Statistical Methods & Applications (2024) 33:153–172
https://doi.org/10.1007/s10260-023-00726-w

ORIGINAL PAPER

Locating γ-ray sources on the celestial sphere via modal clustering

Anna Montin[1] · **Alessandra R. Brazzale**[1] · **Giovanna Menardi**[1] · **Andrea Sottosanti**[2]

Accepted: 31 August 2023 / Published online: 5 October 2023

Abstract

Sky surveys represent the fundamental data basis for detecting and locating as yet undiscovered celestial objects. Since 2008, the Fermi LAT Collaboration has catalogued thousands of γ-ray sources with the aim of extending our knowledge of the highly energetic physical mechanisms and processes that lie at the core of our Universe. In this article, we present a nonparametric clustering algorithm which identifies high-energy astronomical sources using the spatial information of the γ-ray photons detected by the large area telescope onboard the Fermi spacecraft. In particular, the sources are identified using a von Mises–Fisher kernel estimate of the photon count density on the unit sphere via an adjustment of the mean-shift algorithm which accounts for the directional nature of the collected data and the need of local smoothing. This choice entails a number of desirable benefits. It allows us to bypass the difficulties inherent on the borders of any projection of the photon directions onto a 2-dimensional plane, while guaranteeing high flexibility. The smoothing parameter is chosen adaptively, by combining scientific input with optimal selection guidelines, as known from the literature. 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. We calibrate and test our algorithm on simulated count maps provided by the Fermi LAT Collaboration.

Keywords Directional data · Kernel density estimator · Man-shift algorithm · Tree-based classification

Extended author information available on the last page of the article

1 Motivation and rationale

1.1 High-energy astrophysics

The past 3 decades have been a golden era for Astronomy. Pioneering technology has driven remarkable acceleration in the rate of detection and characterization of celestial objects, and new space missions will have more and better quality data to help find and characterize these objects. Discoveries in this field are of utmost relevance as they contain a wealth of information about the history of the Universe, and impact on the understanding of our Galaxy and our own Solar system. An important example is high-energy astrophysics, which acts at the interface between particle physics and astronomy to study the multitude of extreme phenomena which inhabit the Cosmos. To date, the observation of γ-ray photons, that is, of quanta of light in the highest energy range, has provided the basis for a large number of astronomical discoveries. γ-rays are usually generated from accelerated charged particles, such as electrons or protons, boosted by extreme celestial objects such as supermassive black holes, supernova remnants, pulsars and active galactic nuclei, to name a few. The study of these γ-ray emitting sources improves our understanding of high-energy astrophysical phenomena, and might even resolve the mystery of the fundamental nature of dark matter.

The Fermi Gamma-ray Space Telescope[1] is an international and multi-agency space mission launched in June 2008 which studies the Cosmos in the energy range 10 keV–300 GeV. The primary instrument onboard the Fermi spacecraft is the Large Area Telescope (LAT), a wide field-of-view pair-conversion telescope which was designed to perform an all-sky survey aimed at discovering and locating high-energy emitting sources. The data provided by the Fermi LAT Collaboration typically consist of an event list which gives the direction in the sky of each detected photon together with additional information, the primary one being its energy content and the so-called event type which expresses the quality of the measurement. This information is used to determine the number of the emitting extra-galactic sources, measure their intensities, and assign to them the corresponding individual photon counts.

A major challenge of trying and detecting high-energy phenomena from astronomical data is to separate the signal of the emitting sources from noise. The Fermi LAT data, in particular, are characterized by two types of noise: (1) measurement error associated with the components of the LAT (tracker, calorimeter etc.) and (2) the diffuse γ-ray background which spreads over the entire area observed by the telescope. The former is expressed through the LAT's *point spread function* (Ackermann et al. 2013), which is typically included into the model. Different phenomena contribute to the residualγ-ray background (Acero et al. 2016). Broadly speaking, its origins can be brought under two headings:

[1] https://fermi.gsfc.nasa.gov/.

galactic interstellar emission (GIE), that is, the interaction of galactic cosmic rays with gas and radiation fields, and a residual all-sky emission. The latter is commonly called the isotropic diffuse gamma-ray background (IGRB), and includes theγ-ray emission from faint unresolved sources and any residual galactic emission which is approximately isotropic.

In addition, astronomical data typically come in the form of *big data*, whose volumes have increased over the past years from gigabytes into terabytes and petabytes. The Fermi LAT database currently holds over 1 billion photons in the energy range from about 20 MeV to more than 300 GeV collected in over a decade of operation. Efficient tools to account for the computational burden required to analyse these huge amounts of data, possibly on the entire sphere, are in high demand.

1.2 Statistical models for astronomical source detection

The discovery of celestial objects is an intrinsically interdisciplinary field which combines both, statistical and astrophysical methodology. The standard procedure of the Fermi LAT Collaboration for point-like source detection relies on so-called *single-source* models (Hobson et al. 2009, par. 7.4), which require the sky map to be split into small regions. The presence of a possible new source is assessed on a pixel-by-pixel basis: Poisson regression is used to model the number of photons associated with each pixel and likelihood ratio tests assess the significance of the source (Mattox et al. 1996). See also van Dyk et al. (2001) for a Bayesian treatment with application to low-count X-ray data collected by the Chandra X-Ray Observatory.

Conversely, *variable-source-number*models address the problem from a more global perspective, as they simultaneously identify and locate all possible sources in a given sky map (Hobson et al. 2009, par. 7.3). Since point-like sources present themselves as spatially concentrated photon emissions, the problem can naturally be recast as a clustering problem. A recent example within this latter approach is Costantin et al. (2020) who reconstruct the spatial distribution of theγ-ray counts by means of a finite mixture of parametric component densities. The model is estimated via maximum likelihood and associates each component of the finite mixture with a putative high-energy emitting source. In the same guise, yet according to a Bayesian perspective, Jones et al. (2015), Costantin et al. (2020) and Meyer et al. (2021) model X-ray or γ-ray photon counts using finite mixtures which may include auxiliary information to properly account for the background contamination. Costantin et al. (2020), for instance, translate the simulation-based background model developed by Acero et al. (2016) into a workable parametric formulation. However efficient, these proposals rely on a simplified specification of the background distribution. Two*infinite*mixture models are used in Sottosanti et al. (2021), the first to pinpoint the sources and the second to reconstruct the background via a smooth Bayesian nonparametric model based on B-splines. This provides a substantially flexible, yet highly computationally demanding solution.

In this paper, we advocate the use of nonparametric, or modal, clustering for γ-ray source detection. This less widespread formulation, which, for instance, is reviewed in Menardi (2016), associates clusters with the domains of attraction of the modes of the density underlying the data, which are usually reconstructed by suitable nonparametric estimators. The considered framework presents various elements of novelty. First of all, unlike Jones et al. (2015), Costantin et al. (2020), Sottosanti et al. (2021), and Meyer et al. (2021), who use the longitude and latitude of the data points as if they were *rectangular* coordinates located in a 2-dimensional rectangular map, we directly work with the spherical coordinates of the 3-dimensional directions placed in a Cartesian system. As in Costantin et al. (2020), this allows us to bypass the difficulties inherent on the borders of any 2-dimensional projection of the photon directions. Targeting the identification of γ-ray sources by means of modal clustering is itself a further novelty worth to be explored in addition to the use of flexible nonparametric methods, which Sottosanti et al. (2021) started exploring yet to only model the background. Indeed, the accounted notion of cluster as modal region complies with the spatial distribution of photons from a source, which may spread into an arbitrary shape.

The use of modal clustering for directional data has been recently reviewed by Saavedra-Nieves and Fernández-Perez (2023) according to a level set-based approach. We rather follow a*mode bumping* approach, which allows us to explicitly identify the direction of the point-like sources. This is performed via an adjustment of the mean-shift algorithm to account for the directional nature of the Fermi LAT data and the need of local smoothing, required by the high heterogeneity of the sources with respect to size and spread. The issue of selecting the smoothing parameter is thereby addressed adaptively, by linking optimal selection guidelines, as known from the literature, to the use of auxiliary information from scientific input. We furthermore present an automatic way to pinpoint sound candidate sources and to quantify their significance by skimming off the γ-ray emitting diffuse background using suitably adjusted results known from bootstrap-based hypothesis testing and classification built on previous knowledge provided by the Fermi LAT Collaboration. A direct consequence of this approach is that we no longer need to specify the background component, as it was the case in Jones et al. (2015), Costantin et al. (2020), Meyer et al. (2021) and Sottosanti et al. (2021).

The paper is organized as follows. Section 2 sets the methodological background of kernel density estimation for directional data. Bandwidth selection is discussed in Sect. 3. Section 4 presents our proposal of modal detection on the unit sphere and, in particular, discusses how to evaluate the significance of a candidate source and how to skim off the photons emitted by the γ-ray background. Section 5 illustrates the performance of our proposal when applied to simulated data of high-energy photon emission provided by the Fermi LAT Collaboration. The paper closes with the concluding remarks of Sect. 6.

This paper is an extended and improved version of the paper presented at the 51st Scientific Meeting of the Italian Statistical Society on June, 2022 (Montin et al. 2022).

2 Kernel density estimation for astrophysical count maps

2.1 Directional data in astronomy

The position in the sky of a celestial object can be specified in different ways, depending on the coordinate system which is used. If the distance to the object is not known, we place its *direction* onto a 3-dimensional celestial sphere with a suitably chosen center and unit radius. This is shown in the left panel of Fig. 1, which traces the incoming γ-ray photons collected by the LAT over a 5-year period. Directions are expressed through *spherical* coordinates, which emulate how the geographic coordinate system works on Earth's surface. That is, the longitude $l \in (-180, +180)$ and latitude $b \in (-90, +90)$ of a direction correspond to special angles which are usually expressed in degrees. The different coordinate systems differ in the choice of the physical reference at the center of the sphere and of the plane used to divide it into two equal hemispheres. The *Galactic* coordinate system places the Sun at its center and aligns the fundamental plane with the Galactic plane, that is, with the plane on which the majority of the mass of our Galaxy, the Milky Way, lies.

Sometimes, it is useful to project the data points onto a 2-dimensional plane using an equal-area projection as when a whole sky view is aimed at. Different map projections exist, depending on how they distort the surface of the sphere. For instance, in the Mollweide projection shown in the right panel of Fig. 1, meridians form ellipses.

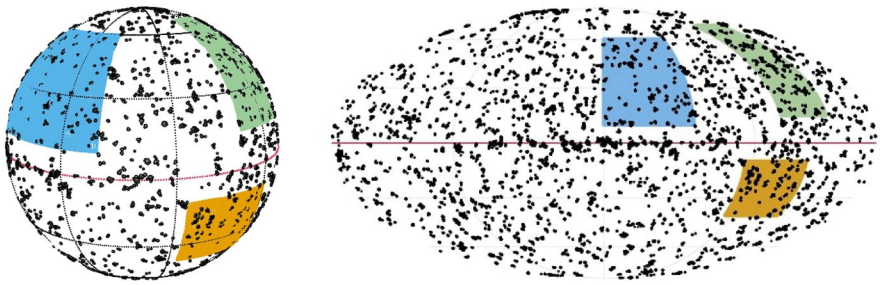


Fig. 1 Simulated Fermi-LAT γ-ray photon count map for a 5-year observation period. Left: Directions of arrival projected onto the 3-dimensional unit sphere. Right: Mollweide projection of the directions of arrival onto the 2-dimensional plane. Yellow: region of size $(l, b) \in [95°, 135°] \times [-40°, -10°]$ analyzed in Sect. 5.3. Blue: region of size $(l, b) \in [0°, 60°] \times [10°, 60°]$ used to select the optimal bandwidth h. Green: region of size $(l, b) \in [120°, 150°] \times [15°, 70°]$ used to train the post-processing classifier. Red: Galactic plane. Longituide is expressed as $[-180°, 180°]$ (color figure online)

2.2 The von Mises–Fisher distribution

Directions in a d-dimensional space, with $d \geq 2$, can be represented using Cartesian coordinates as unit vectors $\mathbf{x}$, that is, as points on the sphere

$$\Omega_{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = x_1^2 + \cdots + x_d^2 = 1\}$$

with unit radius and centre at the origin. These can be retrieved from Galactic coordinates, that is, from the longitude l and the latitude b of a given data point, by

$$\mathbf{x} = [\cos l \cos b,\ \sin l \cos b,\ \sin b]^\top.$$

See Section 9.1 of Mardia and Jupp (2000). A popular choice to model such type of data is the von Mises–Fisher (vMF) distribution

$$f_{vMF}(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = C_{d-1}(\kappa) \exp\{\kappa \boldsymbol{x}^\top \boldsymbol{\mu}\}, \tag{1}$$

where $\boldsymbol{\mu} \in \Omega_{d-1}$ represents the mean direction, while $\kappa \geq 0$is a concentration parameter (Mardia and Jupp (2000), Section 9.3.2). As such, the von Mises–Fisher distribution describes observations which scatter simmetrically around their mean direction $\boldsymbol{\mu}$. The normalizing constant

$$C_{d-1}(\kappa) = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} \mathcal{I}_{\frac{d}{2}-1}(\kappa)}$$

includes the modified Bessel function of the first kind of order ν

$$\mathcal{I}_\nu(z) = \frac{\left(\frac{z}{2}\right)^\nu}{\pi^{1/2}\Gamma(\nu + \frac{1}{2})} \int_{-1}^{1} (1 - t^2)^{\nu - \frac{1}{2}} e^{zt} dt.$$

The von Mises–Fisher distribution for the random d-dimensional unit vector $\boldsymbol{x}$ can be obtained from the d-dimensional normal distribution $N_d(\boldsymbol{\mu}, \kappa^{-1}\mathbf{I}_d)$, with $\mathbf{I}_d$ being the $d \times d$ diagonal unit matrix and $\boldsymbol{\mu}$ a vector of length r, by conditioning on $\|\boldsymbol{x}\|_2 = r$ (Mardia and Jupp 2000, p. 173). If$d = 2$, the von Mises–Fisher distribution reduces to the von Mises distribution on the unit circle, while if $d = 3$

$$C_2(\kappa) = \frac{\kappa^{\frac{1}{2}}}{(2\pi)^{\frac{3}{2}} \mathcal{I}_{\frac{1}{2}}(\kappa)}.$$

Furthermore, if $\kappa = 0$, the density (1) reduces to the uniform density on Ω_{d-1}, while if $\kappa \to \infty$, it tends to a point mass at $\boldsymbol{\mu}$.

2.3 Kernel density estimator

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \Omega_2$ be a random sample of n observations generated by a distribution with density $f(\mathbf{x})$ defined on the unit sphere Ω_2 such that

$$\int_{\Omega_2} f(\mathbf{x})\omega_2(d\mathbf{x}) = 1,$$

where ω_2 is the Lebesgue measure on Ω_2. We can estimate the density $f$using the kernel density estimator proposed by Bai et al. (1988) for directional data,

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

where $K(\cdot)$ is a suitable kernel function which decreases on $[0, \infty)$, and $h > 0$ is the smoothing parameter. The normalizing constant $c_h(K)$, is defined by

$$c_h(K)^{-1} = \int_{\Omega_2} K\left(\frac{1 - \boldsymbol{x}^\top \boldsymbol{x}_i}{h^2}\right)\omega_2(d\boldsymbol{x}) = h^2 \tilde{c}_h(K),$$

where $\tilde{c}_h(K) = \int_0^{2/h^2} K(u)du$. Using the von Mises–Fisher kernel, expression (2) becomes

$$\begin{aligned} \hat{f}_h(\boldsymbol{x}) &= \frac{1}{n} \sum_{i=1}^{n} f_{vMF}\left(\boldsymbol{x}; \boldsymbol{x}_i, \frac{1}{h^2}\right) \\ &= \frac{1}{(2\pi)^{\frac{3}{2}} \mathcal{I}_{\frac{1}{2}}(h^{-2})} \frac{1}{hn} \sum_{i=1}^{n} \exp\left(\frac{\boldsymbol{x}^\top \boldsymbol{x}_i}{h^2}\right). \end{aligned} \tag{3}$$

That is, the kernel density estimator for direction data on the unit sphere is a mixture of 3-dimensional von Mises–Fisher distributions with $\kappa = h^{-2}$.

3 Bandwidth selection

3.1 Automatic selection based on error minimization

A major issue when using a kernel density estimator is the selection of the bandwidth, h. Being able to specify a suitable amount of smoothing is crucial for the reliable identification of the sources. If the kernel function is too concentrated, false peaks may emerge from the background. Conversely, we may miss some faint sources when a too large bandwidth is selected. This is illustrated in Fig. 2, which plots the estimated density for a same sky region using two different values of h.

Also for the directional setting, selecting an optimal bandwidth generally entails minimization of a suitable measure of error. A first option is the likelihood *Cross-Validation* selector h_{LCV} which minimizes the Kullback–Leibler loss by maximising

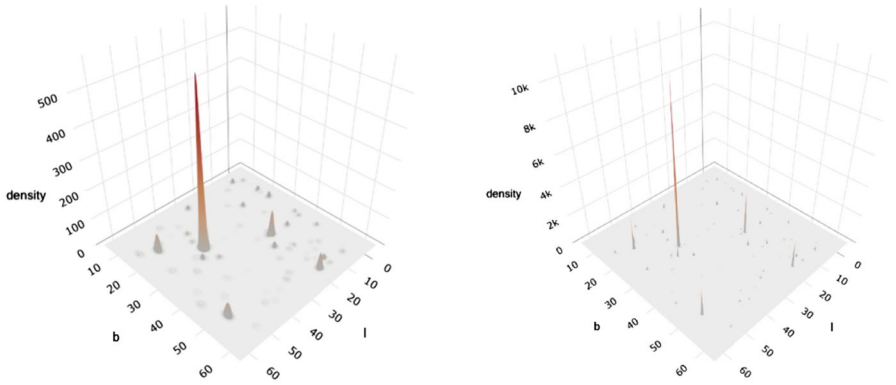


Fig. 2 Two examples of von Mises–Fisher kernel density estimates using the high-energy photons simulated for the validation region $(l, b) \in [0°, 60°] \times [10°, 60°]$ (blue area in Fig. 1) for the two values $h = 0.01$ (left) and $h = 0.001$ (right) which over- and undersmooth, respectively (color figure online)

$$CV(h) = \sum_{i=1}^{n} \log \hat{f}_{h,-i}(\boldsymbol{x}_i).$$

Here, $\hat{f}_{h,-i}(\boldsymbol{x}_i)$ is the kernel density estimate we obtain after having omitted observation i, evaluated at $\boldsymbol{x}_i$. A further natural way to tackle the bandwidth selection problem is to minimize the *Asymptotic Mean Integrated Squared Error* (AMISE) which is readily obtained from the bias and variance approximations provided for the directional setting by Hall et al. (1987). See also the discussion in Klemelä (2000) and Zhao and Wu (2001). This leads to the generalization of García-Portugués (2013)'s*rule of thumb* to spherical data,

$$h_{THUMB} = \left[\frac{8 \sinh^2(\hat{\kappa})}{\hat{\kappa}[(1 + 4\hat{\kappa}^2) \sinh(2\hat{\kappa}) - 2\hat{\kappa} \cosh(2\hat{\kappa})]n} \right]^{\frac{1}{6}},$$

where the concentration parameter $\hat{\kappa}$ is estimated by maximum likelihood.

The above two bandwidth selection rules are known to, respectively, undersmooth and oversmooth the true density. As the empirical evidence reported in Sect. 5.1 will show, this is especially true for the current astronomical context. Here, the density structure embraces thousands of sources which are highly heterogeneous in size and spread so as to require the smoothing parameter to adapt to the local behaviour of f. To the best of our knowledge, the issue of *adaptive* smoothing didn't receive much attention by the literature on directional data. To have the bandwidth h to depend on the current location $\boldsymbol{x}_i$of the estimator, we hence decided to adapt some results which are well-established in the linear case. A first possibility is to use Abramson's (1982) rule, whereh_i varies proportionally with the inverse of the square root of $\hat{f}_h(\boldsymbol{x}_i)$,

$$h_{i,P}^{A} = h_P \left[\hat{f}_{h_P}(\boldsymbol{x}_i)\right]^{-\frac{1}{2}}.$$

Here, $\hat{f}_{h_P}(\boldsymbol{x}_i)$ is a winsorized (or clipped) version of a suitably constructed pilot kernel density estimate with fixed bandwidth h_P, which may be, for example, h_{THUMB} or h_{LCV}. A further possibility is to use the modification proposed by Silverman (1986, Section 5.3), where the values of $\hat{f}_{h_P}(\boldsymbol{x}_i)$ are scaled by a factor which depends on their geometric mean. In the following, we refer to this criterion as $h_{i,P}^{S}$.

3.2 Using scientific input

A valid alternative for determining the smoothing parameter h is to use scientific input. As mentioned in Sect. 1.1, the spatial scattering of the photons around the source direction $\boldsymbol{\mu}$ is modelled by the LAT's *point spread function* (PSF). This function depends on the energy of the incoming photon, on its inclination angle, and on the quality of the recorded event (Ackermann et al. 2013). The latter is expressed by the PSF event type, that is, an event-level quantity which indicates how well the LAT managed to reconstruct the direction of the incoming photon and which assumes four values, from the lowest quality (PSF0) to the best quality (PSF3). Most importantly, the PSF depends on the scale factor

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

which describes the uncertainty of the event as a decreasing function of the energy E_i, expressed in Mega electron Volt (MeV), and of the two parameters $c_{0,i}$ and $c_{1,i}$, which are given distinct values for the different event qualities and can be retrieved from the Fermi LAT web site.[2] The first constant, $c_{i,0}$, represents multiple scattering while $c_{1,i}$ represents the spatial resolution of the LAT tracker. On this basis, we may specify a variable bandwidth as

$$h_{i,SE} = \sqrt{\left(c_{0,i}\left(\frac{E_i}{100\text{MeV}}\right)^{-0.8}\right)^2 + c_{1,i}^2}. \tag{4}$$

The left panel of Fig. 3 illustrates how the precision of the measurements depends on the energy content. The right panel plots the values we obtain for $h_{i,SE}$ for the four different event types. Note how proposal (4) is coherent with optimal selection guidelines of adaptive bandwidth selection in the linear setting, as h_i varies inversely with the photon scattering and hence with f.

[2] https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_LAT_IRFs/IRF_PSF.html.

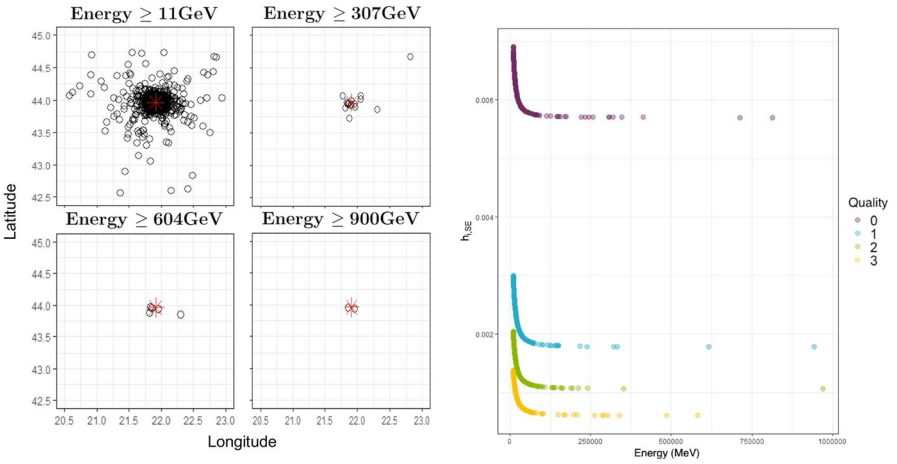


Fig. 3 Left: Photon scattering as a function of their energy content. Right: Values of $h_{i,SE}$ as a function of energy and event quality, where PSF0 represents the worst event type. The higher the energy and quality of the event, the smaller is the smoothing parameter

4 Modal clustering on the unit sphere

4.1 Mode hunting

Modal clustering associates clusters with the domain of attraction of the modes of the underlying density f. Recent insight into the theoretical foundations of modal clustering based on Morse theory (Milnor et al. 1969) is provided by Chacón (2015). Operationally, two main strands are usually pursued to identify the modal regions, depending on whether the modes are detected explicitly or not (Menardi 2016). A first strand follows the route of Hartigan (1975) and identifies clusters with high-density regions of the sample space, defined by the density level sets

$$L_c(f) = \{\mathbf{x} \in \Omega_2 : f(\mathbf{x}) \geq c\}, \quad 0 \leq c \leq \max f.$$

An estimate of the unknown $L_c(f)$ is obtained by replacing $f(\mathbf{x})$ by its non-parametric estimate $\hat{f}(\mathbf{x})$. The rationale behind this class of methods is that any connected component of $L_c(f)$ includes at least one mode of the density function, while, on the other hand, for each mode of the density function, there exists c for which one of the connected components of the associated $L_c(f)$ includes this mode at most. The major drawback is that identifying the connected components of a multidimensional set is not straightforward. See the recent accounts of Saavedra-Nieves and Crujeiras (Saavedra-Nieves and Crujeiras 2022) and Saavedra-Nieves and Fernández-Perez (2023) for a discussion of density-level set estimation and modal clustering for directional data following this route.

As our aim is to discover and identify unknown γ-ray emitting sources, we rather address the modal clustering problem according to a bump-hunting approach, which allows us to associate the modes of the unknown density $f$with the direction of the emitting sources. Yang et al. (2014) adapted the*mean-shift*algorithm developed by Fukunaga and Hostetler (1975) to be used with the directional kernel estimator (3) and fixed bandwidth h. Straightforward calculations allow us to extend their proposal to varying h_i, that is, for adaptive kernel density estimation on the unit sphere.

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, $(s+1)$, we have

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

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 $w_i(\mathbf{x})$ involve the derivative $K'(h_i^{-2}(1-\mathbf{x}^\top\mathbf{x}_i))$ of the kernel function, which leads to the weighted average

$$\hat{\boldsymbol{x}}^{(s+1)} = -\frac{\sum_{i=1}^{n} \boldsymbol{x}_i K'\left(\frac{1-\hat{\boldsymbol{x}}^{(s)\top}\boldsymbol{x}_i}{h_i^2}\right)}{\left\|\sum_{i=1}^{n} \boldsymbol{x}_i K'\left(\frac{1-\hat{\boldsymbol{x}}^{(s)\top}\boldsymbol{x}_i}{h_i^2}\right)\right\|_2},$$

where $\|\cdot\|_2$ is the Euclidean norm and the minus sign is due because $K(\cdot)$ is a decreasing function.

If we replace the kernel function $K(\cdot)$ by the von Mises–Fisher kernel, the above expression becomes

$$\hat{\boldsymbol{x}}^{(s+1)} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i \exp\left(\frac{\hat{\boldsymbol{x}}^{(s)\top}\boldsymbol{x}_i-1}{h_i^2}\right)}{\left\|\sum_{i=1}^{n} \boldsymbol{x}_i \exp\left(\frac{\hat{\boldsymbol{x}}^{(s)\top}\boldsymbol{x}_i-1}{h_i^2}\right)\right\|_2}.$$

As in the linear setting, the resulting sequence follows the gradient ascent path of the associated kernel density estimate of the data. The set of data points whose paths converge to the same local mode hence approximate the pertaining modal region and produce a partition of the sample space.

4.2 Post-processing

As mentioned in Sect. 1.1, the incoming photons were either emitted from a high-energy source or are part of the diffuse γ-ray background which spreads over the entire area observed by the telescope. The directional kernel density estimator (3) tries and reconstructs the corresponding mixture distribution. Hence, the small peaks which emerge as modes may identify true sources, but they may equally well represent a false

signal generated by the irregularly shaped background radiation. To separate the true signal emitted by a source from the background, we developed a post-processing procedure that combines the findings of two parallel quests. One establishes the significance of a candidate mode using a suitable statistical test. The second skims off the photons emitted by the γ-ray background using a suitable classifier build on previous knowledge provided by the Fermi LAT Collaboration. By super-imposing the findings from these two quests, we identify candidate sources which are both, statistically significant and qualified as such according to a set of relevant features. Furthermore, we are now able to distinguish photons emitted by a candidate source from those pertaining to the background radiation.

4.2.1 Statistical significance of the detected sources

Mathematically, we can verify whether a function reaches a local maximum by checking whether all eigenvalues of the Hessian matrix evaluated at the candidate mode are negative. Statistically, developing a suitable test to verify the existence of a mode and deriving its null distribution using eigenvalues is tricky, as these are not continuously differentiable functions of the Hessian. This invalidates resampling-based methods such as the bootstrap and asymptotic expansion by the delta method, which we may use to reconstruct the finite-sample null distribution of the test statistic. In the linear setting, Genovese et al. (2016) suggest to use data splitting to separate the process of finding candidate modes from the process of hypothesis testing. They furthermore propose to base inference on confidence intervals, rather than on$p$ values. The potential modes are hence estimated on the first half of the data, while the second half is used to construct asymptotically valid bootstrap confidence intervals for the eigenvalues of the Hessian matrix, which can be used for hypothesis testing.

The extension of this idea to directional data requires some care, as working on the unit sphere sets some constraints. To calculate the Hessian matrix $\mathcal{H}\hat{f}_h(\boldsymbol{x})$, we first need the total gradient

$$\nabla\hat{f}_h(\boldsymbol{x}) = \frac{C_2(h^{-2})}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x}_i}{h^2}\exp\left(\frac{\boldsymbol{x}^\top\boldsymbol{x}_i - 1}{h^2}\right),$$

where ∇ represents suitable differentiation. The Hessian matrix hence is

$$\begin{aligned}\mathcal{H}\hat{f}_h(\boldsymbol{x}) &= (\mathbf{I}_3 - \boldsymbol{x}\boldsymbol{x}^\top)\Big(\nabla\nabla\hat{f}_h(\boldsymbol{x}) - \nabla\hat{f}_h(\boldsymbol{x})^\top\boldsymbol{x}\mathbf{I}_3\Big)(\mathbf{I}_3 - \boldsymbol{x}\boldsymbol{x}^\top)\\ &= (\mathbf{I}_3 - \boldsymbol{x}\boldsymbol{x}^\top)\left[\frac{C_2(h^{-2})}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{h^4}\exp\left(\frac{\boldsymbol{x}^\top\boldsymbol{x}_i - 1}{h^2}\right) + \right.\\ &\quad \left. - \frac{C_2(h^{-2})}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x}^\top\boldsymbol{x}_i\mathbf{I}_3}{h^2}\exp\left(\frac{\boldsymbol{x}^\top\boldsymbol{x}_i - 1}{h^2}\right)\right](\mathbf{I}_3 - \boldsymbol{x}\boldsymbol{x}^\top).\end{aligned}$$

Likewise, we may obtain the Hessian matrix associated with an adaptive kernel density estimator with variable bandwidth h_i. The tricky part is that the eigenvalue of $\mathcal{H}\hat{f}_h(\boldsymbol{\mu})$, when $\hat{f}_h(\boldsymbol{x})$ is evaluated at $\boldsymbol{\mu}$, is always zero, whether $\boldsymbol{\mu}$ corresponds to a

Table 1 Performance metrics for different choices of the bandwidth h of the von Mises–Fisher kernel density estimator applied to the blue sky region plotted in Fig. 1 of size $(l, b) \in [0°, 60°] \times [10°, 60°]$: ARI = adjusted Rand index; $\bar{d}(s, \hat{s})$ = median angular distance (in degrees) between the directions of true sources (s) and candidate sources ($\hat{s}$) identified by the algorithm; n_s = number of identified sources

h	ARI	$\bar{d}(s, \hat{s})$	n_s
$h_{i,SE}$	0.9976	0.0004	86
h_{THUMB}	0.6841	0.0079	10
$h^A_{i,THUMB}$	0.6805	0.0139	18
$h^S_{i,THUMB}$	0.8524	0.0063	25
h_{LCV}	0.9777	0.0092	142
$h^A_{i,LCV}$	0.9777	0.0092	142
$h^S_{i,LCV}$	0.9777	0.0092	142

The number of true sources is 68

true source or not. This entails that inference has to be based on the remaining two eigenvalues. We hence construct a $1 - \alpha$ level confidence interval for the largest non null eigenvalue using bootstrap resampling. The candidate mode is validated if the interval includes only negative values.

A second possibility, which is operationally equivalent, is to reparametrize the von Mises–Fisher kernel in terms of the Galactic coordinates (l, b). This workaround allows us to directly apply the results by Genovese et al. (2016).

4.2.2 Background filtering

A post-processing step is added to further skim off the photons emitted by the diffuse background and identify those which originate from the extra-galactic sources. This is very much in line with the semi-supervision approach commonly used in particle physics searches to combine the quest for an as yet undetected signal with former knowledge about the phenomenon under study. See, for instance, Casa and Menardi (2022). Our background filter uses features which can be extracted at the various steps of the mean-shift algorithm, such as the number of photons assigned to a mode, the density estimates for the signal and the background models and various types of distances between the photons and their mode, in addition to the information already provided by the Fermi LAT Collaboration which includes the energy content of the photons and their incoming direction. A tree-based classifier is trained and tested on a suitable area of the sky. The final classifier is then pruned so as to assign any cluster with a single photon to the background.

Table 2 Standard deviation (Column 2) of photon scattering for 5 selected sources of varying size (Column 1) and average bandwidths computed using the scale factor of the PSF (Column 3) or selected by Abramson's or Silverman's rules (Columns 4–7)

Source	sd	$\bar{h}_{i,SE}$	$\bar{h}^{A}_{i,LCV}$	$\bar{h}^{A}_{i,THUMB}$	$\bar{h}^{S}_{i,LCV}$	$\bar{h}^{S}_{i,THUMB}$
$n_s = 7$	0.0019	0.0017	3.2958×10^{-06}	0.1053	2.8623×10^{-07}	0.0611
$n_s = 19$	0.0048	0.0028	3.2225×10^{-06}	0.0221	2.7986×10^{-07}	0.0128
$n_s = 31$	0.0042	0.0027	2.8184×10^{-06}	0.0501	2.4477×10^{-07}	0.0290
$n_s = 79$	0.0030	0.0028	2.1721×10^{-06}	0.0314	1.8864×10^{-07}	0.0182
$n_s = 151$	0.0068	0.0027	2.0684×10^{-06}	0.0215	1.7963×10^{-07}	0.0125

5 Application to Fermi LAT data

We benchmarked our algorithm on a simulated sample of γ-ray photon emission for a 5-year period of observation provided by the Fermi LAT Collaboration[3] based on the *Fourth Catalog of Fermi-LAT Sources* (4FGL) (Abdollahi et al. 2020).

5.1 Optimal bandwidth selection

The first step was to select the optimal bandwidth h using a simulated sample of 2335 photons emitted by the 68 sources which are present in the blue validation region $(l, b) \in [0°, 60°] \times [10°, 60°]$ of Fig. 1. Table 1 compares the selection criterion based on scientific input discussed in Sect. 3.2, with the different proposals for bandwidth selection listed in Sect. 3.1. Three performance metrics are used, which are the adjusted Rand index (ARI), the median angular distance (in degrees) between the directions of true sources and candidate sources, $\bar{d}(s, \hat{s})$, and the number n_s of identified sources. The three proposals based on the rule of thumb oversmooth the true photon density, leading to rather low ARI values. Likelihood cross validation, on the other hand, tends to over adapt the true density yielding too many candidate sources: 142 in place of the 68 present. The best partition of the selected sky region is obtained when using the variable bandwidth $h_{i,SE}$, that is, the scale factor of the LAT's point spread function.

Further support to the choice of $h_{i,SE}$ is provided by Table 2, which contrasts the selected optimal bandwidths (Columns 3–7) with the true photon scattering, as measured by its standard deviation (Column 2), for 5 selected sources of varying size, that is, which emit from a minimum of $n_s = 7$ photons up to a maximum of $n_s = 151$ photons. Again, $h_{i,SE}$ is the best performing choice.

5.2 Post-processing

We next trained and tested our tree-based classifier on the sample of 35,365 simulated photon emissions of the green sky region of size $(l, b) \in [120°, 150°] \times [15°, 70°]$ in

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

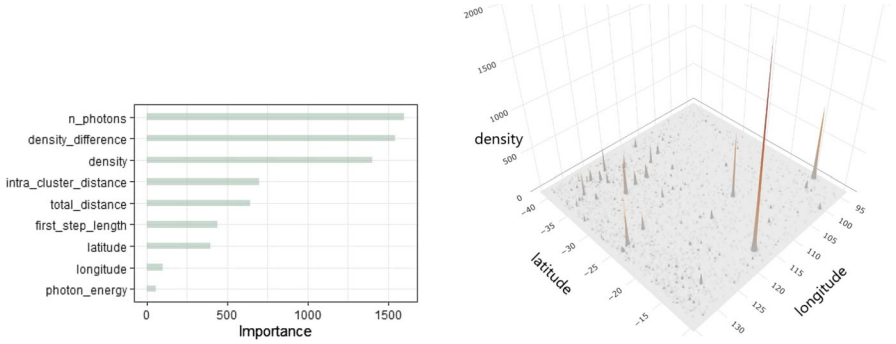


Fig. 4 Left: Feature importance plot for the tree-based photon classifier used to discriminate between source and diffuse γ-ray background emission. Right: Kernel density estimate using a von Mises–Fisher kernel for the 3849 γ-ray photon counts simulated for the LAT in a 5-year period in the region of the Southern sky identified by $(l, b) \in [95°, 135°] \times [-40°, -10°]$ and shown in yellow in Fig. 1. About 26% of the photons were emitted by the 44 sources present in the area (color figure online)

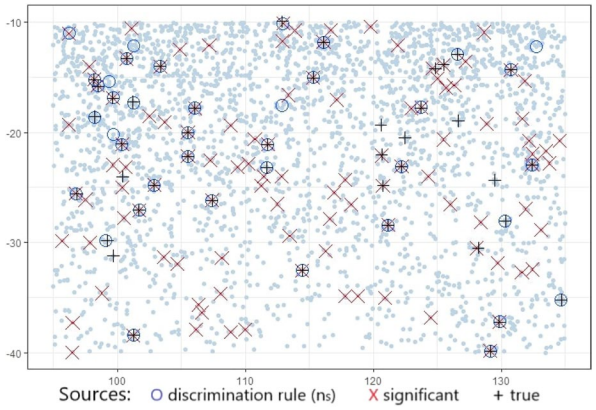
Fig. 1. This area covers a large portion of the Northern sky to account for the rather prominent variability of the diffuse γ-ray background as we move away from the Galactic plane. The classifier was estimated on the first 2/3 of the sample, for a total of 24,573 photons, and tested on the remaining 11,062 photons, where it selected a total of $n_s = 86$ sources. In both sets, about 85% of the photons were emitted from the background. The average sensibility, computed on the candidate sources identified by the classifier, was 90.5%, while the average specificity was 99.5%. The ARI is 0.9752 and the median angular distance between the true sources and the identified ones is 0.0005 degrees.

The importance of the selected predictor variables is shown in the left panel of Fig. 4. The most discriminating features are the number of photons assigned to a cluster (*n_photons*), the difference between the two photon densities for, respectively, the all sky and background counts only (*density_differences*), and the density observed for each photon (*density*). The final classifier was furthermore pruned so as to assign any cluster with a single photon to the background.

5.3 Source detection

The yellow region in Fig. 1 shows a portion of the Southern sky of size $(l, b) \in [95°, 135°] \times [-40°, -10°]$ with 3849 simulated photon counts for a 5-year period of observation. 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 Fig. 4 plots the estimated kernel density (3) using a von Mises–Fisher kernel. Here, the bandwidth parameter h was set according to scientific input, as described in Sect. 3.2. In all, the mean-shift algorithm identified 876 modes. To further refine the list of candidate sources we proceeded in two steps as outlined in Sect. 4.2.

On one side, we used the tree-based classifier trained in Sect. 5.2 to discriminate between source and background emission. This reduced the original 876 modes to 39 candidate sources, which are shown as blue circles in the left panel of Fig. 5. The table



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

Fig. 5 Left: Simulated Fermi-LAT γ-ray photon count map for the analysed 5-year observation period with superimposed the true sources (black crosses) and our candidate sources. A red cross pinpoints a candidate source which is statistically significant at the 5% level, while a blue circle identifies a candidate source on the basis of its features. Right: Performance measures of the tree-based classifier (color figure online)

on the right reports the performance of our classifier in terms of ARI and median angular 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. In parallel, we tested all the 555 clusters which contain two or more photons at a significance level of 5% as outlined in Sect. 4.2.1 and applying Bonferroni's correction. This skimmed off 448 modes, for a total of 107 remaining candidate sources, shown in the left panel of Fig. 5 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%, while the ARI exceeds 0.96.

For the sake of comparison, Costantin et al. (2020), Table 2)—which is the only contribution we can to a certain extent compare to—achieved remarkably lower ARIs and median distances on similar cuts, and this on photon emission counts which didn't consider the background.

6 Concluding remarks

The recent literature on density estimation has boasted novel approaches for general complex domains; see, for example, Arnone et al. (2022) and Baldi et al. (2009) and references therein. But, it has also refueled the interest in well-established techniques by generalizing their usability as, for instance, done by

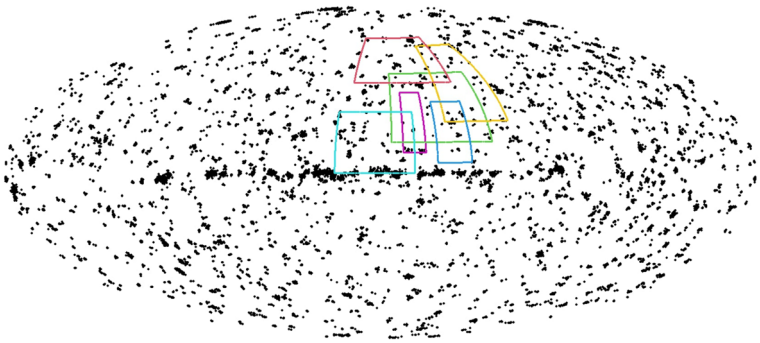


Fig. 6 Selection of overlapping regions for *consensus* clustering

Di Marzio et al. (2019). Our contribution intends to follow the latter route by taking advantage of the natural link between kernel density estimation and the mean-shift algorithm for mode identification. Our proposal represents a step forward towards the implementation of a fast and scalable computational tool to efficiently and effectively extract knowledge from today's large astronomical databases. Here, the widely used model-based approach to multivariate classification, which involves maximizing the likelihood of the mixture model using, for instance, the expectation maximization (EM) algorithm or Markov chain Monte Carlo (MCMC) simulation, is computationally impractical. Nonparametric methods, on the other hand, allow us to maintain the high flexibility required by the complex structure of astronomical data characterized by a sometimes low signal-to-noise ratio, while relying on sound theoretical basis.

We aimed at providing a highly performing tool for the identification of astronomical sources in terms of both, detection performance and computational speed. But, the focus was also on enriching the statistical toolbox of astrophysicists by considering a modeling approach which substantially differs to what done so far. In particular, using scientific input proved to be a winning choice for the selection of the optimal bandwidth. Furthermore, a set of features linked to the mean-shift algorithm turned out to useful to discrimante background photon emission from true source emission. Our final goal is to analyze the whole sky maps in one go. We are currently fine-tuning our algorithm by including a *consensus clustering* step. This will allow us to aggregate results from multiple runs, while guaranteeing more stable and robust results (Monti et al. 2003; Vega-Pons and Ruiz-Shulcloper 2011). More precisely, borrowing from Nordhaug Myhre et al. (2018), we form a *clustering ensemble* consisting of separate and bootstrapped runs of the mean-shift algorithm on a given number of overlapping regions of the sky, as shown in Fig. 6. The size and location of these regions varies on a random basis. The final modes are identified by selecting the cluster configuration which was observed most of the times. This way of proceeding guarantees robustness with respect to the choice of the smoothing

parameter h, while at the same time allowing us to work with tremendous amount of data.

Acknowledgements We would like to acknowledge the many useful comments and suggestions provided by two anonymous referees which greatly helped us improving a previous version of the manuscript. Andrea Sottosanti was supported by PNRR (National Recovery and Resilience Plan) Mission 4—Investment 1.3, financed by European Union—NextGenerationEU, Project AGE-IT—A Novel Public-Private Alliance to Generate Socioeconomic, Biomedical and Technological Solutions for an Inclusive Italian Ageing Society, CUP C93C22005240007.

Funding Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interest to disclose.

References

Abdollahi S, Acero F, Ackermann M, Ajello M, Atwood WB, Axelsson M et al (2020) Fermi large area telescope fourth source catalog. Astrophys J Suppl Ser 247(1):33

Abramson IS (1982) Arbitrariness of the pilot estimator in adaptive kernel methods. J Multivar Anal 12(4):562–567

Acero F, Ackermann M, Ajello M, Albert A, Baldini L, Ballet J, Barbiellini G (2016) Development of the model of galactic interstellar emission for standard point source analysis of Fermi Large Area Telescope data. Astrophys J Suppl Ser 223(2):26

Ackermann M, Ajello M, Allafort A (2013) 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. Astrophys J 765(1):54

Arnone E, Ferraccioli F, Pigolotti C, Sangalli LM (2022) A roughness penalty approach to estimate densities over two-dimensional manifolds. Comput Stat Data Anal 174:107527

Bai ZD, Rao CR, Zhao LC (1988) Kernel estimators of density function of directional data. J Multivar Anal 27(1):24–39

Baldi P, Kerkyacharian G, Marinucci D, Picard D (2009) Asymptotics for spherical needlets. Ann Stat 37(3):1150–1171

Casa A, Menardi G (2022) Nonparametric semi-supervised classification with application to signal detection in high energy physics. Stat Methods Appl 31(3):531–550

Chacón JE (2015) A population background for nonparametric density-based clustering. Stat Sci 30(4):518–532

Costantin D, Menardi G, Brazzale AR, Bastieri D, Fan JH (2020a) A novel approach for pre-filtering event sources using the von Mises–Fisher distribution. Astrophys Space Sci 365:53

Costantin D, Sottosanti A, Brazzale AR, Bastieri D, Fan JH (2020b) Bayesian mixture modelling of the high-energy photon counts collected by the Fermi Large Area Telescope. Stat Model 22(3):1–37

Di Marzio M, Fensore S, Panzera A, Taylor CC (2019) Kernel density classification for spherical data. Stat Probab Lett 144:23–29
Fukunaga K, Hostetler LD (1975) The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory 21(1):32–40
García-Portugués E (2013) Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. Electron J Stat 7(1):1655–1685
Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L (2016) Non-parametric inference for density modes. J R Stat Soc 78(1):99–126
Hall P, Watson GS, Cabrera J (1987) Kernel density estimation with spherical data. Biometrika 74(4):751–762
Hartigan JA (1975) Clustering algorithms. Wiley, Hoboken
Hobson M, Jaffe A, Liddle A, Mukherjee P, Parkinson D (eds) (2009) Bayesian methods in cosmology. Cambridge University Press, Cambridge
Jones DE, Kashyap VL, Van Dyk DA (2015) Disentangling overlapping astronomical sources using spatial and spectral information. Astrophys J 808(2):137
Klemelä J (2000) Estimation of densities and derivatives of densities with directional data. J Multivar Anal 73(1):18–40
Mardia KV, Jupp PE (2000) Directional statistics. Wiley, Hoboken
Mattox JR, Bertsch DL, Chiang J, Dingus BL (1996) The likelihood analysis of EGRET data. Astrophys J 148:148–162
Menardi G (2016) A review on modal clustering. Int Stat Rev 84(3):413–433
Meyer AD, van Dyk DA, Kashyap VL, Campos LF, Jones DE, Siemiginowska A, Zezas A (2021) eBASCS: disentangling overlapping astronomical sources II, using spatial, spectral, and temporal information. Mon Not R Astron Soc 506(4):6160–6180
Milnor J, Spivak M, Wells R (1969) Morse theory (AM-51), vol 51. Princeton University Press, Princeton
Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn 52:91–118
Montin A, Brazzale AR, Menardi G (2022) Locating γ-ray sources on the celestial sphere via modal clustering. In: Balzanella A, Bini M, Cavicchia C, Verde R (eds) Book of the short papers—SIS 2022, pp 1582–1587
Nordhaug Myhre J, Øyvind Mikalsen K, Løkse S, Jenssen R (2018) Robust clustering using a kNN mode seeking ensemble. Pattern Recognit 76:491–505
Saavedra-Nieves P, Crujeiras RM (2022) Nonparametric estimation of directional highest density regions. Adv Data Anal Classif 16(3):761–796
Saavedra-Nieves P, Fernández-Perez M (2023) Directional density-based clustering. arXiv:2302.09897
Silverman BW (1986) Density estimation for statistics and data analysis. Stat Appl Probab
Sottosanti A, Bernardi M, Brazzale AR, Geringer-Sameth A, Stenning DC, Trotta R, van Dyk DA (2021) Identification of high-energy astrophysical point sources via hierarchical Bayesian nonparametric clustering. arXiv:2104.11492
van Dyk DA, Connors A, Kashyap VL, Siemiginowska A (2001) Analysis of energy spectra with low photon counts via Bayesian posterior simulation. Astrophys J 548(1):224–243
Vega-Pons S, Ruiz-Shulcloper J (2011) A survey of clustering ensemble algorithms. Int J Pattern Recognit Artif Intell 25(3):337–372
Yang MS, Chang-Chien SJ, Kuo HC (2014) On mean shift clustering for directional data on a hypersphere. In: Rutkowski L, Korytkowski M, Scherer R, Tadeusiewicz R, Zadeh LA, Zurada JM (eds) Artif Intell Soft Comput. Springer, Cham, pp 809–818
Zhao L, Wu C (2001) Central limit theorem for integrated square error of kernel estimators of spherical density. Sci China Ser A Math 44:474–483

Authors and Affiliations

Anna Montin[1] · Alessandra R. Brazzale[1] · Giovanna Menardi[1] · Andrea Sottosanti[2]

✉ Alessandra R. Brazzale
alessandra.brazzale@unipd.it

Anna Montin
anna.montin@studenti.unipd.it

Giovanna Menardi
giovanna.menardi@unipd.it

Andrea Sottosanti
andrea.sottosanti@unipd.it

[1] Department of Statistical Sciences, University of Padova, Via C. Battisti 241, 35121 Padua, PD, Italy

[2] Department of Medicine, University of Padova, Via Giustiniani 2, 35128 Padua, PD, Italy