Mining for Spectra – The Dortmund Spectrum Estimation Algorithm

Tim Ruhe,[1] Tobias Voigt,[2], Max Wornowizki[2], Mathis Börner[1], Wolfgang Rhode[1] and Katharina Morik[3]

[1] *Lehrstuhl Experimentelle Physik 5, TU Dortmund, Dortmund, Germany;* `tim.ruhe@tu-dortmund.de`

[2] *Lehrstuhl Statistik in den Biowissenschaften, TU Dortmund, Dortmund, Germany;*

[3] *Lehrstuhl für künstliche Intelligenz, TU Dortmund, Germany;*

Abstract. Obtaining energy spectra of incident particles such as neutrinos or gamma-rays is a common challenge in neutrino- and Air-Cherenkov astronomy. Mathematically this corresponds to an inverse problem, which is described by the Fredholm integral equation of the first kind. Several algorithms for solving inverse problems exist, which are, however, somewhat limited. This limitation arises from the limited number of input observables and the fact that information on individual events is lost and only the unfolded distribution is returned. In this paper we present the Dortmund Spectrum Estimation Algorithm (DSEA), which aims at overcoming the aforementioned obstacles by treating the inverse problem as a multinominal classification task. DSEA is a modular and highly flexible algorithm that can easily be tailored to a problem at hand. To avoid a potential bias on the class distribution used for the training of the learner, DSEA can be applied in an iterative manner using a uniform class-distribution as input.

1. Introduction

Solving inverse problems is a common challenge in neutrino- (Aartsen et al. 2015; Adrián-Martínez et al. 2013) and imaging Air-Cherenkov astronomy (Albert et al. 2007; Aharonian et al. 2006). In both cases the energy spectrum $f(x)$ cannot be accessed directly, but has to be inferred from the distribution $g(y)$ of other observables, e.g. energy losses of secondary particles. The task is further made difficult by the fact that the production of secondaries, e.g. in a neutrino-nucleon interaction is governed by stochastical processes. Additional smearing of the observables is introduced by the limited acceptance of the detector.

Mathematically, $f(x)$ and $g(y)$ are connected by the so-called response function $A(x, y)$ in the Fredholm integral equation of the first kind. Several algorithms for solving inverse problems exist (Milke et al. 2013; D'Agostini 1995, 2010; Höcker & Kartvelishvili 1996; Feindt 2004), which are, however, somewhat limited, for example in the number of input variables or in the sense that only the unfolded distribution is returned and the information on individual events is lost.

This paper presents the Dortmund Spectrum Estimation Algorithm (DSEA), which overcomes these limitations by treating the inverse problem as a multinominal classification task. This classification task can then be solved by an arbitrary learning algorithm. By design DSEA is flexible, highly modular and allows for the use of any learning algorithm, as long as it returns the confidences of the individual classes. To avoid a potential bias on the class distribution used for the training of the learner, DSEA can be used iteratively using a uniform class-distribution as input.

The paper is organized as follows: Section 2 describes the algorithm itself. Its convergence and performance by means of a short simulation study on artificial data generated using Gaussian smearing are addressed in Sec. 3. A summary concludes the paper in Sec. 4.

2. Algorithm

Within DSEA a binned version $\vec{f}(x)$ of the sought after spectrum $f(x)$ is estimated by iteratively carrying out following steps:

1. **Discretization:** $f(x) \mapsto \vec{f}(x) = (f_1, ..., f_m)$. **(Initalize)**

2. **Classifier Training:** A dataset $(Y, W, F) = \{(\vec{y}, w, f)_1; ...; (\vec{y}, w, f)_n\}$ of n examples is used to train a model $M(Y, W, F)$. Each example consists of h attributes $\vec{y} = (y_1, ..., y_h)$, a weight w and a label f. In the first iteration all weights are initialized with $w_i = 1$.

3. **Model Prediction:** The model $M(Y, W, L)$ is applied to a set of $\tilde{n}$ unlabeled examples $\tilde{Y} = (\vec{\tilde{y}}_1, ..., \vec{\tilde{y}}_{\tilde{n}})$, yielding a confidence $c_{ijk} = g(M(Y, W, L), \vec{\tilde{y}})$ for the i-th example to belong to f_j.

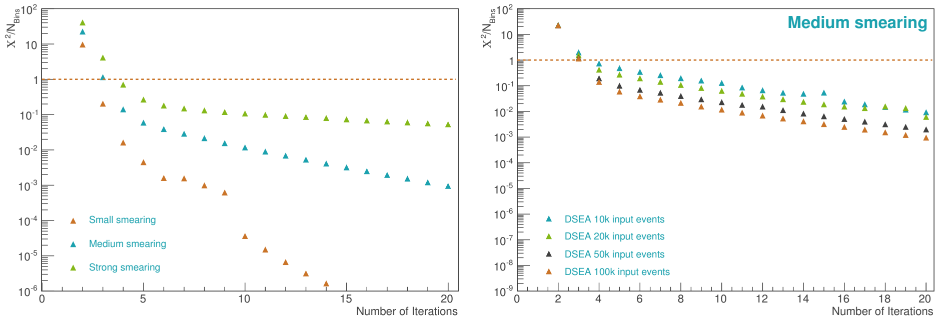
4. **Spectral Reconstruction:** The bin content $\hat{f}_{j,k}$ of the j-th bin obtained in the k-th iteration is estimated as $\hat{f}_{j,k} = \sum_{i=1}^{\tilde{n}} c_{ij}$.

5. **Reweighting:** The weights for the training data in the $(k + 1)$-th iteration are updated according to $w_{i,k+1} = \frac{\hat{f}_{l_i,k}}{\tilde{n}}$. **(Continue with step 2).**

In case a Naive Bayes classifier is used, the confidences become $c_{i,j} = p(f_j|\vec{a}_i)$ and DSEA has some overlap with D'Agostinis Iterative Bayesian Unfolding (D'Agostini 1995, 2010), but is different in two important points. Firstly, DSEA estimates the $p(f_j|\vec{a}_i)$ on an event-by-event rather than on a bin-by-bin basis. This property is desirable as events may be reconstructed with similar values in an attribute a_i although originating from different f_j, due to different event topologies. Secondly, DSEA allows to use the complete event topology, as the number of input variables is arbitrary, which results in an increase of information available in the reconstruction process.

3. Convergence and Performance

A Naive Bayes classifier implemented in the data mining toolkit RapidMiner (Mierswa et al. 2016) was chosen as a learning algorithm for the studies presented in this paper.



(a) Convergence of DSEA for different amounts of smearing using 10^5 training events.

(b) Convergence of DSEA using medium smearing evaluated for different numbers of training events.

The convergence of DSEA was investigated using the convergence criterion from (D'Agostini 1995), defined as $\chi^2/m \leq 1$, with $\chi^2 = \sum_{j=1}^{m} \left(\frac{\hat{f}_{j,k-1} - \hat{f}_{j,k}}{\sqrt{\hat{f}_{j,k-1}}} \right)^2$, where $\hat{f}_{j,k}$ is the content of the j-th bin obtained in the k-th iteration.

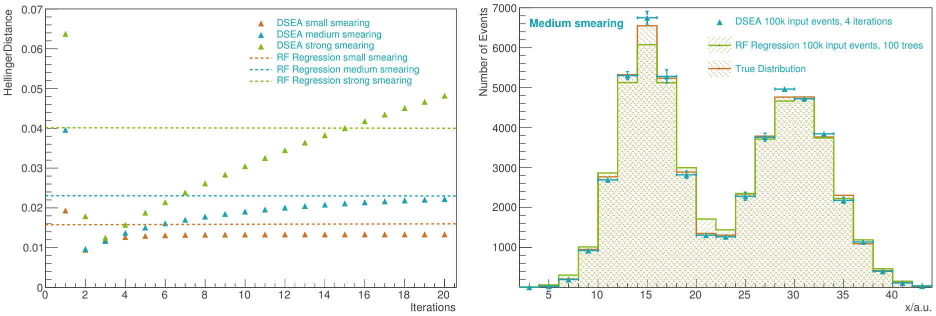
Figure 1a shows the convergence of DSEA for small smearing (orange), medium smearing (blue) and strong smearing (green), obtained using 10^5 events to train and $5 \cdot 10^4$ events to test the classifier. A uniform distribution of events was used for training, whereas a distribution with a two peak structure was used for testing.

One finds that the algorithm converges faster, for smaller smearing, which corresponds to a simpler classification task. For all three levels of smearing considered in this paper, however, the convergence criterion (orange dotted line) is met after three to five iterations. The structure observed for small smearing in the range between five and ten iterations does not affect the overall convergence of DSEA as in this case χ^2/m is already on the order of 10^{-3} and the convergence criterion is met after three iterations.

Figure 1b shows the convergence for medium smearing, evaluated for different numbers of training events. One finds that DSEA converges faster for larger numbers of input events. The performance criterion (orange dotted line), however, is met after four iterations, independent of the number of input events.

Figure 2a depicts the agreement of the reconstructed spectrum with the underlying distribution as a function of the number of iterations for three different levels of smearing. The Hellinger distance is used to quantify the agreement. The dashed lines indicate the agreement achieved by reconstructing the spectrum performing a random forest regression with 100 trees on datasets with small- (orange), medium- (blue) and strong smearing (green). For the classifier training in DSEA as well as for the random forest regression 10^5 events were randomly selected according to a uniform distribution and used as training sample for the learning algorithm.

One finds that the agreement of the spectra reconstructed using DSEA is significantly better compared to spectra obtained with a random forest regression, in case the iteration is stopped after two to ten iterations. For strong smearing the performance of the regression exceeds the performance of DSEA after 16 iterations. As already argued above, however, three to five iterations are sufficient to reach the convergence criterion.



(a) Hellinger distance as a function of the number of iterations for three levels of smearing in DSEA. The dashed lines indicate the agreement obtained using a random forest regression.

(b) Spectra obtained for medium smearing with DSEA and a random forest regression compared to the true distribution. Error-bars on the random forest regression result are omitted for better visibility.

Figure 2b shows the spectra obtained with DSEA terminated after four iterations (light blue) and the random forest regression (green) compared to the true distribution of examples (orange) for medium smearing. In general, one finds that the spectrum obtained with DSEA matches the underlying distribution of examples significantly better than one obtained using a random forest regression.

4. Summary

In this paper we presented the Dortmund Spectrum Estimation Algorithm (DSEA) and its performance on artificial data. By treating the inverse problem as a multinominal classification task, DSEA overcomes the obstacles generally associated with their solution. To avoid a potential bias a uniform distribution of examples was used as input to train the classifier, and the algorithm was used iteratively. DSEA was found to converge after three to five iterations, depending on the amount smearing. Furthermore, DSEA outperformed a random forest regression for all three levels of smearing studied in the scope of this paper.

Acknowledgments. This work has been supported by the DFG, Collaborative Research Center SFB 876, project C3 (http://sfb876.tu-dortmund.de).

References

Aartsen, M., et al. 2015, Eur. Phys. J. C, 75, 1
Adrián-Martínez, S., et al. 2013, European Physical Journal C, 73, 2606
Aharonian, F., et al. 2006, Astronomy and Astrophysics, 457, 899
Albert, J., et al. 2007, Nuclear Instruments and Methods in Physics Research A, 583, 494
D'Agostini, G. 1995, Nuclear Instruments and Methods in Physics Research A, 362, 487
— 2010, ArXiv e-prints. `1010.0632`
Feindt, M. 2004, ArXiv Physics e-prints. `physics/0402093`
Höcker, A., & Kartvelishvili, V. 1996, Nuclear Instruments and Methods in Physics Research A, 372, 469
Mierswa, I., et al. 2016, "`https://rapidminer.com/`",
Milke, N., et al. 2013, Nuclear Instruments and Methods in Physics Research A, 697, 133