

COMPARISON OF MACHINE LEARNING APPROACH TO OTHER COMMONLY USED UNFOLDING METHODS*

PETR BAROŇ

Joint Laboratory of Optics of Palacký University
and
Institute of Physics AS CR, Faculty of Science, Palacký University
17. listopadu 12, 771 46 Olomouc, Czech Republic
`petr.baron@upol.cz`

(Received March 31, 2021; accepted May 4, 2021)

Unfolding in high energy physics represents the correction of measured spectra in data for the finite detector efficiency, acceptance, and resolution from the detector to particle level. Recent machine learning approaches provide unfolding on an event-by-event basis allowing to simultaneously unfold a large number of variables and thus to cover a wider region of the features that affect detector response. This study focuses on a simple comparison of commonly used methods in RooUnfold package to the machine learning package OmniFold.

DOI:10.5506/APhysPolB.52.863

1. Introduction

The equation of unfolding can be written as

$$p = \frac{1}{\epsilon} M^{-1} \eta (D - B), \quad [1] \quad (1)$$

where D is the data spectrum from which the background spectrum B is subtracted followed by multiplication of acceptance correction η , so the main input to the unfolded procedure is prepared. The unfolding is here schematically given by a so-called migration matrix M^{-1} which maps one-to-one events from the detector to particle level. Behind the symbol M^{-1} , one could also imagine not necessarily the algebraic matrix inversion, but rather different unfolding methods, because the treatment of unfolding input differs. However, the aim of this study is not to fully describe all the methods

* Presented at XXVII Cracow Epiphany Conference on *Future of Particle Physics*, Cracow, Poland, January 7–10, 2021.

separately, but rather to make a comparison between them. The result of unfolding has to be corrected by the detector efficiency ϵ to obtain the unfolded “truth” spectrum p ideally close to the truth (particle) level. The background B is not considered in this study. Figure 1 provides insight to the ingredients on the example of the transverse momentum spectrum of the hadronically decaying top quark in the $pp \rightarrow t\bar{t}$ process [1].

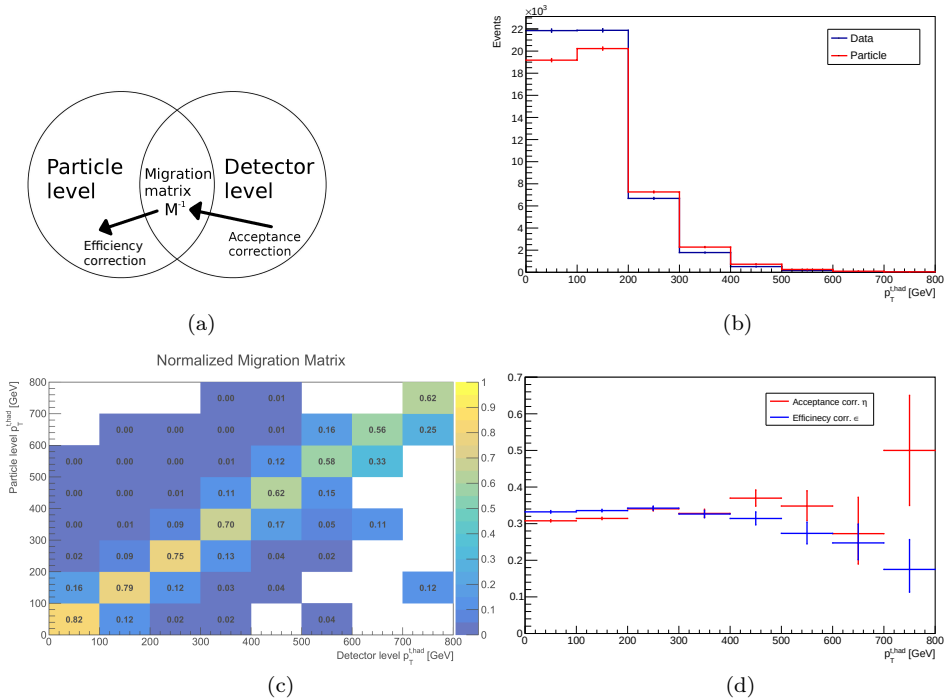


Fig. 1. (Colour on-line) Unfolding inputs. (a) Unfolding procedure diagram. (b) Detector-level (black/blue) and particle-level (grey/red) spectra. (c) Migration matrix between particle and detector levels. (d) Efficiency (black/blue) and acceptance (grey/red) corrections as a function of the transverse momentum of the hadronically decaying top quark [2].

2. Machine learning approach

The machine learning unfolding is similar to *e.g.* face recognition problem. The idea is to take all the possible information from the detector in a similar way as a photo, and train some neural network to classify what process returns such a signature in the detector as the photo of the face is classified to one particular person.

In both cases, machine learning needs the truth information to train the neural network, *e.g.* the face on the photo belonging to a particular person or the detector signature belonging to the $pp \rightarrow t\bar{t}$ process. The model

describing the physics is still necessary. Authors of OmniFold [3] call the truth process and truth detector signature natural and the model is called synthetic, see figure 2.

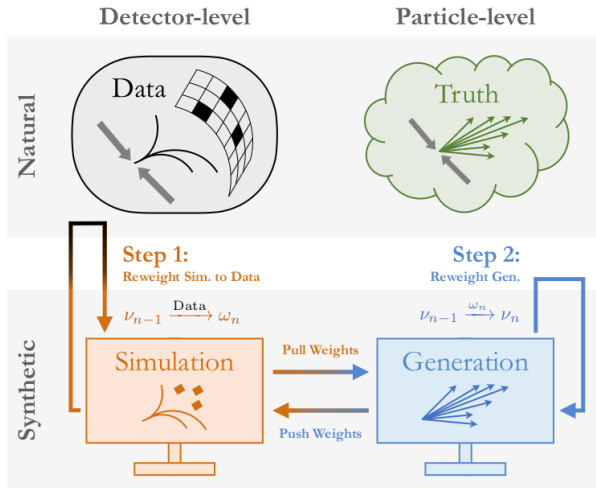


Fig. 2. An illustration of the OmniFold method applied to a set of synthetic and natural data. As a first step, starting from prior weights ν_0 , the detector-level synthetic data (“simulation”) is reweighted to match the detector-level natural data (simply “data”). These weights ω_1 are pulled back to induce weights on the particle-level synthetic data (“generation”). As a second step, the initial generation is reweighted to match the new weighted generation. The resulting weights ν_1 are pushed forward to induce a new simulation, and the process is iterated [3].

The authors extended the idea of Iterative Bayes unfolding to continuous form and with machine learning concept enabled to perform unfolding event-by-event, so the trained network returns a set of weights for each event or a function which can be applied to measured data. The detailed description of the algorithm can be found in Appendix of [3].

3. Performing closure test

The closure test of the unfolding method is to unfold not the measured data, but rather the generated particle-level spectrum. If the method is consistent, their ratio should be close to unity. To avoid other systematical uncertainties from efficiency and acceptance corrections, the events are chosen only from the overlap of the particle and detector-level phase spaces, see the intersection of the circles in figure 1.

As the process of study, the process of top-quark pair production in proton–proton collisions $pp \rightarrow t\bar{t}$ in $\ell + \text{jets}$ channel was chosen, see figure 3, simulated using MadGraph [4] software with a generation of events using

PYTHIA 8 [5] with the detector-level and the pseudo-data simulated using Delphes [6] with ATLAS detector card. The basic selection and cuts were applied to obtain spectra with a similar shape to those measured at the Large Hadron Collider (LHC) in the real ATLAS experiment, although comparison of unfolding methods could be performed with an arbitrary measured process.

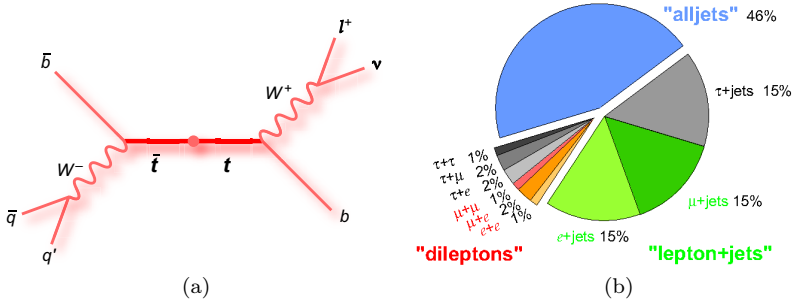


Fig. 3. (a) Final-state diagram of the process $pp \rightarrow t\bar{t}$; (b) Top-quark pair production branching ratios.

The input data set was divided into two statistically-independent sets, so the classical unfolding methods were using migration matrix built of one set and the input to unfolding procedure from the other set. The same exclusive sets were used for training the neural network and to perform the machine learning unfolding.

4. Results

The following spectra of interest were chosen: the transverse momentum, mass, energy, and pseudo-rapidity of the hadronically-, leptonically-decaying top quark, and also the $t\bar{t}$ system. The distribution ϕ was omitted due to its flat shape. In total, $4 \times 3 = 12$ binned spectra were unfolded classically using the RooUnfold package [7] with Bayes (3 iterations) [8], SVD ($k = 5$) [9], and Ids ($k = 1$) [10] methods. The variables were used event-by-event in the neural network with 100 epochs to later perform simultaneous unfolding.

Even though results from the machine learning approach could be shown as continuous spectra, for comparison purposes, particular fine binning was chosen.

Particle-level spectra as the input into the unfolding procedure were chosen to perform the closure test. Thus, the ratio between input particle-level spectrum and unfolded spectrum should be ideally close to unity as it was already discussed in Section 3.

The metric of comparison is χ^2 divided by the number of degrees of freedom NDF which is equal to the number of bins in the spectrum [1].

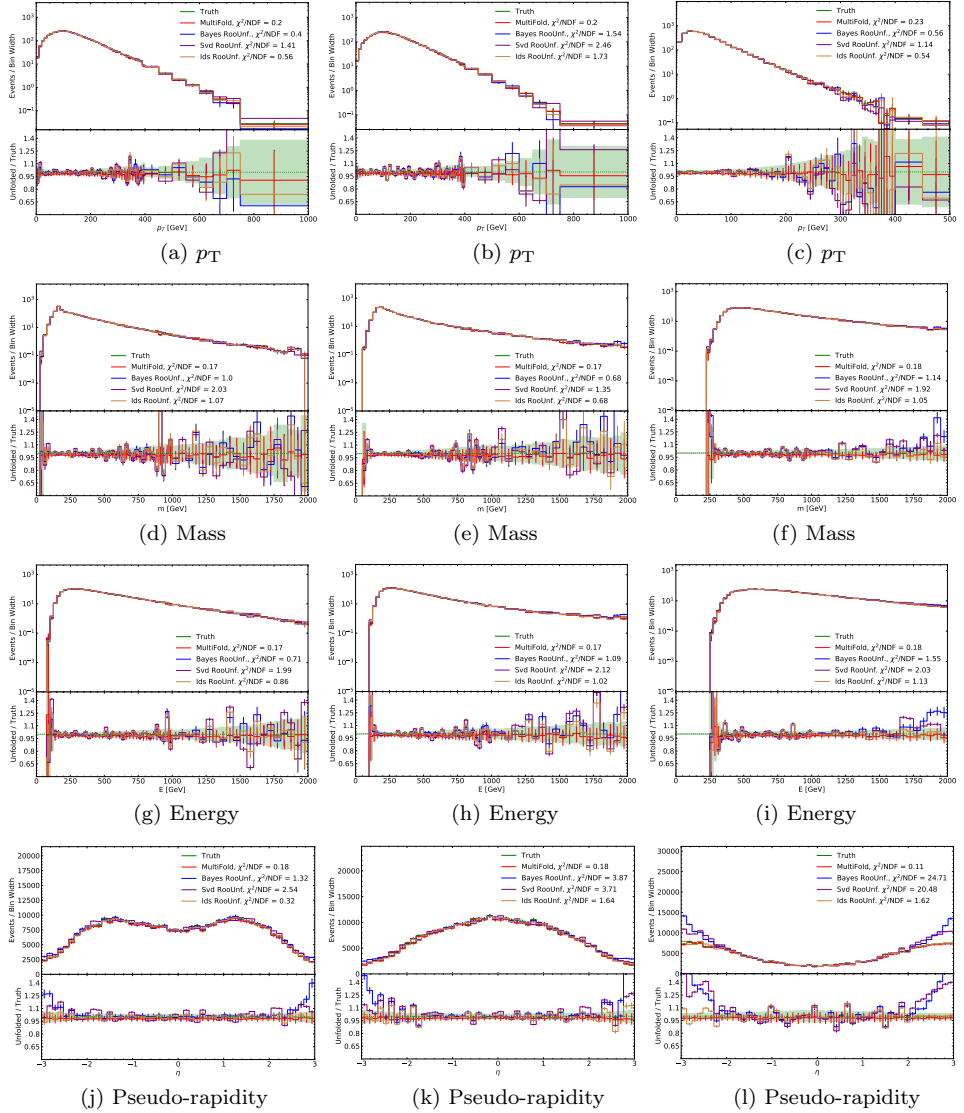


Fig. 4. Closure test of the spectra of the hadronically- (left), leptonically- (middle) decaying top quark and of the $t\bar{t}$ system (right) of the p_T , mass, energy and pseudo-rapidity with the lower pads showing the ratio of the unfolded to the truth spectrum.

5. Conclusion

Table I summarizes the χ^2/NDF results and proves that machine learning approach at this particular study performed the best results compared to values in Tables II, III and IV.

The study aimed to demonstrate the possible potential of the machine learning methods on four typical spectra used in high-energy physics. A slight disadvantage of the machine learning method might be its initial CPU time needed to train the neural network. Although more complex tests of **OmniFold** unfolding need to be performed in the future, the χ^2/NDF presents promising results.

TABLE I

Results of the closure test, χ^2/NDF between the truth and unfolded spectra using the **OmniFold** method.

	Tr. momentum	Mass	Energy	Pseudo-rapidity
Hadronic top	0.20	0.17	0.17	0.18
Leptonic top	0.20	0.17	0.17	0.18
$t\bar{t}$ system	0.23	0.18	0.18	0.11

TABLE II

Results of the closure test, χ^2/NDF between the truth and unfolded spectra using the **Bayes RooUnfold** method.

	Tr. momentum	Mass	Energy	Pseudo-rapidity
Hadronic top	0.40	1.00	0.71	1.32
Leptonic top	1.54	0.68	1.09	3.87
$t\bar{t}$ system	0.56	1.14	1.55	24.71

TABLE III

Results of the closure test, χ^2/NDF between the truth and unfolded spectra using the **Svd RooUnfold** method.

	Tr. momentum	Mass	Energy	Pseudo-rapidity
Hadronic top	1.41	2.03	1.99	2.54
Leptonic top	2.46	1.35	2.12	3.71
$t\bar{t}$ system	1.14	1.92	2.03	20.48

TABLE IV

Results of the closure test, χ^2/NDF between the truth and unfolded spectra using the Ids RooUnfold method.

	Tr. momentum	Mass	Energy	Pseudo-rapidity
Hadronic top	0.56	1.07	0.86	0.32
Leptonic top	1.73	0.68	1.02	1.64
$t\bar{t}$ system	0.54	1.05	1.13	1.62

The author gratefully acknowledges the support from the project IGA_PrF_2021_004 of the Palacký University as well as grant GACR 19-21484S of MSM, Czech Republic.

REFERENCES

- [1] P. Baroň, J. Kvita, «Extending the Fully Bayesian Unfolding with Regularization Using a Combined Sampling Method», *Symmetry* **12**, 2100 (2020).
- [2] P. Baroň, «Fully Bayesian Unfolding in High-energy Physics», *Acta Phys. Pol. B* **51**, 1241 (2020).
- [3] A. Andreassen *et al.*, «OmniFold: A Method to Simultaneously Unfold All Observables», *Phys. Rev. Lett.* **124**, 182001 (2020), [arXiv:1911.09107 \[hep-ph\]](#).
- [4] J. Alwall *et al.*, «The automated computation of tree-level and next-to-leading order differential cross-sections, and their matching to parton shower simulations», *J. High Energy Phys.* **1407**, 079 (2014), [arXiv:1405.0301 \[hep-ph\]](#).
- [5] T. Sjöstrand *et al.*, «An Introduction to PYTHIA 8.2», *Comput. Phys. Commun.* **191**, 159 (2015), [arXiv:1410.3012 \[hep-ph\]](#).
- [6] DELPHES 3 Collaboration (J. de Favereau *et al.*), «DELPHES 3, A modular framework for fast simulation of a generic collider experiment», *J. High Energy Phys.* **1402**, 057 (2014), [arXiv:1307.6346 \[hep-ex\]](#).
- [7] T. Auye, «Unfolding algorithms and tests using RooUnfold», in: Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland, 17–20 January 2011, (Eds.) H.B. Prosper L. Lyons, CERN-2011-006, pp. 313–318.
- [8] G. D’Agostini, «A multidimensional unfolding method based on Bayes’ theorem», *Nucl. Instrum. Methods Phys. Res. A* **362**, 487 (1995).
- [9] A. Hoecker, V. Kartvelishvili, «SVD approach to data unfolding», *Nucl. Instrum. Methods Phys. Res. A* **372**, 469 (1996), [arXiv:hep-ph/9509307](#).
- [10] B. Malaescu, «An iterative, dynamically stabilized method of data unfolding», [arXiv:0907.3791 \[physics.data-an\]](#).