Nuclear Inst. and Methods in Physics Research, A 867 (2017) 222–230

Contents lists available at ScienceDirect

Nuclear Inst. and Methods in Physics Research, A

journal homepage: www.elsevier.com/locate/nima

A Bayesian on–off analysis of cosmic ray data

Dalibor Nosek [a,*], Jana Nosková [b]

[a] Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
[b] Czech Technical University, Faculty of Civil Engineering, Prague, Czech Republic

A R T I C L E I N F O

Keywords:
Bayesian inference
On–off problem
Source detection
Cosmic rays

A B S T R A C T

We deal with the analysis of on–off measurements designed for the confirmation of a weak source of events whose presence is hypothesized, based on former observations. The problem of a small number of source events that are masked by an imprecisely known background is addressed from a Bayesian point of view. We examine three closely related variables, the posterior distributions of which carry relevant information about various aspects of the investigated phenomena. This information is utilized for predictions of further observations, given actual data. Backed by details of detection, we propose how to quantify disparities between different measurements. The usefulness of the Bayesian inference is demonstrated on examples taken from cosmic ray physics.

1. Introduction

The search for new phenomena often yields data that consists of a set of discrete events distributed in time, space, energy or some other observables. In most cases, source events associated with a new effect are hidden by background events, while these two classes of events cannot be distinguished in principle. Such a search can be accomplished with an on–off measurement by checking whether the same process of a constant but unknown intensity may be responsible for observed counts in the on-source region, where a new phenomenon is searched for, and in the reference off-source region, where only background events contribute. Any inconsistency between the numbers of events collected in these zones, when they are properly normalized, then indicates the predominance of a source producing more events in one explored region over the other.

In this study, we focus on the problems which are often encountered when searching for cosmic ray sources while detecting rare events. Characteristics of possible sources are usually proposed based on analysis of a test set of observed data. Then, further observations are to be conducted in order to examine the presence of a source or to improve conditions for its verifications. But, due to unknown phenomena, the outcome is always uncertain which calls, first, for as less as possible initial assumptions about underlying processes and, second, for the quantification of disparities between observations with the option to correct for experimental imperfections.

In order to satisfy the first condition, we follow our previous analysis of on–off measurements formulated within the Bayesian setting [1]. Unlike other Bayesian approaches [2–9], we handle the source and background processes on an equal footing. This option provides us with solutions that are minimally affected by external presumptions. In order to track the behavior of a signal registered in a selected on-source region, we utilize variables with the capability to assess the consistency between on–off measurements. Specifically, giving the net effect, the difference variable [1] is well suited for estimating source fluxes if exposures are known. In case of stable or at least predictable background rates, we eliminate the effect of exposures by using fractional variables which reveal relatively the manifestation of a source. For example, the time evolution of a given source, if still observed in the same way, is easily examined by the ratio of the on-source rate to the total rate. In a more general case, we employ the on-source rate expressed in terms of the rate deduced from the background. In summary, we receive posterior distributions of different variables that include what is available from measurements, while providing us with all kinds of estimates, as traditionally communicated, and allowing us to make various observation-based predictions.

Related to the on–off issue, the Bayesian inference provides solutions in the case of small numbers, including the null experiment or the experiment with no background, when classical methods based on the asymptotic properties of the likelihood ratio statistic [10–13] are not easily applicable. Also, there are no difficulties with the regularity conditions of Wilks' theorem, with unphysical likelihood estimates or with the discreteness of counting experiments, in general, see e.g. Refs. [14–17]. On the other hand, the subjective nature of Bayesian reasoning, often

* Corresponding author.
E-mail address: nosek@ipnp.troja.mff.cuni.cz (D. Nosek).

http://dx.doi.org/10.1016/j.nima.2017.06.034
Received 24 August 2016; Received in revised form 1 May 2017; Accepted 21 June 2017
Available online 27 June 2017
0168-9002/

mentioned as its disadvantage, may be at least partially eliminated by using a family of uninformative prior options.

The proposed method is suitable for experiments searching for rare events in which the observational conditions may not be adjusted optimally, with little opportunity for repeating measurements conducted under exactly the same conditions. Besides searches for possible sources of the highest energy cosmic rays, see e.g. Refs. [18–22], examples include observations of peculiar sources which exhibit surprising temporal or spectral behavior. Another class of observations comprises searches for events accompanying radiation from transient sources that have been identified in different energy ranges. The identification of the properties of very-high-energy γ-rays associated with observed gamma-ray bursts belongs to this class of problems [1–3].

The structure of this paper is as follows. Our formulation of the Bayesian approach to the on–off problem is described in Section 2, complemented by five Appendices. Further details about our approach can be found in Ref. [1]. In Section 2.1 we summarize how to store experimental information by using appropriate on–off variables. Two ways to examine possible inconsistencies in independent observations are proposed in Sections 2.2 and 2.3. Several realistic examples taken from cosmic ray physics are presented and discussed in Section 3. The paper is concluded in Section 4.

2. Bayesian inferences from on–off experiment

In the on–off experiment, two kinds of measurements are collected in order to validate a source signal immersed in background. The number of on-source events, $n_{\rm on}$, is recorded in a signal on-source region, while the number of off-source events, $n_{\rm off}$, detected in a background off-source zone serves as a reference measurement. The on- and off-source counts are modeled as discrete random variables generated in two independent Poisson processes with unknown on- and off-source means, $\mu_{\rm on}$ and $\mu_{\rm off}$, i.e. $n_{\rm on} \sim {\rm Po}(\mu_{\rm on})$ and $n_{\rm off} \sim {\rm Po}(\mu_{\rm off})$. The relationship between the on- and off-source zone is ensured by the ratio of on- and off-source exposures $\alpha > 0$.

In the Bayesian approach, for on- and off-source means we adopted a family of prior distributions conjugate to the Poisson sampling process [1]. This family consists of Gamma distributions, i.e.

$$\mu_{\rm on} \sim {\rm Ga}(s_p, \gamma_p - 1), \qquad \mu_{\rm off} \sim {\rm Ga}(s_q, \gamma_q - 1), \tag{1}$$

where $s_p > 0$ and $s_q > 0$ are prior shape parameters, and the prior rate parameters $\gamma_p > 1$ and $\gamma_q > 1$. It includes several frequently discussed options, i.e. scale invariant, uniform, as well as Jeffreys' prior distributions. After the on–off measurement has been conducted, when $n_{\rm on}$ and $n_{\rm off}$ counts were registered independently in the on- and off-source regions, using Eq. (1) we obtain independent posterior distributions

$$(\mu_{\rm on} | n_{\rm on}) \sim {\rm Ga}(p, \gamma_p), \qquad (\mu_{\rm b} | n_{\rm off}) \sim {\rm Ga}\left(q, \frac{\gamma_q}{\alpha}\right), \tag{2}$$

where $\mu_{\rm b} = \alpha\mu_{\rm off}$ denotes the expected background rate in the on-source zone and $p = n_{\rm on} + s_p$ and $q = n_{\rm off} + s_q$. For more details see Ref. [1].

We recall that our next steps diverge from the traditional treatment. In order to assess what is observed, we define suitable on–off variables by combining the on- and off-source means, assuming that the underlying processes are independent. From the Bayesian perspective, this choice is motivated by the fact that, according to Jeffreys' rule, the joint prior distribution is separable in the on- and off-source means [1,2]. Furthermore, as in classical statistical approaches [10–16], the proposed option allows us to obtain adequate results regardless of in which of the two zones the source effects are revealed [1,7].

2.1. On–off variables

In our previous work [1], we focused on the properties of the difference between the on-source and background means, $\delta = \mu_{\rm on} - \mu_{\rm b}$, using maximally uninformative joint distributions, as dictated by the principle of maximum entropy. In this section, we briefly recapitulate our previous result and introduce other on–off variables that equally well describe the on–off problem.

Under the transformation $\delta = \mu_{\rm on} - \mu_{\rm b}$, with a real valued domain, while keeping $\mu_{\rm b} = \alpha\mu_{\rm off}$ unchanged and marginalizing over $\mu_{\rm b}$, the probability density function of the difference is (for details of our notation see Ref. [1])

$$f_\delta(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\Gamma(p)}\, e^{-\gamma_p x} x^{p+q-1}\, U(q, p+q, \eta x), \quad x \geq 0, \tag{3}$$

$$f_\delta(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\Gamma(q)}\, e^{\frac{\gamma_q}{\alpha} x} (-x)^{p+q-1}\, U(p, p+q, -\eta x), \quad x < 0, \tag{4}$$

where $p = n_{\rm on} + s_p$, $q = n_{\rm off} + s_q$, $\eta = \gamma_p + \frac{\gamma_q}{\alpha}$, $\Gamma(a)$ stands for the Gamma function and $U(a, b, z)$ is the Tricomi confluent hypergeometric function [23]. Exhaustive discussion concerning this distribution can be found in Ref. [1], where also some special cases ($\gamma_p = \gamma_q \to 1$) based on uninformative prior distributions, scale invariant ($s_p = s_q \to 0$), Jeffreys' ($s_p = s_q = \frac{1}{2}$) and uniform ($s_p = s_q = 1$) options, are described.

The difference δ yields information about the source flux. The posterior distribution of the source flux is obtained by a scale transformation, i.e. $j = \delta/a$ where $a = \frac{\alpha}{1+\alpha} A$ is the exposure of the on-source zone and A denotes the integrated exposure of the on–off experiment, both considered as constants.

A similar picture is obtained with the ratio of the on-source and background means ($\mu_{\rm b} = \alpha\mu_{\rm off}$)

$$\beta = \frac{\mu_{\rm on}}{\mu_{\rm b}}, \quad \beta \geq 0. \tag{5}$$

This variable represents the intensity registered in the on-source region expressed in terms of the background intensity, i.e. $\beta \leq 1$ when no source is present in the on-source zone. The ratio β obeys the generalized Beta distribution of the second kind [24], $\beta \sim {\rm B}_{\rm g2}(p, q, \rho)$ where $p = n_{\rm on} + s_p$, $q = n_{\rm off} + s_q$ and $\rho = \alpha\gamma_p/\gamma_q$, with the probability density function

$$f_\beta(x) = \frac{\rho^p}{B(p,q)} \frac{x^{p-1}}{(1+\rho x)^{p+q}}, \quad x \geq 0, \tag{6}$$

where $B(a, b)$ is the Beta function [23]. This posterior distribution was obtained after the transformation $\beta = \mu_{\rm on}/\mu_{\rm b}$ while treating $\mu_{\rm on}$ and $\mu_{\rm b}$ as independent variables (see Eq. (2)) and keeping $\mu_{\rm b}$ unchanged, with the Jacobian $J = \mu_{\rm b}$, and marginalizing over $\mu_{\rm b}$.

In a special case, using the uniform prior distributions for the on- and off-source means, i.e. $\gamma_p = \gamma_q \to 1$ and $s_p = s_q = 1$, and assuming that the on–off data were registered in the regions of the same exposure, when $\rho = \alpha = 1$, the posterior distribution for the ratio β written in Eq. (6) reduces to the result given originally in Ref. [5]. Assuming $\gamma_p = \gamma_q \to 1$ and $\alpha = 1$, i.e. $\rho = 1$, the result presented in Eq. (13) in Ref. [6] is obtained.

In some cases, it may be appropriate to use a variable

$$\omega = \frac{\mu_{\rm on}}{\mu_{\rm on} + \mu_{\rm off}}, \quad \omega \in \langle 0, 1 \rangle, \tag{7}$$

that represents the fraction of the total intensity registered in the on-source zone. Considering that $\omega = \alpha\beta/(1+\alpha\beta)$, we recover from Eq. (6) that the probability density function of the proportion ω is

$$f_\omega(x) = \frac{\kappa^p}{B(p,q)} \frac{x^{p-1}(1-x)^{q-1}}{[1 + (\kappa - 1)x]^{p+q}}, \quad x \in \langle 0, 1 \rangle, \tag{8}$$

where $p = n_{\rm on} + s_p$, $q = n_{\rm off} + s_q$ and $\kappa = \gamma_p/\gamma_q$ is the ratio of the prior rate parameters. In this case, equally intensive on- and off-source processes ($\mu_{\rm on} = \mu_{\rm b}$) are described by a balance value of $\omega = \frac{\alpha}{1+\alpha}$.