Efficient Disentangling γ-Ray Sources from Diffuse Background in the Sky Map

Francesco Freni[1] Giovanna Menardi[1]

[1]University of Padova

Framework and Motivation

Discovering and locating γ-ray sources in the whole sky map is a declared target of the Fermi LAT Collaboration [1].

The light emitted by celestial bodies is the primary carrier of astronomical information from the Universe to us. Hence, the identification of γ-ray sources allows us to gain a deeper insight on their origins and the physical mechanisms that power them.

The Large Area Telescope (LAT) onboard the Fermi spacecraft, designed to perform an all-sky survey, detects γ-rays in the galactic and extra-galactic space.

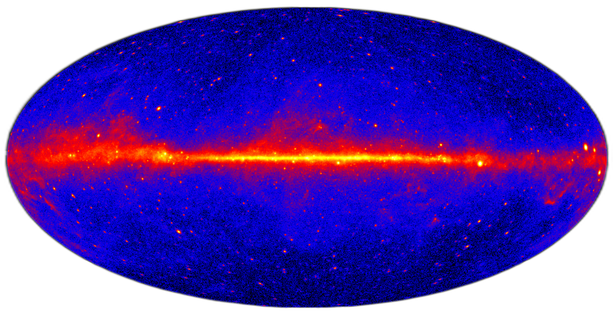

Figure 1. Image map of the γ-ray counts at energies larger than 1 GeV accumulated by the Fermi Large Area Telescope over five years of operation.

State of the art

The problem of detecting γ-ray sources has been widely issued in literature and it offers a wide field of application for different statistical techniques, from parametric to non-parametric approaches, both in the Frequentist and in the Bayesian paradigms [3, 5, 7].

Main goal and Statistical Framework

Since γ-ray sources are intended as peaks of energy arising from the background, their identification can be recast to a **clustering problem**, which is treated in a **non-parametric framework**, while tackling obstacles dictated by the specific nature of the data available, that pose both conceptual and computational limits to the study.

Data and Complications

Statistical units are **photons** detected by the LAT. The event is reconstructed using the tracker and the calorimeter. The former reconstructs the direction of each event, while the latter reconstructs the photon energy. Thus, together with the direction of the emission, additional information is provided, such as photon energy and quality.

- **The geometry of the data**: directions in three-dimensional space can be represented by Cartesian coordinates as vectors $\boldsymbol{x}$ of unit norm, i.e. points on the sphere (Figure 3)

$$\Omega_2 = \{\boldsymbol{x} \in \mathbb{R}^3 : \|\boldsymbol{x}\|_2 = x_1^2 + x_2^2 + x_3^2 = 1\}.$$

  → Modal clustering for directional data.

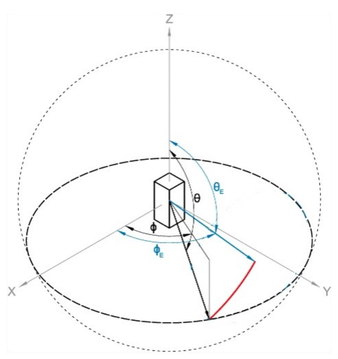

Figure 2. Representation of the coordinate system.

- **The clustering structure**: hundreds of sources, heterogeneous in size and variability, usually leptokurtic.
  → Non-parametric framework.
- **The background noise**: most of photon emission origins from a background noise. The diffuse γ-ray background spreads over the entire area observed by the telescope and is concentrated at the Galactic plane.
  → It is required to identify the sources, locate them, and distinguish them from the diffuse background.
- **Computational complexity**: problems are due to highly concentrated sources and huge moles of data.
  → Binned kernel density estimation and sphere partition into independent subregions.

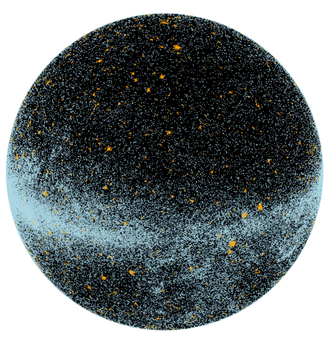

Figure 3. Source data from the 3FHL catalog of the Fermi LAT collaboration (orange), and the diffuse background (light blue).

Methods

Modal Clustering

- The data $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$ are sampled from a probability density function f.
- Clusters correspond to the domains of attraction of the modes of the underlying density [2].
- Two main strands can be identified, depending on whether the modes are given explicitely or not [4].
- Modes are implicitly found via the identification of the connected components of density level sets.
  → For $0 \leq \lambda \leq \max f$, define the level set $L(\lambda)$ as:

$$L(\lambda; f) = \left\{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) \geq \lambda\right\}.$$

  → Identify the maximum connected components of $L(\lambda)$.
  → When λ varies, the number of connected components of $L(\lambda)$ varies and a hierarchical tree structure, the cluster tree, is generated.
- For each mode there exists some λ for which one of the connected components of the associated $L(\lambda)$ includes that mode at most and identifies the excess mass of that mode [6].
- Clusters are not bounded to have a particular shape and the cluster tree naturally defines different levels of cluster resolution.

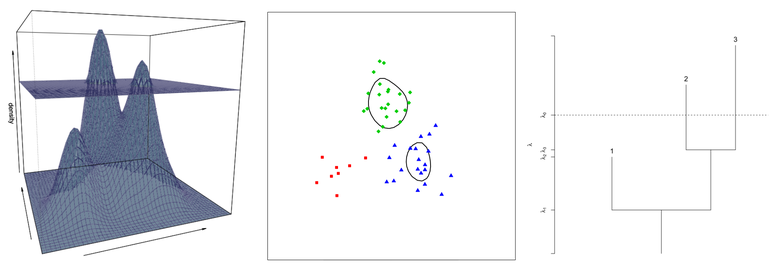

Figure 4. A section of a three-modal density function and the identified level set (middle panel) formed by two disconnected regions. In the right panel, the associated cluster tree.

Mesh creation

- Due to the huge mole of available data, streamlining is firstly pursued via data discretization.
- The sphere is partitioned into a thick triangular mesh.
  → An icosahedron is recursively subdivided and projected onto a spherical surface.

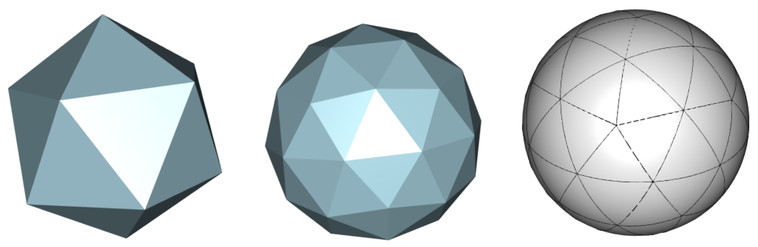

Figure 5. Representation of an icosahedron (left), the figure obtained by dividing each of its edges in two (center) and the projection onto the sphere (right).

- Working with fewer observations makes the density estimation phase and cluster identification less computationally burdensome.

Density estimation

- Each of the B bins of the mesh is associated with the count n_b of its inner photons.
- Density of photon emissions is estimated nonparametrically, via binned directional kernel methods:

$$\hat{f}(x) = \frac{1}{n}\sum_{b=1}^{B} n_b K_h(x - m_b)$$

  where $K_h(\cdot)$ is the von Mises-Fisher kernel with concentration parameter $1/h^2$, n is the sample size, and m_b is the centroid of the b^{th} bin. This already produces by itself a computational gain in efficiency.
- Photons emitted by sources are more concentrated along the direction of emission.
  → Variable bandwidth parameter: $h_i = h_0 \left[\frac{1}{g}\tilde{f}_{h_0}(\boldsymbol{x}_i)\right]^{-0.5}$, where $\tilde{f}_{h_0}$ is a pilot density obtained with a normal-scale parameter and g is the geometric mean of $\tilde{f}_{h_0}$.

Cluster identification

- Detection of the connected components of each level set is performed by finding the connected components of a suitable graph, whose nodes are the bins' centroids.
- Grid cells are grouped in place of sample observations.
- Outskirt photons are labelled as background, since a mode (a source) is present where a higher concentration of probability mass is identified.

Computational aspects

- The use of binned kernel density estimate must be supported by finding the right trade-off between computational gain and estimation bias.
- To avoid multiple sources falling into a single bin, a finer grid is used.
- Empty cells between regions occupied by photons are removed in order to split the sphere into independent portions.
  → Remarkable gain in efficiency.

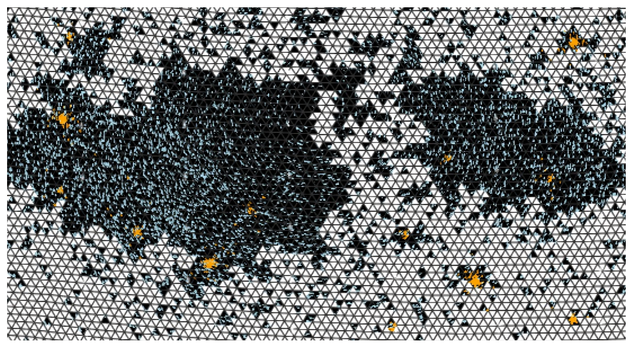

Figure 6. Portion of the sphere where empty cells are deleted in order to work on independent regions.

- Since photons are more concentrated at the Galactic plane, the chosen grid is finer at latitude 0°, while cells are bigger everywhere else.

Results

We applied the proposed procedure to a set of data drawn from the 3FHL catalog of the Fermi LAT Collaboration (Figure 3) and spread on the whole sky map, along with the diffuse background. The data set include 469784 photons, among which 73318 are emitted by 1529 sources, whose size range from 4 to 3572 photons. Since the data are drawn from a catalogue of already detected sources, we may evaluate the performance of the procedure with respect to the knowledge of the pertaining source of each photon emission.

- Results show a good performance in terms of source detection.
- Several spurious clusters are identified.
  → This is due to the sphere partition: many subregions only contain photons emitted from the background noise, which are classified as emitted from a source.
  → High False Positive Rate.
  → The number of individual misclassifications increases.

	background	sources
$\widehat{background}$	310430	7239
$\widehat{sources}$	86036	66079
MED	0.1985	
TPR	0.8947	
FPR	0.8515	

Discussion and Future Work

Promising results, yet much room for improvement.

- Choosing a finer grid leads to multiple small regions containing photons emitted from the background.
  → Several spurious clusters are identified.
  → The sphere should be split in a different way.
- Need to focus on bandwidth selection.
- Need to evaluate the significance of the identified clusters.

References

[1] Fermi - gamma-ray space telescope. `https://fermi.gsfc.nasa.gov/`.

[2] José E. Chacón. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4), nov 2015.

[3] Denise Costantin, Giovanna Menardi, Alessandra Brazzale, D. Bastieri, and J. Fan. A novel approach for pre-filtering event sources using the von mises–fisher distribution. *Astrophysics and Space Science*, 365, 03 2020.

[4] Giovanna Menardi. A Review on Modal Clustering. *International Statistical Review*, 84(3):413–433, December 2016.

[5] Anna Montin, Alessandra R. Brazzale, and Giovanna Menardi. Locating γ-ray sources on the celestial sphere via modal clustering, 2023.

[6] D. W. Muller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415):738–746, 1991.

[7] Andrea Sottosanti, Mauro Bernardi, Alessandra Brazzale, Alex Geringer-Sameth, David Stenning, Roberto Trotta, and David van Dyk. Identification of high-energy astrophysical point sources via hierarchical bayesian nonparametric clustering. 04 2021.

Contact information

Francesco Freni, BSc
University of Padova
francesco.freni@studenti.unipd.it

Giovanna Menardi, Associate Professor
University of Padova
menardi@stat.unipd.it