EFFICIENT DISENTANGLING γ-RAY SOURCES FROM DIFFUSE BACKGROUND IN THE SKY MAP

Francesco Freni [1] and Giovanna Menardi[1]

[1] Department of Statistical Science, University of Padova,
(e-mail: francesco.freni6@gmail.com, menardi@stat.unipd.it)

ABSTRACT: Searching for as yet undetected γ−ray sources is a major target of the Fermi LAT Collaboration. We address the problem by clustering the directions of the high-energy photon emissions detected by the telescope onboard the Fermi spacecraft. Putative sources are identified as the excess mass of disconnected high density regions on a sphere mesh, which allows for their joint discrimination from the diffuse γ−ray background spreading over the entire area. Density is estimated nonparametrically via binned directional kernel methods. The identification is accomplished by breaking the problem into independent subregions of the sphere separated by empty bins, thus leading to a remarkable gain in efficiency.

KEYWORDS: astrostatistics, directional data, modal clustering

1 Introduction

The Large Area Telescope (LAT) is an imaging γ−ray detector onboard the Fermi spacecraft, designed to perform an all-sky survey and gain a deeper comprehension of the processes responsible for generating and boosting γ−ray particles discharged by celestial bodies. Discovering and locating such sources is one of main purposes of the survey, and a declared target of the Fermi LAT collaboration. A main challenge, however, is the to separate the signal of the putative emitting sources from the diffuse γ−ray background which spreads over the entire area observed by the telescope. Furthermore, it is required to handle a remarkable computational burden, due to the huge amount of data recorded by the LAT.

Since γ−ray sources shall be intended as peaks of energy arising from a diffuse background, the underlying intuition complies with the *nonparametric*, or *modal* formulation of a clustering problem, which is here efficiently adapted to the considered framework. Modal clustering relies on the assumption that a probability density underlies the data, and clusters are defined as the domains of attraction of the density modes. With respect to most clustering methods,

relying on heuristic ideas of similarity between objects, the modal formulation is built on a probabilistic framework, which allows, for instance, a natural application of inferential tools. Additionally, the number of clusters is inherent to the data density and determined itself within the estimation process.

In this work we discuss a nonparametric method specifically conceived for high-energy $\gamma-$ray sources detection and discrimination from the background noise (Section 2). Its application is illustrated on a set of data drawn from one of the catalogues released by the Fermi LAT Collaboration (Section 3).

2 Nonparametric clustering for $\gamma-$ray sources detection

Nonparametric, or *modal* clustering hinges on the assumption that the data $(x_1,\dots,x_n)'$ are sampled from a probability density function f. The modes of f represent the archetypes of the clusters, in turn described by the surrounding regions. An indirect route to identify clusters, without attempting the explicit task of mode detection, is through disconnected (upper) density level sets of the sample space. Specifically, any section of f, at a level λ, singles out the set

$$L(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}, \quad 0 \leq \lambda \leq \max f$$

which may be connected or disconnected. In the latter case, it consists of a number of connected components, associated with a cluster at the level λ.

While there may not exist a single λ catching all the modal regions, any connected component of $L(\lambda)$ includes at least one mode of the density. On the other hand, 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 (Müller & Sawitzki, 1991). Hence, all the modal regions may be detected as the connected components of $L(\lambda)$ by varying λ (Figure 1). Points belonging to the surrounding regions are usually allocated to the clusters subsequently. See Menardi, 2016 for a detailed review.

Within the framework of $\gamma-$ray source detection, the data typically consist of an event list which gives the direction in the sky of each detected photon along with additional information. If the distance to the emitting source is not relevant, the data points are placed on the celestial sphere with Earth at its center and unit radius, as shown in the left panel of Figure 2. Directions are expressed in polar coordinates, that is, co-latitude (θ) and longitude (ϕ) in geographical terms, which can easily be back-transformed to Cartesian coordinates $x = (\cos\theta, \sin\theta\cos\phi, \sin\theta\sin\phi)$ on the unit sphere.

Due to the huge mole of available data, streamlining is firstly pursued via data discretization: rather than considering single photon emissions, the sphere

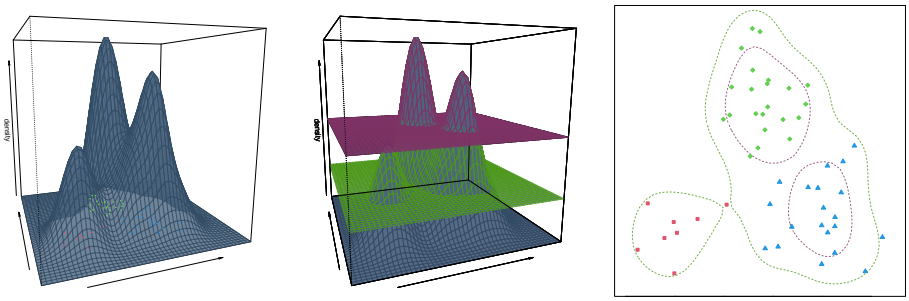


Figure 1. *A set of data from a trimodal density function (left), and the two sections of the density (center) for which the connected components of the associated $L(\lambda)$ identify the excess mass of the three modes (right).*

is partitioned into a thick triangle mesh, by recursively subdividing an icosahedron. Each of the B bins of the mesh is then associated with the count n_b of its inner photons. Density of photon emissions is then estimated nonparametrically, via binned directional kernel methods:

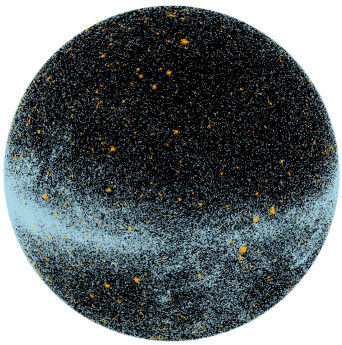
$$\hat{f}(x) = \frac{1}{n} \sum_{b=1}^{B} n_b K_h(x - m_b)$$

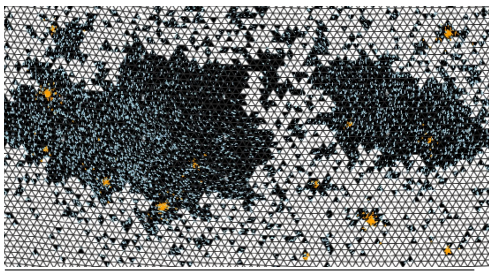
where $K_H(\cdot)$ is von Mises-Fisher kernel with concentration parameter $1/h^2$, n is the sample size, and m_b is the centroid of the b^{th} bin. This already produces by itself a computational gain of efficiency. Clustering is then built by identifying, for varying λ, the connected components of the upper level sets of the binned estimate, as union of edge-connected bins of the mesh. Once again, the mesh structure allows to accomplish the task efficiently on the whole sphere, by breaking the problem into independent subregions separated by empty bins.

The specific λ identifying the excess mass of each modal region allows for defining a source as the set of photons lying within the associated upper density level set. Outskirt photons are labeled as background.

3 Empirical analysis

We applied the proposed procedure to a set of data drawn from the 3FHL catalog of the Fermi LAT collaboration and spread on the whole sky map, along with the diffuse background. The sky distribution of the data, illustrated in Figure 2, is quite heterogeneous, with most of photon emission (around 84.4%) originated from a diffuse background noise, which mostly concentrates





Number of detected sources	
TPR	0.9013
FPR	0.2170

Figure 2. *Left: source data from the 3FHL catalog of the Fermi LAT collaboration (yellow), and the diffuse background (light blue). Right: a cut of the sphere highlighting the mesh built on the sphere and the opportunity of working on small separated regions separated by empty bins. The table reports the results of the proposed method.*

at the Galactic plane; in the same area, overlapping sources of various sizes arise while in the extragalactic sky sources are rather separated. 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. As a summarizing measure of the quality of the association, we compute the True Positive Rate (TPR) and the False Positive Rate (FPR). The former index is defined as the proportion of true sources correctly detected, while the latter one corresponds to the proportion of estimated components formed in fact by background photons.

Results, summarised in Figure 2, show an overall good performance with respect to both the detection of sources, and the discrimination between photons emitted by sources and background. Future research will focus on providing the detected sources with a significance measure, as well as reducing the spread of the detected sources, since not reported results show a non negligible quote of misclassified background photons, lying nearby the sources.

References

MENARDI, G. 2016. A review on modal clustering. *Int.Stat.Rev.*, **84**, 413–433.

MÜLLER, D.W., & SAWITZKI, G. 1991. Excess mass estimates and tests for multimodality. *J. Am. Stat. Ass.*, **86**, 738–746.