Regularization and error assignment to unfolded distributions

Günter Zech
Universität Siegen, Germany

Abstract

The commonly used approach to present unfolded data only in graphical form with the diagonal error depending on the regularization strength is unsatisfactory. It does not permit the adjustment of parameters of theories, the exclusion of theories that are admitted by the observed data and does not allow the combination of data from different experiments. We propose fixing the regularization strength by a p-value criterion, indicating the experimental uncertainties independent of the regularization and publishing the unfolded data in addition without regularization. These considerations are illustrated with three different unfolding and smoothing approaches applied to a toy example.

1 Introduction and general considerations

Unfolding is a difficult mathematical problem, because independent of the amount of data the solution of the Fredholm equation $f'(x') = \int_{-\infty}^{\infty} t(x,x')f(x)dx$ does not lead to a stable solution; here $f(x)$ is the function of interest, $t(x,x')$ is the response- or smearing function and $f'(x')$ is the known distorted function. Therefore regularization methods have been developed.

In particle physics, distributions are usually presented in the form of histograms. Histogramming is a first regularization step. With a wide enough binning, unfolding is reduced to a simple inference problem where the parameters, i.e. the content of the bins, have to be estimated. These parameters and their errors can be determined by a standard fitting procedure. As is common practice in parameter fitting, the compatibility with the data can be expressed by a p-value. So far, the situation is rather clear and not controversial. The regularization by binning has the nice property that the interpretation of the unfolded distribution is straight forward, the point estimates and the error matrix fully document the result, there are no hidden parameters and all true distributions that are compatible with the observed data are admitted by the result.

Problems arise as soon as we choose a binning that is narrow compared to the smearing and try to represent unfolding results graphically. There are usually strong fluctuations between adjacent bins, these are then negatively correlated and the non-diagonal error matrix elements are huge which makes a graphical representation unreadable. But while the point estimates are badly known, the error limits are rather precisely derived from the data. To avoid the unpleasant oscillations, usually a second regularization step is introduced. Contrary to the implicit regularization by binning, here the kind of smoothing and its strength are hidden and their effect is difficult to assess from the published data. Furthermore, the explicit regularization introduces constraints that eliminate high frequency contributions, reduce the errors assigned to the histogram bins and thus exclude solutions that are admitted by the observed data. (We call the errors obtained in a regularized fit nominal errors to distinguish them from the errors defined by the measurement alone.) Theories that are compatible with the data but where their distributions contains narrow structures may be rejected.

As has been discussed by Blobel at this conference, for a given unfolding problem, an effective number of degrees of freedom can be estimated. This is the number of independent significant parameters of the unfolded distribution, i.e. the minimum number of parameters that is required to describe the observed data within their errors. This number depends on the width of the smearing function and on the available statistics. Thus a possible and also sensible solution would be to eliminate the second smoothing step and to publish data with a number of rather wide bins, a number not much larger than the effective degrees of freedom. The correlations would be relatively mild and could be documented with a correlation matrix.

One common objection to wide bins is that choosing wide bins for the estimated distribution introduces a strong dependence of the smearing matrix on the input distribution used in the Monte Carlo simulation. However, this dependence can be avoided, by combining bins of the unfolded histogram. Another reason for choosing histograms with many bins is that they look much more impressive than a crudely binned histogram (see Fig. 7 in Blobel's contribution) and that they indicate better the anticipated shape of the true distribution. More scientific arguments for a not too wide binning are the following: i) When we increase the statistics by combining the data of different experiments, the spectral resolution should increase but can only be taken advantage of if the bin size is not too large. ii) Normally we choose bins of equal size. Then the minimum number of bins may not be adequate to describe the functional dependence.

All problems can be avoided if a theoretical prediction of the true distribution is available. Then we can fold the theoretical distribution and compare it to the observed data. We do not have to construct a response matrix; binning is only required in x' and not in x, the distorted distribution is simulated and compared to the data. Unknown parameters can be estimated by re-weighting the simulated events [1]. It does not make sense to pass through a histogram with many parameters to finally determine a few parameters of interest. More important, the direct fit, for instance of the amplitude of a narrow peak with given width superposed to a uniform background, would produce a correct result, while the true distribution might be incompatible with an unfolded distribution where high frequencies are filtered out.

In the situation where no generally accepted theoretical description is available, a sensible solution to the stated unfolding problems is to separate the graphical representation of the result from its documentation. We will first sketch the way the data can be documented and then turn to the explicit regularization. We discuss how to fix the regularization strength, illustrate the problems with a simple example and three different unfolding approaches and end with a summary and recommendations.

2 Documentation of the result

To avoid the exclusion of distributions admitted by the data and to permit the combination and comparison of measurements of different experiments the data have to be published without explicit regularization.

There are at least three possibilities to document the full information: i) Unfold without regularization and provide an error matrix. The number of bins has to be low, not much higher than the effective number of degrees of freedom. ii) Publish the raw data together with the response matrix. iii) Publish the eigenvectors of the matrix C (see Sect. 4.1), their weights and the uncorrelated errors of the weights.

There might be statistical or technical difficulties applying any of these approaches which cannot be discussed in a short paper. It should just be mentioned that proposal i) requires enough data to approximate the error distributions of the unfolded histogram bins by Gaussians. Method ii) leaves all the work to the user of the data. To follow point iii), the correlations between the contents of the histogram bins have to be eliminated by diagonalization.

3 Fixing the regularization strength

In addition to the documentation of the data in form of tables we want to illustrate the result of our experiments graphically and in most cases have to add an explicit smoothing step. There is no optimal regularization algorithm, and smoothing is partially subjective. The only obvious requirement is that the smoothing step does not destroy the compatibility of the result with the observed data. A very attractive method favored by statisticians is to suppress small eigenvalue contributions of the least square matrix or equivalently in the SVD decomposition (see Blobel's and Kartvelishvili's contributions to this conference) but there are other methods that work equally well. The performance of a regularization

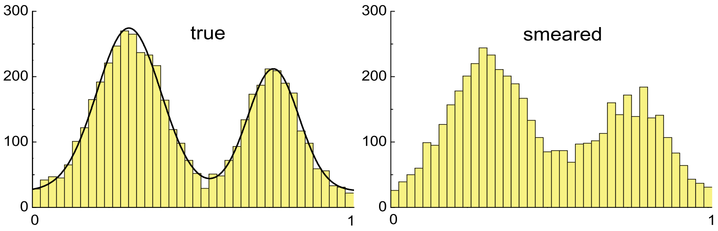

Fig. 1: Distributions used in the unfolding example. Here and in all following graphs the axes correspond to measured variable $0 < x < 1$ (horizontal) and to the number of entries (vertical).

method depends to a certain extent on the specific problem to be solved.

There are several methods to fix the regularization strength, i.e. from the kink of the L-curve [2], from minimum norm or vanishing global correlation. These are mathematical concepts. From a physicist's point of view, the essential criterion is the compatibility of the unfolded histogram with the observed data which can be measured with a p-value derived from the χ^2 statistic. What is important is not the absolute value of χ^2 of the fit of the bin contents of the unfolded histogram but its change $\Delta\chi^2$ caused by the regularization. The p-value is $p = \int_{\Delta\chi^2}^{\infty} u_N(x)dx$ where u_N is the χ^2 distribution for N degrees of freedom, i.e. N bins of the unfolded histogram. Assuming ideal Gaussian errors, this means that the regularized solution is located at the border of a $1-p$ confidence ellipsoid. When we change the regularization strength, rather independent on how we define it, the p-value initially remains close to one as long as we include high frequency contributions that have little effect on the unfolded data. As soon as we start to affect the unfolded histogram, by definition the p-value drops dramatically.

We propose to fix p to 90% or even to a higher value. This means that the parameter set of the regularized histogram is within a 10% confidence interval of the minimum χ^2 point which corresponds to the undistorted measurement. There is no necessity to suppress all fluctuations, we accept them also in standard measurements that are not affected by resolution effects. The proposed p-value is arbitrary, but it should be large, as we want to cut only insignificant features of the data. Its choice is motivated by experience.

It has been argued [3] that $\Delta\chi^2$ should not be referred to N but to the difference between N and the effective degrees of freedom N_{eff}, i.e. $\Delta\chi^2 \approx N - N_{eff}$. ($N_{eff}$ is the number of independent parameters that are necessary to describe the data or the effective rank of the response matrix, see Blobel's contribution.) The additional $N - N_{eff}$ parameters are expected to contribute 1 to χ^2 each. Indeed, for the specific case $N = N_{eff}$ in the framework of SVD one would not regularize at all. On the other hand smoothing N bins of an arbitrary distribution one would usually allow for a change of χ^2 proportional to N. Because of the sharp kink in the distribution of χ^2 as a function of the regularization strength, the two different methods usually will not lead to very different results. Further studies should clarify this issue.

It is debatable what the best value for the number of degrees of freedom (NDF) should be for converting $\Delta\chi^2$ to a p-value. However, since the cut on the p-value is arbitrary, the choice for NDF is not crucial.

It would be very helpful if the community could agree on a common scheme. This would make the comparison of different methods and of unfolded results much easier than it is now.

4 Toy example and three unfolding approaches

These abstract considerations are now illustrated with a simple toy example and 3 different regularization procedures. The true distribution is a superposition of two Gaussians and a uniform distribution: 2500

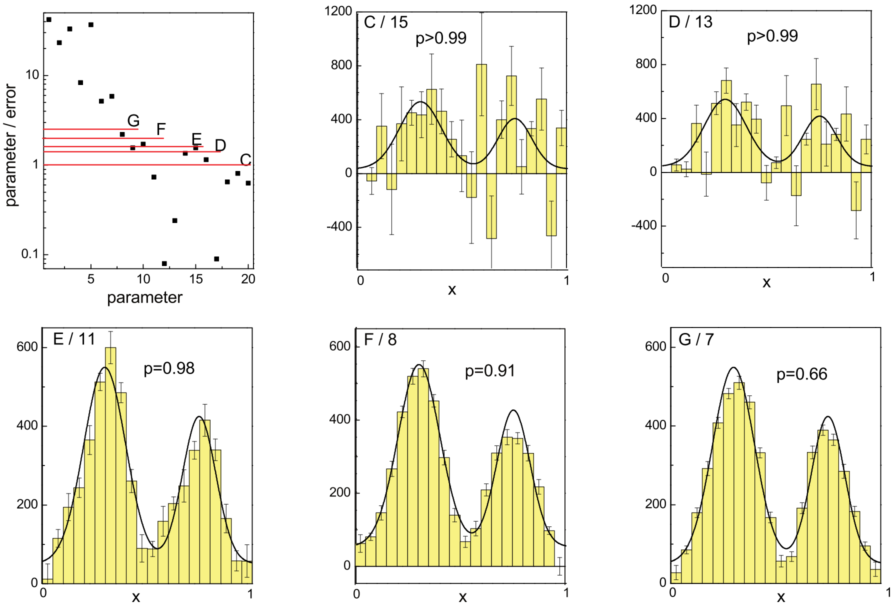

Fig. 2: Significance s of parameters (parameter / error) ordered according to decreasing eigenvalues (top left) and unfolded distribution for different cuts in the significance. The number of contributing eigenvectors is indicated in the top left corner of the plots.

events $N(0.3; 0.1)$, 1500 events $N(0.75; 0.08)$, 1000 events uniform in the range 0 to 1. The resolution function is again a Gaussian, $N(0; 0.07)$. To demonstrate the problems, a relatively large smearing was chosen. The effective number of degrees of freedom is around 8. In Fig. 1 the corresponding true distribution (curve), the original and the smeared histograms from the Monte Carlo simulation are shown. For the unfolding 20 or 10 bins for the unfolded histogram and 40 bins for the observed data were chosen. Our notation is the following: The bin content of the histogram corresponding to the true distribution is represented by the vector θ, the histogram of the observed data by $\hat{d}$, and the response matrix is A. The expectation value d is related to θ by $d = A\theta$.

4.1 Method 1: Truncation of the eigenvalue sequence

The square matrix $C = A^T V^{-1} A$, with V the error matrix of the observed data vector $\hat{d}$, is decomposed into eigenvectors u_i. The true distribution vector θ can be expressed as a sum $\theta = \sum_i c_i u_i$, where the coefficients c_i of the eigenvectors are to be determined. The coefficients c_i are uncorrelated. With decreasing eigenvalues the components of the eigenvectorss oscillate more and more but the coefficients become less significant and their contributions can be eliminated. In theory this is very attractive but in practice the situation is not always simple, as is shown in Fig. 2. The upper left hand plot shows the significance $s = |c_i/\delta c_i|$ for the parameters ordered by eigenvalue. We realize that the parameter with the lowest eigenvalue is not necessarily the one with the smallest significance. The significance of each component is proportional to the square root of the number of events. With increasing statistics more components would become significant and correspondingly the spectral resolution would increase. However, since the sequence of ordered eigenvalues decreases rapidly, statistics in most cases has only a small effect on the spectral resolution. The dominant effect comes from the width of the response function. In the literature cutting or damping low eigenvalue is proposed instead of removing low significant

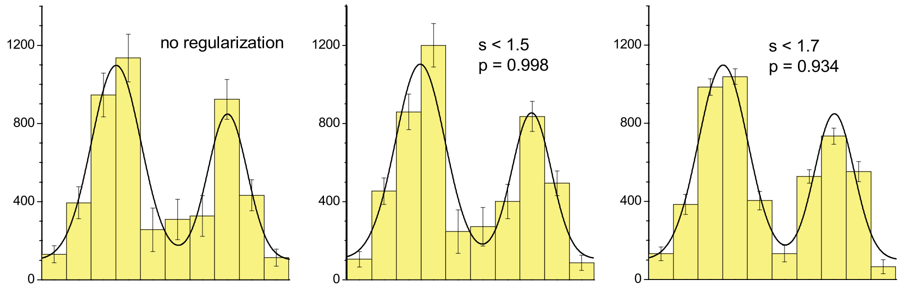

Fig. 3: Unfolded distributions for different cuts in significance.

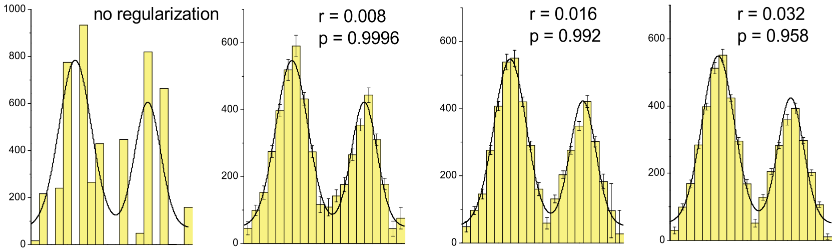

Fig. 4: Distributions unfolded with a Poisson likelihood fit and a curvature penalty.

components, but then the shape of the observed distribution which may favor certain eigenvectors is not taken into account.

In Fig. 2 the unfolded histograms for different cuts in significance are displayed and compared to the true distribution (curves). How should we cut? In principle we should choose the cut without looking at the unfolding results. Cutting contributions that are compatible with zero within their error, $s < 1$, produces unacceptable results. We have to find a compromise between loss in information and smoothing. Retaining between 8 and 11 components seems to be reasonable. The plot labeled E would be the preferred one, as it does not show excessive fluctuations and at the same time corresponds to a large p-value.

The regularization leads to a smooth unfolded distribution but, as expected, the nominal diagonal errors decrease strongly with increasing regularization strength. Because of this dependency, the errors, which are more or less arbitrary, are unreliable indicators of the precision of the result. We show them, to highlight the problem.

The situation becomes more reasonable when we turn to a smaller number of bins as illustrated in the Fig. 3. With 10 bins, which is about the effective number of degrees of freedom, negative bins disappear. The calculated errors become more reasonable because the correlations are smaller. Here we could even renounce explicit smoothing and document the errors with a simple error matrix.

The results presented have been obtained with a least square fit (LSF) and the simple matrix formalism. It would be better to apply a Poisson maximum likelihood fit; negative entries are then suppressed, but the qualitative features and the conclusions would remain the same. Also a smooth cutoff of the low eigenvalue contributions would not alter the result by much.

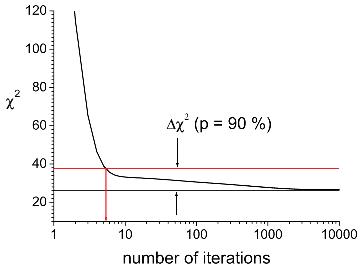

Fig. 5: χ^2 as a function of the number of iterations. Stopping at 5 iterations corresponds to a p-value of about 0.9.

4.2 Method 2: Poisson maximum likelihood fit with penalty regularization

Here we fit the contents of the true histogram by maximizing the log-likelihood

$$\begin{aligned}\ln L &= \ln L_{stat} - R \\ \ln L_{stat} &= \sum_{i=1}^{M} \hat{d}_i \ln\left(\sum_{j=1}^{N} A_{ij}\theta_j\right) - \sum_{j=1}^{N} A_{ij}\theta_j \\ R &= r\sum(2\theta_i - \theta_{i-1} - \theta_{i+1})^2\end{aligned}$$

which consists of the statistical term and a regularization term which is related to the local curvature. The χ^2 statistic is evaluated for different regularization constants r using Gaussian errors.

Fig. 4 shows some results of this method. Negative entries are excluded but the dependence of the nominal errors on the regularization strength of course remains. The p-values are indicated. A value around 90% corresponds to a reasonable smoothing. As in the general case, we do not know the true distribution, it is difficult to fix the constant r if it cannot be done with a p-value.

Curvature regularization which favors linear distributions is popular, but other penalty functions can be chosen. For a nearly exponential distribution, deviations from linearity of $\log\theta$ could be penalized. The application of penalty functions smooths the distribution even when the smearing matrix is diagonal.

4.3 Method 3: Iterative unfolding, stopping the iteration

Iterative unfolding has become popular because there exists a simple to use program by D'Agostini. Iterative unfolding is an old concept [4]. D'Agostini has re-invented it and interpreted it in the context of Bayesian statistics [5]. The iteration procedure can however also be viewed as a simple mathematical algorithm to invert the matrix equation $d = A\theta$. The method is explained in the Appendix. It is so simple that it can be coded in a few lines. The convergence is very fast at the beginning and suddenly slows down (see Fig. 5). The computation is fast; 10000 iterations were obtained in 1 minute with a standard laptop computer. They produced results almost indistinguishable from an least square or maximum likelihood fit (see Fig. 6). Quite good agreement of the result with the data is usually obtained after a few steps. As with other smoothing methods it is reasonable to fix the regularization strength, e.g. to stop the iteration with the p-value criterion. To compute the p-value we have to determine the minimal χ^2 either by a separate fit or by estimating its value from a large number of iterations.

The nominal errors of the unfolded distribution could in principle be obtained by error propagation, but this is extremely tedious as the analytic relation due to the iteration sequence is extremely complicated. Anyway, the nominal errors are rather useless.

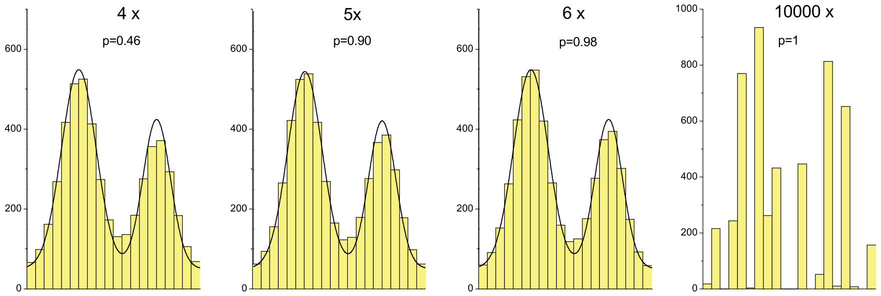

Fig. 6: Iteratively unfolded distribution for different numbers of iterations.

5 Error assignment to the graphical representation

The fitting methods produce error estimates automatically. For other methods the uncertainties can be obtained by the usual error propagation, but these nominal errors depend on the strength of the regularization which on the other hand is unrelated to the statistical accuracy of the data. A strongly regularized distribution may exhibit even smaller diagonal errors than the distribution before the convolution, i.e. smaller than the square root of the number of entries. This is unacceptable, and misleading (see the fake bump in the Alice experiment shown in the contribution to this conference by Gross-Oetringhaus) as we loose information by smearing. We should present errors which are useful in that they indicate whether functions are compatible with the data or not and do not depend on data manipulations.

A sensible graphical presentation of the unfolding result where the values but not the errors depend on the regularization is the following [1]: For each unfolded bin j we attribute a relative statistical error $\delta\theta_j/\theta_j$ equal to one over the square root of the number of observed events associated to that bin. This is equal to one over square root of θ_j if there are no acceptance corrections. A horizontal bar indicates the experimental resolution [1].

6 Summary and recommendations

Experimental results should be published such that 1. the data can be compared to and combined with those of other experiments in such a way that the combined result exhibits smaller statistical uncertainties and superior resolution in the smeared variable, 2. theoretical predictions can be tested, 3. all predictions that are compatible with the data are admitted. These requirements can only be satisfied when the experimental data are published without explicit regularization. When we compare a theory to the measurement, we should fold the theoretical prediction and compare it to the raw data.

A graphical representation with wide bins, such that oscillations are damped, is recommended. In any case the number of bins should be less than twice the effective number of degrees of freedom. If an explicit regularization is applied, we have to be aware that smoothing introduces constraints and modifies the experimental information. We propose to fix the regularization strength using a p-value criterion, which guarantees that the regularized distribution is compatible with the observed data. The errors that are assigned to the unfolded histogram have to be independent of the regularization.

All smoothing approaches that fulfill these requirements are acceptable, however, efficient methods reduce essentially the fluctuations between adjacent bins. All three approaches studied above in a simple example produce sensible and very similar results. Regularization with a penalty term is especially transparent. Iterative unfolding is simple and the smoothing prejudice is included in a flexible way. The truncation of singular value components is mathematically attractive and singular value decomposition (SVD) provides insight into the problem of unfolding.

Appendix: Iterative least square fit

The relation $\hat{d} = A\hat{\theta}$ can be solved iteratively, provided the response matrix A is positive definite. The idea behind the iteration algorithm is the following [1]. Starting with a preliminary guess of $\hat{\theta}^{(0)}$, the corresponding prediction for the observed distribution $d^{(0)}$ is computed. It is compared to $\hat{d}$ and for a bin i the ratio $\hat{d}_i/d_i^{(0)}$ is formed which ideally should be equal to one. To improve the agreement, all true components are changed in proportion to their contribution $A_{ij}\theta_j^{(0)}$ to $d_i^{(0)}$ This procedure when iterated corresponds to the following equations: The prediction $d^{(k)}$ of the iteration k is obtained in a *folding step* from the true vector $\theta^{(k)}$:

$$d_i^{(k)} = \sum_j A_{ij}\theta_j^{(k)} \ .$$

In an *unfolding step*, the components $A_{ij}\theta_j^{(k)}$ are scaled with $\hat{d}_i/d_i^{(k)}$ and added up into the bin j of the true distribution from which it originated:

$$\theta_j^{(k+1)} = \sum_i A_{ij}\theta_j^{(k)} \frac{\hat{d}_i}{d_i^{(k)}} / \varepsilon_j \ .$$

Dividing by the efficiency ε_j corrects for acceptance losses.Empirically, it has been shown that with increasing number of iterations, the result converges to the maximum likelihood fit result for Poisson distributed errors [4].

In D'Agostini's Bayesian approach (see, e.g., Bierwagen, these Proceedings), the same iteration sequence is applied, however, between each iteration the unfolded distribution is smoothed by a polynomial fit. The details of the smoothing step are left to the user. Convergence is implied by the method. The degree of smoothing depends on the intermediate smoothing algorithm. Furthermore, prior densities are introduced for the parameters of the multinomial and the Poisson distributions that are used in the evaluation of the uncertainties. It is not obvious that the priors have a noticeable effect. The reliability of the error estimates is unclear.

Acknowledgment

I would like to thank the organizers and especially Louis Lyons for the invitation to this conference with an interesting program in a pleasant atmosphere. I would like to thank Gerhard Bohm, my referee Gero Flucke and Louis Lyons for a careful reading of the manuscript and for valuable comments.

References

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

[2] P.C. Hansen, *Discrete Inverse Problems – Insight and Algorithms*, SIAM Fundamentals of Algorithms series, Philadelphia (2010).

[3] L. Lyons, private communication.

[4] L.B. Lucy, *An iterative technique for the rectification of observed distributions* Astronomical Journal 79 (6) (1974) 745; 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 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 (1987) 371.

[5] G. D'Agostini, *A multidimensional unfolding method based on Bayes' theorem*, Nucl. Instr. and Meth. A 362 (1995) 48, G. D'Agostini, *Improved iterative Bayesian unfolding*, arXiv:1010.632v1 (2010).