SIS | 2022

51st Scientific Meeting of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

SIS Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

800 ANNI 1222-2022 Università degli Studi di Padova | Dipartimento di Scienze Statistiche

Stata

sas

Università degli Studi del Sannio Benevento

Pearson

Matilde Bini (Chair of the Program Committee) - *Università Europea di Roma*
Rosanna Verde (Chair of the Local Organizing Committee) - *Università della Campania "Luigi Vanvitelli"*

PROGRAM COMMITTEE

Matilde Bini (Chair), Giovanna Boccuzzo, Antonio Canale, Maurizio Carpita, Carlo Cavicchia, Claudio Conversano, Fabio Crescenzi, Domenico De Stefano, Lara Fontanella, Ornella Giambalvo, Gabriella Grassia - Università degli Studi di Napoli Federico II, Tiziana Laureti, Caterina Liberati, Lucio Masserini, Cira Perna, Pier Francesco Perri, Elena Pirani, Gennaro Punzo, Emanuele Raffinetti, Matteo Ruggiero, Salvatore Strozza, Rosanna Verde, Donatella Vicari.

LOCAL ORGANIZING COMMITTEE

Rosanna Verde (Chair), Antonio Balzanella, Ida Camminatiello, Lelio Campanile, Stefania Capecchi, Andrea Diana, Michele Gallo, Giuseppe Giordano, Ferdinando Grillo, Mauro Iacono, Antonio Irpino, Rosaria Lombardo, Michele Mastroianni, Fabrizio Maturo, Fiammetta Marulli, Paolo Mazzocchi, Marco Menale, Giuseppe Pandolfi, Antonella Rocca, Elvira Romano, Biagio Simonetti.

ORGANIZERS OF SPECIALIZED, SOLICITED, AND GUEST SESSIONS

Arianna Agosto, Raffaele Argiento, Massimo Aria, Rossella Berni, Rosalia Castellano, Marta Catalano, Paola Cerchiello, Francesco Maria Chelli, Enrico Ciavolino, Pier Luigi Conti, Lisa Crosato, Marusca De Castris, Giovanni De Luca, Enrico Di Bella, Daniele Durante, Maria Rosaria Ferrante, Francesca Fortuna, Giuseppe Gabrielli, Stefania Galimberti, Francesca Giambona, Francesca Greselin, Elena Grimaccia, Raffaele Guetto, Rosalba Ignaccolo, Giovanna Jona Lasinio, Eugenio Lippiello, Rosaria Lombardo, Marica Manisera, Daniela Marella, Michelangelo Misuraca, Alessia Naccarato, Alessio Pollice, Giancarlo Ragozini, Giuseppe Luca Romagnoli, Alessandra Righi, Cecilia Tomassini, Arjuna Tuzzi, Simone Vantini, Agnese Vitali, Giorgia Zaccaria.

ADDITIONAL COLLABORATORS TO THE REVIEWING ACTIVITIES

Ilaria Lucrezia Amerise, Ilaria Benedetti, Andrea Bucci, Annalisa Busetta, Francesca Condino, Anthony Cossari, Paolo Carmelo Cozzucoli, Simone Di Zio, Paolo Giudici, Antonio Irpino, Fabrizio Maturo, Elvira Romano, Annalina Sarra, Alessandro Spelta, Manuela Stranges, Pasquale Valentini, Giorgia Zaccaria.

Copyright © 2022

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

Contents

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

Individuazione di sorgenti di raggi γ sulla sfera mediante clustering non parametrico

Anna Montin, Alessandra R. Brazzale, Giovanna Menardi

Abstract Searching for as yet undetected γ-ray sources is a major target of the Fermi LAT Collaboration. In this paper, we present an algorithm capable to identify such type of sources by non-parametrically clustering the directions of arrival of the high-energy photons detected by the telescope. Using statistical tools from hypothesis testing and classification, we furthermore present an automatic way to skim off sound candidate sources from the γ-ray emitting diffuse background and to quantify their significance. The algorithm was calibrated on simulated data provided by the Fermi LAT collaboration and will be illustrated on a real Fermi LAT case-study.

Abstract *L'individuazione di sorgenti di raggi gamma è uno degli obiettivi dichiarati della Collaborazione Fermi LAT. Presentiamo qui un algoritmo per l'individuazione di queste sorgenti basato sul clustering non parametrico delle direzioni di arrivo dei fotoni ad alta energia rilevate dal telescopio. Sfruttando risultati della teoria dei test statistici e dell'apprendimento supervisionato, presentiamo, inoltre, come scremare le sorgenti candidate dalla componente diffusa della radiazione di fondo e attribuire loro una misura della significatività. L'algoritmo è stato calibrato su dati simulati forniti dalla Collaborazione Fermi LAT e sarà illustrato su un caso di studio reale.*

Key words: directional data, kernel density estimator, man-shift algorithm, tree-based classification

Anna Montin
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: anna.montin@studenti.unipd.it

Alessandra R. Brazzale
Dipartimento di Scienze Statistiche, Università degli Studi di Padova, e-mail: alessandra.brazzale@unipd.it

Giovanna Menardi
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: giovanna.menardi@unipd.it

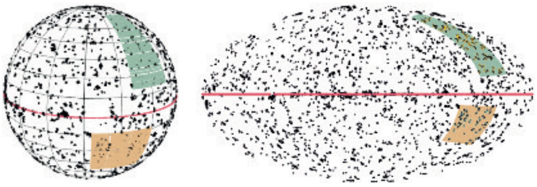


Fig. 1 Fermi-LAT γ-ray photon count maps for a 5-year observation period. Left: in polar coordinates. Right: in Galactic coordinates. Yellow: analysed region. Green: training set for post-processing classifier. Red: Galactic plane.

1 Motivation and rationale

In γ-ray astronomy, the data typically consist of an event list which gives the direction in the sky of each detected photon together 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 1. Directions are often expressed in *galactic coordinates*, which place the origin of the Cartesian system in the center of our galaxy — the Milky Way — and align the x-axis with the galactic plane (right panel of Figure 1). To overcome mismatches due to projecting data on the 2-dimensional sky map, we rather express directions through *polar coordinates*, that is, co-latitude (θ) and longitude (ϕ) in geographical terms, which can easily be back-transformed to Cartesian coordinates $\mathbf{x} = (\cos\theta, \sin\theta\cos\phi, \sin\theta\sin\phi)^\top$ on the unit sphere.

Discovering and locating high-energy emitting sources in the whole sky map is a declared target of the Fermi Gamma-ray Space Telescope collaboration. An astronomical source is an object in outer space which, in our case, emits γ-ray photons, that is, quanta of light in the highest energy range. Traditionally, analyses are based on so-called *single-source models* [4, § 7.4], which require the whole sky map to be split into small regions. The presence of a possible new source is assessed on a pixel-by-pixel basis using Poisson regression and likelihood ratio testing. Conversely, *variable-source-number models* address the problem from a more global perspective, as they simultaneously model and locate all sources in a sky map [4, § 7.3]. A most recent example of application to the γ-ray count maps accumulated by the LAT — the principal scientific instrument on board the Fermi spacecraft — is [7].

In this paper, we present a flexible algorithm for the efficient identification of γ-ray sources. In particular, we address the problem from the global perspective of variable-source-number models while working on the sphere. From the modeling point of view, the sources will be represented by highly concentrated clusters. [2] provide an illustration of directional *model-based* clustering of Fermi LAT data using a finite mixture of von Mises-Fisher distributions. Our approach uses *modal clustering*, which combines the advantages of both, model-based clustering and non-parametric methods, to guarantee the required flexibility. The corresponding methodological background is reviewed in Section 2. We will illustrate our proposal through a case-study of Fermi LAT data in Section 3.

2 Modal clustering on the sphere

General framework. Allocating objects to an unknown number of groups according to a set of observed attributes or features is a natural activity of any science. A surge of techniques has been proposed over the years, which differ significantly in their definition of what a "group" is. The non-parametric formulation, referred to as *modal* clustering, associates clusters with the domain of attraction of the modes of the underlying density, which are usually estimated non-parametrically [6]. Modal clustering can be recast into the frame of a standard statistical problem by considering the observed data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ as a sample of n realisations of a d-dimensional random vector $\mathbf{X}$ from a probability density function $f : \mathscr{X} \subseteq \mathbb{R}^d \to \mathbb{R}^+$. The modes of $f(\cdot)$ represent the archetypes of the clusters, which are in turn described by the surrounding regions.

Mode hunting. In our setting, directions in $\mathbb{R}^3$ are represented as unit vectors $\mathbf{x}$, that is, as points 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. Density estimation is performed by a suitable extension of a kernel density estimator to the directional setting [1]

$$\hat{f}_n(\mathbf{x}) = \frac{c_h(K)}{n} \sum_{i=1}^{n} K\left(\frac{1 - \mathbf{x}^\top \mathbf{x}_i}{h^2}\right), \tag{1}$$

where $K(\cdot)$ is a kernel function, $c_h(K)$ the associated normalizing constant, and the bandwidth parameter $h > 0$ controls the smoothness of the estimator.

To account for the rugged nature of the data, which exhibit highly heterogeneous levels of concentration over the sphere, the estimator (1) is extended to account for a variable bandwidth $h = h(\mathbf{x}_i)$, selected according to scientific input. In particular, we use the scale parameter of the *point spread function*, which describes the response of the LAT to the point source. Grossly, this amounts to associating smaller values of h with precise events characterized by higher energies and which are usually disclosed around the direction of photon emission. The choice is consistent with the requirement of reducing the amount of smoothing nearby the high-density regions, as it is usually acknowledged by variable kernel density estimators.

A popular choice for the directional kernel is linked to the von Mises-Fisher (vMF) distribution [5]

$$f_{\mathrm{vMF}}(\mathbf{x}; \mu, \kappa) = c_3(\kappa) e^{\kappa \mathbf{x}^\top \mu},$$

where c_3 is a normalizing constant, $\mu \in \Omega^2$ is the mean direction, and κ is a concentration parameter around the mean. Here, κ is set to vary inversely with the bandwidth, i.e. $K(\cdot) = f_{\mathrm{vMF}}(\cdot; \mathbf{x}_i, 1/h)$. This distribution describes observations which scatter symmetrically around their mean value and can be regarded as the generalization of the normal distribution to spherical data.

Sources are aimed to be identified by pursuing the explicit task of mode detection, and modal regions are formed by sets of points along the steepest ascent path towards a mode. This is achieved by adapting the *mean-shift* algorithm [6, § 2.2] to be used with the directional kernel estimator (1). Starting from a generic point

$\mathbf{x}^{(0)}$, the algorithm recursively shifts it to a local weighted mean, until convergence. Denoted by $w_i(\mathbf{x}^{(s)})$ the vector of weights of the components of $\mathbf{x}_i$ at step s, at the next step

$$\mathbf{x}^{(s+1)} = \sum_{i=1}^{n} w_i(\mathbf{x}^{(s)})\mathbf{x}_i = \mathbf{x}^{(s)} + M(\mathbf{x}^{(s)}),$$

where $M(\mathbf{x}^{(s)}) = \sum_{i=1}^{n} w_i(\mathbf{x}^{(s)})\mathbf{x}_i - \mathbf{x}^{(s)}$ denotes the mean shift. Up to a normalising factor, the weights $w_i(\mathbf{x})$ are specified as $\nabla K\left(h^{-2}(1-\mathbf{x}^\top\mathbf{x}_i)\right)$, where $\nabla K(\cdot)$ is the gradient of the kernel function.

Post-processing. To separate the signal of the supposed emitting source from the diffuse γ-ray background, which spreads over the entire area observed by the telescope, we propose a post-processing procedure that combines the findings of two parallel quests.

On one hand, we supervise a suitable classifier, based on a training sample drawn from the available LAT catalogue, for which information on the earlier detected sources is available. The classifier integrates additional information on the photons such as their energy, position, the number present in the same cluster, the density estimates for the signal and the background model and various types of distances to the detected mode. This allows us to skim off the photons emitted from high-energy emitting sources from those which originate from the diffuse background.

In parallel, we evaluate the significance of each candidate mode. Here, we consider an adaptation of [3] who verify whether the maximum eigenvalue of the Hessian matrix of the kernel density, evaluated at the mode, is negative. In the directional setting, due to the constraints induced by working on the surface of the sphere, one eigenvalue is necessarily zero. An $1-\alpha$ level confidence interval for the second largest eigenvalue is hence constructed using bootstrap resampling. The mode is considered as such if the interval includes only negative values. The same allows us to infer the significance of the candidate source.

By super-imposing the findings from the two quests, we can identify candidate sources which are both, statistically significant and qualified as such by the nonparametric classifier. A by-product of our post-processing algorithm is the assignment of the photons to their assumed emitting source.

3 A Fermi LAT case study

Mode hunting. The two yellow regions in Figure 1 show a portion of the Southern sky of size $(l,b) \in [95°, 135°] \times [-40°, -10°]$ for which the LAT accumulated 3,849 photon counts over a five-year period of observation.[1] Of these, about 26% were emitted by the 44 sources present in the area, while the remaining 74% originated from the diffuse γ-ray background. The left panel of Figure 2 plots the estimated kernel density (1) using a von Mises-Fisher kernel. Here, the bandwidth parame-

[1] https://fermi.gsfc.nasa.gov/ssc/data/access/

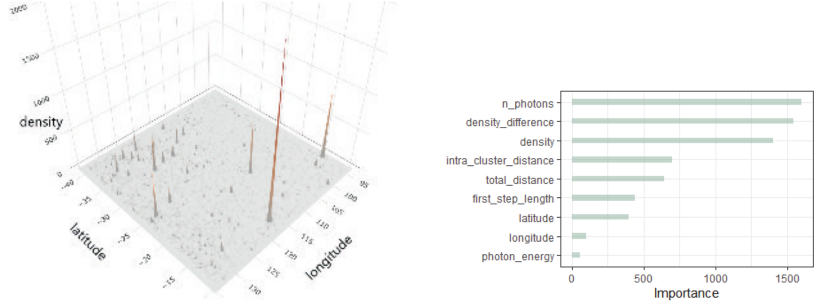


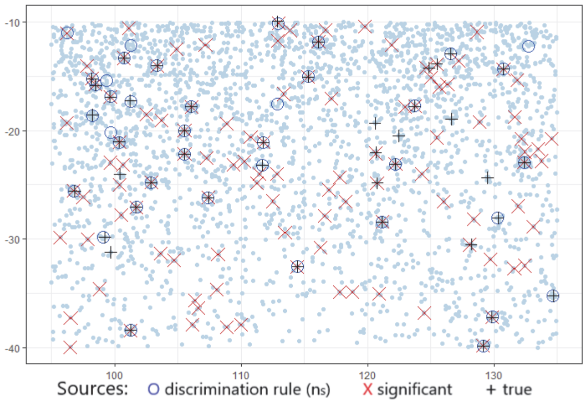
Fig. 2 Left: Kernel density estimate using a von Mises-Fisher kernel of the γ-ray photon counts accumulated by the LAT in a 5-year period. Right: Feature importance plot for the tree-based photon classifier.

ter h was set according to scientific input, as described in Section 2. This choice revealed to be the most performing one in terms of adjusted Rand index (ARI), average distance between the true source direction and the reconstructed one ($\bar{d}(s,\hat{s})$) and number of identified sources (n_s), according to an extensive numerical investigation (results not shown here) we carried out. In all, the mean-shift algorithm identified 876 modes.

Post-processing. To further refine the list of candidate sources we proceeded in two steps as outlined in Section 2.

1. A tree-based classifier to discriminate between source and background photons was trained on the 6,814 photon counts highlighted in green in the right panel of Figure 1 using as predictor variables those listed in the right panel of Figure 2. The most discriminating features are the number of photons assigned to a cluster, the difference between the two photon densities for, respectively, the all sky and background counts only, and the density observed for each photon. This reduces the original 876 modes to 39 candidate sources, which are shown as blue circles in the left panel of Figure 3. The table on the right reports the performance of our classifier in terms of ARI and average distance $\bar{d}(s,\hat{s})$. The true positive rate for single photon classification is 98.5% rate, while the percentage of false positives is 22.9%. Indeed, the five missed sources are the less photon emitting ones.
2. In parallel, we tested all the 555 clusters which contain two or more photons at a significance level of 5% while applying Bonferroni's correction. This skimmed off 448 modes, for a total of 107 remaining candidate sources, shown in the left panel of Figure 3 as red crosses. Here, the true positive rate for single photon classification is 85.0% and the false positive rate is 11.2%.

By super-imposing these two findings, we obtain in all 27 sources which are both, statistically significant and qualified as such by the non-parametric classifier. The global true positive rate for single photon classification is 94.6% while the false positive rate is 14.1%.



Results	
ARI	0.961
$\bar{d}(s,\hat{s})$	0.001
n_s	39
True sources	44

Fig. 3 Left: Fermi-LAT γ-ray photon count map (in Galactic coordinates) for the analysed 5-year observation period with superimposed the true and candidate sources. Right: Performance measures of the tree-based classifier.

To analyse the whole sky map, *consensus clustering* [8] 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. Bai Z.D., Rao C.R., Zhao L.C.: Kernel estimators of density function of directional data. Journal of Multivariate Analysis **27**, 24–39 (1988)
2. Costantin D., Menardi G., Brazzale A.R., Bastieri D., Fan J.H.: A novel approach for pre-filtering event sources using the von Mises-Fisher distribution. Astrophysics and Space Science **365** (2020)
3. Genovese C.R., Perone-Pacifico M., Verdinelli I., Wasserman L.: Non-parametric inference for density modes. Journal of the Royal Statistical Society Series B **78**, 99–126 (2016)
4. Hobson M.P., Jaffe A.H., Liddle A.R., Mukherjee P., Parkinson D.: Bayesian Methods in Cosmology. Cambridge University Press (2009).
5. Mardia K.V., Jupp P.E.: Directional Statistics. John Wiley & Sons (2000)
6. Menardi G.: A review on modal clustering. International Statistical Review **84**, 413–433 (2006)
7. Sottosanti A., Bernardi M., Brazzale A.R., Geringer-Sameth A., Stenning D.C., Trotta R., van Dyk D.A: Identification of high-energy astrophysical point sources via hierarchical Bayesian nonparametric clustering. arXiv:2104.11492 (2021)
8. Vega-Pons S., Ruiz-Shulcloper J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence **25**, 337–372 (2011)