PROCEEDINGS OF SCIENCE

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

Excess estimation in On/Off measurements including single-event variables

Giacomo D'Amico,[a,*] **Tomislav Terzić,**[b] **Jelena Strišković,**[c] **Michele Doro,**[d] **Marcel Strzys**[e] **and Juliane van Scherpenberg**[f]

[a] *Department for Physics and Technology, University of Bergen, Bergen NO-5020, Norway*
[b] *University of Rijeka, Department of Physics, 51000 Rijeka, Croatia*
[c] *Josip Juraj Strossmayer University of Osijek, Department of Physics, 31000 Osijek, Croatia*
[d] *Università di Padova and INFN, I-35131 Padova, Italy*
[e] *Institute for Cosmic Ray Research (ICRR), The University of Tokyo, Kashiwa, 277-8582 Chiba, Japan*
[f] *Max-Planck-Institut für Physik, D-80805 München, Germany*
E-mail: damico@mppmu.mpg.de, giacomo.damico@uib.no

Signal estimation in the presence of background noise is a common problem in many scientific disciplines. An "On/Off" measurement is when the background itself is imprecisely measured, which is the case for instance of observations performed in astronomy. We propose a new method for estimating the signal rate based on the Bayesian formalism. It uses information on single-event variables and their distribution for the signal and background population. Events are thereby weighted according to their likelihood of being a signal or a background event and background suppression can be achieved without performing data selection cuts. Simulating "On/Off" measurements from imaging atmospheric Cherenkov observations, we conclude that this new method is capable of increasing the resolution of the signal estimation, in particular for background dominated observations.

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

*Presenter

PoS(ICRC2021)702

1. Introduction

In an "On/Off" measurement the number of counts N_{on} and N_{off} are independent positive numbers that follows a Poisson distribution with expected counts $s + \alpha b$ (the signal rate plus the background rate in the On region) and b (the background rate in the Off region), respectively. The variable α is the normalization factor between the On and Off exposure.

Assuming flat priors $p(s)$ and $p(b)$ (with $s > 0$ and $b > 0$) and by applying the Bayes theorem, we get that the Probability Distribution Function (PDF) for the signal rate s is

$$p(s \mid N_{on}, N_{off}; \alpha) = \frac{\int db\ p(N_{on}, N_{off} \mid s, b; \alpha) p(b)\, p(s)}{\int ds\ db\ p(N_{on}, N_{off}, s, b; \alpha)} \propto \int db\ p(N_{on}, N_{off} \mid s, b; \alpha). \quad (1)$$

Thus the PDF of the signal rate s is proportional to the likelihood function in which the background rate b is integrated out, leaving a marginal distribution of s.

The likelihood function can be expressed in the following way:

$$p(N_{on}, N_{off} \mid s, b; \alpha) = p(N_{on} \mid s, \alpha b) \cdot p(N_{off} \mid b) = \frac{(s + \alpha b)^{N_{on}}}{N_{on}!} e^{-(s+\alpha b)} \cdot \frac{b^{N_{off}}}{N_{off}!} e^{-b}, \quad (2)$$

It can be shown that the integral in Eq. (1) is

$$p(s \mid N_{on}, N_{off}; \alpha) \propto \sum_{N_s=0}^{N_{on}} \frac{(N_{on} + N_{off} - N_s)!}{(1 + 1/\alpha)^{-N_s} (N_{on} - N_s)!} \cdot \frac{s^{N_s}}{N_s!} e^{-s}, \quad (3)$$

where N_s is the number of signal events in the On region with its Probability Mass Function (PMF) given by

$$p(N_s \mid N_{on}, N_{off}; \alpha) \propto \frac{(N_{on} + N_{off} - N_s)!}{(1 + 1/\alpha)^{-N_s} (N_{on} - N_s)!}. \quad (4)$$

For more details about the derivation of Eq. (3) and for a comparison with the frequentist approach for estimating the signal rate, one can see Ref. [1].

The main goal of this work is to extend Eq. (4) by including the information of the individual events without limiting ourselves with a "global" method that makes use only of the number N_{on} and N_{off}. In Sec. 2, we will explain how to introduce single event information in Eq. (4), and in Sec. 3 we will investigate the effects on the precision in the estimation of the number of the signal rate, using as an example real data and simulations from the MAGIC Imaging Atmospheric Cherenkov telescopes (IACTs).

2. Probability density function of the signal rate including single-event observables

In order to increase the Signal to Noise Ratio (SNR) it is common to select events based on signal extraction cuts on one or more event variables.

Cutting data has the disadvantage of excluding from the analysis also a fraction of the signal events, which translates to a reduced exposure on the target. Moreover, normally after the selection, all events surviving a specific set of cuts are treated as equally probable signal (or background) events, regardless their "distance" from the cuts.

We aim instead to fully exploit the information on how single-events variables distribute for a signal or a background population and replacing fixed signal extraction cuts with a statistical weighting of the events. We call this novel method Bayesian Analysis including Single-event Likelihoods `BASiL`.

We start by including the information about the variables $\mathbf{x}$, which we have observed for each event, in the inference of the signal rate s. The variable $\mathbf{x}$ might be a single observable (like a discriminating variable obtained by a classification algorithm) or a set of observables. It can be shown (see. Ref. [1]) that by including $\vec{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{N_{on}}\}$ in Eq. (1) we got that the PDF for the signal rate is give by:

$$p(s \mid \vec{\mathbf{x}}, N_{on}, N_{off}; \alpha) \propto \sum_{N_s=0}^{N_{on}} \frac{(N_{on} + N_{off} - N_s)!}{(N_{on} - N_s)!(1 + 1/\alpha)^{-N_s}} \frac{C(\vec{\mathbf{x}}, N_s)}{\binom{N_{on}}{N_s}} \cdot \frac{s^{N_s}}{N_s!} e^{-s}, \tag{5}$$

where one can recognize the PMF of N_s

$$p(N_s \mid \vec{\mathbf{x}}, N_{on}, N_{off}; \alpha) \propto \frac{(N_{on} + N_{off} - N_s)!}{(N_{on} - N_s)!(1 + 1/\alpha)^{-N_s}} \frac{C(\vec{\mathbf{x}}, N_s)}{\binom{N_{on}}{N_s}}. \tag{6}$$

The function C represents the combinatorial term:

$$C(\vec{\mathbf{x}}, N_s) = \sum_{A \in F_{N_s}} \prod_{i \in A} p(\mathbf{x}_i|\gamma) \cdot \prod_{j \in A^c} p(\mathbf{x}_j|\bar{\gamma}) \tag{7}$$

with F_{N_s} being the set of all subsets of N_s integer numbers that can be selected from $\{1, \ldots, N_{on}\}$, while the terms $p(\mathbf{x} \mid \gamma)$ and $p(\mathbf{x} \mid \bar{\gamma})$ are the PDFs of observing the variables $\mathbf{x}$ from a signal or background population, respectively. Depending on the kind of variable or problem under study, $p(\mathbf{x} \mid \gamma)$ and $p(\mathbf{x} \mid \bar{\gamma})$ can be estimated from MC simulations, a different data set or be based on a theoretical model.

With the introduction of the combinatorial term in Eq. (6) we have devised the method to include event-by-event information for the computation of N_s. The power of this method clearly depends on the specifics of the datasets in which it is applied, and in turn, it depends on (i) the event parameters that are used, (ii) how they distribute for the signal and background population, and (iii) how performing is the signal extraction method that relies on a fixed fiducial cut. However, in order to be predictive and define a framework to assess the performance of the BASiL method, we apply it to a specific case, that of gamma-ray observation. For this purpose, we analyze real data from the Major Atmospheric Gamma Imaging Cherenkov (MAGIC) Collaboration[1]. Results reported in Ref. [2] will be used as a benchmark case.

3. The case of Imaging Atmospheric Cherenkov telescopes

IACTs image the Cherenkov light emitted in the atmosphere by extended atmospheric showers generated by cosmic gamma rays (or cosmic rays) when entering the atmosphere. An irreducible

[1] https://magic.mpp.mpg.de/

background survives all possible image selection criteria and the signal estimation is performed through an "On/Off" comparison based, in which the Off sample is taken from a region in the sky where no signal is expected. For steady point-like sources two variables are generally further used to suppress the background: the squared[2] angular distance from the source θ^2, and a particle identification variable, which in the case of MAGIC is computed by means of a Random Forest (RF) algorithm, and is dubbed *Hadronness* (h) [3].

Therefore, the individual-event variables to consider are

$$\mathbf{x} = (\theta^2, h)\,.$$

For brevity and consistency with Ref. [2], we consider only events with estimated energy between 189 and 300 GeV.

The optimization of the SNR can be done in several ways. The MAGIC collaboration elaborated a set of cuts specific for each energy bin, according to an "efficiency" parameter ϵ defined as the fraction of Monte Carlo signal events surviving a certain cut. In the following, we elaborate on this, and compare the outcome with the novel method which we propose.

Assuming a signal rate s and background rate b, we simulate MAGIC On/Off observations and estimate the signal rate $\hat{s}$ using only the information about the total counts N_{on} and N_{off} and the single-event variables $\mathbf{x} = (h, \theta^2)$. This estimation is done using two different approaches, referred to as the "standard" and "BASiL" approach, which are descibed respectively below:

1. The estimated signal rate is $\hat{s} = N_{on} - \alpha N_{off}$ where N_{on} and N_{off} are the numbers of events surviving the cut in θ^2 and/or *Hadronness* for the On and Off measurement, respectively. Cut values are obtained assuming a given γ-ray efficiency ϵ computed from the signal distributions.

2. The estimated signal rate $\hat{s}$ is defined from the mode of the PMF in Eq. (6) where $\mathbf{x}$ can be either θ^2 and *Hadronness*, or only one of them. The combinatorial term in Eq. (7) will be obtained using signal and background likelihood values from the signal distributions and background distribution.

Assuming a background intensity in the On region $\alpha b = 1000$ and a SNR of 10%, i.e. $s = 100$, we got the signal estimation precision and bias reported in Fig. 1 for the θ^2 and *Hadronness* case separately.

We therefore conclude that the BASiL method, by including the likelihood of each event of being a signal or background, estimates the signal rate more precisely, while keeping the bias comparably low: for a SNR of 10% the improvement in precision is about $\sim 15\%$ in both *Hadronness* and θ^2.

After having evaluated the performance of the method by using MC simulations of events observed by the MAGIC telescopes, we now apply the method on a real data set. For this purpose we used the data[3] released by the MAGIC collaboration. The standard data analysis (whose results

[2]Signal events spread around the region of interest and for a point-like source they distribute according to a 2-dimensional Gaussian distribution. Such a 2-dimensional Gaussian in the θ_x and θ_y space will correspond to an exponential function for the distribution of $\theta^2 = \theta_x^2 + \theta_y^2$.

[3]The corresponding data in FITS format are publicly available in https://github.com/open-gamma-ray-astro/joint-crab/tree/master/data/magic

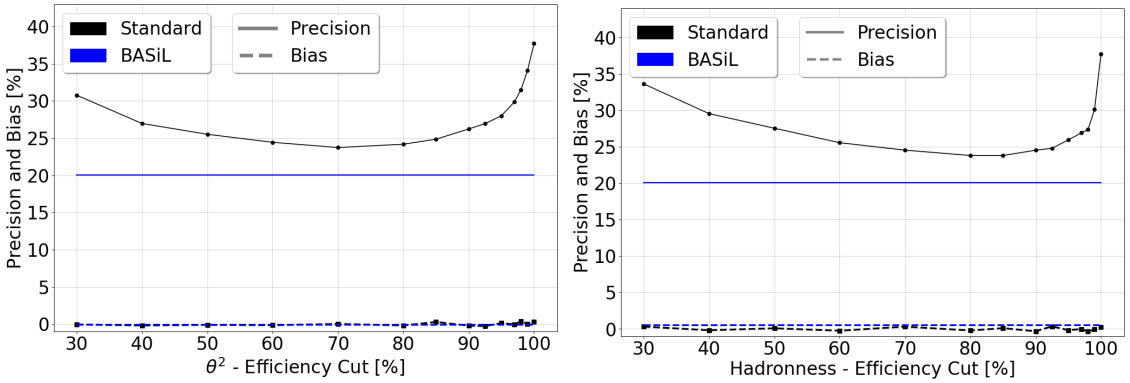

Figure 1: Comparison between the standard (black) and BASiL (blue) approach for the evolution of the precision (full line) and bias (dashed line) in the signal-rate estimation assuming different gamma-ray efficiency cut in θ^2 (left) and *Hadronness* (right). Note that for the BASiL approach the precision and bias do not depend on the efficiency, being $\epsilon = 1$ in such case. Nonetheless, for a visual comparison these values are shown as horizontal lines. Observations are simulated assuming $s = 10^2$ and $\alpha b = 10^3$, with $\alpha = 1/3$.

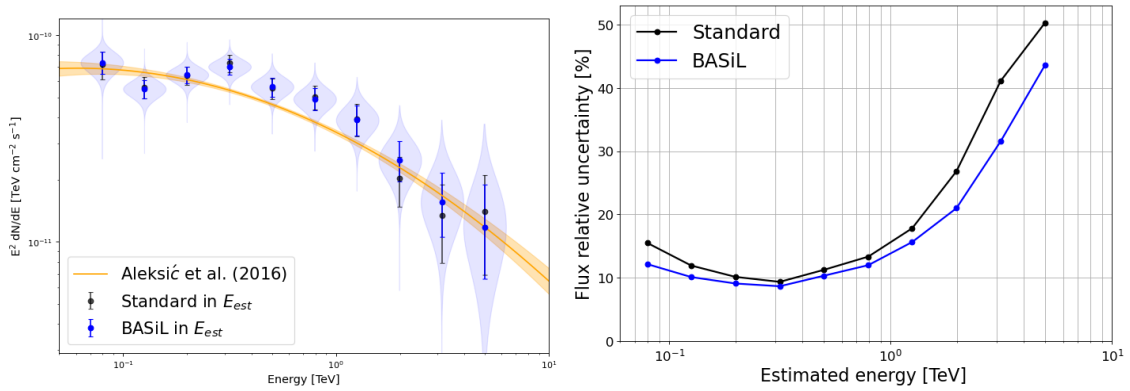

Figure 2: *Left:* SED in estimated energy of the Crab Nebula (in blue) obtained by processing 0.66 hours of data released by the MAGIC collaboration with the BASiL method. For comparison also the results (in black) obtained from the same data sample using the standard analysis procedure are reported in which efficiency cuts are applied. "Violin" plots around each blue point represent the flux PDF. The obtained results are also compared to the Crab Nebula SED (in orange) from Ref. [2]. *Right:* Relative uncertainty in the flux estimation for the standard (black) and BASiL (blue) approach.

are shown in black in the left plot of Fig. 2) has been performed using the MAGIC Analysis and Reconstruction Software (MARS) [4] where a *Hadronness* and θ^2 cut according to a high γ-ray efficiency (90% and 75% respectively) is applied. For the BASiL analysis instead (shown in blue in the left plot of Fig. 2) no cut is applied on the data set. Only a global $\theta^2 < 0.08\text{deg}^2$ is considered to define four identical non-overlapping regions from the center of the camera: one for the On region and three for the Off regions. In the right plot Fig. 2 we report the relative uncertainties in the flux estimation using the standard (in black) and BASiL (in blue) approach. One can see that relative uncertainties in the flux estimation are smaller in the BASiL approach, especially at higher energies where the signal rate is weaker. More details about the BASiL performances on On/Off measurements and on MAGIC data can be found in Ref. [1].

4. Conclusion and outlook

In this proceeding we introduced a novel method for estimating the signal rate in experiments with imprecisely measured background. Such method has been described in further details by the same Authors in Ref. [1].

The BASiL method, as we dubbed it, relies on the Bayesian, rather than the, more common, frequentist approach. Its main feature is that it weights events according to their individual likelihood of being signal or background, considering all the information available. This weighting is best summarized by the PMF of the number of signal events in Eq. (6), in which the novelty of the method, i.e. the combinatorial term defined in Eq. (7), shows up. By doing so, BASiL avoids cutting data according to some (or a combination of) variable to suppress the background, which inevitably discards a part of the signal. Moreover, the new method, while yielding results consistent with the standard data analysis method, improves the precision of the signal estimation.

5. Acknowledgments

We would like to thank the MAGIC Collaboration for permitting the use of proprietary Monte Carlo simulations and astronomical data. We particularly would like to thank A. Moralejo and J. Sitarek for useful discussions on this method. G.D. acknowledges funding from the Research Council of Norway, project number 301718. M.D. acknowledges funding from Italian Ministry of Education, University and Research (MIUR) through the "Dipartimenti di eccellenza" project Science of the Universe. T.T. and J.S. acknowledge funding from the University of Rijeka, project number 13.12.1.3.02. T.T. also acknowledges funding from the Croatian Science Foundation (HrZZ), project number IP-2016-06-9782. J. v. S. acknowledge funding from the International Max PlanckResearch School for Elementary Particle Physics.

References

[1] G. D'Amico, T. Terzić, J. Strišković, M. Doro, M. Strzys and J. van Scherpenberg, *Signal estimation in on/off measurements including event-by-event variables*, *Physical Review D* **103** (2021) 123001.

[2] J. Aleksić, S. Ansoldi, L.A. Antonelli, P. Antoranz, A. Babic, P. Bangale et al., *The major upgrade of the MAGIC telescopes, Part II: A performance study using observations of the Crab Nebula*, *Astroparticle Physics* **72** (2016) 76 [1409.5594].

[3] J. Albert, E. Aliu, H. Anderhub, P. Antoranz, A. Armada, M. Asensio et al., *Implementation of the Random Forest method for the Imaging Atmospheric Cherenkov Telescope MAGIC*, *Nuclear Instruments and Methods in Physics Research A* **588** (2008) 424 [0709.3719].

[4] R. Zanin, E. Carmona, J. Sitarek, P. Colin, K. Frantzen, M. Gaug et al., *MARS, The MAGIC Analysis and Reconstruction Software*, in *International Cosmic Ray Conference*, vol. 33 of *International Cosmic Ray Conference*, p. 2937, Jan., 2013.