

UNFOLDING WITH SYSTEM IDENTIFICATION

N.D. GAGUNASHVILI

*University of Akureyri, Faculty of Information Technology, Borgir, v/Nordurhlóð, IS-600 Akureyri, Iceland
E-mail: nikolai@unak.is*

A procedure for unfolding the true distribution from experimental data is presented. Methods of system identification are applied for the creation of a model of a transformation of the true distribution to the experimentally measured distribution. A priori information about the true distribution which is known from theory or previous experiments is used. The stability of the result of the unfolding is obtained by a sensible binning and by application of D-optimization. In this paper it is shown how to decrease the bias of the unfolded distribution by introducing the X^2 selection criteria for distributions used for system identification. Application of the D-optimization and the Least Squares Method allow us to minimize the statistical errors of the unfolded distribution. The unfolding procedure may be applied for detectors with a linear or nonlinear transformation of a true distribution into the experimentally measured one. The dimensionality of the solved problem can be arbitrary. The procedure can be applied for solving the unfolding problem with both smooth and non-smooth solutions. This method does not require a large amount of Monte-Carlo simulations of the experiment.

1 Introduction

An experimentally measured distribution differs from the true physical distribution due to limited acceptance and finite resolution of a set-up. To obtain a physical distribution an unfolding procedure is applied^{1 2 3 4}. The unfolding problem is an under-specified problem. Any approach to solve the problem requires a priori information about the solution. Different methods of unfolding differ, directly or indirectly, through the use of this a priori information.

In⁵ an approach to an unfolding problem related to methods of system identification is presented. To obtain a stable solution of an unfolding problem, information about the shape of the distribution to be measured is used for system identification. This paper further develops the ideas presented in⁵. D-optimization which is used in the theory of experimental design is applied to minimize the statistical errors of the unfolded distribution. The X^2 selection criterion is introduced for a set of distributions used for system identification; this criteria minimizes the bias of the solution.

2 Main equation

In this work we will use the linear model for a transformation of a true distribution to the measured one

$$\mathbf{f} = \mathbf{P}\boldsymbol{\phi} + \boldsymbol{\epsilon}, \quad (1)$$

where \mathbf{f} is an m -component column vector of an experimentally measured histogram content, \mathbf{P} is an $m \times n$ matrix, with $m \geq n$, $\boldsymbol{\phi}$ is an n -component

vector of some true histogram content and $\boldsymbol{\epsilon}$ is an m -component vector of random residuals with expectation value $\mathbf{E}\boldsymbol{\epsilon} = \mathbf{0}$ and a diagonal variance matrix $\boldsymbol{\Sigma} = \text{Var}\boldsymbol{\epsilon} = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$, where σ_i is the statistical error of the measured distribution for the i th bin. The linear model (1) is reasonable for the majority of set-ups. It is only an approximate model for set-ups with non-linear transformation from a true distribution into the measured distribution.

A Least Squares Method can give an estimator for the true distribution $\boldsymbol{\phi}$,

$$\hat{\boldsymbol{\phi}} = (\mathbf{P}'\boldsymbol{\Sigma}^{-1}\mathbf{P})^{-1}\mathbf{P}'\boldsymbol{\Sigma}^{-1}\mathbf{f} \quad (2)$$

where $\hat{\boldsymbol{\phi}}$, the estimator, is the *unfolded distribution*, and the full matrix of errors of the unfolded distribution is given by

$$\text{Var}\hat{\boldsymbol{\phi}} = (\mathbf{P}'\boldsymbol{\Sigma}^{-1}\mathbf{P})^{-1}. \quad (3)$$

3 System identification and regularization

To realize the scheme described in the previous section, the matrix \mathbf{P} must be defined. This problem can be solved using system identification methods⁷. System identification may be defined as the process of determining a model of a dynamic system using observed input-output data. In our case it is the model of transformation of a true physical distribution into the experimentally measured distribution, represented by the matrix \mathbf{P} . The Monte-Carlo simulation of a set-up can be used to get input-output data. Control input signals are used for system iden-

tification . The most popular choice is to use impulse control signals.

An impulse input control signal is a generated distribution in which the histogram has only one bin with non-zero content. For the model (1) there are n different impulse inputs that can be presented as the diagonal matrix $\Phi^c = \text{diag}(\phi_{11}^c, \dots, \phi_{nn}^c)$, where each row contains the content from a generated histogram. Let us denote corresponding values of i th component of the reconstructed vector (output) as $\mathbf{f}_i^c = (f_{i1}^c f_{i2}^c \dots f_{im}^c)'$. Each element of the i th row of the matrix

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ \dots & \dots & \dots & \dots \\ p_{i1} & p_{i2} & \dots & p_{in} \\ \dots & \dots & \dots & \dots \\ p_{m1} & p_{m2} & \dots & p_{mn} \end{pmatrix},$$

can be found from the equation

$$\mathbf{f}_i^c = \Phi^c \mathbf{p}_i, \quad (4)$$

where $\mathbf{p}_i = (p_{i1} p_{i2} \dots p_{in})'$, and $p_{ij} = f_{ij}^c / \phi_{jj}^c$. Equation (2), with the matrix \mathbf{P} calculated this way, gives a highly fluctuating unfolded function with large statistical errors. Also, it is possible that the matrix $\mathbf{P}'\Sigma^{-1}\mathbf{P}$ is singular, in which case a solution does not exist.

To regularize the solution of the unfolding problem, let us use for system identification not an impulse control distribution, but rather a priori distributions that may be known from theory, or from some other experimental data.

Assume we have q control generated distributions, and now present them as a $q \times n$ matrix

$$\Phi^c = \begin{pmatrix} \phi_{11}^c & \phi_{12}^c & \dots & \phi_{1n}^c \\ \phi_{21}^c & \phi_{22}^c & \dots & \phi_{2n}^c \\ \dots & \dots & \dots & \dots \\ \phi_{q1}^c & \phi_{q2}^c & \dots & \phi_{qn}^c \end{pmatrix},$$

where each row represents a generated histogram content. For each i th row of the matrix \mathbf{P} we can write the equation

$$\mathbf{f}_i^c = \Phi^c \mathbf{p}_i + \boldsymbol{\xi}_i, \quad (5)$$

where $\mathbf{p}_i = (p_{i1} p_{i2} \dots p_{in})'$, \mathbf{f}_i^c is a q -component vector of reconstructed i th bin content for different generated control distributions, and $\boldsymbol{\xi}_i$ is a q -component vector of random residuals with expectation value $E \boldsymbol{\xi}_i = \mathbf{0}$ and a diagonal variance matrix $\Delta_i = \text{Var} \boldsymbol{\xi}_i = \text{diag}(\delta_{i1}^2, \dots, \delta_{iq}^2)$, where δ_{ij} is the

statistical error of the reconstructed distribution for the i th bin and the j th control generated distribution. A Least Squares Method gives an estimator for $\mathbf{p}_i, i = 1, \dots, m$

$$\hat{\mathbf{p}}_i = (\Phi^{c'} \Delta_i^{-1} \Phi^c)^{-1} \Phi^{c'} \Delta_i^{-1} \mathbf{f}_i^c. \quad (6)$$

Columns of the matrix Φ^c can correlate with each other. This means that transformation of the control generated distribution to the i th bin of reconstructed distribution can be parametrized by the subset of elements of the row \mathbf{p}_i . Elements of the row that do not belong to the subset are set to 0. Moreover, there can be more than one subset that describes this transformation in a sufficiently good manner. Thus for each i th reconstructed bin we will have the set of N_i candidate rows, and for all reconstructed bins the set of $N_1 \times N_2 \times \dots \times N_m$ candidate matrices \mathbf{P} . We need to choose a matrix \mathbf{P} that is good, or optimal, in some sense. The most convenient criterion in our case is D-optimality⁸ that is related to the minimization of

$$\det(\mathbf{P}'\Sigma^{-1}\mathbf{P})^{-1} = \det(\text{Var}(\boldsymbol{\phi})). \quad (7)$$

There are many algorithms and programs of minimization (7). The matrix \mathbf{P} that minimizes function (7) gives us a stable solution of the unfolding problem (2) with minimal volume of the confidence ellipsoid. Further improvement of the quality of the solution can be achieved by introducing the selection criteria described below.

A control generated distribution has a corresponding reconstructed control distribution that can be compared with the experimentally measured distribution using a χ^2 test⁶. Let us take for identification a generated control distribution that has a corresponding reconstructed distribution satisfying a $X^2 < a$ selection criteria. The statistic X^2 is calculated to test the compatibility of the experimental distribution with the reconstructed control distribution⁶. The parameter a defines how close the set of reconstructed control distributions is to the experimental distribution. A decrease in parameter a represents a decrease in systematic and statistical errors of the solution.

4 The unfolding procedure

In this section a description of the complete unfolding procedure is presented. The procedure can be

divided into four parts: initialization, system identification, solution of the basic equation, and test of goodness-of-fit.

Initialization

Define the binning for the experimental data.

The strategy in selecting the size of the bins is to start with large bin sizes, then increase the number of bins incrementally and stop the process when the value of the determinant of the complete matrix of errors of unfolded distribution stops decreasing.

Define the binning for the unfolded distribution.

The way to choose the bin size is to pick a reasonably large size of bin for the first step, then decrease the size of bins on further steps and stop this process before the correlation between adjusted bins becomes too big. The number of bins of an unfolded distribution, n , must be lower than the number of bins for the experimentally measured distribution, m , due to the fact that we use the Least Squares Method for the solution of the main equation.

System identification

Choose a set of control generated distributions.

Control generated distributions for the set must be chosen with lowest possible values of the X^2 statistics. A second iteration can be made to find a better set of control distributions. The method of the re-weighting of events can be used in this case. The number of generated distributions must be greater than the expected number of non-zero elements in any row of matrix P (for reasons related to the use of the Least Squares Method).

Calculate the set of candidates for the matrix

P . A stepwise regression algorithm can be used for this calculation⁹. The first element in the stepwise algorithm can define a candidate row. To obtain as many variants as possible of each row, each element of the matrix is used as a first element in the stepwise algorithm.

Calculate the D-optimal matrix P . On the first step, matrix P is chosen randomly from the set of candidates. After this, optimization can be done by Fedorov's reliable EA algorithm⁸. In the majority of cases this algorithm finds a matrix that has a global minimum for $\det(P'\Sigma^{-1}P)^{-1}$. The optimization procedure can be repeated with another randomly chosen matrix to be sure that the minimum is global.

Solution of the basic equation

Calculate the unfolded distribution Eq. (2) *with the full matrix of errors* Eq. (3). The correlation

matrix calculated from the full matrix of errors can give hints for an improved binning of the unfolding distribution. For example, if the correlation between two adjacent bins is high, they should be combined.

Test of goodness-of-fit

Fit unfolded distribution, and then use the fit to generate a new 'experimental' distribution (including effects of resolution and acceptance), to compare with the real data. This is the only objective test of goodness-of-fit of the unfolding procedure and it should be done with an analysis of the studentised residuals⁶.

5 A numerical example

The method described above is now illustrated with an example taken from¹. We take a true distribution

$$\phi(x) = A_1 \frac{C_1^2}{(x - B_1)^2 + C_1^2} + A_2 \frac{C_2^2}{(x - B_2)^2 + C_2^2} \quad (8)$$

with parameters $A_1 = 2, A_2 = 1, B_1 = 10, B_2 = 14, C_1 = C_2 = 1$; x is defined on the interval $[4, 16]$. An experimentally measured distribution is defined as

$$f(x) = \int_4^{16} \phi(x') A(x') R(x, x') dx' \quad (9)$$

where the acceptance function $A(x)$ is

$$A(x) = 1 - \frac{(x - 10)^2}{36} \quad (10)$$

and

$$R(x, x') = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - x')^2}{2\sigma^2}\right) \quad (11)$$

is the detector resolution function with $\sigma = 1.5$. The true distribution, acceptance and the resolution functions are shown in Fig. 1a. A histogram of the measured distribution f was obtained by simulating 10^4 events with $m = 90$ bins, and is shown in Fig. 1b.

For the true distribution histogram we choose thirty equal size bins, i.e. $n = 30$. We use for the detector identification 100 distributions defined by formula (8) with parameters simulated according to uniform distributions on the intervals :

$[1, 3]$ for A_1 ; $[8, 12]$ for B_1 ; $[0.5, 1.5]$ for C_1 ;
 $[0.5, 1.5]$ for A_2 ; $[10, 18]$ for B_2 ; $[0.5, 1.5]$ for C_2 .

Each distribution is represented by a histogram with 10^4 events. The first example is calculated without X^2 selection cut. Fig. 2a shows 30

of the 100 control distributions used for identification, and Fig. 2b shows the unfolded distribution and the true distribution as a solid line.

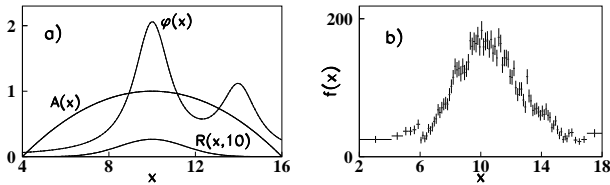


Figure 1. An example of the true distribution $\phi(x)$, the acceptance function $A(x)$, the resolution function $R(x,10)$ and of the measured distribution f (number of events per bin).

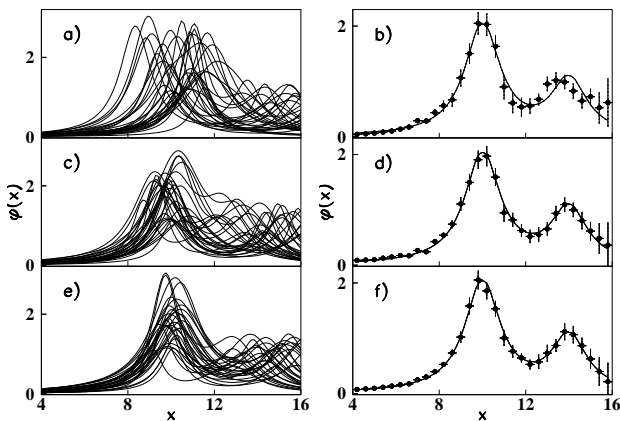


Figure 2. The first 30 control distribution generated for system identification and an unfolded distribution a),b) without X^2 cut; c),d) with $X^2 < 200$ cut; e),f) with $X^2 < 124$ cut.

Two other examples are calculated with $X^2 < 200$ selection criteria (significance level $p = 2.5 \cdot 10^{-10}$) (see Figs. 2c,d) and with $X^2 < 124$ cut (significance level $p = 10^{-2}$) (see Figs. 2e,f). The D-optimal determinant in the first case is equal to $2117^{-1/30}$, in the second case equal to $4445^{-1/30}$ and in the third one $4719^{-1/30}$. The average number of non-zero elements in the rows of matrix P in the last two cases is 4 and for first case 5. Notice that for a lower values of the X^2 cut we have a lower values of the determinant of the full matrix of errors and a lower systematic deviation of the unfolded distribution from true one. The errors seem a bit large in comparison with fluctuations of the unfolded distribution. The reason for these errors is the positive correlation between adjacent bins.

6 Conclusion

The main idea of the method presented in this paper is the use of a set of a priori distributions for system identification, i.e. the construction of the transformation matrix. For this set of distributions, we introduce a X^2 selection criteria, which permits us to decrease the possible bias of the procedure. D-optimization and the application of the Least Squares Method gives the stable solution with minimal statistical errors. The method of identification provides a linear approximation of the transformation of the true distribution into the measured distribution in case this transformation is non-linear. The method does not require a large amount of Monte-Carlo simulation of the experiment, because of the relatively low number of non-zero elements in the transformation matrix and because a re-weighting technique is used. The procedure has no restrictions due to the dimensionality of the problem. The method can be applied for solving the unfolding problem with a non-smooth solution. Based only on a statistical approach, the method has a good statistical interpretation.

Acknowledgments

The author would like to thank Dr. Mark O'Brien and Dr. J. Nystrom for a critical reading of this paper and making comments that improved the text.

References

1. V.P. Zhigunov, *Nucl. Instrum. Meth.* **216**, 183 (1983).
2. V. Blobel, *CERN 85-02*, 1985.
3. G. Zech, *DESY 95-113*, 1995.
4. A. Höcker, V.Kartvelishvili, *Nucl. Instrum. Meth. A* **372**, 469 (1996)
5. N. Gagunashvili, *Nucl. Instrum. Meth. A* **451**, 657 (1994).
6. N. Gagunashvili, χ^2 Test for the Comparison of Weighted and Unweighted Histograms, in these proceedings.
7. L. Ljung, *System Identification: Theory for User*, (Prentice Hall, 1999).
8. V.V. Fedorov, *Theory of Optimal Experiments*, (Academic Press, New York, 1972).
9. G.A.F. Seber and A.J. Lee, *Linear Regression Analysis*, (John Wiley & Sons Inc, 2003).