UNIVERSITÀ DEGLI STUDI DI PADOVA

Locating γ-Ray Sources on the Celestial Sphere via Modal Clustering

Anna Montin, Giovanna Menardi & Alessandra R. Brazzale

COSMOSTATS 2023 - Bridging the Gap: Statistical Modeling of Cosmology Extremes — Poster Session: 13 September 2023

Framework and Motivation

Searching for as yet undetected γ-ray sources is a major target of the *Fermi* LAT Collaboration [1].

- What are γ-ray sources: celestial objects which emit γ-ray photons, i.e. quanta of light in the highest energy range.
- Why is it important to locate them: to study our Universe and to better understand the mechanisms that create and accelerate particles emitted by celestial objects.
- How are γ-ray sources observed: the Large Area Telescope (LAT) on board the *Fermi* spacecraft detects γ-rays in the galactic and extra-galactic space.

State of the art

The problem of identifying new γ-ray sources has been widely covered in literature. Approaches vary in the:

- use of the data (Galactic coordinates on 2-dimensional sky maps or Cartesian coordinates on 3-dimensional sky maps as shown in Figure 1);
- applied statistical methods (parametric, semi-parametric or non-parametric framework) [6, 4].

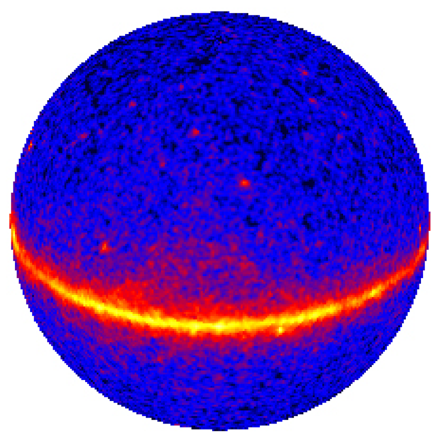

Figure 1. Spherical representation of the γ-ray photon counts detected by the Fermi-LAT. The brighter the colour, the higher the photon count: the Galactic plane shows our galaxy, the Milky Way.

Our goal

To develop an algorithm which treats the problem in a non-parametric framework while accounting for the directional nature of the data.

Data

The units detected by the LAT are photons. Data consist of an event list which gives the direction in the sky of each detected photon together with additional information such as photon energy and quality.

- We work with photon directions expressed by Cartesian coordinates on the 3-dimensional sphere.
- Directions in $\mathbb{R}^3$ are represented as unit vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where n is the number of units in the observed sample and $\mathbf{x}$ is a point on the sphere $\Omega^2 = \{\mathbf{x} \in \mathbb{R}^3 : ||\mathbf{x}||_2 = x_1^2 + x_2^2 + x_3^2 = 1\}$ with unit radius and centre at the origin.
- Photons are emitted by celestial sources and from the background.
- The diffuse γ-ray background spreads over the entire area observed by the telescope.
- Our goal is to identify *point-like sources* characterized by high concentration of photons.

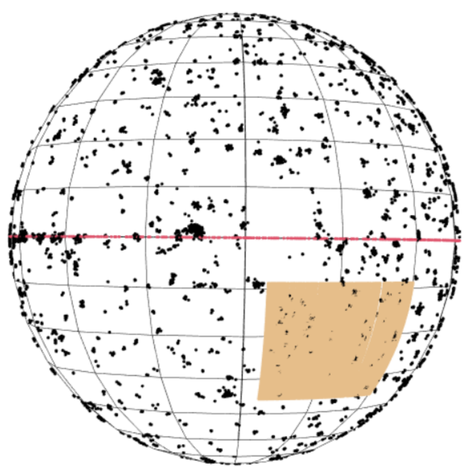

Figure 2. Fermi-LAT γ-ray photon count map for a 5-year observation period. In yellow the analysed region and in red the Galactic plane.

Problems faced and	Proposed solutions
▪ Searching for regions with high photon concentration + Directional nature of the data	→ Modal clustering for directional data
▪ Heavy presence of background emission	→ Non-parametric discriminant rule
▪ Evaluate the significance of each celestial source	→ Resampling-based statistical test

Methods and Models

Modal clustering

- A probability density, estimated in a non-parametric way, underlies the data.
- Clusters are defined as domains of attraction of the density modes.

Since our purpose is to identify *point-like γ-ray sources*, we can match regions of space with a high concentration of photons to the modes of the photon density [3].

Density estimation

- Directional kernel estimator with von Mises-Fisher kernel $K(\cdot)$

$$\hat{f}_{h_i}(\mathbf{x}) = \frac{c_{h_i}(K)}{n} \sum_{i=1}^{n} K\left(\frac{1 - \mathbf{x}^\top \mathbf{x}_i}{h_i^2}\right)$$

where $c_{h_i}(K)$ is the associated normalizing constant.
- Variable bandwidth given by the scale parameter of the *point-spread function* (PSF)[2]

$$h_i = \sqrt{\left[c_{0,i}\left(\frac{E_i}{100\text{MeV}}\right)^{-0.8}\right]^2 + c_{1,i}^2}.$$

↓

Variable bandwidth smaller for:

- precise or high quality events (constants c_0 and c_1 related to the precision of the photon detection);
- events with high energy (E).

→ Reduces the amount of smoothing nearby the high-density regions.

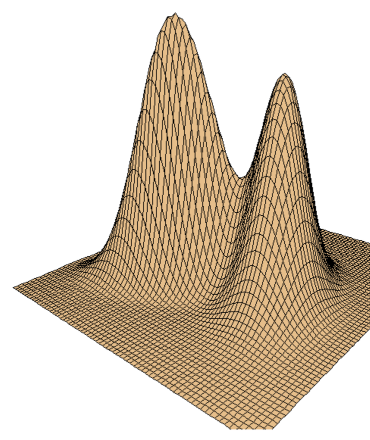

Figure 3. Kernel density estimation.

Mode hunting → Source hunting

- Sources are aimed to be identified by pursuing the task of mode detection.
- Modal regions are formed by sets of points along the steepest ascent path towards a mode.
- The mean-shift algorithm is adapted to be used with the directional kernel estimator, where at step $s+1$

$$\mathbf{x}^{(s+1)} = \mathbf{x}^{(s)} + M(\mathbf{x}^{(s)}),$$

with $M(\mathbf{x}^{(s)})$ the mean-shift [7].

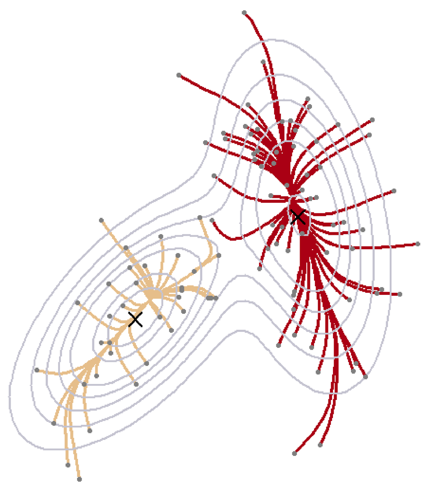

Figure 4. Gradient ascending path for each photon.

⇓

- Space partitioned into regions associated with sources → Clusters include background photons.
- Heavy presence of background photons → Many clusters of only background photons.

Post-processing

To separate the signal of the supposed emitting source from the γ-ray background we propose a post-processing procedure consisting of two parallel quests.

1. We supervise a tree-based classifier
   - based on a training sample drawn from the available LAT catalogue;
   - while integrating additional information on the photons as shown in Figure 5.

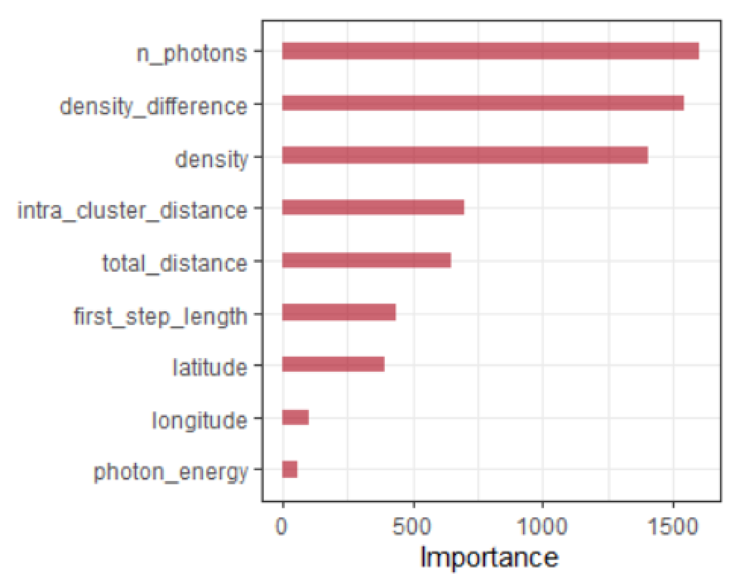

Figure 5. Feature importance plot for the tree-based photon classifier.

2. We evaluate the significance of each candidate mode by considering the second largest eigenvalue of the Hessian matrix of the kernel density evaluated at the mode adapting [5] in the directional setting.

⇒ We construct a confidence interval for each mode using bootstrap resampling: the mode is considered as such if the interval includes only negative values.

Results

Mode hunting

We studied the region of the sphere (in Figure 2) in which the LAT accumulated 3,849 photon counts over a five-year period of observation. The mean-shift algorithm identified 876 modes.

Signal-background separation

We evaluated the goodness of the results by reporting the true positive rate for single photon classification (TPR) and the false positive rate (FPR).

Quest	N. of candidate sources	TPR	FPR
Discriminant rule	39	98.5%	22.9%
Nonparametric statistical test	107	85.0%	11.2%

Final results

By super-imposing the findings of the two quests, we identified in all 27 candidate sources which are both, statistically significant and qualified as such by the non-parametric classifier.

Results	
TPR	94.6%
FPR	14.1%

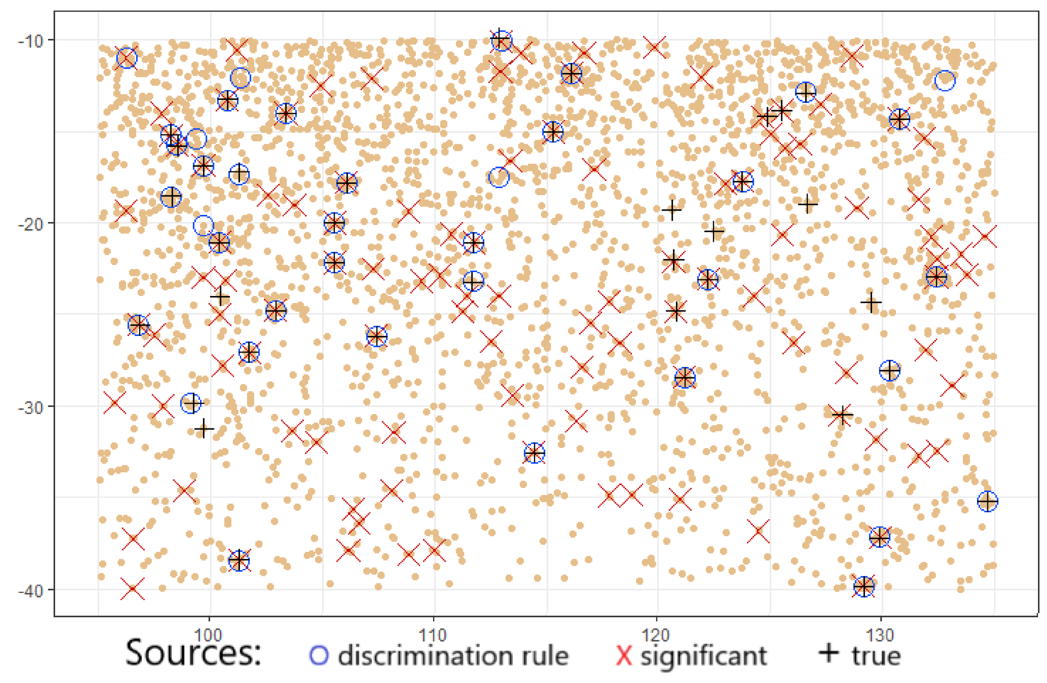

Figure 6. Fermi-LAT γ-ray photon count map (in Galactic coordinates) for the analysed 5-year observation period with superimposed the 44 true sources and the candidate sources.

Discussion and Future Work

- We obtained a partition of the sphere corresponding to the regions with high concentration of photons thus separating signal photons from diffuse γ-ray background photons and assessing the significance of each candidate source.
- Non-parametric statistical test found many significant background-modes.
- Improve the result by studying different and more elaborate classifiers.
- To analyse the whole sky map, consensus clustering may allow us to aggregate results from multiple runs on portions of the sphere whenever computational costs and limited memory won't allow us to do it in one go.

References

[1] Fermi - Gamma-ray Space Telescope. `https://fermi.gsfc.nasa.gov/`.

[2] M. Ackermann and M. Ajello. Determination of the point-spread function for the Fermi large area telescope from on-orbit data and limits on pair halos of active galactic nuclei. *Astrophysical J.*, **765**:1–19, 2013.

[3] J. E. Chacón. A population background for nonparametric density-based clustering. *Statistical Science*, **30**:518–532, 2015.

[4] D. Costantin, A. Sottosanti, A. R. Brazzale, D. Bastieri, and J. H. Fan. Bayesian mixture modelling of the high-energy photon counts collected by the Fermi Large Area Telescope. *Statistical Modelling*, **22**:175–198, 2022.

[5] C. R . Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Non-parametric inference for density modes. *JRSS*, **78**:99–126, 2016.

[6] A. Sottosanti, M. Bernardi, A. R. Brazzale, A. Geringer-Sameth, D. C. Stenning, R. Trotta, and D. A. van Dyk. Identification of high-energy astrophysical point sources via hierarchical Bayesian nonparametric clustering. 2021. URL: `https://doi.org/10.48550/arXiv.2104.11492`.

[7] Y. Zhang and Y.C. Chen. Kernel Smoothing, Mean Shift, and Their Learning Theory with Directional Data. *JMLR*, **22**:1–92, 2020.

Contact information

Anna Montin, MSc
University of Padova
anna.montin@studenti.unipd.it
Anna Montin