33RD INTERNATIONAL COSMIC RAY CONFERENCE, RIO DE JANEIRO 2013
THE ASTROPARTICLE PHYSICS CONFERENCE

ICRC 2013

DSEA: A Data Mining Approach to Unfolding

TIM RUHE[1], MARTIN SCHMITZ[1], TOBIAS VOIGT[2], MAX WORNOWIZKI[2]

[1] *Lehrstuhl Experimentelle Physik 5, Technische Universität Dortmund*
[2] *Fakultät Statistik, Technische Universität Dortmund*

tim.ruhe@tu-dortmund.de

Abstract: Solving inverse problems described by the Fredholm integral equation of first kind is a common challenge in many particle and astroparticle physics experiments. Several algorithms for the solution of these problems exist, most of them aiming at the determination of the response matrix on Monte Carlo simulations. In this paper a novel data mining based approach towards unfolding is given, treating the inverse problem as a multinominal classification task. This approach offers the advantage of an event-by-event unfolding, which retains the full information on any given event. Treating the inverse problem as a classification task further offers the possibility to use additional information on the geometry of the individual events. The algorithm is described and toy Monte Carlo studies on the performance are presented.

Keywords: Inverse Problem, Unfolding, Spectrum Reconstruction

1 Introduction

It is a common challenge in particle and astroparticle physics, that the true distribution $f(x)$ of an attribute x cannot be accessed directly and a second distribution $g(y)$ is measured instead. Due to smearing effects and a limited acceptance of the detector, $g(y)$ cannot be directly converted into $f(x)$. Instead both distributions are connected by the Fredholm integral of first kind:

$$g(y) = \int_a^b A(x,y) f(x) dx, \tag{1}$$

where A(x,y) represents the response function of the detector. This is commonly referred to as an inverse or ill-posed problem. Several algorithms for the solution of inverse problems exist, including regularised unfolding as implemented in TRUEE [1] and $\mathscr{RUN}$ [2].
Most of these algorithms first convert the integral equation in (1) into a matrix equation of the form:

$$\vec{g}(y) = A(x,y)\vec{f}(x), \tag{2}$$

by a discretising operation. In general a solution can be obtained by first determining the matrix on Monte Carlo simulations, where $\vec{g}_{MC}(y)$ and $\vec{f}_{MC}(x)$ are exactly known. The determination of $\vec{f}(x)$ on real data utilizing the estimation of $A_{MC}(x,y)$ obtained on simulations differs between individual unfolding approaches [3]. In general, however, a simple inversion of the response matrix is not feasible, as oscillating solutions might occur [1]. Thus, a solution to equation (2) is obtained by maximizing a log-likelihood expression which includes a regularisation term [1].
The use of the matrix, however, poses a potential problem, as in general only one such response matrix is obtained for the entire detector. This approximation is fully valid for small and homogeneous detectors but might become problematic for large detectors utilising natural media, e.g. large scale neutrino telescopes. In these experiments particles of the same energy will cause significantly different event patterns depending on where they enter the detector. The same challenge emerges from the fact that particles may be produced far outside the detectors. Accordingly events can be starting, stopping or through-going depending on their point of production. The event pattern created by a through-going track, however, is significantly distinct from the one of an event stopping inside the detector.
Thus, geometrical information on the track of the individual particles will provide useful information, which can be utilised in order to improve the unfolding. Unfortunately, the number of input variables is limited in many unfolding procedures, e.g. to three in the most recent version of TRUEE [1]. This limitation is inherent in many unfolding approaches, as the determination of the matrix corresponds to building a density based model. The number of simulated events, required for a reliable determination of a density based model is known to scale exponentially with the number of input parameters.
Furthermore, after the sought distribution $f(x)$ is determined, all information on the individual events that contributed to the distribution is lost. Thus, one is unable to follow individual events through the complete unfolding process, in order to determine where and how much they contributed to the final spectrum. Moreover, physically useful information, e.g on the zenith angle of an event, is lost. Studies on changes of a spectrum with time or zenith angle therefore require a number of individual unfoldings, which all have to be tested and optimised separately.
In this paper a machine learning based approach to unfolding is presented. The algorithm itself is outlined in section 2, whereas a comparison of the results to those obtained using the well known and well tested unfolding software TRUEE is presented in section 3. section 4 discusses the dependency on the distribution used as input for the learner. In section 5 an example on the utilisation of the full event information is given. A summary is given in section 6.

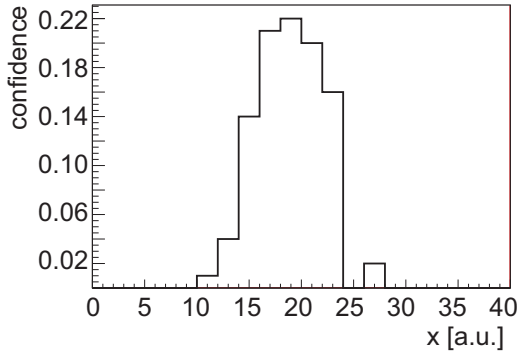

Figure 1: Confidence distribution of a selected event after the application of the forest.

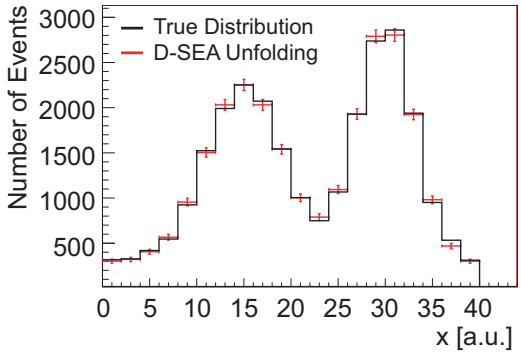

Figure 2: The outcome the DSEA unfolding in red compared to the true distribution in black. The unfolding results are found to agree with the true distribution within statistical uncertainties.

2 DSEA

Solving the matrix equation (2) is sufficient for the application of an unfolding algorithm in a physics experiment, as $\vec{f}(x)$ serves as a reliable approximation of $f(x)$. From the machine learning point of view, however, the individual bins f_j of $\vec{f}(x)$ can be interpreted as different classes of events. The unfolding can thus be treated as a multinominal classification task. Several algorithms for the solution of multinominal classification problems exist. All studies presented in the following were carried out using toy Monte Carlo simulations produced with Gaussian smearing. A Random Forest [4] was used as a learning algorithm. In total 2.6×10^6 examples were evaluated in a 5-fold cross validation. The number of events used for training was limited to 8×10^4. A stable and reliable performance without any signs of overtraining was observed for the forest. Our studies showed, however, that treating an unfolding as a simple classification task does not restore the true distribution. This is due to the fact that similar attribute values are observed for neighbouring classes, due to the detector smearing. Thus, the ordinary classification via the maximum of the confidence distribution leads to not very distinct results.
The confidence distribution of a selected event, as obtained from the application of the forest is shown in Fig. 1. One finds that the differences in confidence are small for neighbouring classes.

Within the Dortmund Spectrum Estimation Algorithm (DSEA) the confidence values obtained for individual events can be interpreted as conditional probability densities. The confidence value c_{ij}, describing the

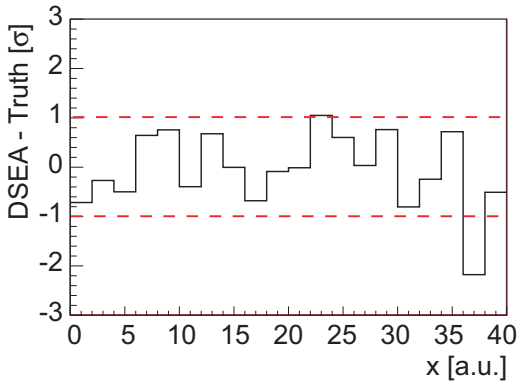

Figure 3: Deviation of the DSEA unfolding results from the true distribution in units of the statistical uncertainty σ. Only two bins were found to lie above the $1\,\sigma$ limit.

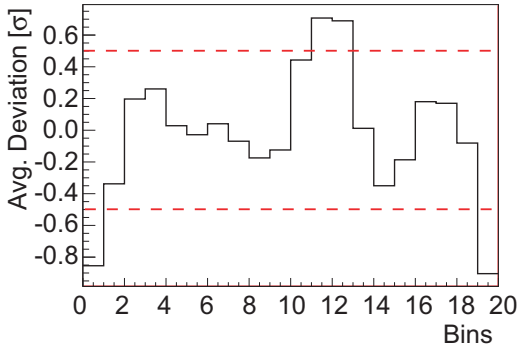

Figure 4: Average deviation of the DSEA result from the true distribution obtained via bootstrapping in units of the statistical uncertainty σ. A fraction of 10% of the events was drawn at random in each iteration and the deviation of the unfolding result to the true distribution was computed in units of the estimated statistical uncertainty. The average deviation of both distributions (pull mean) is shown on the y-axis.

contribution of the i-th observation to the j-th bin, coresponds to $\hat{p}(j|d_i)$, the probability for the event to lie in bin j given the observed data d_i. These conditional probability densities can then be utilised to reconstruct the spectrum in a simple summation. The bin content f_j of bin j, for example is obtained via:

$$f_j = \sum_{i=1}^{N} c_{ij}, \tag{3}$$

where N, represents the number of events, utilized in the unfolding process. The outcome of the unfolding using DSEA is depicted in Fig. 2, where the true distribution (solid black line) is shown for comparison. One finds that both distributions agree within statistical uncertainties. Figure 3 shows the deviation of the unfolding result obtained using DSEA from the true distribution in units of the estimated statistical uncertainty σ. In the current version of DSEA σ is estimated as the square root of the bin content according to a compound poisson model. Only two bins were found to lie above the $1\,\sigma$ limit. No deviations larger than $3\,\sigma$ were observed.

A bootstrapping procedure was used in order to test the statistical reliability of DSEA. In each iteration 10% of the events corresponding to 2.6×10^5 examples, were drawn at random and unfolded accordingly. After each

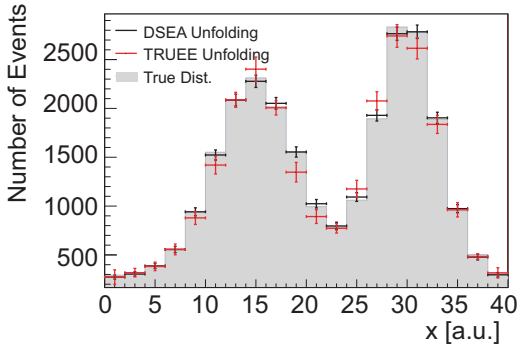

Figure 5: Comparison of the results obtained using D-SEA to results of the TRUEE unfolding. The results of both unfolding algorithms were found to agree within the statistical uncertainties.

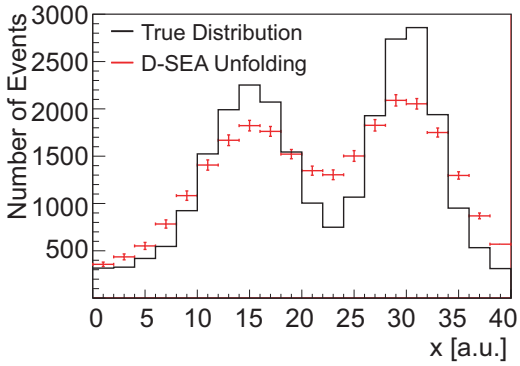

Figure 6: Unfolding result (red) obtained using a uniform distribution as input for the training of the classifier. Although the true distribution (black) is not reconstructed to the full extent, the positions of the two peaks are reconstructed correctly.

unfolding the deviation of the unfolding result from the true deviation was calculated in units of the estimated statistical uncertainty σ. Finally the average deviation was computed. The outcome of the bootstrapping procedure is depicted in Fig. 4. One finds that on average no deviations larger than $1\,\sigma$ are observed for any of the bins, which indicates a stable behaviour of the unfolding using DSEA. Slight oscillations of the unfolding result are observed. Such a behaviour is typical for solutions of inverse problems and in general suppressed by the use of regularisation. Regularisation is not directly implemented in DSEA, but one of the key topics for the future development of the algorithm. One should, however, note that the observed oscillations are well below the $1\,\sigma$ limit and can therefore be tolerated.

3 Comparison to TRUEE

In order to further validate the results obtained with DSEA the toy Monte Carlo simulation was unfolded using TRUEE. Both results are depicted in Fig. 5. The true distribution is shown for comparsion. One finds that the unfolding results obtained with the different algorithms agree within the statistical uncertainties.
No oscillations were observed for the result obtained with TRUEE, which can be attributed to the explicit use of regularisation within the algorithm. Furthermore, smaller statistical uncertainties were obtained for the DSEA result.

4 Dependency on the Input Distribution

In order to study the dependency of the unfolding result on the distribution of events used for the training, a uniform distribution was chosen as input for the learner. This is the most reasonable distribution to commence with, in case no additional information, experimental or theoretical, is available. One should note, however, that in general at least some information on the sought distribution is available.
The outcome of utilising a uniform distribution as input for unfolding is shown in Fig. 6. The sought distribution is depicted in black, whereas the unfolding result is shown in red. Despite the fact that the true distribution is not reconstructed to the full extend, the positions of the two peaks are reconstructed correctly. This implies that certain features of unknown distributions can be correctly reconstructed, even if not directly simulated. Furthermore,

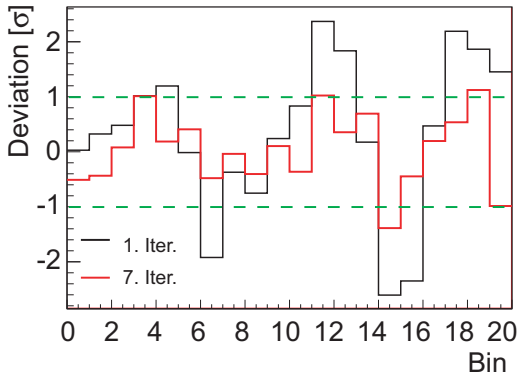

Figure 7: Average deviation of the unfolding result from the true distribution in units of the statistical uncertainty σ, obtained in an iterative unfolding. The first iteration is shown in black, whereas the 7th iteration is depicted in red.. The unfolding result obtained in iteration i, was used to generate Monte Carlo events utilised as input for iteration $i+1$. A uniform distribution was used as input in the first iteration. The oscillating behaviour is found decrease significantly between the first and the 7th iteration.

the unfolded distribution clearly deviates from a uniform distribution.
In addition, the unfolding can be carried out in an iterative procedure. In that case the unfolding result obtained in iteration i is utilised for the generation of Monte Carlo events used for the training of the learner in iteration $i+1$. The pull distributions for such an iterative unfolding are presented in Fig. 7. In these pulls the unfolding is carried out on a randomly drawn subset of events and the deviation from the true distribution is calculated in units of the statistical uncertainty σ. One finds that the observed deviations decrease significantly between the first and the 7th iteration. In fact, the pull distribution of the 7th iteration is found to deviate only marginally from the pull distribution obtained using the correct distribution of events as input for the learning algorithm.

5 Utilising the Additional Information

Compared to most other unfolding approaches DSEA offers the advantage of retaining all information on the individual events. This information can then be utilised in analyses aiming at studying changes of a spectrum with a second

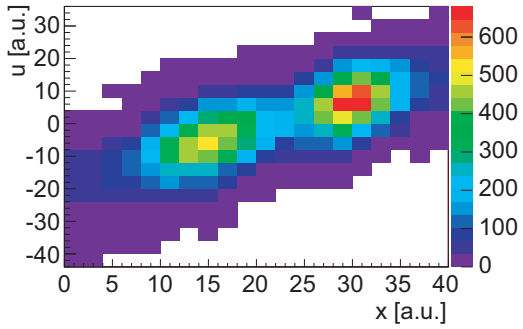

Figure 8: Sought two-dimensional distribution. Two distinct peaks are observed.

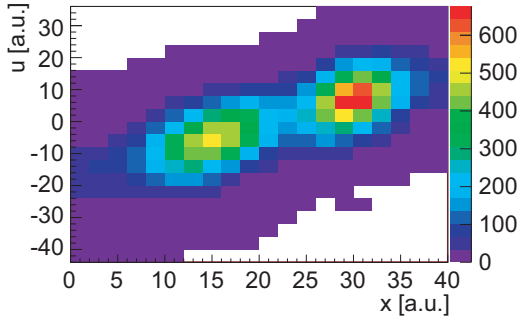

Figure 9: Reconstructed two-dimensional distribution after application of DSEA. Although the resolution of the plot is not comparable to that of the true distribution, the major features are clearly reconstructed.

quantity of interest u. Note that this is not to be confused with a two-dimensional unfolding as no additional smearing in u is introduced.
The true two-dimensional distribution is shown in Fig. 8. Two distinct peaks are observed. The two-dimensional distribution, reconstructed by utilising the DSEA output is shown in Fig. 9. Again the major features of the distribution, the two distinct peaks, are reconstructed correctly. The resolution in x is limited by the binning of the unfolded distribution, while the binning in u can in principle be chosen arbitrarily but was adjusted to match the binning in x. Note that the reconstruction of the distribution shown in Fig. 9 would require a number of different unfoldings using other unfolding algorithms. In DSEA, however, no additional unfolding is required as the information on individual events can be utilised once a properly trained learning algorithm has been applied to the data set.
Figure 10 shows the deviation of the DSEA result from the true distribution in units of the statistical uncertainty σ. No deviations exceeding the 3σ limit are observed. Several bins with discrepancies exceeding 1σ were observed. One should note, however, that the number of bins in Fig. 10 is $n_{\text{bins}} = 400$. Thus, 128 bins are expected with a deviation of 1σ or more.

6 Summary

A novel unfolding approach treating inverse problems as multinominal classification tasks and utilising the output of state of the art machine learning algorithms has been

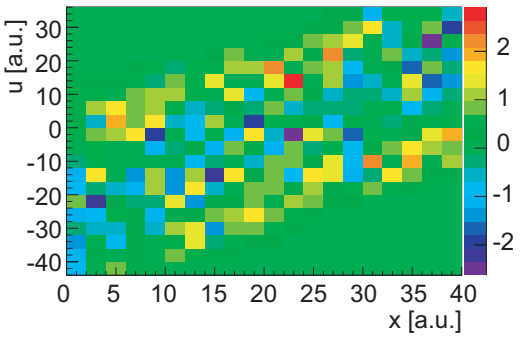

Figure 10: Deviation of the unfolded and the true two-dimensional distribution in units of the estimated statistical uncertainty, depicted as the colour column. No deviations above 3σ were observed. Note that the number of bins is $n_{\text{bins}} = 400$. Deviations exceeding the 1σ limit are therefore expected for as many as 128 bins.

presented. Excellent results were obtained on toy Monte Carlo simulations produced with Gaussian smearing. The unfolding result was found to agree with the true distribution within the obtained statistical uncertainty.
Tests comparing the result of the novel approach to unfolding results obtained with TRUEE showed that both algorithms deliver the same results within the statistical uncertainties.
The DSEA solution was found to show a slightly oscillating behaviour inherent to the solution of inverse problems. These oscillations, however, can be suppressed utilising regularisation algorithms. One should, however, notice that the observed fluctuations were small, as none of the average deviations exceeded the 1σ limit.
No regularisation is implemented for the current version of DSEA but different regularisation methods considered for the use in DSEA are currently under investigation.

Acknowledgement

Part of the work on this paper has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 ”Providing Information by Resource Constrained Analysis”, project C3. We also acknowledge the support from the German Ministry of Education and Research (BMBF).

References

[1] N. Milke et al., Nuclear Instruments and Methods in Physics Research A 697 (2013) 133.
[2] V. Blobel, Technical Note TN361 OPAL (1996) 1.
[3] G. Cowan, Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics (2002) 18.
[4] L. Breiman, Machine Learning 45 (2001) 5.