Unfolding: Introduction

Louis Lyons
Blackett Laboratory, Imperial College, London SW7 2BW, UK
and Particle Physics, Oxford OX1 3RH, UK

Abstract
As a non-expert on unfolding, I wanted to make a few 'obvious' non-controversial remarks about unfolding. It turns out, however, that even such innocuous comments can become the subject of heated debate.

1 The Problem

Given a one dimensional histogram for a particular variable x, obtained in a detector with known experimental resolution, can we estimate the histogram that we would have obtained had the detector not introduced any smearing? We assume that the smearing is specified by a matrix M, whose $(i,j)^{th}$ element is the probability that an event actually in bin j of the true histogram for x appears in bin i for the smeared data. It may be that the matrix M is known exactly. More likely is that its elements have statistical errors from its estimation via a simulation of detector effects; and/or it has systematic uncertainties because it was derived from an approximate model.

Of course we will also need to provide a covariance matrix for our de-smeared histogram. There are some obvious extensions of this example: the original distribution in x could be unbinned rather than in a histogram; the problem could involve more than one dimension, etc.

Thus the High Energy Physics unfolding problem is different from the more common statistics situation, where the issue is to remove the effects of smearing from an optical image in order to sharpen it, and to decide, for example, whether the photograph is of a dog, a cat or a person. In that case, estimates of uncertainties in pixel intensities, or in their correlations, are not of interest.

2 Why Unfold?

Because folding an assumed true distribution in x is simpler than unfolding an observed distribution in an attempt to obtain the true one, it is in general preferable to avoid unfolding. Then a comparison between a predicted theory and observed data is performed at the level of the smeared theory with the actual data, rather than between the pure theory and the unfolded data. This can even be performed for future theories, provided that the smearing matrix is published along with the data. The argument that a theorist, say in 2051, will find it difficult to smear his/her theory is very weak, since multiplying a vector by a matrix is computationally and conceptually no more difficult than comparing the original theory with the unsmeared data, which involves a non-trivial error matrix.

This then raises the question of when it might be necessary to unfold. A few cases are listed:

a) Comparing or combining experimental distributions from experiments with different smearing matrices M.

b) Tuning a Monte Carlo simulation, by fitting the parameters involved in the QCD theory to the data. Apparently this proceeds too slowly if the theory has to be smeared at each step of the iterative fitting.

c) Obtaining a plot for posterity that shows the estimate of the true distribution, rather than including the non-fundamental effects of experimental resolution. However, the unfolded distribution can contain strong bin-to-bin correlations, and physicists are accustomed to making eyeball judgements only about histograms whose bin contents are uncorrelated.

3 Matrix Method

Let d_i be the contents of the i^{th} bin of the data histogram and t_j that of the j^{th} bin of the distribution without smearing then,

$$d_i = \Sigma M_{ij} t_j. \tag{1}$$

Note that there can be fewer bins for the unfolded distribution than for the data.

The maximum likelihood solution for t can have large bin-to-bin oscillations, with estimated bin contents actually being negative. These effects become less serious for wide bins when the off-diagonal elements of the smearing matix are smaller. However, they then become more sensitive to the model assumed for deriving the smearing matrix; any such systematic effect should be taken into account in estimating the error matrix for the unsmeared distribution.

4 Bin-by-bin Correction Factors

This is an easy-to-use method, but unfortunately it has some serious drawbacks. Monte Carlo simulation is used for an assumed true distribution with a_i events in the true histogram, which after smearing becomes p_i events in the pseudo-data histogram. (Note that p_i in a given bin depends on all the a.) Then the correction factor C_i for the i^{th} bin is simply defined by

$$a_i = C_i \times p_i, \tag{2}$$

and depends on the assumed distribution. These factors are used to correct the observed data d_i to the estimated truth e_i by

$$e_i = C_i \times d_i. \tag{3}$$

An example of this approach for a simple 2-bin distribution is shown in Table 2, for the smearing matrix of Table 1. The numbers of true unsmeared events in bins 1 and 2 are 800 and 200, respectively. With the smearing matrix of Table 1, this results in 760 and 240 expected events in the 2 bins of the smeared distribution. We set the observed d_1 and d_2 equal to their expected values.

Each row of Table 2 corresponds to a different assumption about the a_i. ($a_1 + a_2$ is always taken as 1000, as in the assumed real data). The second row of numbers has them set at their true values, 800 and 200, respectively; then the estimates e_1 and e_2 are correct. For all other rows, the estimates are biassed, even for the case where both correction factors are unity; and also when the assumed numbers are set equal to the observed ones. Also their sum is not equal to the total number of observed events[1]. Furthermore, the errors on the bin contents of the unfolded histogram are in general taken as uncorrelated.

Calculating the uncertainties on the unfolded bin contents is problematic. For example, with $d_i = 100 \pm 10$ and $C_i = 0.1$, it is tempting to write $e_i = 10 \pm 1$. This uncertainty is wrong (it is even smaller than $\sqrt{d_i}$), as it is merely the uncertainty on the expected number, and does not include the statistical fluctuation on the observed number. Perhaps more importantly, only diagonal errors are obtained in this method.

Because of the sensitivity of the derived answers to the assumed unfolded distribution, which is needed for calculating the C_i, and because of the problems with obtaining the error matrix for the unfolded distribution, this method is *not* recommended[2].

5 Questions to be Resolved

5.1 Bin size

If the bin size of the unfolded distribution is too narrow, the smearing matrix has large off-diagonal elements. On the other hand, large bin width results in a loss of sensivity to high frequency components

[1] This contrasts with the matrix method.

[2] Though in some cases an iterative approach may work.

Bin	Truth 1	Truth 2
Observed 1	0.9	0.2
Observed 2	0.1	0.8

Table 1: The smearing matrix, whose elements M_{ij} are the probabilities that, as a result of smearing, an event in bin j of the true distribution appears in bin i of the distribution for the actual data.

a_1	a_2	p_1	p_2	C_1	C_2	e_1	e_2	Sum
1000	0	900	100	1.11	0	844	0	844
800	200	760	240	1.05	0.83	800	200	1000
760	240	732	268	1.04	0.90	789	215	1004
667	333	667	333	1.00	1 00	760	240	1000
500	500	550	450	0.91	1.11	691	267	958
200	800	340	660	0.59	1.21	447	291	738
0	1000	200	800	0	1.25	0	300	300

Table 2: Problems with correction factors. The correction factors C_i are calculated from assumed true numbers a_i and the corresponding smeared numbers p_i in each bin. The estimated numbers e_i are calculated as C_i times the actual observed numbers 760 and 240, and are to be compared with the true numbers 800 and 200 respectively. 'Sum' is $e_1 + e_2$, and in general is not equal to the observed sum of 1000.

of the true distribution; and to uncertainties in the elements of the smearing matrix, caused by their sensitivity to the distribution of the variable of interest across a bin. Some recommendation about the choice of bin width would be useful.

5.2 Regularisation

Because the maximum likelihood solution to the unfolding problem can result in an unfolded distribution with large bin-to-bin oscillations, some form of regularisation is generally employed to damp down these oscillations and to produce a smoother solution. This is usually achieved by adding a term to the likelihood which penalises large second derivatives; or by removing high frequency modes from the solution. Alternatively, orthogonal decomposition with suppression of small components can work for all distributions except those with sharp features.

A variety of regularisation methods is available, and again advice would be useful on the best form and the optimal regularisation strength to use in a given problem.

5.3 Size of errors

As already mentioned, calculating the uncertainties in the estimated unfolded distribution is often not trivial. (In cases of difficulty, a bootstrap method or some other method of varying the bin contents may be useful.) In particular, the question arises as to whether the uncertainty on an unfolded number of events e_i can be smaller than $\sqrt{e_i}$; that is, can the estimate in a situation with smearing in x have a smaller uncertainty than if x had been determined precisely? The answer may be *yes* because regularisation provides some form of local averaging, which can reduce the uncertainty on e_i.

5.4 Assessing a solution

It is not obvious how to assess which solution is best out of a series of solutions to an unfolding problem. Some of the problems are:

– A criterion of the largest p-value would favour a solution with large errors.
– Taking an unweighted sum of squared deviations ignores the fact that some bins are determined more precisely than others.
– Would we be satisfied if the minimum χ^2 were for a solution with wild anti-correlated oscillations?

Issues such as these need to be resolved before an Unfolding Challenge, along the lines of Banff Challenges 1 (for intervals) and 2 (for discovery), can meaningfully be set.

Bob Cousins has suggested a test that should be satisfied by any method that is used to unsmear two different data distributions, produced from two known 'true' distributions and specified detector smearing. Then he would expect

$$\Delta\chi^2_s = \Delta\chi^2_u \tag{4}$$

where $\Delta\chi^2$ is the difference in χ^2 between models 1 and 2; and the subscript 's' refers to the χ^2 being calculated by comparing smeared theory with the data, while 'u' refers to the comparison between the de-smeared data and the original theory. The reason for using 2 theoretical models rather than just one is that an unsmearing method could be tuned to work for a specific theory. Opinion is divided as to whether all reasonable theories would pass the test; or whether it is obvious that regularisation would invalidate it.

6 Today's Talks

We are very fortunate to have with us Victor Panaretos, a statistician from Lausanne, who will give the Statisticians' view of our problem, and will be present throughout the day, to encourage us to use statistically acceptable methods, and at very least to prevent us from straying too far afield. The other speakers before lunch (Blobel, Zech, Kartvelishvili and Bierwagen) will talk about the more common methods developed by High Energy Physicists. The afternoon speakers will deal with other HEP methods; a software framework for unfolding methods; and about the methods used in practice in 3 of the LHC collaborations.

7 Postscript

It was hoped that as a result of the meeting we could produce a consensus of points on which the major unfolding programme developers were in agreement. Perhaps as might have been predicted from the particularly lively discussions before, during and after the sessions, this turned out not to be possible. However, it is a goal worth striving for in the near future.

It is a pleasure to thank Volker Blobel for illuminating discussions.