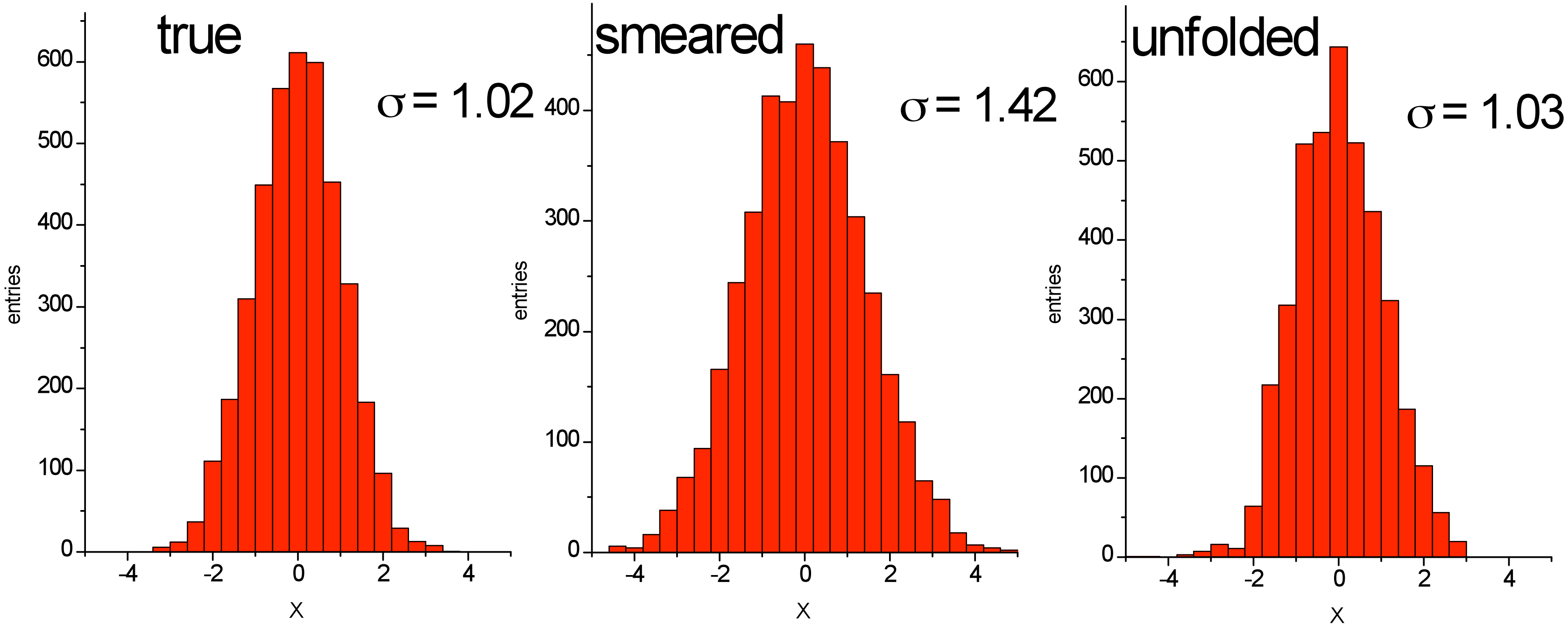


4000 events, side band regularuzation, x projection

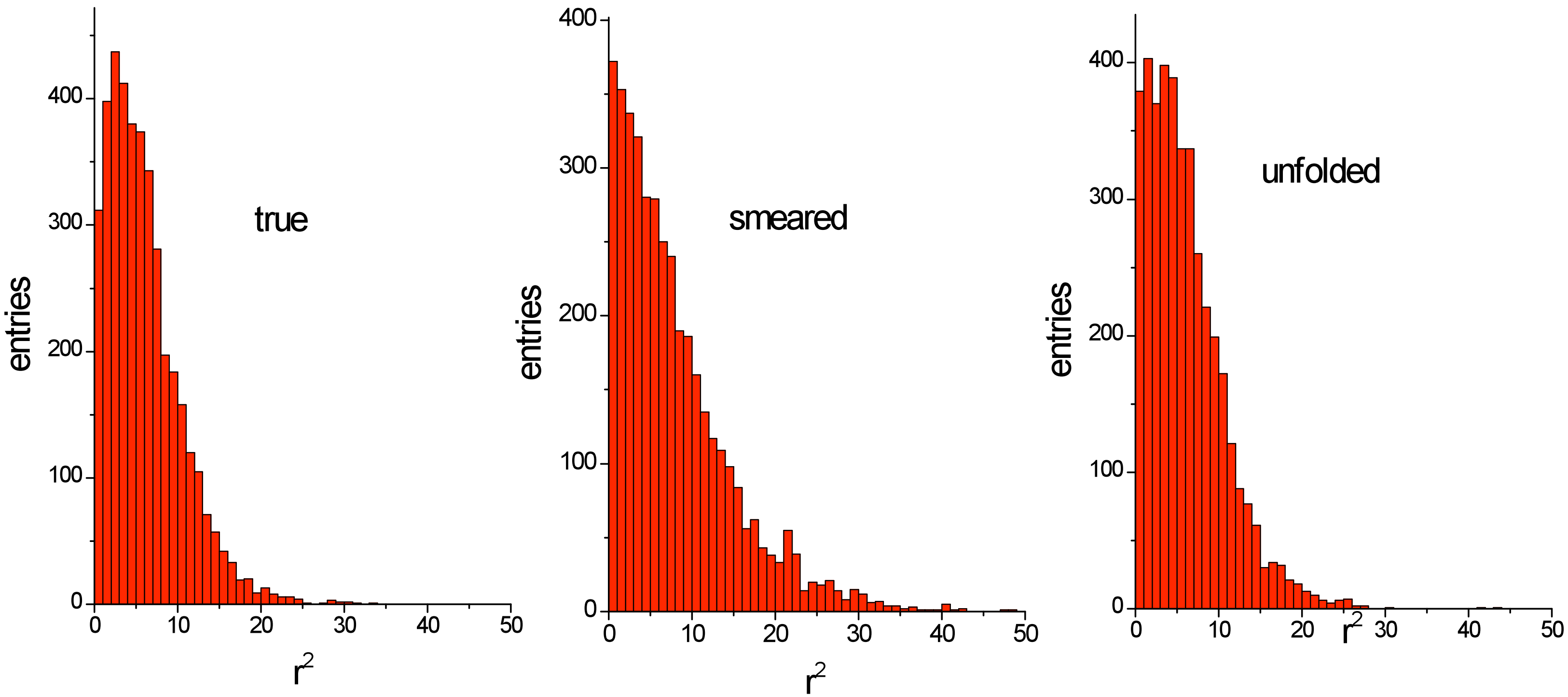


y projection

4000 events, y-projection, cr = 0.0001



Distribution of $r^2 = x^2 + y^2$
(not possible with binning)



$\sigma = 4.5$ $\sigma = 7.0$ $\sigma = 4.7$

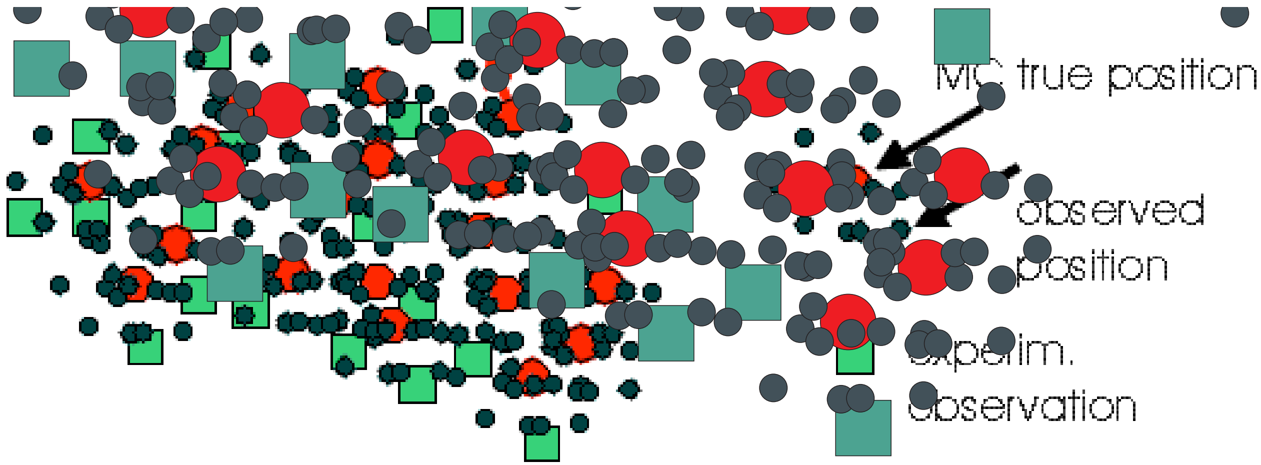
Acceptance and resolution from Monte Carlo

So far we had assumed an analytic resolution function. Normally we know it from a Monte Carlo simulation.

We replace the analytic function by Monte Carlo satellites: Each MC true point is surrounded by *k* observed points (stallites) which are simulated measurements.

We move the true point together with its satellites until the observed points are compatible with the experimental data.

To do so, we need a binning-free goodness-of-fit statistic to measure the aggreement of the simulation with the data: energy test statistic or k nearest neighbor statistic. (see Refs.)



The MC points move until the distribution of the black dots agrees with the distribution of the green boxes

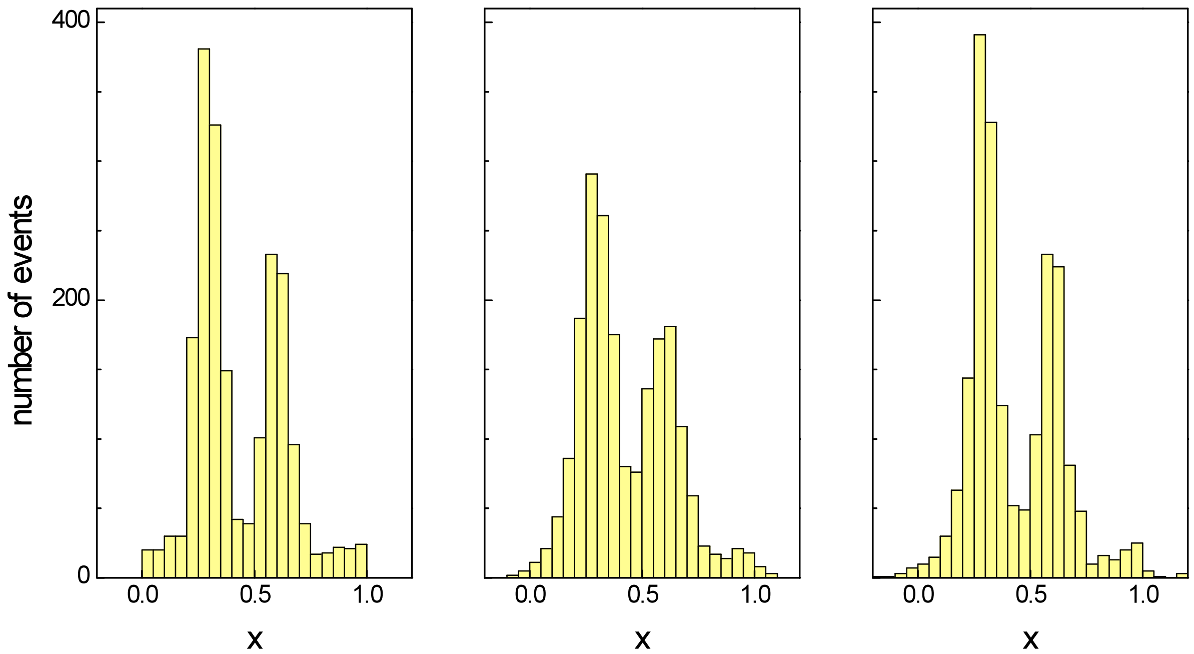
Remarks:

- The smearing is reduced by factor $\sqrt{k}$
- Result is independent of the distance function.
- Result is independent of migration step width.
- Regularization strength depends on k
- Regularization can be steered by stopping the process

Examples

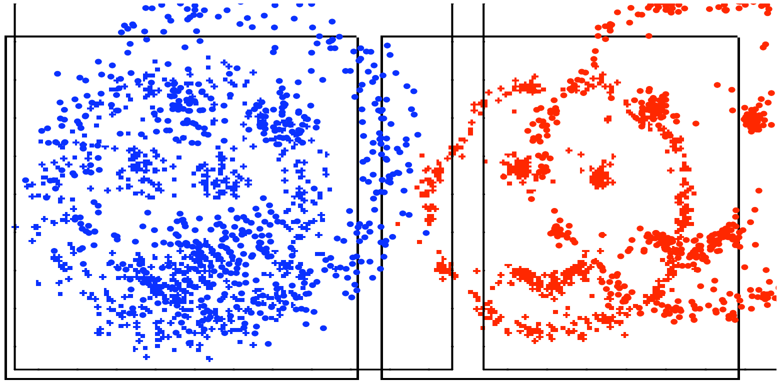
One-dimensional distribution

(unfolded binning-free, presented as histogram)

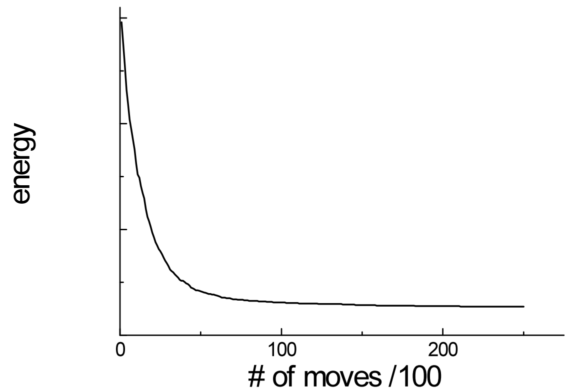


original smeared unfolded

Two-dimensional drawing (not feasable with binning)



600 experimental observations,
$k = 25$ observation per MC true point
20 000 random moves



Some complications

Acceptance losses

Solution: weighting

During generation of observations remember $w_j = k$ / # of trials
→ MC observations are weighted.
After unfolding, weights are included in the error calculation.

Variation of resolution and acceptance with position

(similar problem as in binned case)

Solution: iteration, repeat the simulation periodically

What about speed?

With analytic function, 2 dimensions, N=1000 events + side band regularization, 100000 moves: 100 s
N=4000 → 15 min. $t \sim N^2$
(on a 5 years old slow labtop)

With MC satellites time increases proportional to the number k^2 of satellites

Speed can be increased:

- faster computer,
- migration in two steps. step 1: use approximate analytic function
  step 2: simulate satellites and iprove precision.
- consider only points in neighborhood → $t \sim N$
- increase # of satellites during process

Future improvements

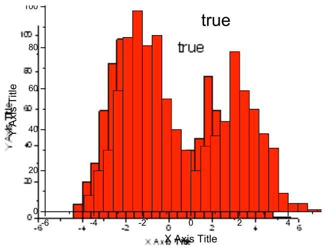
• include side band regularization into MC scheme
• combine analytic and MC approaches.
    step 1: use approximate analytic function
    step 2: simulate satellites and improve precision.
• increase speed by storing addresses of neighboring points
• automatic parameter setting based on data

More details can be found in:

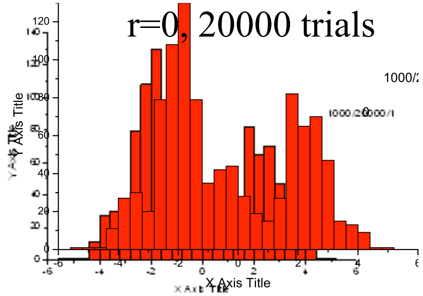
1. G. Bohm, G. Zech, Einführung in Statistik und Messwertanalyse für Physiker , E-book, Desy Library
2. G. Bohm, G. Zech, Introduction to Statistics and Data Analysis for Physicists , E-book, Desy Library (considerably extended w.r.t. to German version) (soon available)
3. B. Aslan and G. Zech, Statistical energy as a tool for binning- free goodness-of-fit tests, two sample comparison and unfolding. NIM A 537 (2005) 626
4. B. Aslan and G. Zech, \emph{New Test for the Multivariate Two-Sample Problem based on the concept of Minimum Energy}, J. Statist.Comput. Simul. 75, 2 (2004), 109

Side-band regularization in x

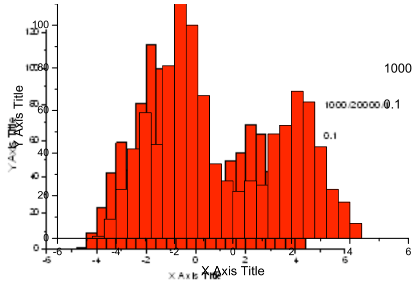
true



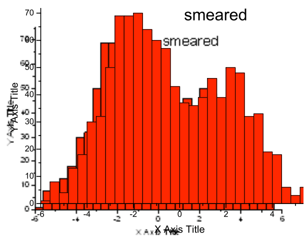
r=0, 20000 trials



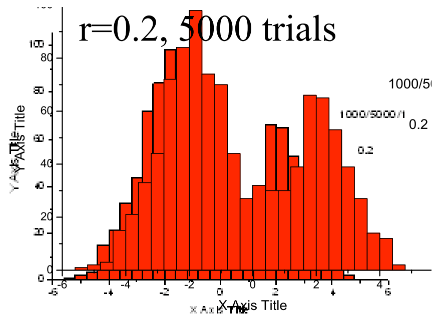
r=0.1, 20000 trials



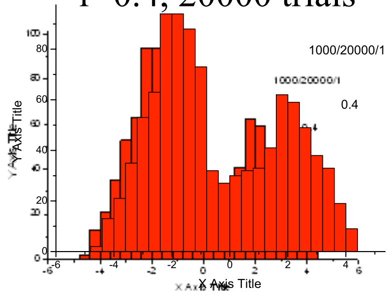
smeared



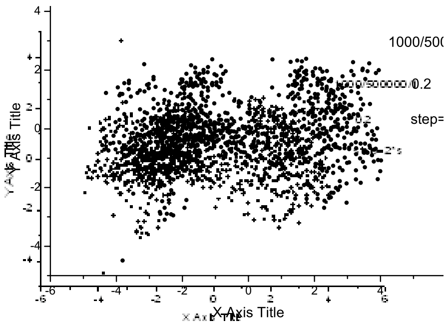
r=0.2, 5000 trials



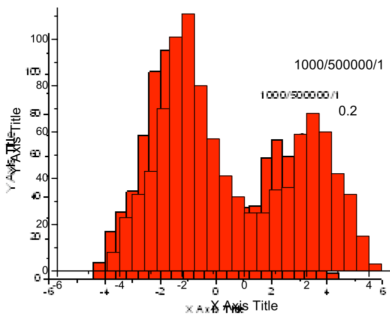
r=0.4, 20000 trials



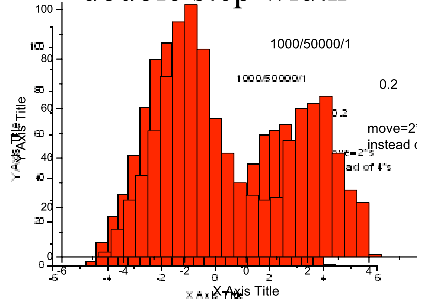
r=0.2, 500000 trials



r=0.2, 500000 trials



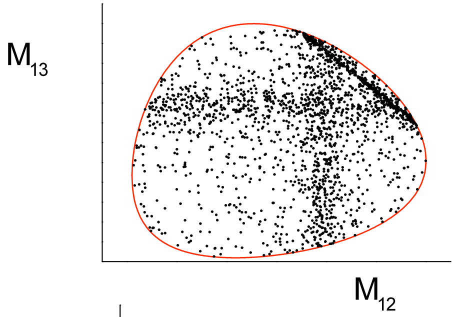
r=0.2, 50000 trials
double step width



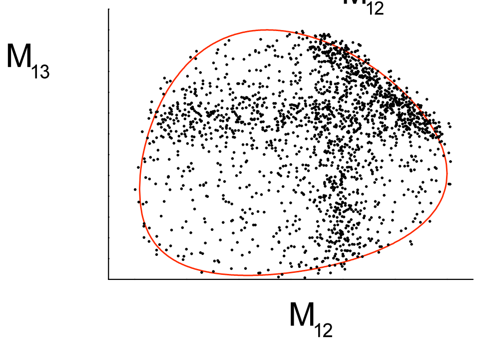
Dalitz plot with 25 satellites

2000 events, K*, ϕ

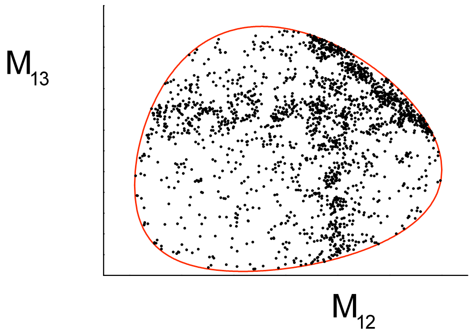
k =25



original data



smeared data



unfolded data