Unfolding in particle physics: a window on solving inverse problems

Francesco Spanò

Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom

DOI: http://dx.doi.org/10.3204/DESY-PROC-2014-02/52

A pedagogical exposition of unfolding techniques in particle physics is presented. Building on example analyses about top quark physics, the origin, the main building blocks and the crucial challenges of inverse, ill-posed problems in particle physics are discussed.

1 Unfolding foundations

In particle physics *unfolding* is the ensemble of statistical techniques used to solve what is defined as the *inverse problem*: infer an unknown distribution $f(y)$ for a variable y from the measured distribution $g(s)$ by using knowledge and/or assumptions on the probability distribution that links the observation to the "true" value.

The mathematical foundations of unfolding are intimately related to the description of the inverse problem provided by the Fredholm integral equation of the first type $g(\mathbf{s}) = \int_\Omega K(\mathbf{s},\mathbf{y}) f(\mathbf{y}) d\mathbf{y}$, where the true $f(\mathbf{y})$ distribution of the variable $\mathbf{y} = (y_1,..,y_J)$ is related to the measured or observed distribution $g(\mathbf{s})$ of the L-dimensional variable $\mathbf{s} = (s_1,..,s_L)$ by the convolution with the *kernel* function $K(\mathbf{s},\mathbf{y})$ over the subspace Ω of the J-dimensional space where $\mathbf{y}$ is defined. An illustrative example, shown in the cartoon of Figure 1, is the inversion of the measuring process for the invariant mass of the pair of top-antitop quarks ($t\bar{t}$, $m_{t\bar{t}}$) produced in proton-proton (pp) collisions at a center-of-mass energy ($\sqrt{s}$) of 7 TeV at the Large Hadron Collider (LHC) and reconstructed by the ATLAS detector.

In the very common one-dimensional case, the measured and the true distributions are approximated by histograms representing the values ν_i or μ_i, the expected number of counts in a given interval of real variables s or y respectively, according to the definitions $\nu_i = \int_{s_{i-1}}^{s_i} g(s) ds$ and $\mu_j = \int_{y_{j-1}}^{y_j} f(y) dy$, where the intervals of definition for s and y are divided in N and M subintervals respectively by a set of $(s_1,...,s_N)$ and $(y_1,...,y_M)$ values. The integral *kernel* $K(s,y)$ form is approximated by a response matrix $R(i,j)$ representing the probability that an event with a value of the y variable in bin j is observed as an event with a value of the s variable in bin i. The extended discretized one-dimensional form of the Fredholm equation is then written [5] as $E[n_i] = \nu_i = \sum_{j=1}^{M} R_{i,j}\mu_j + \beta_i$, whose vectorial compact form is $E[\mathbf{n}] = \boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$, where the response matrix $R(i,j)$ also includes the estimate of the reconstruction efficiency and $\boldsymbol{\beta}$ is the vector of the number of expected background events [1].

[1] Events that pass the selection requirements, but have different origin from the ones of interest.

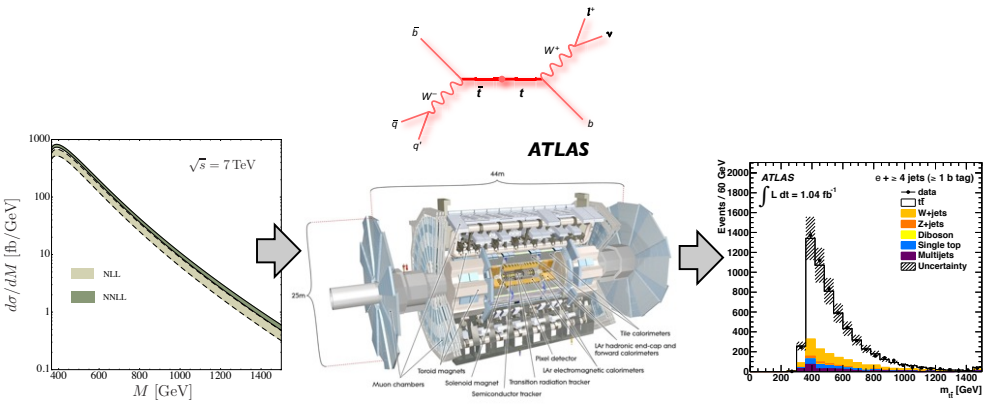

Figure 1: Scheme of evolution of the measurement of $m_{t\bar{t}}$. The predicted $m_{t\bar{t}}$ distribution [1] (left) for $t\bar{t}$ events produced in $\sqrt{s} = 7$ TeV pp collisions at the LHC is reconstructed [2] (right) after the top quark decay products are measured by the ATLAS detector [3] (middle). A Feynman diagram [4] shows the final state partons from the $t\bar{t}$ decay at leading order.

2 The art of matrix inversion: max. likelihood solution

The formal solution to $\boldsymbol{\nu} = R\boldsymbol{\mu}+\boldsymbol{\beta}$ is written as $\boldsymbol{\mu_{est}} = R^{-1}(\boldsymbol{\nu}-\boldsymbol{\beta})$, where R^{-1} is the inverse of R. This estimate for $\boldsymbol{\mu}$ can also be derived from the principle of maximum likelihood (ML). If one assumes (fairly generally) that the data are independent Poisson observations in each histogram bin, the corresponding likelihood is $\mathcal{L} = \nu_i^{n_i}\frac{e^{-\nu_i}}{n_i!}$ where $\boldsymbol{\nu} = \boldsymbol{\nu}(\boldsymbol{\mu})$ according to the discretized unfolding equation and n_i is the observed number of events in bin i. Consequently the maximum likelihood estimator for $\boldsymbol{\nu}$ obtained by imposing $\partial\log\mathcal{L}(\mu_i)/\partial\mu_i = 0\ \forall\ i$ is given by $\boldsymbol{\nu}_{ML} = \mathbf{n}$ and consequently the estimate of $\boldsymbol{\mu}$ is obtained as $\boldsymbol{\mu}_{ML} = R^{-1}(\boldsymbol{\nu}_{ML} - \boldsymbol{\beta}) = R^{-1}(\mathbf{n} - \boldsymbol{\beta}) = \boldsymbol{\mu}_{est}$.

Is this solution always working? An example shown in Ref. [5] reports a double-peaked true distribution for which the resulting ML estimate shows a multi-peaked shape with extremely large variances and very large anti-correlation between neighbouring bins: the estimate turns out to be very different from the known input. The response matrix R for this example is known to have sizable non-diagonal elements and the bin size of the histogram to be "inverted" is smaller than the detector resolution encoded in the model for event migrations. Figure 2 shows the generated "true" histogram $\boldsymbol{\mu}$ and the unfolded estimator $\boldsymbol{\mu_{est}}$. What is happening? The application of R^{-1} aims at restoring the original histogram. If the migrations are properly modelled, the inversion returns the correct values if the input data are the expectation vector $\boldsymbol{\nu}$ of the reconstructed bin contents. However the matrix inversion is applied to one instance vector of the data, $\mathbf{n}$, it is not applied to its expectation value $\boldsymbol{\nu}$. As a consequence, in a suggestively descriptive way, R "assumes" that the fluctuations in $\mathbf{n}$ are the residual of a real original structure diluted by the detection effects (and not of statistical origin) and uses the given input and the available model for migrations to reconstruct $\boldsymbol{\mu}$, i.e. it magnifies the fluctuations back into the result. Independently of the large fluctuations induced by the application of the matrix inversion the ML solution provides the unbiased estimator with the smallest variance [5].

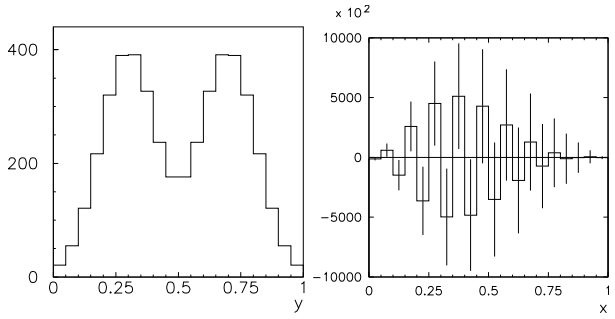

Figure 2: Examples of "true" distribution (left) ($\boldsymbol{\mu}$), the resulting estimate for $\boldsymbol{\mu}_{est}$ using the ML solution for a given assumption on resolution and efficiency (right, see text) [5]. The vectors $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\mathbf{n}$ and $\boldsymbol{\mu}_{est}$ are defined in the text.

3 From inside ill-posed problems to regularization

A detailed two-steps analysis of the discretized unfolding equation is outlined in Section 1.5 of Ref [6] and illustrates the link between fluctuations and instability of the ML solution by exposing the origin of instability in a quantitative manner. A synthetic description is reported here. The likelihood representing the unfolding problem takes the form $\mathcal{L} \propto e^{-\frac{1}{2}\chi^2(\boldsymbol{\mu},\mathbf{d})}$ and the ML solution coincides with a least squares estimate [6]. As a first step a transformation of variables that diagonalizes the generally non-diagonal $\chi^2(\boldsymbol{\mu},\mathbf{d})$ in the form $\frac{1}{2}\chi^2(\boldsymbol{\mu},\mathbf{d}) = (R'\boldsymbol{\mu} - \mathbf{d}')^T(R'\boldsymbol{\mu} - \mathbf{d}')$ provides a new vector $\mathbf{d}'$ and a new R' matrix that are written in terms of significances i.e. variables normalized to their uncertainties [6]. In the second step the resulting ML solution is written in terms of significances and parameters that are sensitive to fluctuations by using a *singular value decomposition* (SVD) of the R' matrix as follows: $\boldsymbol{\mu}_{est} = \sum_{i=1}^{min(N,M)} \frac{1}{\sigma_i}(\mathbf{u}_i^T\mathbf{d}')\mathbf{v_i}$, where $U = (\mathbf{u}_1,..,\mathbf{u}_N)$ and $V = (\mathbf{v}_1,..,\mathbf{v}_M)$ are unitary matrices written in terms of their column vectors and $\Sigma = U^T R' V$ is a diagonal matrix of (generally) dimensions $M \times N$ such that $\Sigma_{i,j} = \sigma_i$ for $i = j$, otherwise $\Sigma_{i,j} = 0$. The diagonal σ_i values are called *singular values* of the matrix R', they are not negative and can always be arranged in non-increasing order. The sensitivity to fluctuations associated with the ML solution can be quantified by the maximum ratio of the relative precision of the estimated solution $\boldsymbol{\mu}_{est}$ to the relative precision of the measured input vector $\mathbf{d} = \mathbf{n}$ - $\boldsymbol{\beta}$, defined as $c = \max_{\mathbf{d},\delta\mathbf{d}} \left(\frac{\delta\boldsymbol{\mu}_{est}/\boldsymbol{\mu}_{est}}{\delta\mathbf{d}/\mathbf{d}}\right)$ A large value for c implies instability due to small fluctuations in the input i.e. sensitivity to "noise" in the measurement. The quantity $c = c(R)$ is called the *condition* of the R matrix associated to the unfolding. It can be shown that $c(R) = \sigma_{max}/\sigma_{min}$ [6], so the *condition* of the matrix R can be read off from its SVD. Once the problem is described in terms of uncertainty normalized variables, the large sensitivity of the ML estimator to small, high-frequency-like fluctuations can be detected in the high condition number $c(R)$. In order to pose the problem more properly, it is then necessary to reduce the impact of the low significance, highly oscillating input components while preserving the information available in the remaining high significance, more stable components. The problem is then said to have been "regularized". As the ML estimator is unbiased, regularization inevitably leads to accepting a certain level of bias in exchange for a reduced variance. The bias is defined as the difference between the expected

value of the unfolded result and the true unmeasured expected value. It should be noted that, despite its large variances, the ML solution is fit for use in testing a model against a measurement, as long as the full corresponding covariance matrix is used.

The likelihood formulation of the unfolding problem quantifies the distance between the data vector $\mathbf{n}$ and the expectation vector $\boldsymbol{\nu}$. In order to filter out a certain amount of the high frequency components of the input and alleviate the sensitivity to large fluctuations (i.e."regularize" the solution), constraints on the initial likelihood can be imposed by adding Lagrange multipliers and describing the regularization as a maximization procedure for a new log-likelihood ϕ written as $\phi = \log\mathcal{L}(\boldsymbol{\mu}) + \tau S(\boldsymbol{\mu})$, where $\mathcal{L}(\boldsymbol{\mu})$ is the initial likelihood, $S(\boldsymbol{\mu})$ is called regularization function, τ is the regularization parameter to tune the strength of the constraints. In this explicit formalism the ingredients for the regularization of a given likelihood $\mathcal{L}(\boldsymbol{\mu})$ are the regularization function $S(\boldsymbol{\mu})$ and a prescription for τ.

A large number of different regularization schemes is available [6]. Examples of schemes used in particle physics include:

- Tikhonov schemes whose constraining function is the mean square of the k^{th} derivative of f(y): $S[f(y)] = \int(\frac{d^k f(y)}{dy^k})^2 dy$. In most applications $k = 2$ is chosen, setting a constraint on the curvature of the one dimensional distribution being unfolded.

- Iterative schemes using steps where an improved estimate at step n for the distribution to be unfolded is obtained by convolving the estimate at step $n-1$ with an updating function that depends on the response matrix, the observed distribution and the estimate $(n-1)$ itself.

- Maximum-entropy schemes whose constraining function is the expected amount of information gained in passing from the the initial ansatz to the best estimate i.e. $S(\boldsymbol{\mu}) = H(\boldsymbol{\mu}) = \sum_i^M \mu_i \log\frac{\mu_i}{\epsilon_i}$, where $\boldsymbol{\mu}$ is the estimator vector for the unknown probability distribution, the index i goes from 1 to M, the number of bins of the distribution, and $\boldsymbol{\epsilon}$ is the best initial knowledge about the true, unknown distribution, assumed to be non-negative.

- Non-iterative Bayes-inspired schemes where the full unfolded spectrum is considered a variable to be obtained by a convolution integral of the probability for the migration model and the observed spectrum.

- Iterative unbinned schemes, using event-by-event weights based on the ratio of expected to observed local densities to derive a new estimate of the distribution to be unfolded at each step. A test function based on a Tikhonov-like distance or with the same analytic form as an electric potential (of the new estimate with respect to the old one) is used to quantify the agreement between the estimate at step n and the one at step $n-1$.

All these schemes provide estimators that result in a reduced statistical variance with respect to the ML solution and inherently add a certain level of bias to the unfolded distribution. The heart of unfolding problems lies in achieving a balance between bias and overall uncertainty.

4 Applied unfolding: the balance of bias and uncertainty

The unprecedentedly large production of top quarks at the LHC allows to use (and explore) unfolding schemes to measure cross sections differentially or to extract parameters from unfolded distributions. These two general classes of analyses provide interesting unfolding examples.

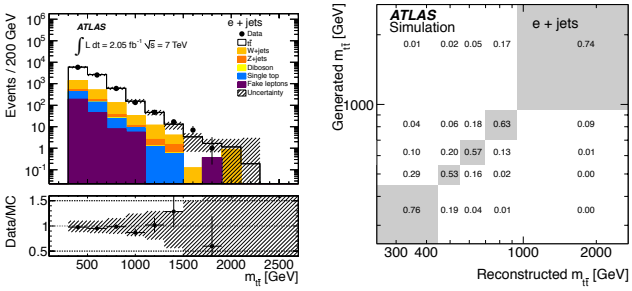

Figure 3: Distribution of reconstructed $m_{t\bar{t}}$ with data compared to predictions (left) and associated migration matrix (right) in the electron plus jets channel. Details of figures are reported in Ref. [7].

In the first analysis class the measurement of the relative differential cross section for $t\bar{t}$ production in LHC pp collisions at $\sqrt{s} = 7$ TeV as a function of $m_{t\bar{t}}$ [7] ($1/\sigma_{t\bar{t}}\, \mathrm{d}\sigma_{t\bar{t}}/\mathrm{d}m_{t\bar{t}}$) shows an example of unregularized unfolding. The $m_{t\bar{t}}$ distribution before unfolding and the corresponding migration matrix are shown in Figure 3 for selected events in the electron plus jets channel. The Tikhonov unfolding scheme with $k = 2$ is tested by reweighting simulated $t\bar{t}$ events to enhance the number of events in a single bin (see the cartoon in Figure 4). The response to this "delta-like" pulse is reduced i.e. biased at least by 30% even for the mildest regularized solution, while it maintains a linearity within 1% for the ML unregularized solution. The increase in statistical uncertainty in the final ML-unfolded result (reported in Figure 4) is tolerable as the systematic component is still dominant and under control with respect to the regularized biased result.

In the second class of analyses, a regularized scheme is used to measure the distribution of the difference between the absolute rapidities ($\Delta|y_t|$) of the reconstructed top quark and antitop quark in a sample enhanced in $t\bar{t}$ events obtained by LHC pp collisions at $\sqrt{s} = 7$ TeV [2]. A Bayesian-inspired iterative technique is used to unfold the distribution of $\Delta|y_t|$, shown in Figure 5 together with the corresponding migration matrix for selected events in the muon plus jets channel.

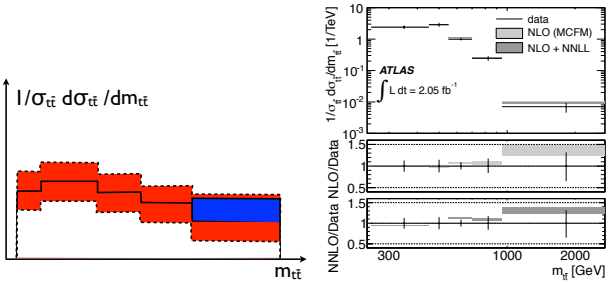

Figure 4: Scheme of "stress" test described in the text (left) and final $1/\sigma_{t\bar{t}}\, \mathrm{d}\sigma_{t\bar{t}}/\mathrm{d}m_{t\bar{t}}$ from Ref. [7] (right).

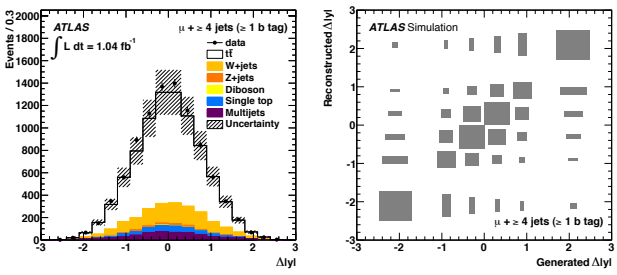

Figure 5: Distribution of reconstructed $\Delta|y_t|$ with data compared to predictions (left) and associated migration matrix (right) in the muon plus jets channel. Details of figures are reported in Ref. [2].

The number of iterations is tuned to get the expected variation of the value for the asymmetry to be stable within 0.1% in simulated $t\bar{t}$ events. Simulated $t\bar{t}$ events are re-weighted to produce samples with different true asymmetry. The analysis is performed on each sample and the input asymmetries are plotted versus the resulting measured asymmetries after unfolding to check the linearity of the unfolding procedure (as illustrated in the cartoon of Figure 6). The small biases observed in the recon-

structed distributions and the extracted asymmetry are quantified by the largest relative deviation over all the bins and the mean uncertainty-normalized relative difference between true and unfolded values from the pull distributions, respectively. Such values are used to assign additional systematic uncertainties to the unfolded distributions (for which an example is shown in Figure 6) and the final asymmetry.

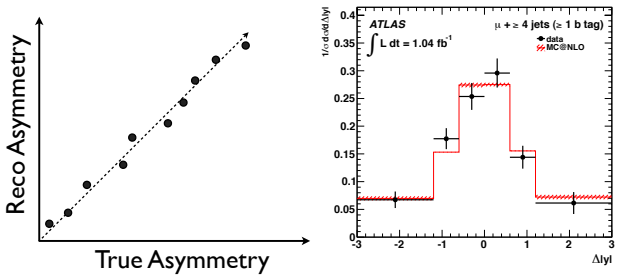

Figure 6: Scheme of the linearity test described in the text (left) and final unfolded distribution of $\Delta|y_t|$ in the muon plus jets channel from Ref. [2] (right).

5 Optimization and good practices

The ideal (infinite-simulation, infinite-manpower, infinite -time) general procedure for unfolding optimization can be generally described as the variation of the parameters of the unfolding scheme and the binning to scan the values of the figures of merit on which the performance is judged. Ideally one should do this for more than one unfolding scheme, then the method that is expected to perform best should be chosen. All the studies should be performed on simulated events. One can either scan a multi-parameter space i.e. have a function that accommodates requirements in one or more regions of phase space or summarize the requirement for the unfolding in one figure of merit that is a function of the parameters describing the different phase space regions. The figures of merit vary depending on the goal to be achieved and they are functions of bias, statistical and systematic uncertainties of the measurements and additional assessment criteria determined by the analyzers.

References

[1] V. Ahrens, A. Ferroglia, M. Neubert, B. D. Pecjak and L. L. Yang, JHEP **1009** (2010) 097, arXiv:1003.5827 [hep-ph].

[2] G. Aad *et al.* [ATLAS Collaboration], Eur. Phys. J. C **72** (2012) 2039, arXiv:1203.4211 [hep-ex].

[3] Joao Pequenao, CERN-GE-0803012 [http://cds.cern.ch/record/1095924/].

[4] http://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages/top_feynman_diagrams.html

[5] G. Cowan, Conf. Proc. C **0203181** (2002) 248.

[6] F. Spanò, EPJ Web Conf. **55** (2013) 03002.

[7] G. Aad *et al.* [ATLAS Collaboration], Eur. Phys. J. C **73** (2013) 2261, arXiv:1207.5644 [hep-ex].