Statistical uncertainty derivation in probabilistic classification with DSEA+

Leonora Kardum
technische universität dortmund

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

Introduction

In astronomy and particle physics, deconvolution or unfolding refers to the process of retrieving the true function $f(y)$ convoluted with a smearing matrix in the stochastic processes of detection, which results in observing the apparent distribution $g(x) = \int A(x|y)\, f(y) dy$. To handle the unstable solutions coming from high condition smearing matrices, advanced approaches to this problem ought to be utilized. Dortmund Spectrum Estimation Algorithm (DSEA+) is a procedure that solves the deconvolution task through employment of commonly used machine learning libraries.

DSEA+ unfolding

DSEA+ reconstructs the convoluted distribution by discretizing the observable space, effectively translating the problem to classification. Approximating $\vec{f}_j = \sum_j^n c_{ij}$ where c_{ij} are confidence values from the employed machine learning classifier, the solution is used to reweight the events and the process is repeated an arbitrary number of times. Therefore, DSEA+ is an iterative attempt at the solution retrieval.

Error propagation

When utilizing the Naive Bayes classifier, confidence values correspond to probabilities of events belonging to some discrete bin, corrected by the bin's efficiency, where the probability is given by Bayes theorem. After l iterations, the solution can be expressed as

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

and the covariance matrix of the solution being

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

The uncertainty spanning from previous reconstructions leaks to subsequent steps through the iterative nature of DSEA+. This mandates a careful consideration of error propagation in reconstruction repetition.

Solution

Consideration of error propagation through iterations leads to the differential on the reconstruction being

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

differential in non-iterative case

additional term enlarging the covariance matrix dependant on the number of iterations

Conclusion

This works presents a refined approach to uncertainty calculation in a single operating mode of DSEA+ resulting in confidence regions with better coverage (shown with pull distributions) in comparison to previous approaches.

inflated uncertainty, but with better coverage in comparison to previous methods was shown for modes with up to six iterations