Binning-Free Unfolding Based on Monte Carlo Migration*

G. Zech and B. Aslan
University of Siegen, 57072 Siegen, Germany

An experimental data sample is compared to a Monte Carlo sample in the observation space. The Monte Carlo events migrate in the true variate space until the two observed samples coincide. The agreement is quantified using a Gaussian or logarithmic weight for the distance of the observations. The approach is binning-free and especially powerful in multidimensional applications where unfolding of histograms suffers from the "curse of dimensionality". The study is preliminary.

1. INTRODUCTION

In order to deconvolute experimental data which are distorted by measurement errors, physicists usually group the data in histogram bins [1]. The histogram of the true distribution can be regarded as a set of unknown parameters which are determined in fitting procedures. Statistically not significant bin-to-bin oscillations are damped by appropriate regularization schemes. Other methods regularize the Monte Carlo matrix which relates observed and true parameter values [2].

Less stringent than fitting methods, but simpler and quite effective is iterative unfolding introduced in 1982 [3] and reinvented in 1994 [4, 5]. Oscillations are suppressed by stopping the iteration process.

An introduction to unfolding and the related statistical problems is given in Ref. [6].

The energy approach can be combined with Monte Carlo methods to perform binning-free unfolding. Unfolding without binning offers several advantages.

- Arbitrary bin boundaries are avoided.
- Selection criteria can be applied to the data after unfolding.
- Variable transformations are possible after unfolding.
- The initial Monte Carlo simulation which is required to relate true and observed histograms can be rather crude.
- Low statistics data can be handled in arbitrarily high dimensions where histogramming is problematic.
- The unfolded sample represents the statistical precision of the measurement, while errors associated with histogram bins often depend strongly on the regularization strength.

*Supperted by Bunderministerium für Bildung und Forschung, Deutschland

Iterative unbinned unfolding has been presented in a previous paper [5]. Here we propose a new approach based on a simple idea: We start with a Monte Carlo sample of the same size as the observed data sample. We then let the Monte Carlo events migrate in the true variate space until the observed samples are compatible. When the process has converged, the true Monte Carlo sample represents the unfolding result.

The concept requires four ingredients, some of which are not as trivial as might seem at first sight:

1. To quantify the agreement between the experimental data sample and the Monte Carlo sample we need an appropriate test statistic. We use the *energy* concept[7].
2. The simulation scheme has to avoid additional statistical fluctuations. Each Monte Carlo true event is accompanied by a cloud of observations.
3. An efficient migration process should be used.
4. Regularization has to be provided.

2. OUTLINE OF THE METHOD

2.1. The test statistic

To compare an experimental sample $\mathbf{x}'_1, ..., \mathbf{x}'_N$ in the n-dimensional space $\mathcal{R}^n$ to the Monte Carlo sample $\mathbf{y}'_1, ..., \mathbf{y}'_M$, we use the following relation:

$$\phi = \frac{1}{N^2}\sum_{j>i} R(|\mathbf{x}'_i - \mathbf{x}'_j|) + \frac{1}{M^2}\sum_{j>i} R(|\mathbf{y}'_i - \mathbf{y}'_j|) - \frac{1}{NM}\sum_i\sum_j R(|\mathbf{x}'_i - \mathbf{y}'_j|)$$

The distance function R is either $R_{\log}$ and R_G:

$$R_{\log}(r) = -\ln(r+\varepsilon) \quad (1)$$

$$R_G(r) = e^{-r^2/(2s^2)} \quad (2)$$

The cut-off parameter ε is introduced to avoid divergencies. Its precise value is unimportant, it should be small compared to the distance of observations in the most dense regions. The test statistic ϕ is minimum in the limit $N, M \to \infty$ if and only if the two samples originate from the same parent distribution [7].

2.2. The Monte Carlo sample

We associate with the data sample $\{\mathbf{x}'_i\}$ of size N a Monte Carlo sample $\{\mathbf{y}_i, \mathbf{y}'_{ik}\}$ where i again runs from 1 to N and $\mathbf{y}'_{ik}$ is a set of K observed values of the true value $\mathbf{y}_i$. Thus each true value is accompanied by a subsample of K observations. The number K is typically of the order of 20. The statistical fluctuations of the simulation are smaller than those introduced by the experimental statistic by the factor $\sqrt{K}$. The starting values $\mathbf{y}_i$ are arbitrary. To avoid large computing times, they should not be too far off from their final position.

2.3. The migration process

So far we have spent little effort to optimize the migration process. We select randomly one Monte Carlo point i and modify $\mathbf{y}_i$ by a random uniform displacement $\mathbf{\Delta}$. At the new position K new observations are generated. The change of energy ϕ is computed and the move is accepted if the energy has decreased and rejected otherwise. The migration process continues until the minimum of ϕ is reached.

The process can be accelerated if those Monte Carlo points are selected preferentially which contribute strongly to the energy. The direction of the move could possibly be optimized moving along the energy gradient. Another possibility is to form substructures by grouping observations replacing them by a single replacement charge which is again dissolved at a later stage. We could also consider including only neighboring charges in the energy calculation for a first crude adjustment. The computing time would then be a linear function of the number of events. We have not investigated these possibilities. For up to a few thousand events the calculations can be performed without refinement on a standard PC.

So far we have also neglected the possible existence of more than one local minimum of ϕ. If this happens, one could try introducing an "annealing" term. Moves increasing the energy by $\Delta\phi$ are accepted with probability $p = 1/(1 + e^{\Delta\phi/T})$ with an appropriate choice of the parameter T.

2.4. The regularization

There are two different choices of regularization: i) The migration process can be stopped before the oscillations become intolerable. ii) The value of the parameter K can be adjusted. Large values provide high resolution but introduce oscillation. The point spread function σ_u after deconvolution for N experimental observations of a single point with resolution σ is expected to follow

$$\sigma_u = \sigma\sqrt{\frac{1}{N}(1 + \frac{\kappa}{K})}$$

where κ is a constant depending on the shape of the distance function.

3. EXAMPLES

Example 1: First we illustrate the new unfolding approach with a one-dimensional distribution. Even though the unfolding has been performed without binning, we present the result in form of histograms in Figure 1. Two Gaussians are superposed to a uniform distribution. The standard deviation of the Gaussians and the assumed experimental resolution were both $\sigma = 0.05$. The Monte Carlo enhancement factor was $K = 16$.

Example 2: In Figure 2 we present a simple example in two dimensions. The original picture of the face consisted of infinitely thin circular lines and of dots for the eyes. The picture contains 600 observations. Each true Monte Carlo point was accompanied by $K = 25$ observations. The unfolded picture was obtained after $20,000$ trials of random moves.

It would have been be quite difficult to convert the pictures of Figure 2 into histograms and to apply the standard deconvolution methods.

Example 3: Finally, we apply the unfolding to a toy PET measurement in two dimensions. A positron and an electron annihilate at rest and the tomograph registers the two back-to-back photons at a circular detector at angles α, β. The emission point has to lie on the line connecting the two positions where the photons are detected. For the simple case that the source consists of a single point, all observations are located on a curve in the two-dimensional α, β space. We have simulated the process for a source consisting of two source points located at $(x_1 = 1, y_1 = 0)$ and $(x_2 = 1, y_2 = 1)$ and a total of 500 observation pairs α_i, β_i. The initial position of the Monte Carlo sources was $(x_{MC} = 0, y_{MC} = 0)$.

Figure 3 shows the result of the deconvolution which was performed with a logarithmic distance function and a K factor of 25. The Monte Carlo source points have moved to the expected locations and their angle pairs ly on the corresponding curves.

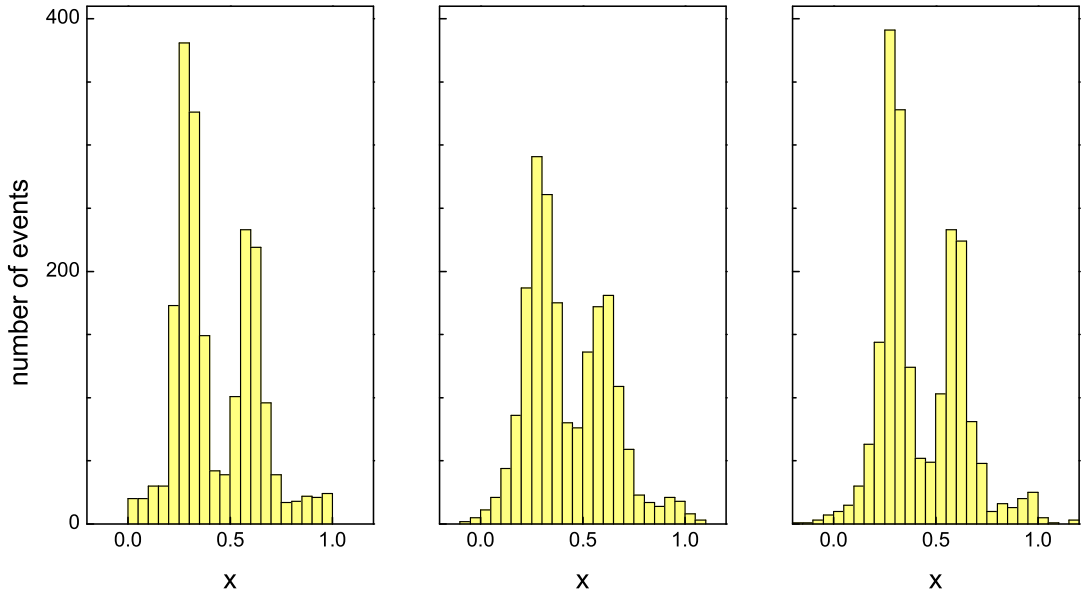

Figure 1: Unfolding of a one-dimensional distribution. The true distribution (left), the smeared distribution (center) and the unfolded distribution (right) are shown in form of histograms.

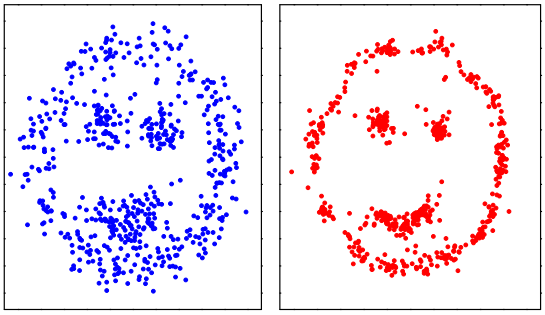

Figure 2: Unfolding of a simple picture.

4. TECHNICAL REMARKS

- The distance function used to compute the energy is only of technical importance. In the limit of large numbers it has no influence on the result, only the speed of convergence depends on it. We suggest using Gaussians with width similar to the resolution or logarithmic distance functions.

- The average migration steps should be larger than the resolution. We propose to generate the steps using uniform random numbers for each dimension. Again, the choice of the migration procedure influences only the speed but not the result.

- After each move the energy has to be recalculated. Only the charge combinations which contain the moving charge have to be evaluated.

- Acceptance losses can be included in our method by weighting the Monte Carlo events. The weight is set equal to the ratio T/K of number T of trials required to generate K observations, the inverse of the acceptance.

- If acceptance and resolution are independent of the location, the K observed Monte Carlo points can migrate together with the true point. Resimulation is not required in this case.

5. DISCUSSION

The performance of the various unfolding procedures on the market is very similar. Without regularization and constraints, histogram fitting methods, likelihood or χ^2, iterative unfolding and matrix in-

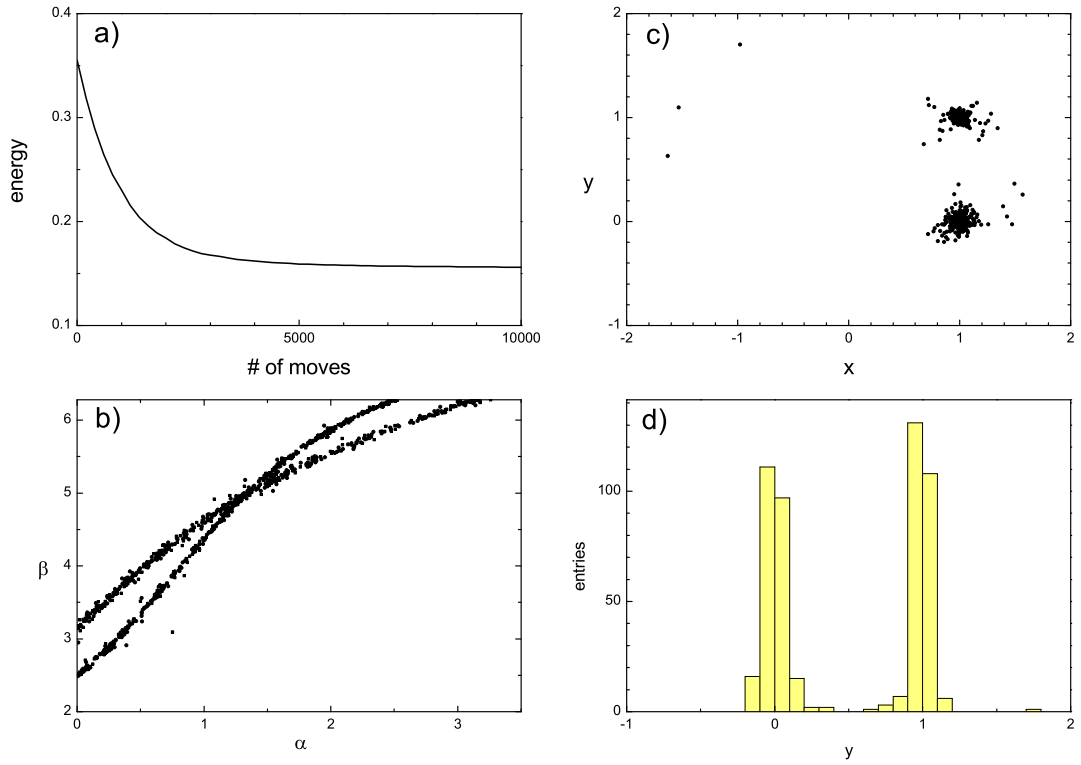

Figure 3: Unfolding of PET measurement: a) convergence of energy, b) observed angles, c) reconstructed source positions, d) corresponding y projection.

version give identical results[1]. The differences of the methods lie in the regularization schemes they apply. Also the binning free approach is exact in the limit of $K >> 1$ up to the necessary smoothing effects. There is no additional loss of information. This is obvious, because we can reproduce the original observed distribution from the unfolded sample up to the regularization effects.

The new approach opens the possibility to solve problems which are not accessible with the conventional unfolding methods. It is especially powerful in multidimensional applications with sharp structures.

In smooth, high statistics distributions, histogramming methods are preferable because they are faster.

Additional work is required to study the effect of local minima, to optimize the migration process and to study cases with a very small number of observations.

[1]There is a difference in rare cases: Matrix inversion can produce negative bin contents which are avoided in the other methods which then yield a biased result.

References

[1] A. N. Tikhonov, "On the solution of improperly posed problems and the method of regularization", Sov. Math. 5 (1963) 1035
V. B. Anykeev, A. A. Spiridonov and V. P. Zhigunov, "Comparative investigation of unfolding methods", Nucl. Instr. and Meth. A303 (1991) 350.

[2] V. Bobel, "An unfolding method for high energy physics experiments", Proceedings of Conf. Advanced Statistical Techniques in Particle physics, ed. M. R. Whalley and L. Lyons, Durham 2002.

[3] L. A. Shepp and Y. Vardi, IEEE trans. Med. Imaging MI-1 (1982) 113.,
A. Kondor, "Method of converging weights - an iterative procedure 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 (1986) 371.

[4] G. D'Agostini, "A multidimensional unfolding method using Bayes' theorem", Nucl. Instr. and Meth. A362 (1995) 487.

[5] L. Lindemann and G. Zech, "Unfolding by weighting Monte Carlo events", Nucl. Instr. and Meth. A354 (1994) 516.

[6] G. Zech, "Comparing statistical data to Monte Carlo simulation - parameter fitting and unfolding", Desy 95-113 (1995).

[7] G. Zech and B. Aslan, "A Multivariate Two-Sample Test Based on the Concept of Minimum Energy", available in these Proceedings on page 97 and at http://arxiv.org/abs/math.PR/0309164.