PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

Statistical uncertainty derivation in probabilistic classification with DSEA+

Leonora Kardum[*]

Technische Universität Dortmund,
August-Schmidt-Straße 4, Dortmund, Germany
E-mail: leonora.kardum@tu-dortmund.de

The Dortmund Spectrum Estimation Algorithm (DSEA+) is a novel approach to unfolding by translating deconvolution tasks into multinomial classification problems, which enables the use of readily available tools. The algorithm is employable with several prebuilt classification models, making it advantageous to other methods due to its generality, simplicity, and broadness. DSEA+, primarily developed for the purpose of reconstructing energy spectra in the field of Cherenkov astronomy, can be therefore applied to other areas of research. The estimation of statistical uncertainties within DSEA mandates a special treatment of the algorithm's iterative nature. Here, we present a full derivation of statistical uncertainties in DSEA+ with probabilistic classification applied to spectral reconstruction.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

[*]Presenter

PoS(ICRC2021)1180

1. Introduction

In astronomy and particle physics, the sought values are distorted through stochastic processes involved in their detection. Often the value of interest can not be measured directly and is deduced from related quantities. The distribution $f(y)$ is said to be $convoluted$ with the smearing matrix $A(x|y)$ resulting in measured distribution $g(x)$:

$$g(x) = \int A(x|y) f(y) dy. \qquad (1)$$

Inferring the distribution $f(y)$ is known as $deconvolution$, or often referred to as $unfolding$ in astronomy and related studies. For non-square smearing matrices with high condition numbers, solving the inverse problem is not straightforward. Ill-defined matrices lead to unstable solutions with large uncertainties. Several approaches to unfolding exist, one of which is presented in this work. Dortmund Spectrum Estimation Algorithm (DSEA) [1] solves the deconvolution task by utilizing common machine learning classifiers. After discretizing the observable space, an arbitrary classifier is trained on Monte Carlo data with known distributions. Initially, all events are weighted uniformly. The trained model is now applied to the unlabeled set yielding the distribution of interest to be

$$\vec{f}_j = \sum_j^n c_{ij} \qquad (2)$$

where c_{ij} is the model's confidence of event i belonging to reconstruction bin j. The events are then reweighted with (2) and the process is repeated for a chosen number of iteration steps. DSEA has several modes of handling step sizes in its search for the proper distribution, as well as ensuring convergence. It can be used with any classifier, most commonly paired with probabilistic models (as in this work) or with the Random Forest Classifier [2] due to shown superior performance. An updated version with optimized regularization is called DSEA+. The iterative nature of this algorithm mandates a revision of the classical approach to uncertainty calculation, which takes uncertainty to be equal to the covariance matrix of the solution. This does not take into account the number of iterations taken before the solution is reached, nor the uncertainty spanning from the assumptions made in the process. Sections 2 and 3 give a detailed look into derivation, while sections 4 and 5 present the implications and results of this approach.

2. DSEA+ probabilistic unfolding

In DSEA, the number of events in some energy bin i is reconstructed by summation of individual events' probabilities of belonging to the referred bin, corrected by its efficiency:

$$n_i^1 = \sum_j^N \frac{1}{\epsilon_i} P(i \mid E_j) \qquad (3)$$

Probability of event j being correctly reconstructed in the energy bin i is given by Bayes' theorem

$$P(i \mid E_j) = \frac{\alpha(i|E_j) n_i^0}{\sum_m^M \alpha(m|E_j) n_m^0} \qquad (4)$$

where $\alpha(i|E_j)$ is the likelihood of event j given that energy is defined inside i, and n_i^0 the $prior\ assumption$ of what part of events belong to i. Therefore, the prior assumption is normalized. Denominator in (4) is the normalization factor over all possible bins $i = 1, 2, ...M$. Each event has a different normalization factor. Probability is then normalized as well, while likelihood can take any non-negative value.

Likelihood is ought to be evaluated for each measurable parameter in the experiment used in unfolding. If the events $j = 1, 2, ...N$ are measured with a value p_j of some measurable p and having energy i (as constructed in the MC sample), the distribution of their likelihood can be inferred and fitted to a Gaussian function. A Gaussian likelihood is described with the mean $\mu_{i,p}$ and standard deviation $\sigma_{i,p}$ for each measurable and each energy bin. We assume then that the likelihood of some event j having a value p_j can be sampled from the Gaussian distribution by

$$\alpha(i|p_j) = \frac{1}{\sqrt{2\pi\sigma_{i,p}^2}} exp(-\frac{(p_j - \mu_{i,p})^2}{2\sigma_{i,p}^2}) \qquad (5)$$

Likelihood of event j is then taken to be the product of the events likelihoods in parameters by introducing the $naive\ asusumption$ per which the measurables are independent

$$\alpha(i|E_j) = \alpha(i|p_{j,1})\alpha(i|p_{j,2})...\alpha(i|p_{j,k}) = \prod_k^P \alpha(i|p_{j,k}) \qquad (6)$$

where $k = 1, 2, ...P$ is the number of parameters used for unfolding.

Number of true events in bin i is

$$n_i = \sum_j^N \frac{1}{\epsilon_i} \frac{\alpha(i|E_j)n_i^0}{\sum_m^M \alpha(m|E_j)n_m^0} \qquad (7)$$

$$= \sum_j^N U_{ij} \qquad (8)$$

where the expression can be written as an $i \times j$ matrix, referred to as the $unfolding\ matrix$, with M and N being the number of possible energies and measured events, respectively.

In the second step of DSEA, the prior assumption is replaced with the reconstructed distribution n_i that we just evaluated. The process is repeated, now yielding the result of second reconstruction to be

$$n_i^{(2)} = \sum_j^N \frac{1}{\epsilon_i} \frac{\alpha(i|E_j)n_i}{\sum_m^M \alpha(m|E_j)n_m} \qquad (9)$$

The process is repeated for an arbitrary number of iterations, usually until convergence is reached. We can define the reconstructed number of events in energy bin i after l steps

$$n_i^{(l)} = \sum_j^N \frac{1}{\epsilon_i} \frac{\alpha(i|E_j)n_i^{(l-1)}}{\varphi_j^{(l-1)}} = \sum_j^N U_{ij}^{(l)} \qquad (10)$$

with φ being the normalization factor, indexed by the prior it depends on. It is straightforward to see that the prior assumptions in steps 2 and larger are dependant on the likelihood, while only the prior n^0 in the first step is independent as it is an arbitrary chosen constant vector.

3. Error propagation

The effect of variables' error on the uncertainty of function dependant on them is referred to as error propagation, and is straightforward to calculate for linear functions. In nonlinear cases, as (7) is, functions are linearized by using the Taylor expansion, approximately

$$f(x) \approx f^0(x) + \sum_i \frac{\partial f(x)}{\partial x_i} \qquad (11)$$

$$\approx \mathbf{f^0} + \mathbf{Jx} \qquad (12)$$

where (12) is the matrix notation, J being the Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f_1}{\partial 1} & \cdots & \frac{\partial f_1}{\partial n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial 1} & \cdots & \frac{\partial f_m}{\partial n} \end{pmatrix} \qquad (13)$$

The covariance matrix of $f(x)$ due to propagation of error of x it depends on is then, in matrix notation,

$$\mathbf{\Sigma^f} = \mathbf{J\Sigma^x J^T} \qquad (14)$$

where $\mathbf{\Sigma^x}$ is the covariance matrix of x.

Matrix $\mathbf{U}$ depends on likelihood and prior but we cannot quantitatively determine the covariance of prior as it is an arbitrary chosen quantity. In subsequent iterations, the prior becomes dependant on the likelihood. Since the reconstructed distribution changes, its covariance matrix changes as well, and uncertainty has to be evaluated for the chosen number of iterations l, yielding the covariance matrix of n to be

$$\Sigma^{(l)}_{i,m} = \sum^{M}\sum^{N} \frac{\partial n_i^{(l)}}{\partial \alpha(k \mid E_j)} \Sigma^{\alpha}_{kj,on} \frac{\partial n_m^{(l)}}{\partial \alpha(o \mid E_n)}. \qquad (15)$$

Subsequent iterations depend on previous reconstructions which carry the error from reconstructions preceding them. We therefore expect additional terms in uncertainty of iterative unfolding coming from the propagation of uncertainty through repetition of steps.

To evaluate (15), we seek an expression for the differential of (10) over likelihood. Several special cases of this expression can be differed, namely the differential of n in the first step, in iterative case, and differentiation over the supporting and non-supporting likelihoods (we say that likelihood $\alpha(i|E_j)$ supports the energy bin i). For the simple case of the first step, the expression (7) can be expanded to

$$n_i = \sum_j^N \frac{1}{\epsilon_i} \frac{\alpha(i|E_j) n_i^0}{\sum_m^M \alpha(m|E_j) n_m^0} = \sum_j^N \frac{1}{\epsilon_i} \frac{\alpha(i|E_j) n_i^0}{\varphi_j^0} =$$

$$= \frac{1}{\epsilon_i} \frac{\alpha(i|E_1) n_i^0}{\sum_m^M \alpha(m|E_1) n_m^0} + \frac{1}{\epsilon_i} \frac{\alpha(i|E_2) n_i^0}{\sum_m^M \alpha(m|E_2) n_m^0} + \ldots + \frac{1}{\epsilon_i} \frac{\alpha(i|E_j) n_i^0}{\sum_m^M \alpha(m|E_j) n_m^0} + \ldots + \frac{1}{\epsilon_i} \frac{\alpha(i|E_N) n_i^0}{\sum_m^M \alpha(m|E_N) n_m^0} \qquad (16)$$

Knowing that

$$\frac{\partial\alpha(i|E_z)}{\partial\alpha(k|E_j)} = 0, \quad \frac{\partial\alpha(i|E_j)}{\partial\alpha(k|E_j)} = 0, \quad \frac{\partial\alpha(k|E_n)}{\partial\alpha(k|E_j)} = 0 \qquad (17)$$

and using the multiplication rule of derivation, all except one term in (16) can be dropped which gives

$$\frac{\partial n_i}{\partial\alpha(k|E_j)} = \alpha(i|E_j)n_i^0\epsilon_i^{-1}(-\varphi_j^{-2}n_k^0) = -U_{ij}\frac{n_k^0}{\varphi_j} \qquad (18)$$

for non-supporting likelihoods, where we used (8) for the short-hand expression. For supporting likelihoods, the result yields

$$\frac{\partial n_i}{\partial\alpha(i|E_j)} = n_i^0\epsilon_i^{-1}\varphi_j^{-1} + \alpha(i|E_j)n_i^0\epsilon_i^{-1}(-\varphi_j^{-2}n_i^0) = \frac{n_i^0}{\epsilon_i\varphi_j} - U_{ij}\frac{n_i^0}{\varphi_j} \qquad (19)$$

where the extra term comes from the dependency of both the numerator and denominator in (7) on given likelihood.

However, subsequent iterations use a dependant prior which mandates a revised calculation. (10) can be expanded as

$$n_i^{(l)} = \frac{1}{\epsilon_i}\frac{\alpha(i|E_j)n_i^{(l-1)}}{\sum_m^M \alpha(m|E_j)n_m^{(1-1)}} + \sum_{n\neq j}^{N}\frac{1}{\epsilon_i}\frac{\alpha(i|E_n)n_i^{(l-1)}}{\sum_m^M \alpha(m|E_n)n_m^{(l-1)}} \qquad (20)$$

where the j-th element has been extracted to showcase the special case of dependency on likelihood of all three constituting parts. Recalling (17), (19) and using the chain and multiplication rules of derivation, differentiation of n results in

$$\frac{\partial n_i^{(l)}}{\partial\alpha(k|E_j)} = \frac{\left(n_i^{(l-1)}\right)_{i=k} + \alpha(i|E_j)\frac{\partial n_i^{(l-1)}}{\partial\alpha(i|E_j)}}{\epsilon_i\varphi_j^{l-1}} - \frac{\alpha(i|E_j)n_i^{(l-1)}\epsilon_i\left[n_k^{(l-1)} + \sum_m^M \alpha(m|E_j)\frac{\partial n_m^{(l-1)}}{\partial\alpha(k|E_j)}\right]}{(\epsilon_i\varphi_j^{l-1})^2} +$$

$$+\sum_{n\neq j}^{N}\left[\frac{\alpha(i|E_n)\frac{\partial n_i^{(1)}}{\partial\alpha(k|E_j)}}{\epsilon_i\varphi_n^{(1)}} - \frac{\alpha(i|E_n)n_i^{(1)}\sum_m^M \alpha(m|E_n)\frac{\partial n_m^{(1)}}{\partial\alpha(k|E_j)}}{\epsilon_i(\varphi_n^{(1)})^2}\right]. \qquad (21)$$

4. Implications

The obtained expression (21) can be regrouped as

$$\frac{\partial n_i^{(l)}}{\partial\alpha(k|E_j)} = -U_{ij}^{(l-1)}\frac{n_k^{(l-1)}}{\varphi_j^{l-1}} + \left(\frac{n_i^{(l-1)}}{\epsilon_i\varphi_j^{l-1}}\right)_{i=k} + \sum_j^N\left[\frac{\alpha(i|E_j)\frac{\partial n_i^{(l-1)}}{\partial\alpha(k|E_j)}}{\epsilon_i\varphi_n^{(l-1)}} - \frac{\alpha(i|E_j)n_i^{(l-1)}\sum_m^M \alpha(m|E_j)\frac{\partial n_m^{(l-1)}}{\partial\alpha(k|E_j)}}{\epsilon_i(\varphi_n^{(l-1)})^2}\right]. \qquad (22)$$

The first two terms in addition correspond to (18) and (19). This is the error of the recosntruction in its first step. As already mentioned, the error in the first step spans only from the likelihood, as the prior is an arbitrary choice. When unfolding enters its second iteration, classifier uses the previous reconstruction as the prior, which now holds a statistical error of exactly as given in (18) and (19). This propagation of error results in an additional term corresponding to the last part of (22). The

dominant part is the second term under summation, which inflates the uncertainty in subsequent steps. However, if the solution converges, the difference between $n^{(l)}$ and $n^{(l-1)}$ goes to zero with rising number of iterations. This also reduces the last term in (22) implying that letting the number of iterations to infinity will not lead to infinite error, as both the error and the solution converge.

For each energy i there are j likelihoods for j measurements. We treat them as i values in j different events. Therefore, each event is a multivariate which can be described by a multinomial distribution. Although multinomials usually deal with probabilities, in this case the events correspond to likelihoods. A multinomial $M(n, p_1...p_k)$ described by n number of trials, p_k event probabilities for k mutually exclusive outcomes has a variance

$$Cov(X_i, X_j)_{i=j} = Var(X_i) = np_i(1 - p_i) \qquad (23)$$

$$Cov(X_i, X_j) = -np_i p_j \qquad (24)$$

where the number of trials corresponds to the number of measured events in the experiment. Then, the covariance matrix of n as given in (15) is

$$\sigma^{2(l)}_{i,m} = \sum^{M}\sum^{N} \frac{\partial n_i^{(l)}}{\partial\alpha(k \mid E_j)}(-N)\alpha(k \mid E_j)\alpha(o \mid E_n)\left(\frac{\alpha(k \mid E_j) - 1}{\alpha(o \mid E_n)}\right)_{o=k,n=j} \frac{\partial n_m^{(l)}}{\partial\alpha(o \mid E_n)} \qquad (25)$$

5. Results

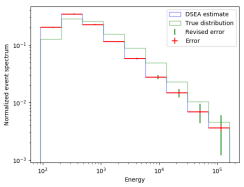
For a toy Monte Carlo unfolded over only seven variables, the revised uncertainty shows a better approach at capturing the error to the true distribution. To test the coverage of uncertainties, we use the pull distribution defined as

$$S_i = \frac{x_i - n_i}{\sigma_i} \qquad (26)$$

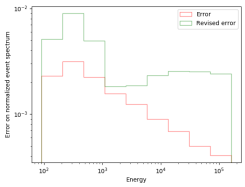
where x_i are individual event contributions, and n_i is the reconstruction. When the given uncertainty explains the data well, the pull distribution resembles a Gaussian, centered around 0 with a standard deviation of 1. Since DSEA+ calculates the mean of contribution in its reconstruction, we expect the mean of pull distribution to be exactly zero, but tend for a smaller variance in comparison to the previous method. A lower spread of the pull distribution points to a a better explanation of data. The results given in Table 1 show a clear reduction of the spread of the pull distribution σ_r on the revised error, as compared to the spread σ_c on the classical error.

6. Conclusion

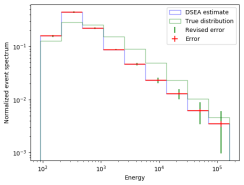
Unfolding is a long-standing problem in physics with many approaches currently used. A precise calculation of statistical errors surrounding the reconstructed distribution is especially important in deducing results and interpreting the nature of the problem. DSEA+ has several modes of operation, each mandating an individually tailored inspection of error propagation. In this work, an improved consideration of uncertainties in event spectrum reconstruction has been given. Results show an advancement in comparison to the classical approach. As future work, generalization of this approach to other models is planned.



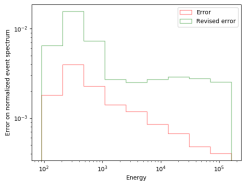
(a) Unfolded event spectrum after one iteration



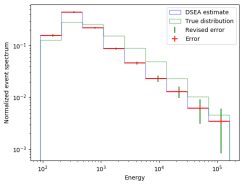
(b) Error magnitudes



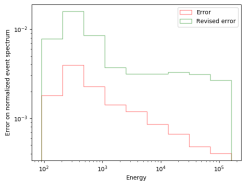
(c) Unfolded event spectrum after two iterations



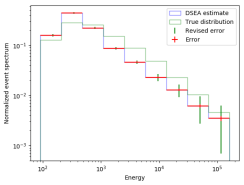
(d) Error magnitudes



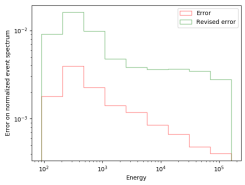
(e) Unfolded event spectrum after three iterations



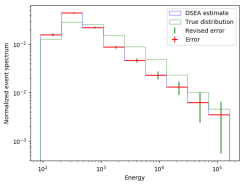
(f) Error magnitudes



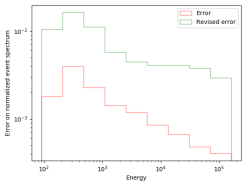
(g) Unfolded event spectrum after four iterations



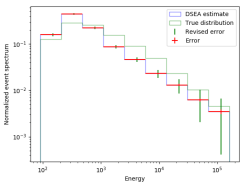
(h) Error magnitudes



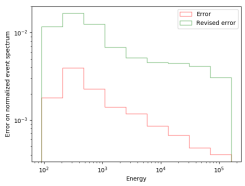
(i) Unfolded event spectrum after five iterations



(j) Error magnitudes



(k) Unfolded event spectrum after six iterations



(l) Error magnitudes

Figure 1: Comparison of classical and revised error on examples of DSEA+ from first up to the sixth iteration on a seven variable dataset

It.		Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Bin 8	Bin 9
	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	σ_c	82.5897	38.7647	75.4029	112.5668	129.4308	198.7757	239.57	328.9727	423.5272
	σ_r	12.6447	9.2268	13.7231	23.3991	29.7588	36.9707	35.7847	38.2404	55.2100
	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	σ_c	82.5897	38.7647	75.4029	112.5668	129.4308	198.7757	239.57	328.9727	423.5272
	σ_r	14.2800	9.3955	15.3905	27.5601	34.1803	41.2739	39.3227	41.68567	57.8545
	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	σ_c	82.5897	38.7647	75.4029	112.5668	129.4308	198.7757	239.57	328.9727	423.5272
	σ_r	16.3765	9.5626	17.4658	33.5201	40.1131	46.6654	43.6167	45.8056	60.7356
	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	σ_c	82.5897	38.7647	75.4029	112.5668	129.4308	198.7757	239.57	328.9727	423.5272
	σ_r	19.1433	9.7269	20.0848	42.7664	48.4621	53.5901	48.9287	50.8169	63.8809
	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	σ_c	82.5897	38.7647	75.4029	112.5668	129.4308	198.7757	239.57	328.9727	423.5272
	σ_r	22.9170	9.8876	23.4130	59.0456	60.9852	62.7464	55.6503	57.0380	67.3209

Table 1: The mean and standard deviations of pull distributions considering classical and revised error approach through iterations 2-6 on different bins in the reconstructed Toy Monte Carlo data

References

[1] T. Ruhe, 'D-SEA: A Data Mining Approach to Unfolding', Part of Proceedings, 33rd International Cosmic Ray Conference (ICRC2013) : Rio de Janeiro, Brazil, July 2-9, 2013

[2] K. Hymon, 'Seasonal Variations of the Unfolded Atmospheric Neutrino Spectrum with IceCube', Part of Proceedings, 37th International Cosmic Ray Conference (ICRC 2021): Berlin, Germany, July 12-23, 2021