Discovering and Locating High-Energy Extra-Galactic Sources by Bayesian Mixture Modelling

A. Sottosanti, D. Costantin, D. Bastieri, A. R. Brazzale

Abstract Discovering and locating gamma-ray sources in the whole sky map is a declared target of the *Fermi* Gamma-ray Space Telescope collaboration. In this paper, we carry out an unsupervised analysis of the collection of high-energy photons accumulated by the Large Area Telescope, the principal instrument on board the *Fermi* spacecraft, over a period of around 7.5 years using a Bayesian mixture model. A fixed, though unknown, number of parametric components identify the extra-galactic emitting sources we are searching for, while a further component represents parametrically the diffuse gamma-ray background due to both, extra-galactic and galactic high-energy photon emission. We determine the number of sources, their coordinates on the map and their intensities. The model parameters are estimated using a reversible jump MCMC algorithm which implements four different types of moves. These allow us to explore the dimension of the parameter space. The possible transitions remove from or add a source to the model, while leaving the background component unchanged. We furthermore present an heuristic procedure, based on the posterior distribution of the mixture weights, to qualify the nature of each detected source.

Andrea Sottosanti
University of Padova, Department of Statistical Sciences, via Cesare Battisti 241, Padova, Italy, e-mail: sottosanti@stat.unipd.it

Denise Costantin
Guangzhou University, Center for Astrophysics, Guangzhou 510006 and Astronomical Science and Technology Research Laboratory, Department of Education, Guangdong Province PRC, e-mail: denise.costantin@gmail.com

Denis Bastieri
University of Padova, Department of Physics and Astronomy "Galileo Galilei", via Belzoni 7, Padova, Italy, e-mail: denis.bastieri@unipd.it

Alessandra R. Brazzale
University of Padova, Department of Statistical Sciences, via Cesare Battisti 241, Padova, Italy, e-mail: alessandra.brazzale@unipd.it

1 Motivation and background

Resolving the γ-ray sky by detecting as yet unidentified sources and accurately measuring the diffuse background emission is a declared key scientific objective of the *Fermi* Gamma-ray Space Telescope collaboration, whose broader aim is to identify and study the nature of high-energy phenomena in the Universe.[1] The target of this contribution is the collection of photon count maps for varying energy bins provided by the Large Area Telescope (LAT), the principal scientific instrument on board the *Fermi* spacecraft, during its almost ten years of operation. In particular, we aim at formulating and fitting a model which allows us to: (i) determine the number of extra-galactic high-energy sources, (ii) measure their intensities, and (iii) pool the individual photon counts into the corresponding clusters.

The discovery of celestial objects is an intrinsically interdisciplinary field which combines both, statistical and astronomical methodology. A main challenge of trying and detecting high-energy phenomena from astronomical data is to separate the signal of the putative emitting source from the diffuse γ-ray background which spreads over the entire area observed by the telescope. Different phenomena contribute to this residual radiation [3]. Broadly speaking, its origins can be brought under two headings: 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 the γ-ray emission from faint unresolved sources and any residual galactic emission which is approximately isotropic.

Traditionally, the analysis is based on so-called *single-source* models, as described in Section 7.4 of [7]. Generally speaking, the application of these models requires 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 significance tests. An illustrative example is given in [11], who employ Poisson regression to model the number of photons at each pixel. Further treatments from both, the frequentist and the Bayesian viewpoints, can be found in [6, 10, 13, 14]. *Variable-source-number* models address the problem from a more global perspective, as they simultaneously estimate the number of sources in the whole map without the need to separate the latter into smaller cells and to work on single pixels [7, Section 7.3]. A recent proposal, which analyses X-ray count maps according to this approach, is made in [8].

Here we propose a Bayesian mixture model with a finite, but unknown, number of components for the known and as yet unidentified extra-galactic high-energy sources plus an additional parametric component to represent the diffuse γ-ray background. The directions of the high-energy photons collected by the *Fermi* LAT over a period of approximately 7.5 years is then used to estimate simultaneously the number of sources in the map, their coordinates and their intensities. As in [8], our algorithm iteratively identifies the sources and pulls the individual photons into the corresponding clusters. It furthermore automatically selects the number of components of the mixture. However, [8] consider only the isotropic diffuse X-ray back-

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

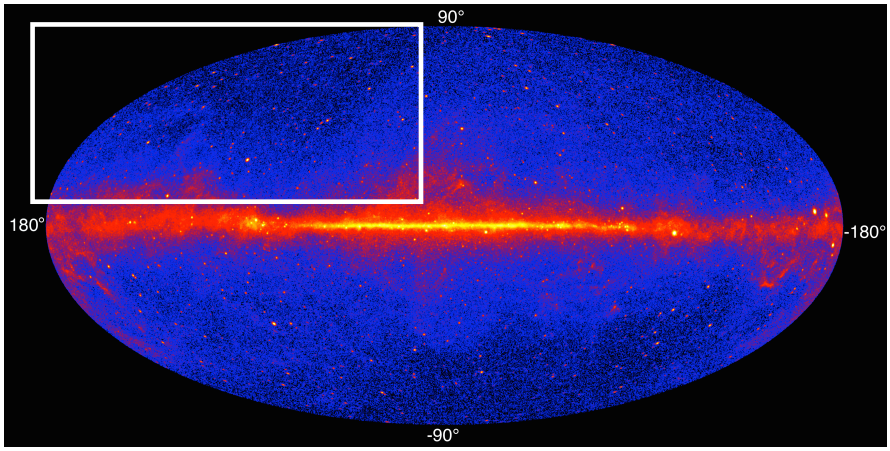

Fig. 1 Whole sky map at γ-ray wavelengths and energies larger than 1 GeV based on data accumulated by the LAT over a period of five years of operation (Image Credit: NASA/DOE/Fermi LAT Collaboration). The region framed in white represents the area analyzed in this paper.

ground, which they model assuming a uniform distribution over the entire map. This assumption is too restrictive if the targets are γ-ray sources, as we cannot neglect the huge contamination due to galactic interstellar emission, but have to suitably model it.

The remainder of the paper is organised as follows. Section 2 presents the *Fermi* LAT data which motivated this contribution. Our proposal of a Bayesian finite mixture model is presented in Section 3 and is fit to the *Fermi* LAT data in Section 4. In this latter section we furthermore discuss the capability of our model to skim off the signal of the sources from the background radiation. The paper closes with the concluding remarks of Section 5.

2 The *Fermi* LAT data

The data collected by the *Fermi* Large Area Telescope (LAT) contribute uniquely to the study of the most extreme phenomena in our Universe such as active galactic nuclei, supernova remnants and pulsar wind nebula. Figure 1 represents the Mollweide projection in galactic coordinates of the entire γ-ray sky at energies larger than 1 GeV and is based on the data collected by NASA's *Fermi* LAT over a five years period.[2] The brighter the grey tone, the larger is the intensity of the γ-ray source. The brilliant horizontal stripe which crosses the middle part of the figure to a huge extent conveys the high-energy photon emission of our Milky Way, at whose center we assume a supermassive black hole. The isotropic diffuse γ-ray background is

[2] http://fermi.gsfc.nasa.gov/ssc/

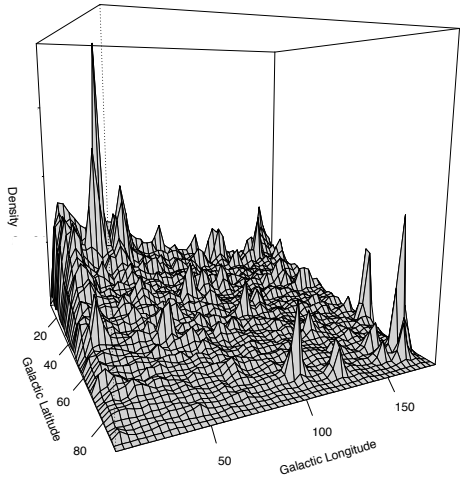

Fig. 2 Nonparametric kernel estimate of the photon density distribution based on the γ-ray count maps accumulated by the *Fermi* LAT over a 7.5 years period. The map is expressed in galactic coordinates and refers to the sky region $[180^\circ, 10^\circ] \times [10^\circ, 90^\circ]$, that is, to the area framed in white in Figure 1. The spikes represent potential high-energy extra-galactic emitting sources. Both components of the γ-ray background, the GIE and the IGRB, are visible.

much less evident, while we can clearly identify extra-galactic point and small-area γ-ray emitting sources.

The dataset used in this paper is the collection of photons, generated by different astrophysical events and collected by the LAT over a period of around 7.5 years of observation, whose energy exceeds 10 GeV. The aim of our analysis is to discriminate the signal of extra-galactic γ-ray emitting sources from the various background phenomena, and to reconstruct their direction in the sky map. In particular, for the reasons we will shortly give below, we restrict our attention to a subregion of the sky whose galactic longitude and latitude lie in the intervals $[180^\circ, 10^\circ]$ and $[10^\circ, 90^\circ]$, respectively.[3] This region is framed in white in Figure 1 and covers broadly the fourth quadrant of the map. In all, 51,000 observations fall in this area. Figure 2 plots the smoothed nonparametric estimate of the photon density distribution. The various spikes identify known and as yet unrevealed high-energy emitting sources.

We decided to test and fine tune our algorithm in a region of the sky map where the diffuse γ-ray background is less prevalent, and possibly dominated by the isotropic diffuse *gamma*-ray background (IGRB) component. Hence, we restricted our analysis to latitudes above 10° to limit the influence of the galactic interstellar emission (GIE). To further reduce and, at least partially remove, the background component radiating from the Galaxy center and from the so-called *Fermi* Bubbles [2], that is, from the two extended regions of excess γ-ray emission located near the galactic center, we only consider longitudes that vary from 180° to 10°. As is evident from Figure 2, the IGRB is still present though less pronounced as compared to Figure 1. In Section 3.1 we will discuss how to suitably model the diffuse background component. The third catalogue of hard Fermi LAT sources (3FHL, for short) reports 288 high-energy γ-ray emitting sources for the outlined region [4].

[3] Here we follow the convention adopted in astronomical whole sky maps to define the longitude on the left at 180° and at -180° on the right.

3 Bayesian source detection

We adopted a flexible Bayesian modelling approach which allowed us to detect catalogued and acknowledged γ-ray sources plus possible new candidates in the sky region of Figure 2. As in [8] we assembled a finite mixture model whose components were automatically identified using the available data and Bayesian computation. That is, in one go we determined both, the number of sources and their directions in space. The main difference to [8] is the presence of the rather intense background radiation which spreads over the entire map and represents a relevant component of our model. Section 3.1 describes the statistical model for the *Fermi* LAT data; Section 3.2 outlines the fitting procedure.

3.1 Statistical model

Let $x_i \in [180, 10]$ and $y_i \in [10, 90]$ represent the galactic longitude and latitude, respectively, of the n photons detected in the area of the extra-galactic space considered by our analysis. We start off by reconstructing the directions of the γ-ray sources by modelling how the photons they emit scatter around their source.

Assume that photon i was generated by source j whose galactic coordinates are $\mu_j = (\mu_{jx}, \mu_{jy})$, $j = 1, \ldots, K$. Here K represents the number of sources present in the map. The direction of photon i can then be modeled as

$$(X_i, Y_i) \mid \mu_j \sim PSF(\mu_j), \quad i = 1, \ldots, n, \tag{1}$$

where $PSF(\cdot)$ represents King's established *Point Spread Function* [9]. This function suitably describes how photons cluster around their emitting source. The corresponding density is

$$f(x_i, y_i \mid \mu) = \frac{C}{[1 + \{d(x_i, y_i \mid \mu)/d_0\}^2]^{\eta_0}}, \tag{2}$$

where

$$d(x_i, y_i \mid \mu) = \sqrt{(x_i - \mu_{jx})^2 + (y_i - \mu_{jy})^2/(1 - \varepsilon_0)^2}.$$

Here $d_0 = 0.6$ is the core radius measured in arcsec, $\eta_0 = 1.5$ is the power-law slope and $\varepsilon_0 = 0.00574$ represents the ellipticity; the normalizing constant C is usually determined numerically. The resulting density essentially characterizes a bivariate Student t distribution. The values of the parameters d_0, η_0 and ε_0 are chosen as in [8]. Actually, the *Fermi* LAT collaboration uses an extended version of King's density [1]. In particular, they assume that photons generated from the same source are not identically distributed, but each is characterized by its own dispersion which, in turn, depends on the energy level of the photon. However, for the energy range considered in this paper (>10 GeV), this variation is negligible and model (1) represents a valid approximation.

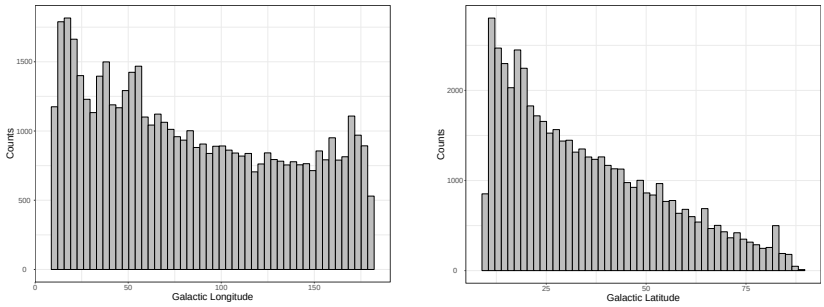

Fig. 3 Distribution of longitudes (left) and of latitudes (right) of the high-energy (>10 GeV) photons detected by the *Fermi* LAT during a 7.5 years period of observation.

A different model needs be specified in case the observed photon was not emitted from a specific source but is part of the background radiation. The authors of [8] assume a uniform distribution over the entire map to model the uniquely present isotropic component. We already discussed in Section 2 that this assumption is too restrictive for γ-ray counts. Model (1) is hence extended by considering a further bi-dimensional component

$$(X_i, Y_i) \mid \sigma_b \;\sim\; Unif(180, 10) \times tExp(\sigma_b). \tag{3}$$

The longitude of a photon stemming from the background is here modelled according to a uniform distribution, while its latitude follows a translated exponential distribution with scale parameter σ_b, that is, an exponential distribution whose support was translated to the interval $[10, +\infty)$. This model well represents the marginal distributions of longitude and latitude for the photons detected by the *Fermi* LAT shown in Figure 3. Suitable values will be chosen for σ_b so as to guarantee that the fitting procedure outlined in the following section generates admissible values for Y_i.

In practice, we have no information whether the photon was emitted from a source or belongs to the background, nor do we know the number of emitting sources and their directions in space. This situation is well represented by a finite mixture model which assumes a fixed, though unknown, number of components to represent the different sources plus an additional component to model the background radiation. This translates into the following marginal model

$$f(x_i, y_i \mid \mu, \sigma_b, \omega) = \omega_0 \; g_b(x_i, y_i \mid \sigma_b) + \sum_{j=1}^{K} \omega_j \; f(x_i, y_i \mid \mu_j), \tag{4}$$

where $g_b(\cdot \mid \cdot)$ represents the distribution of photons from the background as given in (3), while $f(\cdot \mid \cdot)$ models the signal of a specific source according to (1). The vector $\omega = (\omega_0, \omega_1, \ldots, \omega_K)$ contains the mixing proportions ω_j which can be viewed as the intensity ω_0 of the background and of each source, that is, ω_i, $i = 1, \ldots, K$. Our model is hence characterised by a set $\theta_K = \{\mu, \sigma_b, \omega\}$ of $3K + 2$ parameters.

Recall, furthermore, that the number K of undetected sources is itself supposed to be unknown and needs be estimated. So, inference will be made on (θ_K, K).

To write down the likelihood function of the statistical model defined at (4), we augment our data as originally proposed in [12] and also advocated in [8]. That is, for each observation $i = 1, \ldots, n$, we introduce the latent group variable Z_i which assumes values in the discrete set $\{0, 1, \ldots, K\}$ with probabilities given by the components of ω. Though actually never observed, this variable conveys useful information as it indicates the source number for photon i. The full data likelihood is then

$$\begin{aligned} L(\theta_K, K \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) &= p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}; \theta_K, K) p(\mathbf{z} \mid \theta_K, K) \\ &= \left[\prod_{i:z_i=0} g_b(x_i, y_i \mid \sigma_b) \prod_{j=1}^{K} \left\{ \prod_{i:z_i=j} f(x_i, y_i \mid \mu_j) \right\} \right] \prod_{j=0}^{K} \omega_j^{n_j}, \end{aligned} \tag{5}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_m)$ and $\mathbf{z} = (z_1, \ldots, z_n)$ are the vectors of observed and latent data, and $n_j = \sum_{i=1}^{n} I(z_i = j)$. As required by Bayes we complete our model definition by eliciting the a priori distributions for the unknown model parameters θ_K and K. Since there is no prior belief on the direction of the sources, a bivariate uniform distribution is used,

$$\mu_{jx} \sim Unif(180, 10) \quad \text{and} \quad \mu_{jy} \sim Unif(10, 90),$$

while the conjugate gamma distribution

$$\pi(\sigma_b \mid \nu, \beta) = \frac{\beta^{\nu}}{\Gamma(\nu)} \sigma_b^{\nu-1} e^{-\beta \sigma_b},$$

with $\nu = 0.02$ and $\beta = 1$, is chosen for the scale parameter σ_b. We, furthermore, assume that the unknown number of components K distributes as a truncated Poisson

$$K \sim tPoi(\kappa \mid \kappa_{min}, \kappa_{max}),$$

where $\kappa = 288$ equals the number of catalogued sources and $[\kappa_{min}, \kappa_{max}] = [250, 400]$. Lastly, conditionally on K, we let ω follow a Dirichlet distribution of size $K+1$ where the $K+1$ parameters are all set to $\alpha = 1$. This corresponds to assigning a priori equal probability to the K putative sources, or differently stated, to assuming that they have the same intensity.

Applying Bayes' theorem, the posterior joint distribution of the unknown model parameters (θ_K, K), conditionally on the latent group variables $\mathbf{z}$, results in

$$\pi(\theta_K, K \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) \propto L(\theta_K, K \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) \pi(\theta_K, K). \tag{6}$$

This is the function we will use to estimate the parameters. Note that to obtain the posterior distribution of θ_K and K given only the observed data $(\mathbf{x}, \mathbf{y})$, we would have to sum up (6) over all possible combinations of the latent variables $\mathbf{z} = (z_1, \ldots, z_n)$.

Algorithm 1 Reversible jump MCMC – split move

procedure SPLIT j INTO j_1, j_2 WITH PROBABILITY b_k (from k to $k+1$ sources)
  $b_k \leftarrow 0.25$, $d_{k+1} \leftarrow 0.25$
  if $k = \kappa_{min}$ **then** $b_k \leftarrow 0.5$
  $u_1, u_2, u_3 \sim Beta(2,2)$, $v \sim Unif(0,1)$
  $\omega_{j_1} \leftarrow u_1 \omega_j$ and $\omega_{j_2} \leftarrow (1-u_1)\omega_j$
  $\mu_{j_1 x} \leftarrow \mu_{jx} - u_2 \sqrt{\omega_{j_2}/\omega_{j_1}}$ and $\mu_{j_1 y} \leftarrow \mu_{jy} - u_3 \sqrt{\omega_{j_2}/\omega_{j_1}}$
  $\mu_{j_2 x} \leftarrow \mu_{jx} + u_2 \sqrt{\omega_{j_1}/\omega_{j_2}}$ and $\mu_{j_2 y} \leftarrow \mu_{jy} + u_3 \sqrt{\omega_{j_1}/\omega_{j_2}}$
  generate a new vector of labels $\mathbf{z}^*$ using $k+1$ sources
  $p_{k+1} \leftarrow \pi(\theta_{k+1}, k+1 \mid \mathbf{x}, \mathbf{y}, \mathbf{z}^*)$ and $p_k \leftarrow \pi(\theta_k, k \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$
  $g \leftarrow b_{2,2}(u_1) b_{2,2}(u_2) b_{2,2}(u_3)$, where $b_{2,2}(.)$ is the $Beta(2,2)$ density function
  $J \leftarrow \omega_j / [u_1(1-u_1)]$
  $q_k \leftarrow b_k / k$ and $q_{k+1} \leftarrow d_{k+1}/(k+1)$
  if $\arg\min_j \mid\mid \mu_{j_1}, \mu_j \mid\mid = j_2$ **and** $\arg\min_j \mid\mid \mu_{j_2}, \mu_j \mid\mid = j_1$ **then**
    $q_{k+1} \leftarrow 2 q_{k+1}$
  $A \leftarrow (p_{k+1} q_{k+1} J)/(p_k q_k g)$
  if $v \leq \min(1, A)$ **then** accept split

3.2 Model fitting

We by-passed numerical integration of the posterior kernel (6), as would have been required to compute the normalising constant, using Monte Carlo simulation. However, a further aspect considerably challenges the derivation of the posterior distribution of the model parameters: the dimension of θ_K is itself unknown as it depends on the number of sources K. We implemented a reversible jump Markov chain Monte Carlo algorithm, as proposed in [5], thanks to which we were able to both, reconstruct the posterior distributions of the unknown components of the model and to determine how many there are.

Here we present our algorithm. It consists of a two-stage procedure which iterates two steps: given K, we first update the latent group variables $\mathbf{Z}$ and generate values from the posterior distribution of θ_K; in the second step we redetermine the number of components K. Having written $\mathbf{z}^{(t-1)}$, $\theta_K^{(t-1)}$ and $K^{(t-1)}$ for the values generated at iteration $(t-1)$, the two steps can be summarised as follows:

1. generate $(\mathbf{z}^t, \theta_K^t)$ from the full conditional $\pi(\mathbf{z}, \theta_K \mid K^{(t-1)}; \mathbf{x}, \mathbf{y})$;
2. redefine the dimension of the parameter space, that is, specify a new order of the mixture by generating K^t from $\pi(K \mid \theta_K^t, \mathbf{z}^t; \mathbf{x}, \mathbf{y})$.

An alternative is to have the algorithm iterate Step 1 a given number of times, say 5 to 10, before proposing the trans-dimensional jump outlined at Step 2. Let us now have a closer look at the two steps.

Algorithm 2 Reversible Jump MCMC – birth move

procedure GENERATE j^* WITH PROBABILITY b_K (from k to $k+1$ sources)
  $b_k \leftarrow 0.25,\ d_{k+1} \leftarrow 0.25$
  if $k = \kappa_{min}$ **then** $b_k \leftarrow 0.5$
  $\mu_{j^*x} \sim Unif(1-,180),\ \mu_{j^*y} \sim Unif(10,90)$ and $\omega_{j^*} \sim Beta(1,k+1)$
  rescale the weights using $\omega_j \leftarrow \omega_j(1-\omega_{j^*})$
  generate a new vector of labels $\mathbf{z}^*$ using $k+1$ sources
  $p_{k+1} \leftarrow \pi(\theta_{k+1}, k+1 \mid \mathbf{x}, \mathbf{y}, \mathbf{z}^*)$ and $p_k \leftarrow \pi(\theta_k, k \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$
  $g \leftarrow \pi(\mu_{j^*x})\pi(\mu_{j^*y})b_{1,k+1}(\omega_{j^*})$
  $J \leftarrow (1-\omega_{j^*})^{k+1}$
  $q_k \leftarrow b_k/k$ and $q_{k+1} \leftarrow d_{k+1}/(k+1)$
  $A \leftarrow (p_{k+1}q_{k+1}J)/(p_k q_k g)$
  if $v \leq \min(1, A)$ **then** accept birth

Step 1

This step implements a Gibbs sampling scheme to update the model parameters θ_K and the latent variables $\mathbf{Z}$ for a fixed number K of components. Let, as above, the superscripts $(t-1)$ and t identify the values generated at iterations $(t-1)$ and t, respectively, and define as k the number of sources detected at iteration $(t-1)$, that is, $K^{(t-1)} = k$. Step 1 of the algorithm develops as follows.

1. For $i = 1, \ldots, n$, generate z_i^t from a multinomial distribution with probabilities

$$p(z_i^t = 0 \mid \theta_K^{(t-1)}, K^{(t-1)}; \mathbf{x}, \mathbf{y}) \ \propto\ \omega_0^{(t-1)} g_b(x_i, y_i \mid \sigma_b^{(t-1)})$$
$$p(z_i^t = j \mid \theta_K^{(t-1)}, K^{(t-1)}; \mathbf{x}, \mathbf{y}) \ \propto\ \omega_j^{(t-1)} f(x_i, y_i \mid \mu_j^{(t-1)}), \quad j \neq 0.$$

2. Generate a new vector of mixing probabilities ω^t from the Dirichlet distribution $Dir(n_0^t + \alpha, \ldots, n_k^t + \alpha)$, where $n_j = \sum_{i=1}^n I(z_i^t = j)$, $j = 1, \ldots, k$.

3. Generate μ_j^t, $j = 1, \ldots, k$, using a Metropolis-Hastings step applied to the full conditional distribution

$$\pi(\mu \mid \sigma_b^{(t-1)}, K^{(t-1)}; \mathbf{x}, \mathbf{y}, \mathbf{z}^t).$$

   Use as proposal distribution the bivariate normal distribution centered at $\mu_j^{(t-1)}$ and with covariance matrix the identity matrix rescaled by 0.5^2 so as to guarantee a satisfactory overlapping with King's PSF defined in (1).

4. Generate σ_b^t from the gamma distribution with scale parameter $\beta + n_0^t$ and shape parameter $\nu + \sum_{i=1}^n I(z_i = 0) y_i$.

Further examples can be found in [12] and [15].

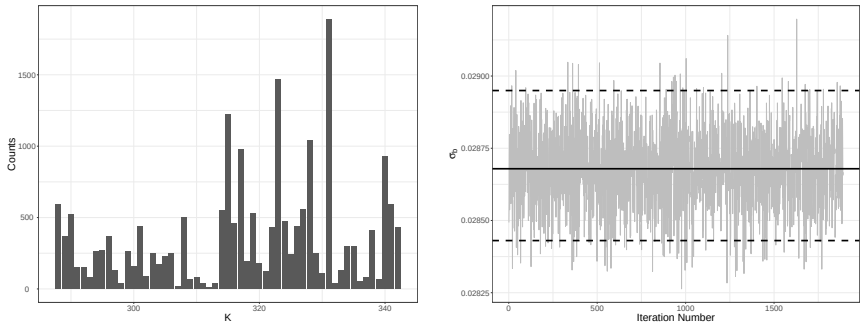

Fig. 4 Summary and diagnostic plots for the fitted model. Left: posterior distribution of K, the putative number of sources present in the analysed sky region. The modal value $K = 331$ is visited 1,892 times out of 20,000. Right: trace plot of the corresponding 1,892 σ_b values. The solid horizontal line at the center represents the posterior mode; the two dashed lines delimit the 0.95% highest posterior density interval.

Step 2

The second step implements the trans-dimensional jump which increases the number of components of the mixture or decreases it by one. The choice is made randomly with equal probabilities. New components are added to the model through either a split or a birth move; a component is removed from the model using a combining or death move [12]. These four steps allow the algorithm to explore the entire map and to search for new sources without affecting the distribution of the background radiation (3). A main difference to [12] is that we allow the algorithm to remove a component from the model using the death move also when it is not empty. This corresponds to delete clusters whose content of information does not qualify them as candidate sources.

The code boxes of Algorithms 1 and 2 list the pseudo code for the split and the birth moves. Note that they also provide the pseudo code for the combining and the death moves we use to down size by one the number of components of the mixture. So, for instance, to evaluate whether to reduce the number of sources from K to $K-1$ by combining two of them, we interchange $K-1$ and K in the split move outlined in Algorithm 1. The acceptance probability is then $\min\{1, 1/A\}$ instead of $\min\{1, A\}$.

4 Modelling the *Fermi* LAT data

We applied model (4) to the *Fermi* LAT data described in Section 2 and shown in Figure 2. The corresponding sky region is framed in white in Figure 1 and covers broadly one fourth of the area observed by the LAT. Recall, furthermore, that the third catalogue of hard *Fermi* LAT sources lists 288 high-energy γ-ray emitting

sources for this sector [4]. The 3FHL catalogue will furthermore be used to benchmark the detection capability of our model. We run our reversible jump MCMC algorithm, as described in the previous section, for a total of 20,000 iterations each. The number and directions of the sources present in the 3FHL catalogue were used as starting points for K and μ, respectively. This way, we acknowledge all the a priori available information. The starting points for the mixture weights ω and of the scale parameter σ_b, which characterises the distribution of the background radiation, were randomly drawn from their a priori distributions.

The left panel of Figure 4 shows the posterior distribution of K, the supposed number of high-energy γ-ray sources present in the analysed region. The posterior mode is $K = 331$, a value which was visited 1,892 times, that is, by around 9.5% iterations. We compared the posterior modes of (μ_{jx}, μ_{jy}), $j = 1, \ldots, K$, for these 331 putative sources with those present in the 3FHL catalogue: appreciably, our algorithm confirmed 255 of the acknowledged ones. The nature of the 76 remaining detections needed be investigated. We will come back to this point shortly. The right panel of Figure 4 traces the 1,892 values generated for σ_b, and shows a good mixing property of the chain. The posterior mode is 0.0287, slightly higher than what expected on average a priori, with 95% highest posterior density (HPD) interval $[0.0284, 0.0289]$. These values are also shown in Figure 4 as solid and dashed horizontal lines, respectively. Most interestingly, however, is the Bayesian estimate of $\omega_0 = 0.9387$ with 95% HPD interval $[0.9364, 0.9407]$. Remember that this value quantifies the intensity of the diffuse background radiation: it results that around 94% of the detected photons originated from it. Differently stated, only 6% of the photons were emitted from around 300 sources whose median intensity is 0.000137.

To further discriminate whether the 76 newly identified clusters correspond to real γ-ray emitting sources, we heuristically used the a posteriori available information on their intensities. Figure 5 shows the asymmetric boxplots of the posterior distributions of the 331 mixing proportions ω_i, $i = 1, \ldots, 331$. The white boxes correspond to the 255 already known sources, while the new candidates are drawn in black. Our ad hoc procedure defines the median of the posterior modes for the 255 catalogued sources as the threshold intensity above which we may expect a γ-ray emitting source. We hence qualified the 33 clusters whose posterior modes for ω_i satisfy this criterion as possible undetected sources. Their coordinates are currently being tested as prescribed by the *Fermi* LAT collaboration [4].

5 Conclusions

The results obtained for our model when applied to the *Fermi* LAT data of the limited sky region described in Section 2 are rather encouraging. We were able to detect 255 already known sources and to pinpoint possible new candidates. Of the 288 catalogued sources 33 were missed because their signal most likely isn't strong enough to be captured by our model but gets confounded with the prominent and irregularly shaped background radiation which pervades the considered area even after the ini-

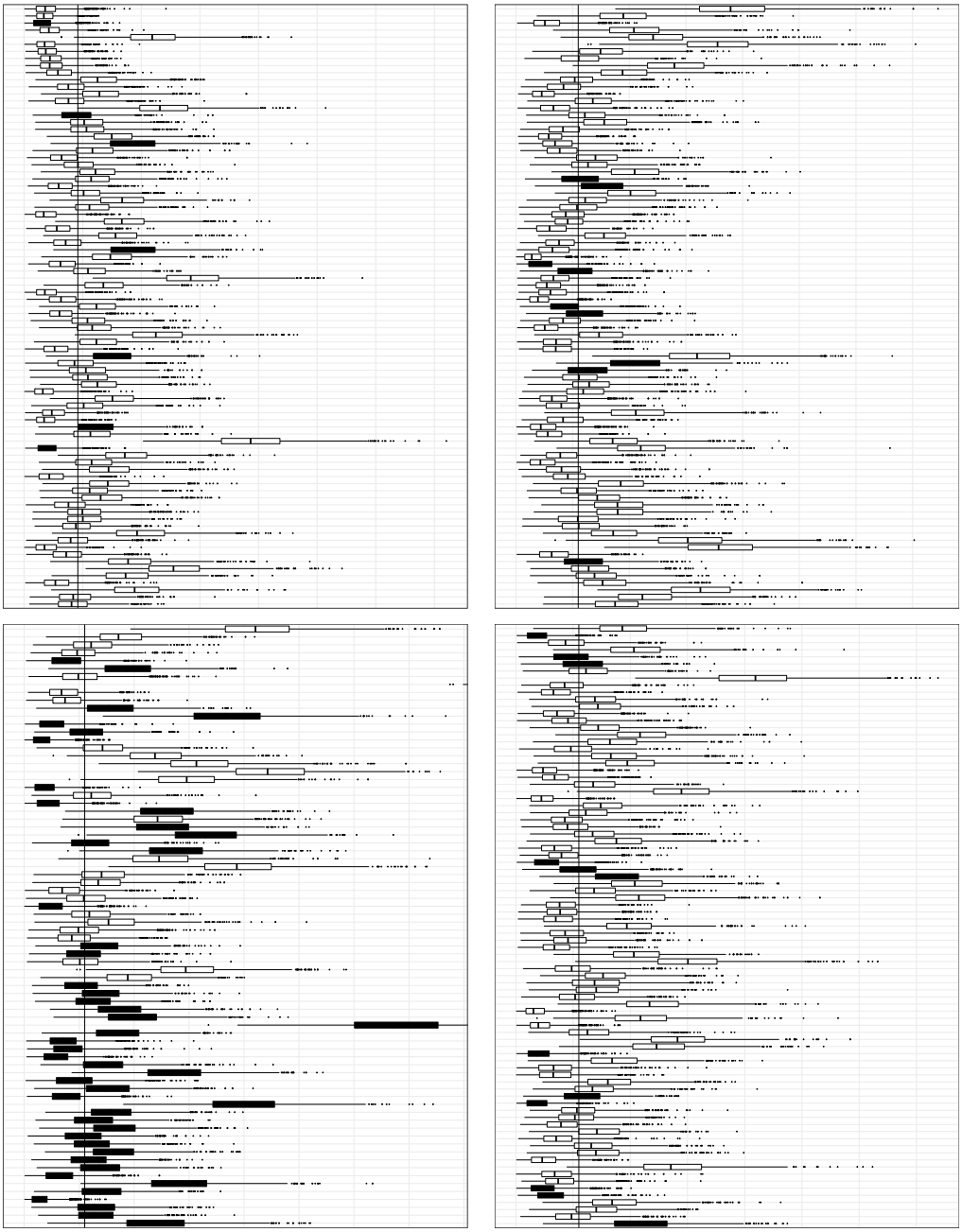

Fig. 5 From left to right and from top to bottom: boxplots of the posterior distribution for the mixing proportions ω_i, $i = 1,\ldots,331$. In white: catalogued high-energy γ-ray sources; in black: new candidate sources. The solid horizontal lines represents the intensity threshold used to further skim the new detections.

tial skimming. The opposite holds for the 43 initially identified and successively declassified sources which probably correspond to small areas of excess background intensity. This aspect represents one of the improvements of our model we are currently working on. The proposed parametric formulation for the diffuse background radiation is, in fact, only partially efficient. Using further data provided by the *Fermi* LAT collaboration we are currently developing a more precise background model.

Further future developments focus on both, theoretical and computational aspects. A first aspect regards the distribution used to describe how photons scatter around their emitting source. King's PSF used as approximation in (2) is currently being replaced by the point spread function proposed in [1]. On the computational side, we are replacing the Metropolis-Hastings step used to generate the values of μ with a more efficient Gibbs sampler. Last but not least, the heuristic approach adopted at the end of the previous section to qualify the newly detected sources needs be replaced by a formal procedure which accounts also for the available, here not used, information on the energy level of each detected photon.

Acknowledgments

We would like to thank Dr. Mauro Bernardi for the most helpful discussion of reversible jump MCMC and its extension. We furthermore express our vote of thanks to an anonymous Referee for her/his most careful reading, especially with respect to our model formulation and estimation. This research was supported by SID 2018 grant "Advanced statistical modelling for indexing celestial object" (BIRD185983) awarded by the Department of Statistical Sciences of the University of Padova. Financial support by Prof. Junhui Fan (grants n. NSFC11733001 and n. NSFC U1531245) is gratefully acknowledged.

References

1. Ackermann Markus *et al.* "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". *The Astrophysical Journal* **765** (2013): 54.
2. Ackermann Markus *et al.* "The spectrum and morphology of the *Fermi* bubbles". *The Astrophysical Journal* **793** (2014): 64.
3. Ackermann Markus *et al.* "The spectrum of isotropic diffuse gamma-ray emission between 100 MeV and 820 GeV". *The Astrophysical Journal* **799** (2015): 86.
4. Ackermann Markus *et al.* "3FHL: The third catalog of hard *Fermi* LAT sources". *The Astrophysical Journal* Supplement Series **232** (2017): 18.
5. Green, Peter J. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". *Biometrika* **82** (1995): 711.
6. Guglielmetti, Fabrizia, Rainer Fischer, and Volker Dose "Mixture modeling for background and sources separation in x-ray astronomical images". *AIP Conference Proceedings* **735** (2004): 111.

7. Hobson, Michael P., Graça Rocha, and Richard S. Savage "Bayesian source extraction". In *Bayesian Methods in Cosmology* (Hobson, Michael P. *et al.* Eds.), Cambridge University Press (2010): 167.
8. Jones, David E., Vinay L. Kashyap, and David A. van Dyk "Disentangling overlapping astronomical sources using spatial and spectral information". *The Astrophysical Journal* **808** (2015): 137.
9. King, Ivan "The structure of star clusters. I. An empirical density law." *The Astronomical Journal* **67** (1962): 471.
10. Kraft, Ralph P., David N. Burrows, and John A. Nousek "Determination of confidence limits for experiments with low numbers of counts". *The Astrophysical Journal* **374** (1991): 344.
11. Mattox, James R. *et al.* "The likelihood analysis of EGRET data". *The Astrophysical Journal* **461** (1996): 396.
12. Richardson, Sylvia, and Peter J. Green "On Bayesian analysis of mixtures with an unknown number of components" (with Discussion). *Journal of the Royal Statistical Society* Series B (Statistical Methodology) **59** (1997): 731.
13. Stein, Nathan M. *et al.* "Detecting unspecified structure in low-count images". *The Astrophysical Journal* **813** (2015): 66.
14. van Dyk, David A. *et al.* "Analysis of energy spectra with low photon counts via Bayesian posterior simulation". *The Astrophysical Journal* **548** (2001): 224.
15. Wiper, Michael, David Rios Insua, and Fabrizio Ruggeri "Mixtures of gamma distributions with applications". *Journal of Computational and Graphical Statistics* **10** (2001): 440.