ZEUS-Note 95-166
December 16, 1995

Comments on the Bayesian unfolding

G. D'Agostini

Abstract

Since one year some people are using the Bayesian unfolding described in the NIM paper **A362** (1995) 487. This note summarizes some answers to typical questions, as well as some comments to the use of the method.

NOTE added to the distribution file (16/12/95)

- the new distribution file can be found in
  http://zeus.roma1.infn.it/pub/bayes_distr.txt
  or via the DIS WG www page (ZEUS members only).

- "Bugs?" Until now no bug in the program has been reported, a part for a small FORTRAN problem in the example which may have disturbed some compilers: -> fixed;

- "Mistakes in the paper?" Essentially not, a part from some misprintings, most of them corrected in the NIM paper (A362(1995)487). The only one left in the NIM paper and which could generate some (initial) confusion is located 4 lines before formula (3): P(C_i|E_j) should have been P(E_j|C_i), as obvious from the text.

- "The program takes much time." If the calculation of the covariance matrix is required and there are too many bins, the CPU time could diverge. Remember to ask for the evaluation of the uncertainties only after the last step (unless they are really needed for the smoothing, see below). Eventually choose the option of Poisson approximation, which is reasonable in most of the cases. In very complicate situations ask only for the diagonal terms (better than nothing ...) or try to evaluate the uncertainties in different ways (see below).

- "The unfolded distribution is not correct." This kind of objections, apart from the cases of trivial mistakes, sounds metaphysical. One has to remember that an important experiment is performed only ONCE, that statistical fluctuations exist, and that "einmal ist keinmal" (cfr M. Kundera). A judgement on the quality of the unfolding is valid only if the program is used on ''simulated data'' with a reasonable statistics, AND repeated several times in order to study the fluctuations of the estimators around the simulated true values.

- "Some bins of the smearing matrix are zero." "Hic sunt leones" ("Here are the lions") used to write the ancient romans

in maps, outside the world known to them: if the detector is not sensitive to a cause (=kinematical region) the experimenter cannot pretend to give a result about that region.
Perhaps other kind of unfoldings do provide some results in the "forbidden" region just as an analytical forbidden from the good one. You may do it anyhow by yourself using some kind of extrapolation, but the program does not take such a responsibility.

- "The iterative procedure is against the Bayesian spirit, since the same data are used many times for the same inference."
  I absolutely agree with this statement, BUT in practice this technique is just a "trick" to give to the experimental data a weight (an importance) larger than that of the priors.
  A more rigorous procedure which took into account uncertainties and correlations of the initial distribution would have been much more complicated (I must confess of having tried several approaches of this kind without any real success ...).

- "How many iterations?" This should be studied case by case on simulated events. The simulated "data" should have the same statistics of the experimental data, and the behaviour of the unfolding on MANY "data" sets should be studied (see above point on "unfolded distribution not correct").
  Then the same criteria should be applied "blindly" to the real data. From the experience of many people it comes out that for "normal" problems 3+-1 iterations is a kind of optimum.
  NEVERTHELESS I recommend to use the SMOOTHING:
    - the procedure is consistent with the Bayesian spirit, in which the knowledge follows from a combination of prejudices and experimental data (see DESY-95-242, hep-ph/9512295, for an introduction, and the discussion at pag. 496 of the NIM paper);
    - fast convergence is ensured;
    - the sensitivity on the particular function is generally very weak;
    - the result is dominated anyhow by the data as an effect of the LAST iteration (notice instead that the final distribution should not be smoothed anymore).

- "Best function for the smoothing?" For "usual" application in 1D any low-order polynomial (in many cases even a straight line!) does correctly the job. Nevertheless, if you know a function which is more suited for the physics case and which can be parameterized in order to accomodate a large variation of results, this should be preferable (it also allows to make a simultaneous unfolding & fit!).

  In most of the cases the smoothing can be done even without taking into account the difference of weights of the different bins (particularly true if all bins contain similar numbers of events). However this is not true in case of low statistics with consequent large uncertainty on the unfolded numbers. In this case it is recommended to take into account also the different weights, using for example MINUIT instead than a simple regression routine.

- "Other ways to dump the oscillations?" In principle yes, but I believe that smoothing is more consistent with the spirit of the method and to the physics case (e.g. structure functions are regular, independently if they rise at low-x or if they saturate).

- "Separation into acceptance and smoothing?" No, please! I think that this is just a way on complicate the life. The smearing matrix should include both effects at once. I have the same criticism also to the use of "bin-to-bin" or "parameterized" (x_corr = f(x_meas)) corrections followed then by the unfolding.

- "Acceptance cut on the physical region". The PHYSICAL region is that before the smearing effects, or after the unfolding. The measured values may have a domain different than the true values, BUT they could carry anyhow a reasonable information about the true values. (For example one could measure in DIS x >> 1 and nevertheless there could be no reasons to discard this data from the analysis if their migration is well undertood).

  For this reason the number of cells in the measured value should be larger than that of the true value.

  This also means that arguments based on "purity" should not be considered: in some cases one could have for all bins purity = 0, without any problem for the unfolding. (Imagine, for example, if there is a large systematic shift of all measured quantities. One may argue that in this case it is better to make a correction before and then the unfolding. I will come back to this again in the point "separation into acceptance and smoothing".)

- "How many bins?". One has to match somehow the experimental resolution (and not only the instrumental one). It is recommended to give a look at the correlation matrix of the results: if adjacent bins have a very high degree of correlation, one should consider to enlarge the bin size (not necessary everywhere, but only where there are very high correlations). This is not really mandatory if the correlations are taken into account (as they should always be!) in the subsequent analysis. The bins of the real data should always contain a reasonable amount of events so that the usual approximations are valid (The minimum? ?? 10, 15, 8, 6, ...)

- "Number of bins of the true value larger than that of the measured value?" For obvious reasons the opposite is recommended. However the method allows such a possibility, but then the degree of correlation between the unfolded numbers, as well as the dependence on the initial distribution, increases. "Unfortunately" the program is very stable and will not complain even if one has only 1 measured bin (e.g. the total number of events): but then the result is exactly the initial distribution, as it is reasonable to be
  The very reason why no control has been introduced in the program is that there may be situations where the user is really interested to unfold a number of bin larger that number of data points, but then he must be very careful in treating the correlations of the results.

- "Unfold background-subtracted distributions?" I find more correct and easier to include the treatment of the background in the unfolding program, as described in sec. 5 of the paper.

(Think, for example, to what it would happen if negative numbers of events arise from the subtraction, just because of statistical fluctuations.)

- "How to merge several samples of MC events?" They can just be summed up, as long as they are independent. (The technique is used when some kinematical regions are not enough populated.) I remind that it may be convenient to use a MC where the events are generated flat in phase space instead than according to the differential cross section. If one has generated several event samples in different regions according to the differential cross section (BUT obviously with the same physical assumptions) the number of events measured in a cause cell and measured in an effect cell can be simply added.

- "Weighted MC events." They can be used to calculate the smearing matrix (to be done by the user), but the program, as it is, is unable calculate their uncertainty. The program needs to be modified in the point where

  Cov[P(E_r|C_u), P(E_s|C_u)]

  is evaluated (see end of section 4 of the paper). Essentially the number of events used to evaluate the covariance matrix should be replaced by the "equivalent numbers of event" (thanks to Roberto Sacchi for this observation and see e.g. the Guenter Zech report DESY 95-113 for the definition of "equivalent number of events").

- "The uncertainties are smaller or larger than expected." There is no reason why the uncertainties should be sqrt of the unfolded numbers. For example, if the smearing matrix was diagonal with all elements much smaller than 1, then the resulting uncertainties would all be much larger than sqrt of the unfolded numbers. In the general case the situation is even more complicate and one has to rely to the complete propagation of uncertainties, besides personal prejudices.
  A way to check the correctness of the uncertainty propagation is to use simulated events. This has been done for example in the NIM paper (fig. 8 and 9). In the figures one may easily see that SAME numbers of unfolded events

have different standard deviations, and that, moreover, the standard deviations depend on the true distribution AS WELL as on the the smearing matrix (though assumed to be known without uncertainties!).
Another way to check the result is to make MANY unfoldings (i.e. following the complete procedure) varying the number of data events randomly around the observed numbers, according to the assumed distribution (typically: Poisson -> Normal). From the set of unfoldings one can then calculate averages, variances and covariances.
Instead, a way to understand what is really going on and why different regions get different statistical significance is to look at the unfolding matrix and follow the flow of "inverse-migration" of the information: observed data --> unfolded numbers.

- "Overall normalization." OVERLOOKED! Sorry. Writing the program I was much concentrated on the shape of the unfolded distribution (P(C_i)) and an overall normalization uncertainty was not considered (I thank Jose' del Peso for having pointed it out). The effect is important for small total number of events which enter in the analysis.
  The present version of the program gives separately the covariance matrix of the shape and the covariance matrix of the unfolded numbers.
  CAUTION: one has to be very careful in performing fit with the covariance matrix if there is an overall normalization uncertainty: see e.g. NIM A346 (1994) 306.

- "Other unfoldings?" Yes please! I only list here those of which I am aware and that are enough "professional", i.e. they at least take into account of the correlations (the names are the ones I use colloquially):
  "Blobel" : see NIM paper.
  "Zech" : DESY 95-113.
  "Sinkus" : used by Ralph Sinkus (ZEUS) in his PhD thesis: see Anykeyev, Spiridonov and Zhigunov, NIM A303(1991)350.
  "SVD" : Hoecker and Kartvelishvili, MC-TH-95/15, LAL-95/55, hep-ph/9509307
  "Weise" : 'the fully Bayesian unfolding?': K. Weise, PTB-N-24, Braunschweig, July 1995 (see also

Weise and Matzke, NIM A280(1989)103, and Weise and Woeger, Meas. Sci. Techn. 4(1993)1.

- I would like to conclude with a citation from the ISO "Guide to the expression of uncertainty in measurement", although referred to uncertainties, as an invitation to think to the problems, instead of seeking for magic formulae:
  ``Although this {\it Guide} provides a framework for assessing uncertainty, it cannot substitute for critical thinking, intellectual honesty, and professional skill.
  The evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement.
  The quality and utility of the uncertainty quoted for the result of a measurement therefore ultimately depend on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value''.

- "Other comments, questions, criticisms, etc?" Please don't hesitate to contact me. I am still very interested to learn about this problem and I replay almost instantly to all questions.

- This note has been e_mailed to those who have asked directly for the FORTRAN code, with the kind request of spreading it among those to which the program has been further distributed.