

SLUO Lectures on Statistics and Numerical Methods in HEP

Lecture 9: Unfolding

Roger Barlow
Manchester University
30th August 2000

1. The problem

You are interested in the distribution of events in some quantity x . Unfortunately x is not directly measurable, the best you can obtain is some ‘smeared’ quantity y . This can be written in terms of functions or matrices

$$g(y) = \int A(y, x) f(x) dx \quad g_i = \sum_j A_{ij} f_j \quad (1)$$

A is the folding of the wanted $f(x)$ distribution into the observed $g(y)$ distribution. It includes acceptance and resolution. There are lots of examples.

- Smearing of acoplanarity in $Z^0 \rightarrow \mu^+ \mu^-$.
- Observed and true charged multiplicity in events.
- Measured and true mass of W particles decaying to jets
- measured mass of two pions (in the ρ region)
- Visible energy and true energy in photon-photon collisions.

The ‘unfolding problem’ is the problem of getting from the observed histogram of g_i values – call it \vec{g} – to an estimate of the original values \vec{f} .

A is completely understood (probably from Monte Carlo). You might think this was therefore an easy problem. Dream on!

1.1 Correction Factors – a disaster

A simple approach is to evaluate correction factors from the Monte Carlo. If the progress from ‘true’ quantities to fully simulated ones multiplies the content of bin i by C_i , this is recovered by dividing the observed data bin i by C_i .

This is horrible. The data will tend to follow the MC that gave you the correction factors. It can only be justified if the smearing process is due to losses (acceptance) and there is no bin-to-bin (resolution) movement.

Example 1: Suppose there are just 2 bins. Your (Standard model, perhaps?) MC gave 75 events in bin 1 and 25 in bin 2 at the true level, changing to 50 and 50 after detector simulation. You observe 5 and 5 in the real data, so you ‘correct’ to 7.5 and 2.5.

The fact that bin 2 is corrected upwards shows that this is not just a variable efficiency. What is probably happening is that your detector is smearing the information so much that the bin is completely random. So your data really tells you nothing - yet after ‘correction’ it gives precise detail, consistent with the MC.

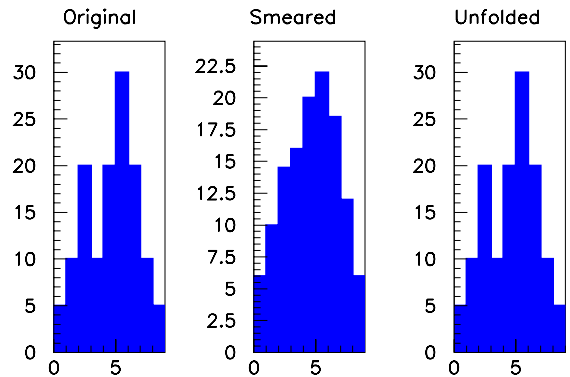


Figure 1: Folding and unfolding

Figure 1 shows a ‘typical’ ideal distribution, the effects of this smearing, and the result of multiplying this distribution by the inverse smearing matrix, getting back where one started.

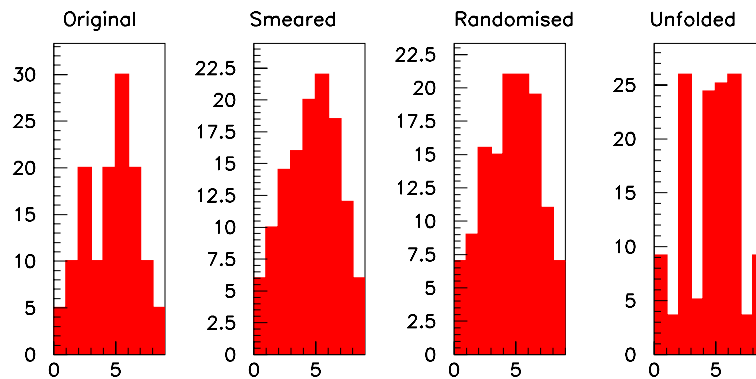


Figure 2: Folding and unfolding (continued)

Figure 2 starts with the same distribution and smears it. It is then ‘randomised’ as would happen in reality (actually more tamely: ± 1 was added to each bin alternately) and the resulting histogram unfolded using the inverse of the matrix. The horrifying output is worse than useless.

1.3 The right approach

Why not? Because we know something more that we haven’t told the fit: that these are bins on a physically meaningful scale and we have good reason to believe that the contents should not oscillate wildly from one bin to the next. Including this information in the fit (in some way) is a process called *regularization*.

This proceeds by saying: there is a term which expresses the disagreement between the prediction $A\vec{f}$ and the data \vec{g} . This is often a χ^2 but in general is the log of a likelihood. There is also a term S which expresses the ‘spikiness’ of the distribution. We minimise

$$-\ln L + \alpha S \tag{3}$$

where α is the ‘regularisation term’ to be chosen by you. If it is zero then we can get spiky distributions that fit the data perfectly. Moderate α results in smoother distributions that don’t agree 100% with the observed data but are very close. Large α will smooth out everything. So part of the problem is the proper choice of α .

2. Matrix Methods

If the log likelihood is taken as a χ^2 , then the first term can be written (dropping a factor of 2)

$$\chi^2 = (A\vec{f} - \vec{g})V^{-1}(A\vec{f} - \vec{g}) \tag{4}$$

The bin contents will be independent so the matrix V is diagonal, however they are in general different (the errors are like \sqrt{n}) which matters in a minimisation. This can be taken care of by a suitable rescaling of the data (see e.g. Ref [1] section 5) so for present purposes we will write

$$\chi^2 = (A\vec{f} - \vec{g})(A\vec{f} - \vec{g}) \tag{5}$$

The matrix A can be diagonalised. It has a set of eigenvalues λ_i and orthogonal eigenvectors. Figure 3 shows the 9 eigenvectors for the previous smearing matrix, with their eigenvalues.

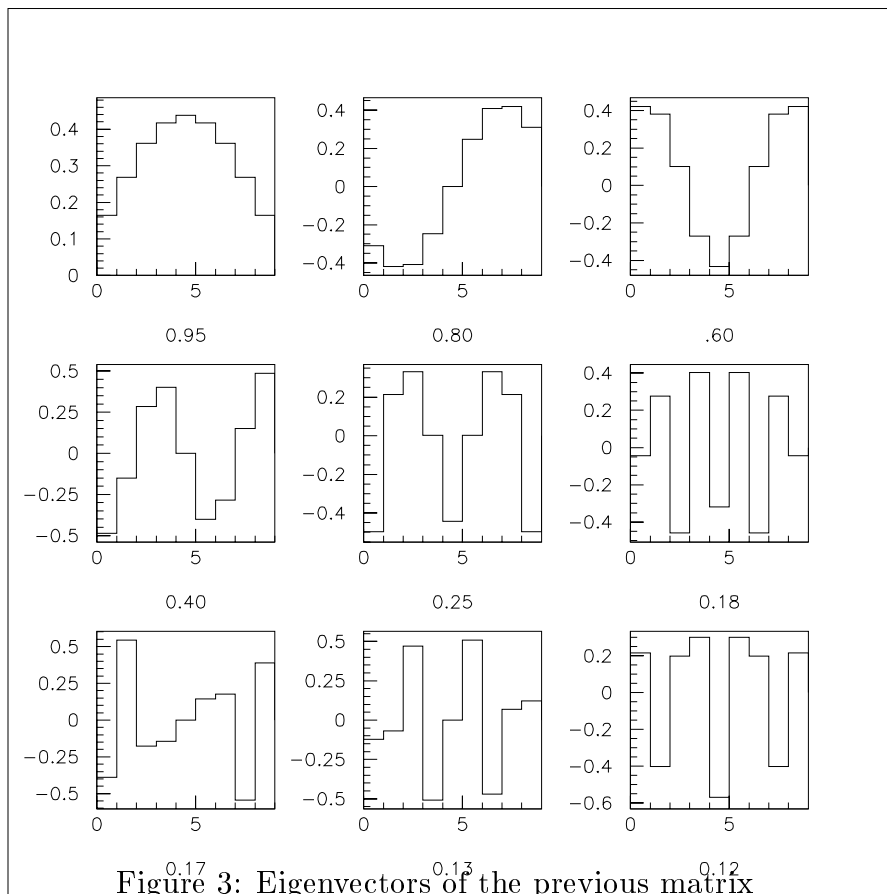


Figure 3: Eigenvectors of the previous matrix

These are the distributions whose shape is unaffected by the smearing, they just get multiplied by their eigenvalue. The unfolding $\hat{\vec{f}} = A^{-1}\vec{g}$ is equivalent to decomposing \vec{g} into the eigenvectors, dividing each by its appropriate eigenvalue, and adding them up again to give $\hat{\vec{f}}$. Clearly it is the eigenvectors with small eigenvalues that are giving trouble; they have rapid fluctuations and large contributions to $\hat{\vec{f}}$ because of division by small λ_i .

It is tempting to throw away the contributions from the low-eigenvalue eigenvectors. However this leads to periodic fluctuations (the ‘Gibbs Phenomenon’).

2.1 Finding Eigenvalues and Eigenvectors.

You were taught to find eigenvalues by finding the N solutions to the N^{th} order secular equation. The easy way is to use *Jacobi Rotation*.

- 1 Find the off-diagonal element of largest magnitude. Say it's A_{ij}
- 2 Calculate $z = \frac{A_{jj}-A_{ii}}{2A_{ij}}$
- 3 If z is positive take $\theta = \tan^{-1}(z - \sqrt{z^2 + 1})$, if negative $\theta = \tan^{-1}(z + \sqrt{z^2 + 1})$,
- 4 Rotate the matrix by an angle θ in the ij plane. i.e. postmultiply by a matrix which is the unit matrix except $R_{ii} = R_{jj} = \cos\theta$, $-R_{ij} = R_{ji} = \sin\theta$, and premultiply by its transpose. (A little thought can save a lot of arithmetic here as only the elements of rows and columns i and j are affected.) The rotated element A'_{ij} is zero.
- 5 Repeat until the largest off-diagonal element is negligible.

When this has converged – and it will – the matrix has the eigenvalues down the trace. If you keep a running product of the rotation matrices you have the eigenvectors.

Of course, there are lots of excellent library routines that will do this for you.

2.2 Tikhonov regularization

Our requirement that the distribution be smooth can be translated into a requirement that the total squared derivative (of some order) be small. For example, the first derivative gives (for equal bin widths)

$$S_1 = \sum_{i=1}^{N-1} (f_i - f_{i+1})^2$$

the second

$$S_2 = \sum_{i=2}^{N-1} (2f_i - f_{i+1} - f_{i-1})^2$$

and the third

$$S_3 = \sum_{i=2}^{N-2} (-f_{i-1} + 3f_i - 3f_{i+1} + f_{i+2})^2$$

S_2 is often chosen: distributions with large first derivatives do occur and are not objectionable.

This can be written

$$S_2 = \widetilde{C} \vec{f} C \vec{f} \tag{6}$$

with

$$C = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots \\ 1 & -2 & 1 & 0 & \dots \\ 0 & 1 & -2 & 2 & \dots \\ 0 & 0 & 1 & -2 & \dots \\ \vdots & & & & \end{pmatrix} \tag{7}$$

We are trying to estimate \vec{f} by minimising

$$(\widetilde{A} \vec{f} - \vec{d})(\vec{f} - \vec{d}) + \alpha \widetilde{C} \vec{f} C \vec{f}$$

and the result will depend on α . However you don't have to repeat the minimisation for every values of α that might be of interest. Suppose we work with $\vec{f}' = C\vec{f}$ and $A' = AC^{-1}$. The function becomes

$$(A'\vec{f}' - \vec{g})(A'\vec{f}' - \vec{g}) + \alpha f'^2$$

Now we diagonalise the A' matrix with some matrix R . $A'' = RA\tilde{R}$ is a diagonal matrix with eigenvalues λ_i down the diagonal. With $\vec{g}' = R\vec{g}$ and $\vec{f}'' = R\vec{f}'$ the χ^2 term becomes

$$(A''\vec{f}'' - \vec{g}')(A''\vec{f}'' - \vec{g}')$$

which (being diagonal) is just $\sum_i (\lambda_i f''_i - g'_i)^2$. Because this diagonalisation is just a rotation, the length of \vec{f}' and \vec{f}'' are the same, and the quantity to be minimised is just

$$\sum_i (\lambda_i f''_i - g'_i)^2 + \alpha f''^2$$

giving solutions

$$f''_i = \frac{g'_i \lambda_i}{\alpha + \lambda_i^2} \tag{8}$$

You can see from this how the low-eigenvalue components are suppressed by α , like a low-pass filter. $\alpha = 0$ gives the standard components - over - eigenvalue solution. The desired f_i are obtained from the f''_i by reversing the rotation and applying C^{-1} .

Don't try this at home, folks! If you want to do this seriously, you are strongly advised to use a package, written by folk who've taken care of all the difficulties I've told you about, and a whole lot more that I havn't.

2.3 Blobel

Blobel's classic unfolding method (Ref [2]) has been used for many years, particularly in two photon physics. The package is available and widely used, though most people treat it as a black box. Some of the technology is such as to make it unnecessarily complicated (e.g. he fits using B-splines to ensure smoothness.)

2.4 Guru

Höckler and Kartvelishvili (Ref [1]) have extended and simplified the method and generally made it more user-friendly. They consider the possibility of having input and output distributions with different numbers of bins (i.e. A non-square), and have a neat method that automatically incorporates the errors in A due to Monte Carlo statistics.

They suggest that the g''_i - the components in the data of the eigenvectors - be examined, and α chosen such as to suppress 'small' values, where 'small' means 'not statistically different from zero.' (In their notation these are the d_i and τ .) Other approaches and checks are also possible.

The reference contains several plots showing how the apparatus works, and an example of its use in practice can be found in Ref [3].

3. Maximum Entropy methods

We take quite a long step backwards.

3.1 Information

Suppose a system is giving you information. It comes in the form of characters, and to keep things simple the probability P of any particular character E is independent of what came previously. (It is a discrete memoryless source.)

Some characters are more common than others. These are not very interesting: they do not contain much information. The low-probability scarcer characters are more informative.

Example 3: A fire alarm sends either a 0 or a 1 at regular intervals. ‘0’ indicates no fire. ‘1’ indicates there is a fire. The character ‘1’ is low-probability and informative.

We would like to define a quantity ‘information’ such that the information of low-probability events is higher, and also that the total information of two separate messages is the sum of the information carried by each separately.

$$I(E_1 + E_2) = I(E_1) + I(E_2)$$

This is fulfilled by

$$I(E) = -\ln P(E) \tag{9}$$

This is the *Shannon Information* of an event: minus the log of the probability.

3.2 Mean information and Entropy

If you have a whole set of possible characters $E_1...E_N$, low-probability events carry more information, but occur less often. The mean information carried by an event is

$$\langle I \rangle = S = \sum_{i=1}^N -P_i \ln P_i \tag{10}$$

Now suppose there is a system - a bunch of particles which can all be in particular states. If I measure the state a particle is in, that gives me information. If I set up the system precisely (say, put them all in the 1st excited state) and then measure it straight away, this yields no information because I know what state every particle is in. The mean information is 0 because P is 0 for all the states except the first excited state, for which $\ln P$ is 0.

If I then let the system interact with itself for a bit, particles will gain and lose energy, I will lose my omniscience and a measurement will convey information.

It seems fair to assume that this process will continue as far as it can. The system will tend to a configuration in which the average result of a measurement will contain the maximum (possible) information.

$$\sum_i -P_i \ln P_i \quad \text{increases to a maximum}$$

hence the mean information is also known as the Entropy.

You can actually derive the whole of Statistical Mechanics from this starting point. You maximise Entropy - subject to the constraints that the total probability is 1, and that the mean energy $\sum_i P_i \mathcal{E}_i$ is prescribed as a constant.

3.3 MaxEnt

The ‘Maximum Entropy’ principle uses the entropy as a regularisation term. The probability of an event being in bin i is f_i/N . So we minimise (following Eq. 3 and Eq 10)

$$-\ln L + \alpha \sum_i f_i/N \ln(f_i/N) \tag{11}$$

where N is the total number of entries in the histogram, and L is still the χ^2 or other likelihood measure expressing the agreement between observed data and the prediction from the unfolded \vec{f} .

Effectively this is a term which prefers all the bin contents to be the same. It does not contain any prescription about adjacent bins, so it is more suitable than Tikhonov regularisation in instances where the true distribution really does have sharp gradients. Typical examples occur in astronomy (where bright stars stand out from black backgrounds) and image processing (everyday objects have sharp edges). The MaxEnt term prefers all bins to have the same number of entries, but the preference is (for reasonable α) not too strong and the occasional blip can be tolerated. It generalises trivially to distributions with more than one dimensiona.

3.4 The Bayesian Motivation

In this language, Bayes’ theorem runs

$$P(\vec{f}|\vec{g}) \propto P(\vec{g}|\vec{f}) P(\vec{f})$$

where $P(\vec{g}|\vec{f})$ is the same resolution-folded matrix term as earlier, or equivalent, and $P(\vec{f})$ is our prior knowledge about the ideal distribution.

If we assume complete ignorance, then we suppose that N entries are going to be placed in this histogram completely randomly between the M bins. The probability of a particular histogram distribution \vec{f} is thus proportional to the number of ways that this set of values could have been produced by these entries: the number of permutations of this particular distribution

$$\frac{N!}{f_1!f_2! \dots f_M!}$$

Taking the logarithm and using Stirling’s approximation gives

$$-\sum_i f_i \ln(f_i/N)$$

This actually gives the probability for any distribution \vec{f} which is more than we can handle. We settle for the distribution with the maximum probability

Minimise

$$-\ln L + \sum_i f_i \ln(f_i/N) \tag{12}$$

This is in accord with Eq. 11, with α prescribed as N ,

Experience shows that this is a little too strong. It tends to flatten everything. (Perhaps insisting that the prior is uniform is a bit strong.) In practice you use the Bayesian motivation to get this far and then allow yourself to be flexible by including a variable α .

3.5 Cross Entropy

The entropy can be extended to include cases where there is prior knowledge of the distribution – either on theoretical grounds, or because the bin widths are different. Equation (10) becomes

$$S = - \sum_i P_i \ln(P_i/Q_i) \tag{12}$$

where Q_i is the prior knowledge about bin i . The normalisation of the Q_i , and a factor of M which appears in the denominator with some authors but not others, really don't matter as they just give an additive constant. (There is also an author-dependent minus sign, which does matter.) This is also known as the Shannon-Jaynes Entropy or the Kullback number.

3.6 Use in practice

The entropy is unamenable to differentiation, so unlike the Matrix Methods, direct solution can't be found. One normally works iteratively, starting at the maximum-entropy point (i.e. all bins having equal contents) and then proceeds to minimise (or maximise) the sum of both terms. This presents the usual problems of multidimensional fitting.

The normalisation generally has to be strictly enforced. For Tikhonov regularisation this is not generally necessary meaning that the 'solution' may correspond to a different number of observed events from those you actually observed, but this effect is small and not a problem. For MaxEnt, because any increase in the numbers drives that entropy term up, this introduces a bias and you have to take pains to work specifically in the subspace that gives the right total. On the plus side, the MaxEnt term ensures that the bin contents are all non-negative: Tikhonov regularisation can give negative bin contents, though the values are usually small and can be discarded.

It has been used in reconstructing photon energies and directions from the OPAL calorimeter, and there are no doubt many other potential applications.

4. Final thought

“You can get it wrong, and still you think it's all right”

Lennon and McCartney, quoted by Blobel in ref [2]

[1] A. Höcker and V. Kartvelishvili. SVD approach to data unfolding. Nucl. Instr. & Meth. A372 (1996) 469

[2] V. Blobel, Unfolding methods in HEP experiments. DESY 84-118 (1984)

[3] ALEPH (R. Barate et al). Measurement of the Spectral Functions of Vector Current Hadronic Tau decays. CERN-PPE 97-013 (1997)

[4] Glen Cowan. Chapter 11 of *Statistical Data Analysis*, OUP 1998