EP4130 – PROJECT REPORT

Aryan Sharan Reddy – BT21BTECH11002

Manikanta Uppulapu – BT21BTECH11005

The On-Off Problem: An Objective Bayesian Analysis

Introduction

Counting experiments are commonly used to measure discrete sets of events, often modelled using the Poisson distribution [1, 2]. While the Poisson distribution can be approximated by a normal distribution for large event counts, this approximation is inadequate when dealing with rare data occurrences. Figure 1 illustrates a typical example of low count data, exemplified by an observation of a Gamma Ray Burst (GRB) using the Fermi-LAT instrument. This scenario raises the fundamental question: How should one proceed when obtaining additional data is unfeasible?

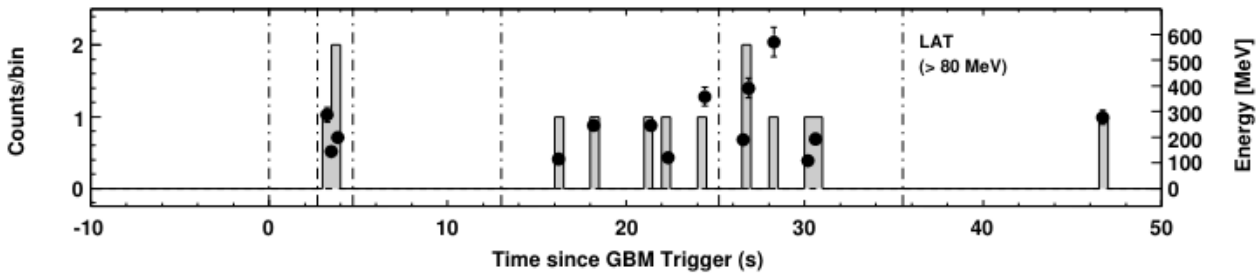


Figure 1: A typical low count high-energy astrophysics data set. It shows gamma rays measured from GRB080825C as observed by Fermi-LAT. Black dots represent the energy measurement, the gray bars represent the number of photons. Figure reproduced from [3]

The On-Off Problem

The On-Off problem, also known as the Li-Ma problem, involves figuring out the signal rate when we're not sure about the background rate. We look at two areas: one where there might be a signal (called "on" region) and another where we know there's no signal (called "off" region). We also consider a third parameter, α, which is the ratio

of exposures between these two regions. In gamma ray astronomy, Berge et al. provide a good explanation of this problem.

Common methods used to solve this problem often assume normal distribution of numbers, which doesn't work well with low-count data and can be problematic at the edge of the parameter space.

There are two main types of Bayesian solutions: subjective Bayesian and tail-area methods. Subjective Bayesian methods introduce a subjective parameter, making the result somewhat dependent on the person doing the analysis. Tail-area methods avoid specifying an alternative hypothesis altogether, but they tend to overestimate probabilities.

None of the existing methods in the literature fully cover all aspects of the problem: we need to calculate the probability that the observed counts are only from the background and estimate the signal contribution.

In this study, I propose a two-step objective Bayesian solution to address these challenges in a unified way. This solution is inspired by previous work and aims to provide a more comprehensive approach to the On-Off problem.

Methods Development

Objective Bayesian analysis uses simple priors, called "flat" priors, to represent our lack of prior knowledge. These priors are often improper, meaning they don't add up to one. But when combined with Bayes' theorem, they help us generate proper results and answer important questions: What if our data were more dominant?

One common objective Bayesian prior is Jeffreys's rule, which was developed to ensure that our results remain consistent regardless of how we re-parameterize our problem. This rule is crucial for our analysis.

Our method follows the approach outlined by previous work and involves two main steps. First, we calculate the likelihood that our observed counts come from the background model. If this likelihood is below a certain threshold, we conclude that we've detected a signal. Second, depending on whether we've detected a signal or not, we estimate either the signal's contribution or its upper limit. The first step uses objective Bayesian hypothesis testing with Bayes factors, while the second step uses objective Bayesian estimation.

First Step: Hypothesis Testing

In objective Bayesian analysis, one tricky aspect is that the priors we use are only defined up to a certain constant, which matters in our calculations. There isn't a clear agreement on how to handle this in objective Bayesian hypothesis testing.

To solve this problem, I suggest using a method called the "minimal sample device." This method helps us deal with the issues related to these constant factors. We've evaluated these assumptions specifically for the On-Off problem. In the end, we compute the odds of the background model versus the signal model.

$$B_{01}=\frac{c_0\,\gamma}{c_1\,\delta} \tag{3.1}$$

where γ and δ are defined using the Gamma function Γ(x) and the hypergeometric function 2 F1 (a,b;c;z):

$$\gamma := (1+2\,N\,off)\,\alpha^{\left(\frac{1}{2}+N_{on}+N\,off\right)}\,\Gamma\left(\frac{1}{2}+N_{on}+N\,off\right)$$

$$\delta := 2(1+\alpha)^{N_{on}+N_{off}}\,\Gamma(1+N_{off}+N_{on})\,{}_2F_1\left(\frac{1}{2}+N_{off},1+N_{off}+N_{on};\frac{3}{2}+N_{off};-\frac{1}{\alpha}\right)$$

$$\frac{c_0}{c_1}=\frac{2\arctan\left(\frac{1}{\sqrt{\alpha}}\right)}{\sqrt{\pi}} \tag{3.2}$$

Code:

```python
from numpy import frompyfunc, arange
from scipy.optimize import fminbound
from scipy.special import hyp2f1, gamma, rgamma
from scipy.integrate import quad
from scipy.special import erfinv
import matplotlib.pyplot as plt
```

```python
def bayes_factor(Non, Noff, alpha):
    """
    Calculate the Bayes factor, representing the odds of background hypothesis over the signal hypothesis.
    """
    Nges = Non + Noff
    gam = (1 + 2 * Noff) * alpha**0.5 * gamma(0.5 + Nges)
    delta = 2 * (1 + alpha)**Nges * gamma(1 + Nges) * hyp2f1(0.5 + Noff, 1 + Nges, 1.5 + Noff, -1 / alpha)
    c1_c2 = 2 * (pi / 4) ** 0.5
    return gam / (c1_c2 * delta)
```

To determine if we've detected a signal, we use Equation 3.1 and check if the odds of the background model are low. I suggest using a "Bayesian z-value," similar to what's outlined in a previous study.

$$S_b=\sqrt{2}\,erf^{-1}(1-B_{01}) \tag{3.3}$$

For example, if B_{01} equals 5.7×10^{-7}, it corresponds to Sb = 5 or "5 sigma." This makes it easy to compare with frequentist significance methods. However, it's important to note that the odds of a model and the frequency of an outcome are different. B_{01} explicitly considers alternative models, while frequentist methods don't.

Code:

```python
def bayesian_z_value(Non, Noff, alpha):
    """
    Calculate the Bayesian z-value via the Bayes factor.
    """
    bayes_factor_val = bayes_factor(Non, Noff, alpha)
    buf = 1 - bayes_factor_val
    return (2 ** 0.5) * erfinv(max(min(buf, 1), -1))
```

After we've determined the Bayes factor comparing the background model to the signal model, we move on to estimating the signal contribution using objective Bayesian estimation.

If the data indicate a significant detection, it suggests that the signal model is likely true. In this case, we calculate the most probable value of the signal parameter and establish a physical error interval.

If the data don't show a significant detection, we calculate an upper limit on the signal parameter, assuming the signal is present but too weak to measure.

In both cases, we need the conditional probability $P(\lambda_s \mid N_{on}, N_{off}, H_1)$ of the signal parameter λ_s, given the counts and the signal model H_1.

The use of an improper prior is acceptable here because the proportionality constant cancels out, resulting in a proper posterior. After integrating out the background parameter λ_{bg}, the result obtained is:

$$P(\lambda_s|N_{on},N_{off},H_1)=P_p(N_{on}+N_{off}|\lambda_s)\frac{U[\frac{1}{2}+N_{off},1+N_{off}+N_{on},(1+\frac{1}{\alpha})\lambda_s]}{{}_2\widetilde{F}_1(\frac{1}{2}+N_{off},1+N_{off}+N_{on};\frac{3}{2}+N_{off};-\frac{1}{\alpha})} \tag{3.4}$$

The posterior distribution, which holds all the information about the signal parameter, is expressed using three functions: the Poisson distribution $P_P(N \mid \lambda)$, the regularized hypergeometric function ${}_2\widetilde{F}_1(a,b;c;z)=\frac{{}_2F_1(a,b;c;z)}{\Gamma(c)}$, and the Tricomi confluent hypergeometric function $U(a, b, z)$.

To determine a flux, one should identify the mode λ_s^*, of the posterior distribution P (λ_s | N_{on} , N_{off} , H_1) as the signal estimator.

The error on this signal estimator can be calculated numerically from the cumulative distribution function. One notable option is the highest posterior density interval (HPD) [λ_{min} , λ_{max}] containing 68% probability, which is determined by:

$$\int_{\lambda_{min}}^{\lambda_{max}} P(\lambda_s|N_{on},N_{off},H_1)d\lambda_s=0.68, \tag{3.5}$$

together with the constraint:

$$P(\lambda_{min}|N_{on},N_{off},H_1)=P(\lambda_{max}|N_{on},N_{off},H_1) \tag{3.6}$$

In case an upper limit should be calculated one can solve the cumulative distribution function, for instance, for a 99% probability limit λ_{99} on the signal parameter λs as:

$$\int_{0}^{\lambda_{99}} P(\lambda_s|N_{on},N_{off},H_1)d\lambda_s=0.99. \tag{3.7}$$

In a Bayesian approach, we can naturally calculate key results, which can be challenging with frequentist methods. Frequentist approaches often struggle with certain calculations and encounter difficulties at the edges of parameter space.

With Bayesian methods, we handle all potential number counts uniformly, whether there are zero counts or thousands. This ensures that probability intervals for the signal are always physically meaningful.

As an example, let's consider gamma rays from Gamma-Ray Bursts (GRBs), brief flashes of gamma rays from outer space. Understanding how GRBs produce high-energy gamma rays is important in astrophysics. Due to their short duration and intensity, we typically detect only a few events during a burst or shortly after.

For instance, let's look at GRB080825C observed by Fermi-LAT. We had 15 "on" events and 19 "off" events, with a certain exposure ratio. Calculating the odds, we found strong evidence favoring the background model, indicating the detection of the GRB. Next, we estimated the signal parameter, which aligned well with published values.

In another example, GRB080330 observed by the VERITAS Cherenkov telescope showed no "on" events and 15 "off" events. Here, the odds favored the null hypothesis due to the absence of any "on" events. However, assuming the source is present, we determined an upper limit to the signal parameter, which contrasted with a frequentist method's result. Further analysis indicated that the frequentist method tends to overestimate, particularly at certain boundaries, making it limited in such cases. These limitations are overcome by the Bayesian method.

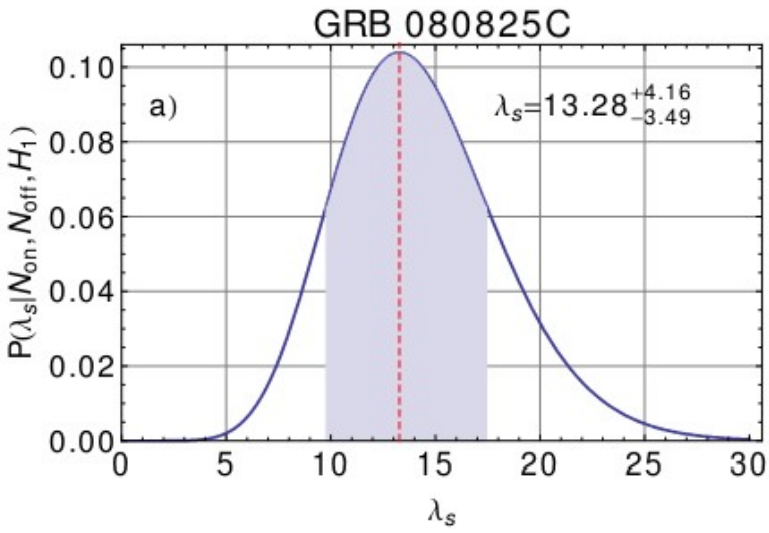


Figure 2: The conditional probability P(λs |Non,Noff,H1) of the signal λs , given the Fermi-LAT number counts of GRB080825C. The blue band indicates the HPD interval for the signal parameter posterior probability. Figure reproduced from [8].

Validation and Discussion

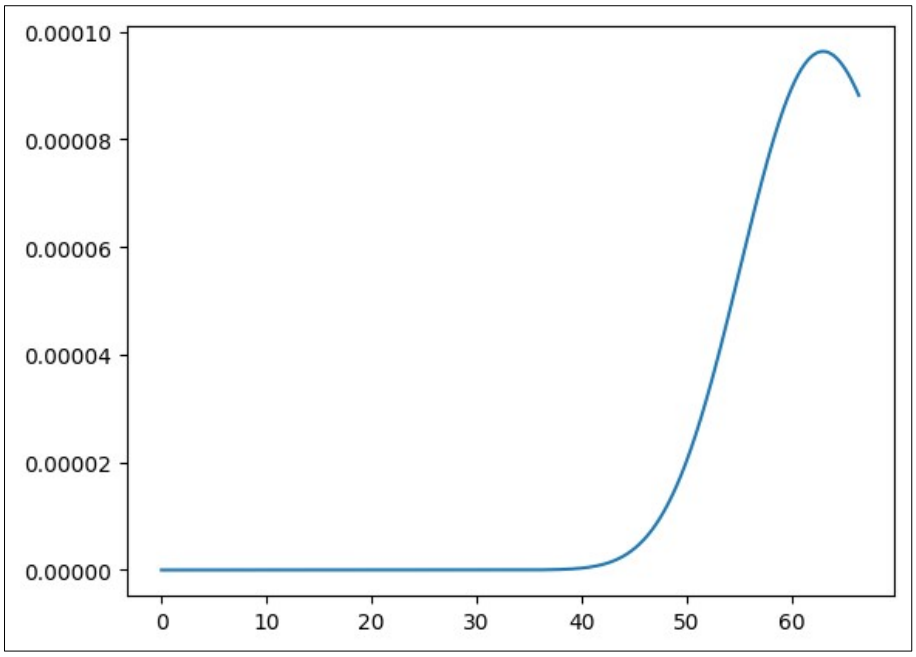
To make sure our method is reliable and the assumptions we've made are reasonable, we conducted thorough validation tests. These tests showed that our two-step method works well across different scenarios, especially when the number of observed events is close to the expected ratio of "on" to "off" events. Our objective Bayesian hypothesis testing matches closely with results from other methods when dealing with a large number of events. Furthermore, our objective Bayesian signal estimation accurately estimates the true signal parameter while providing a dependable error estimate.

```python
if __name__ == "__main__":
    n_on = 29
    n_off = 34
    alpha = 1 / 5

    l_99_buf = credible_interval(n_on, n_off, alpha, 0.99)
    significance = bayesian_z_value(n_on, n_off, alpha)
    if significance > 3.0:
        l_star_buf = find_local_maximum(n_on, n_off, alpha)
        l_84_buf = credible_interval(n_on, n_off, alpha, 0.84)
        l_16_buf = credible_interval(n_on, n_off, alpha, 0.16)
        print("Measurement: N_on, N_off, alpha, B_01, S_B_01, signal_estimate")
        print(f"{n_on}, {n_off}, {alpha:.3f}, {bayes_factor(n_on, n_off, alpha):.3e}, {significance:.3f}, "
              f"{l_star_buf:.3f} + {l_84_buf - l_star_buf:.3f} - {l_star_buf - l_16_buf:.3f}")
    else:
        print("Measurement: N_on, N_off, alpha, B_01, S_B_01, signal_estimate")
        print(f"{n_on}, {n_off}, {alpha:.3f}, {bayes_factor(n_on, n_off, alpha):.3e}, {significance:.3f}, "
              f"<{l_99_buf:.3f}")

    # Plot the PDF
    x_range = arange(0, 2 * l_99_buf, 0.1)
    y_range = signal_posterior(x_range, n_on, n_off, alpha)
    plt.plot(x_range, y_range)
    plt.show()
```

```
Measurement: N_on, N_off, alpha, B_01, S_B_01, signal_estimate
29, 34, 0.200, 1.020e+37, -inf, <33.220
```



Source code used to reproduce the results can be found in this **github** repository

References

1. On the On-Off Problem: An Objective Bayesian Analysis – *https://arxiv.org/abs/1508.05855*
2. Code used in the [1]: *https://bitbucket.org/mknoetig/obayes_onoff_problem/src/master/*