Spatial field reconstruction with INLA:
Application to simulated galaxies

Majda Smole1*, João Rino-Silvestre2, Santiago González-Gaitán2 and Marko Stalevski1,3

1 Astronomical Observatory, Volgina 7, 11060 Belgrade, Serbia
2 CENTRA, Instituto Superior Técnico, Av. Rovisco Pais, Lisboa, 1049-001, Portugal
3 Sterrenkundig Observatorium, Universiteit Gent, Krijgslaan 281-S9, Gent, 9000, Belgium

November 7, 2022

ABSTRACT

Aims. Monte Carlo Radiative Transfer (MCRT) simulations are a powerful tool for understanding the role of dust in astrophysical systems and its influence on observations. However, due to the strong coupling of the radiation field and medium across the whole computational domain, the problem is non-local and non-linear and such simulations are computationally expensive in case of realistic 3D inhomogeneous dust distributions. We explore a novel technique for post-processing MCRT output to reduce the total computational run time by enhancing the output of computationally less expensive simulations of lower-quality.

Methods. We combine principal component analysis (PCA) and non-negative matrix factorization (NMF) as dimensionality reduction techniques together with Gaussian Markov random fields and the Integrated nested Laplace approximation (INLA), an approximate method for Bayesian inference, to detect and reconstruct the non-random spatial structure in the images of lower signal-to-noise or with missing data.

Results. We test our methodology using synthetic observations of a galaxy from the SKIRT Auriga project - a suite of high resolution magneto-hydrodynamic Milky Way-sized galaxies simulated in cosmological environment by 'zoom-in' technique. With this approach, we are able to reproduce high photon number reference images ~5 times faster with median residuals below ~20%.

Key words. Galaxies: general, dust, extinction, radiative transfer, Methods: numerical, Techniques: image processing

1. Introduction

Cosmological simulations provide a powerful tool for understanding galaxy formation, their evolution, and cosmic structure on large scales. State-of-the-art hydrodynamical simulations use subgrid physics for processes as radiative cooling, star formation, metal enrichment, and active galactic nuclei (AGN) feedback. Examples of such hydrodynamical cosmological simulations include Illustris (Vogelsberger et al. 2014a,b; Genel et al. 2014), EAGLE (Crain et al. 2015; Schaye et al. 2015) and IllustrisTNG projects (Weinberger et al. 2017; Pillepich et al. 2018).

Both observational and simulated data can be fully exploited only by their detailed comparison. Observations are required to test the results of numerical simulations. On the other hand, simulations can support the interpretation of observations and constrain predictions for the future instruments and observational missions. One of the challenges includes the possibility that the observed radiation may have been altered by the medium between the emitting source and the observer. Given the high abundance of dust in the interstellar medium, dust grains have considerable importance in shaping the observed signal, absorbing and scattering ultraviolet and optical radiation, and then re-emitting the absorbed energy at infrared wavelengths. Thus, the observed signal provides information about the emitting source and at the same time reveals the properties of the medium on its path.

Synthetic observations can be used to fill a gap between properties derived from numerical simulations and real observations. They are generated, for example, starting from a snapshot of an output of a hydrodynamical simulation and solving the radiative transfer (RT) problem by simulating how radiation propagates from an astrophysical source to the observer, optionally taking into account the degradation of the image due to transfer through the atmosphere and optics of the instrument. RT simulations play an important role in understanding the effects of dust on astrophysical observations. However, realistic dusty media are often complex and inhomogeneous, and calculating the propagation of radiation in such systems requires the implementation of special numerical methods, such as Monte Carlo radiative transfer (MCRT). These simulations use a large number of photon packages and emulate the relevant physical processes to mimic the propagation of real photons in dusty environments. We refer to Steinacker et al. (2013) for a review of numerical solution techniques for 3D dust RT problems, and Noebauer & Sim (2019) for a recent review of MCRT codes.

Different studies offer publicly available synthetic observational data for cosmological simulations, using the zoom-in technique. This technique involves a process when an additional zoom-in simulation with higher mass and/or time resolution is performed for a selected part of the original simulation. Initial conditions are taken directly from a cosmological simulation, with further implementation of subgrid physics of relevant processes that cannot be resolved directly in the simulations. Recent examples include studies by Kapoor et al. (2021) and Camps et al. (2022) that provide high resolution broadband images of galaxies in the Auriga (Grand et al. 2017) and ARTEMIS (Font et al. 2020) projects, respectively.

* email: msmole@aob.rs
However, even with state-of-the-art MCRT codes, simulating high resolution synthetic images proves to be a computationally challenging task (Steinacker et al. 2013). Memory requirements and simulation running times increase with the volume grid density and with the amount of photon packages used. In this work our goal is to develop a MCRT post-processing methodology able to achieve the quality of high photon number images, using low photon number images as an input. We explore the potential of Integrated nested Laplace approximation (INLA; Rue et al. 2009; Rue et al. 2017), an approximate method for Bayesian inference, coupled to Gaussian Markov Random Fields to detect and reconstruct the non-random spatial structure of simulated galaxies. The capability of this INLA method to reveal the underlying spatial structure of astronomical data has already been tested (González-Gaitán et al. 2019). Applied to integral field units (IFU) galaxy data, INLA is able to recover structures otherwise hidden, even with highly sparse spatial information. The INLA method has been successfully applied to various research fields, few examples include identifying the climatic drivers of honey bee disease in England and Wales (Rowland et al. 2021), forecasting deforestation in the Brazilian Amazon (Jaffé et al. 2021) and mapping air pollution (Gong et al. 2021).

Alongside the work presented here, we developed an alternative post-processing methodology for MCRT codes, combining an auto-encoder neural network and INLA (Rino-Silvestre et al. 2022). Here we offer another methodology combining principal component analysis (PCA) and non-negative matrix factorization (NMF) as dimensionality reduction techniques together with INLA, and test the performance using high resolution broadband images of galaxies from the Auriga project (Grand et al. 2017).

In section 2 we introduce the employed techniques and codes, together with datasets used in this work, and describe their implementation in our methodology. We apply our method to a galaxy from the Auriga project. Results are presented and discussed in section 3. In section 4 we summarise the main findings of this work and draw conclusions.

2. Methods

2.1. Simulated Data

2.1.1. SKIRT code

The SKIRT code has been widely used in studying dust properties in spiral galaxies (Baes et al. 2010; De Loore et al. 2012a,b; De Geyter et al. 2015), sub-millimetre galaxies (Dudzevičiūtė et al. 2020), dust-driven properties of high redshift galaxies (Behrens et al. 2018; Vijayan et al. 2022), galaxy evolution (Whitney et al. 2021; Deelley et al. 2021; Zanella et al. 2021), dust attenuation law in high redshift quasars (Di Mascia et al. 2021) and investigating AGN structure (Stalevski et al. 2016; Stalevski et al. 2017; Stalevski et al. 2019), to name a few applications.

2.1.2. Auriga galaxies

The Auriga project provides a set of cosmological magnetohydrodynamic zoom-in simulations of Milky Way (MW) type galaxies (Grand et al. 2017). Simulations were performed using the hydrodynamic moving mesh code AREPO (Springel 2010), with a complete treatment of galaxy formation and evolution process such as gas cooling and heating, star formation, black hole and stellar feedback, etc (Vogelsberger et al. 2013; Genel et al. 2014).

The SKIRT Auriga project offers high resolution synthetic images of 30 simulated MW type galaxies from the Auriga project (Kapoor et al. 2021). This synthetic observational data, obtained with SKIRT, represents a sample of galaxies in the broad wavelength range from UV to sub-mm. For each galaxy, sets of star particles and gas cells are extracted from the Auriga simulation snapshots. Star particles with an age above 10 Myr are assigned a spectral energy distribution (SED) using Bruzual & Charlot (2003) stellar template library, based on metallicity and age. Star particles with an age below 10 Myr are assumed to be surrounded by dust in star-forming regions, and are assigned SED from the MAPPINGS III (Groves et al. 2008) template library. The initial dust distribution is calculated based on the metallicity and gas density extracted from the Auriga snapshots, using the THEMIS dust model (Jones et al. 2017). We refer to Kapoor et al. (2021) for detailed recipes for the transformation from the set of particles extracted from Auriga snapshot to the set of particles used in SKIRT simulation.

Synthetic images from the SKIRT Auriga project are publicly available in the flexible image transport system (FITTS) format composed of 2D spatial distributions at 50 different wavelength bins. Three different inclinations of the galaxy i.e. the angle between the angular momentum vector and the direction towards the observer comprise the dataset of each galaxy: `face-on' (i = 0°), 'edge-on' (i = 90°) and 'intermediate' (i = 115°). 'Intermediate' viewing angle corresponds to the original orientation of the galaxy in the simulation box.

2.2. INLA

The Integrated nested Laplace approximations (INLA; Rue et al. 2009) is a computational method for approximate Bayesian inference of Latent Gaussian fields. Bayesian inference is a process of deducing a probability distribution from data using Bayes theorem. Bayesian inference (equation 1) calculates a posterior distribution \( \pi(\theta | y) \), for the given model likelihood \( \pi(y | \theta) \), prior distribution \( \pi(\theta) \) and marginal likelihood \( \pi = \int \pi(y | \theta) \pi(\theta) \, d\theta \).

\[
\pi(\theta | y) = \frac{\pi(y | \theta) \pi(\theta)}{\pi(y)}.
\]

Most techniques for calculating posterior distribution rely on Markov chain Monte Carlo (MCMC) methods (Collins et al. 2021).

1. http://www.skirt.ugent.be

2. For a full list, see https://skirt.ugent.be/root/releases/publications_all_date.html

3. https://www.mpa-garching.mpg.de/auriga/

4. https://www.auriga.ugent.be/Home---SKIRT-Auriga-Project.html
In this class of sampling-based numerical methods, the posterior distribution is obtained after many iterations, which is often computationally expensive. INLA provides a novel approach for faster Bayesian inference. While MCMC methods draw a sample from the joint posterior distribution, the Laplace approach is a method that approximates posterior distributions of the model parameters to Gaussians, which is computationally more effective.

Within the INLA framework, the posterior distribution of the latent Gaussian variables \( x \) and hyperparameters of the model \( \theta \) is:

\[
\pi(x, \theta | y) = \pi(y | x, \theta) \pi(x, \theta) \pi(\theta) \propto \pi(y | x, \theta) \pi(x, \theta),
\]

where \( y = (y_1, \ldots, y_n) \) represents set of observations i.e. a pixel matrix. Every pixel is treated with a latent Gaussian effect, so that each \( x_i \) (which is just mean value \( \mu_i \) and standard deviation \( \sigma_i \) of a Gaussian for a pixel \( y_i \)) corresponds to observation \( y_i \), where \( i \in [1, \ldots, n] \). The observations are conditionally independent given the latent effect \( x \) and the hyperparameters \( \theta \), so the model likelihood is:

\[
\pi(y | x, \theta) = \prod_i \pi(y_i | x_i, \theta).
\]

The joint distribution of the latent effects and the hyperparameters, \( \pi(x, \theta) \), can be written as \( \pi(x | \theta) \pi(\theta) \), where \( \pi(x | \theta) \) represents the prior distribution of hyperparameters \( \theta \). It is assumed that the spatial information can be treated as a discrete sampling of an underlying continuous spatial field, a latent Gaussian Markov random field (GMRF), that takes into account the spatial correlations, and whose hyperparameters are inferred in the process. For a GMRF, the posterior distribution of the latent effects is:

\[
\pi(x | \theta) \propto |\Omega(\theta)|^{1/2} \exp[-\frac{1}{2} x^T\Omega(\theta)x],
\]

where \( \Omega(\theta) \) represents a precision matrix, or inverse of a covariance matrix, which depends on a vector of hyperparameters \( \theta \). This kernel matrix is what actually treats the spatial correlation between neighbouring pixels. Using equation 2 the joint posterior distribution of the latent effects and hyperparameters can be written as:

\[
\pi(x, \theta | y) \propto \pi(\theta) |\Omega(\theta)|^{1/2} \exp[-\frac{1}{2} x^T\Omega(\theta)x] \prod_i \pi(y_i | x_i, \theta) \pi(\theta) \propto \pi(\theta) |\Omega(\theta)|^{1/2} \exp[-\frac{1}{2} x^T\Omega(\theta)x + \sum_i \log(\pi(y_i | x_i, \theta))].
\]

Instead of obtaining the exact posterior distribution from equation 5, INLA approximates the posterior marginals of the latent effects and hyperparameters. Thus, the key feature of INLA methodology is to use appropriate approximations for the following integrals:

\[
\pi(x_i | y) = \int \pi(x_i | \theta, y) \pi(\theta | y) d\theta,
\]

\[
\pi(\theta | y) = \int \pi(\theta | y) d\theta_{-j},
\]

where \( \theta_{-j} \) is a vector of hyperparameters \( \theta \) without element \( \theta_j \). INLA constructs nested approximations:

\[
\tilde{\pi}(x_i | y) = \int \tilde{\pi}(x_i | \theta, y) \tilde{\pi}(\theta | y) d\theta
\]

\[
\tilde{\pi}(\theta | y) = \int \tilde{\pi}(\theta | y) d\theta_{-j},
\]

where \( \tilde{\pi}(\cdot | \cdot) \) is an approximated posterior density.

Using the Laplace approximation, the posterior marginals of hyperparameters \( \pi(\theta | y) \) at a specific value \( \theta = \theta_j \) can be written as:

\[
\tilde{\pi}(\theta_j | y) \propto \pi(x, \theta_j | y) \propto \pi(y | x, \theta_j) \pi(x | \theta_j) \pi(\theta_j) \propto \frac{\pi(y | x, \theta_j) \pi(x | \theta_j) \pi(\theta_j)}{\pi_G(x | \theta_j, y)} |x-x^{\star}(\theta_j)|,
\]

\[
\tilde{\pi}_G(x | \theta, y) \propto \exp[-\frac{1}{2} x^T\Omega(\theta)x + \sum_i g_i(x_i)]
\]

where \( \tilde{\pi}_G(x | \theta, y) \) is the Gaussian approximation to the full conditional of \( x \) and \( x^{\star}(\theta_j) \) is the mode of the full conditional \( x \) for given \( \theta \). (Rue et al., 2009).

The posterior marginals of the latent effects are numerically integrated as following:

\[
\tilde{\pi}(x_i | y) \propto \sum_{j} \tilde{\pi}(x_i | \theta_j, y)\tilde{\pi}(\theta_j | y)\Delta_j,
\]

where \( \Delta_j \) represents integration weight. However, a good approximation for \( \pi(x_i | \theta, y) \) is required and INLA offers three different options: Gaussian approximation, Laplace approximation and simplified Laplace approximation (Rue et al., 2009). In this work we used the simplified Laplace approximation, which is the default one and represents a compromise between the accuracy of the Laplace approximation and the reduced computational cost achieved with the Gaussian approximation.

As described above INLA uses spatial correlations between data points to reconstruct missing and/or noisy data. We refer to Rue et al., 2009 and Gómez-Rubio, 2021 for more details on the mathematical background of INLA and the methods it employs.

INLA is available as R-INLA package\(^5\) designed for modeling spatial data (Rue et al., 2017). In conjunction with INLA, we use the PARDISO package (Van Niekerk et al., 2021) which represents a high-performance software for solving sparse symmetry that arises in the R-INLA approach to Bayesian inference, and thus reduces INLA’s computational time. Since synthetic observations simulated with the SKIRT code represent non-random spatial structures, we explore the potential of INLA method as a tool for enhancing the MCRT images, and thus, optimizing the total computational run time invested in obtaining high quality images.

2.3. Dimensionality reduction

In order to achieve additional reduction in computational time, we implement dimensionality reduction techniques in our methodology for SKIRT code post-processing. Dimensionality reduction methods diminish the number of attributes in a original dataset while attempting to preserve as much of the variation as possible.

\(^5\) \url{http://www.r-inla.org/}
2.3.1. PCA

Principal component analysis (PCA) is an unsupervised method that analyses the feature space of the training data and creates orthogonal vectors (linear combinations of the initial variables) whose direction indicates the most variability (for derivations see Pearson 1901; Hotelling 1933, for a modern review see e.g. Jolliffe & Cadima 2016). These new vectors in the transformed data set are called eigenvectors, or principal components, while eigenvalues represent the coefficients attached to eigenvectors, which give the relative amount of variance carried in each principal component. Principal components are uncorrelated and most of the information is compressed into the first components. Even though the transformed dataset has the same number of dimensions as the original dataset, by discarding components with low information, dimensionality reduction is achieved.

Presently, within the astronomy and cosmology fields of research, PCA is mostly used in algorithms and/or pipelines that implement different combinations of learning methods (Krone-Martins & Moutinho 2014; Logan & Fotopoulou 2020). In this work we use \texttt{prcomp}\textsuperscript{6} function.

2.3.2. NMF

Non-negative matrix factorization (NMF) is another unsupervised dimensionality reduction method widely used in various fields (e.g. Lee & Seung 1999; Tandon & Sra 2010). Similar to PCA, NMF decomposes the feature space of the training dataset into the product of two matrices with smaller ranks. One matrix represents the NMF features that are linear combinations of the original dataset features, while the second NMF matrix contains coefficients i.e. weights associated with NMF features. NMF is initiated by attributing guess values to the elements of the reduced rank matrices which are then iteratively updated until both NMF matrices are stable and their product approaches the original dataset features, while the second NMF matrix contains coefficients for less prominent features and noise, and can be discarded by sampling the input map data, instead of using complete dimensions: \(2150 \times 2150 \times 50\) for 'face-on'; \(2400 \times 900 \times 50\) for 'edge-on' and \(1800 \times 2200 \times 50\) for 'intermediate' cube.

Full data cube reconstruction requires 50 individual spatial reconstructions, one for each of the wavelength bins. Such reconstructions are labelled 'pure INLA' in the following text. We employ PCA and NMF techniques to reduce the number of dimensions, i.e. number of individual spatial maps to be reconstructed with INLA. Both methods transform the original dataset in the spectral dimension, such that these spectra are described with a new orthogonal basis (in order of decreasing variability) instead of the 50 wavelengths. In this transformed space, the first components will carry most of the information, while the rest are responsible for less prominent features and noise, and can be discarded in order to trade some of the precision for efficiency. After performing PCA/NMF transformation on the initial variables (wavelength bins), these are replaced by a reduced number of principal components, whose maps are used as input to INLA. INLA then reconstructs the spatial maps of the coefficients attached to our selected set of principal components, instead of flux densities from the original data cube. Reconstructed spatial maps are then multiplied by the transpose of the principal components, resulting in the reconstructed data cube with the original dimensions, labelled 'PCA/NMF+INLA' reconstructions.

In Appendix A we show spatial maps of PCA/NMF components before and after INLA reconstructions. In Appendix B we compare PCA/NMF reconstructions before and after applying INLA. Pure PCA/NMF reconstructions show higher residuals compared to PCA/NMF+INLA reconstructions.

Additionally, the INLA spatial reconstruction time can be reduced by sampling the input map data, instead of using complete spatial information. INLA shows optimal performance when applied to sparse data, the execution time is notably reduced while the quality of reconstructions is influenced only to a lesser degree (González-Gaitán et al. 2019). This is further studied in section 3.1, where we perform INLA reconstructions using different sampling percentages. We also test different number of

---

\textsuperscript{6} https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp

\textsuperscript{7} https://www.rdocumentation.org/packages/NMF/versions/0.24.0/topics/nmf

\textsuperscript{8} https://www.auriga.ugent.be/DataZenodoRedirect.html

---

\textsuperscript{8} \textsuperscript{8}
PCA/NMF components in order to find an optimal compromise between the quality of reconstructed maps and the required running times.

3. Results and discussion

3.1. Field cuts optimization

We test the optimal pixel sampling percentage using the 600x600x50 cut of ‘face-on’ cube. The quality of INLA reconstructions, compared to HPN reference images, is quantified by the normalized residuals, calculated as:

\[
\text{Residuals} \, (\%) = \left| \frac{X' - X}{X} \right| \times 100\%
\]

where \(X'\) and \(X\) refer to INLA reconstruction and HPN reference images, respectively.

Figure 2 presents the median of normalized residuals of pure INLA reconstructions as a function of the sampling percentage used (5%, 10%, 25%, 50% and 100%), for the 600x600x50 cut. The total running times for INLA reconstructions are represented with different colours. Horizontal lines refer to the median residuals of the LPN inputs in relation to the HPN reference image without applying INLA, for \(3 \times 10^8\) (circles) and \(3 \times 10^9\) (triangles) photon number simulations.

Increasing sample size improves the quality of reconstructions at the cost of a higher computational time. In terms of balance between residuals and the running times, INLA shows the optimal performances when applied to sparse data. Figure 2 shows that using more than 25% of the spatial data as INLA input significantly increases the running times with only a lesser (1 – 2%) improvement in the quality of reconstructions. In the following sections, we present reconstructions using 10% and 25% of the spatial information.

Next, we explore the optimal number of PCA/NMF components. PCA and NMF analyses are performed using 100% of data within 600x600x50 cut, followed by sampling a percentage (10% or 25%) of spatial maps of PCA/NMF coefficients as an input to INLA. Figure 3 represents the median of normalized residuals as a function of the number of PCA/NMF components used for the reconstruction. The figure shows a clear improvement in reconstructions when more than two components are used, followed by a more gradual change. This behaviour suggests that the first components describe the overall structure and major features, while others represent e.g. diffuse emission and low flux regions. In the case of PCA, a small decrease in residuals is noticed up to the tenth component while using more than ten PCA/NMF components does not improve the results. However, NMF does not show such a clear trend and using between
4 and 12 components results in only $\sim 0.5\%$ change in median residuals.

Application of our methodology to this field cut shows that using dimensionality reduction techniques prior to INLA drastically reduces the total computation time of the proposed post-processing technique without loss of precision. Additionally, even the quality of reconstructions is improved compared to pure INLA (horizontal lines in Figure 3), if more than two components are used.

Next, we move to the full field processing. Reconstructions of full cube images shown in the following section are performed using 10 PCA/NMF components, with data sampling of 10% and 25%.

### 3.2. Full field processing

In this section, we present the results for full field reconstructions using pure INLA and PCA/NMF+INLA techniques. We test our method using LPN SKIRT simulations ($3 \times 10^8$ and $3 \times 10^9$ photon packages) with different inclination angles: ‘face-on’, ‘edge-on’ and ‘intermediate’. INLA reconstructions are performed using 10% or 25% of the available spatial information. All figures shown in this section refer to reconstructions using 25% of LPN ($3 \times 10^9$) input data, while the complete statistics of full field processing are summarised in Table 1 and Table 2.

#### 3.2.1. Spectral reconstructions

In this section, we explore how the reconstructed images follow the expected SED of the HPN reference.

First, we compare the spatially integrated SEDs, obtained from integrating all flux at each wavelength bin. Figure 4 shows integrated SEDs for HPN reference ($\odot$), pure INLA ($\bigcirc$), PCA+INLA ($\bigstar$) and NMF+INLA ($\bigtriangleright$) (upper panels) with the associated residuals (lower panels), for ‘face-on’ (left), ‘edge-on’ (middle) and ‘intermediate’ (right) cubes. For each of the employed techniques integrated SEDs provide a good match to the HPN reference, throughout the whole wavelength range. PCA+INLA reconstructions display the highest deviations from the expected SEDs at wavelengths around 100 $\mu$m, up to $\sim 1.5\%$ for ‘face-on’ and ‘edge-on’ cubes, i.e. $\sim 3\%$ for the intermediate cube.

Table 1 summarises the statistics for all of the employed realizations: ‘face-on’, ‘edge-on’ and ‘intermediate’ cubes, simulated using $3 \times 10^8$ and $3 \times 10^9$ photon packages. Median of the normalized residuals percentages are shown for pure INLA and PCA/NMF+INLA integrated SEDs, compared to HPN reference. Predicted integrated SEDs are in excellent agreement with the expectations, with the median of the normalized residuals $\lesssim 0.1\%$ for pure INLA and NMF+INLA reconstructions, and $\lesssim 0.3\%$ for PCA+INLA.

Next, we explore how the quality of our predictions changes at various regions throughout the galaxy plane. We inspect SEDs for individual spaxels whose spatial positions are marked in Figure 4. The chosen spaxels occupy different regions of galaxy morphology, such as central parts, strong spiral arms, low flux density regions between spiral arms and galaxy outskirts. Figures 5, 6 and 7 show a single spaxel SEDs for ‘face-on’, ‘edge-on’ and ‘intermediate’ cubes, respectively, sampling 25% of $3 \times 10^9$ photon number realization.

Both pure INLA (green) and PCA/NMF+INLA (blue and cyan lines) reconstructions are in general in good agreement with HPN reference SEDs (black lines) for most pixel positions. The quality of reconstructions is correlated with the spaxel position and the flux density. At high flux density regions, such as central parts and prominent spiral arms, LPN input SEDs (red lines) closely follow the HPN reference, and each of the employed techniques provides accurate predictions (spaxels at positions 1 and 2). At lower flux densities along spiral arms (spaxels 3 and 4), LPN input becomes noisy, however the quality of our reconstructions is not affected. However, reconstruction of faint, outermost parts of the galaxy (spaxels 5 and 6) becomes challenging given the LPN input SEDs with both noisy and missing

---

**Table 1: Median of the normalized residuals for integrated SEDs from pure INLA and PCA/NMF+INLA reconstructions, using different sampling percentages (10% and 25%) of $3 \times 10^8$ and $3 \times 10^9$ photon number cubes.**

| Angle      | Photon Number Input | Sample | Median of res. (%) |
|------------|---------------------|--------|--------------------|
|            | pure INLA | PCA+INLA | NMF+INLA | pure INLA | PCA+INLA | NMF+INLA |
|            | $3 \times 10^7$      | 25%    | 0.02   | 0.15   | 0.03 |
| face-on    | $3 \times 10^8$      | 10%    | 0.04   | 0.15   | 0.07 |
|            | $3 \times 10^9$      | 25%    | 0.06   | 0.32   | 0.13 |
| edge-on    | $3 \times 10^8$      | 25%    | 0.07   | 0.30   | 0.11 |
|            | $3 \times 10^9$      | 25%    | 0.03   | 0.07   | 0.03 |
|            | $3 \times 10^8$      | 25%    | 0.10   | 0.13   | 0.06 |
|            | $3 \times 10^9$      | 25%    | 0.07   | 0.23   | 0.05 |
|            | $3 \times 10^8$      | 25%    | 0.17   | 0.21   | 0.06 |
|            | $3 \times 10^9$      | 25%    | 0.03   | 0.22   | 0.02 |
|            | $3 \times 10^8$      | 10%    | 0.04   | 0.25   | 0.03 |
| intermediate| $3 \times 10^9$    | 25%    | 0.05   | 0.39   | 0.05 |
|            |                     | 10%    | 0.08   | 0.35   | 0.08 |

9 Spaxel refers to spectral pixel, with a spectrum associated to each spatial pixel.
3.2.2. Spatial reconstructions

Figures in this section show spatial distribution of randomly sampling 25% of the LPN input cube spaxels, together with pure INLA and PCA/NMF+INLA reconstructions (upper panels). Each of these cubes is compared to the HPN reference and normalized residuals are shown on the lower panels. Figures [8][9] show spatial reconstructions of the ‘face-on’ cube at four wavelength bins: 0.23 μm, 7.88 μm, 100.08 μm and 515.36 μm. Figure [10] and Figure [11] show ‘edge-on’ and ‘intermediate cubes’ at wavelength bin 7.88 μm.

By using a sample of LPN input cube our methods are able to recover underlying spatial information and reveal structures seen on HPN reference images (Figure [1]). The quality of reconstructions is quantified by the median of normalized residuals (%), calculated for each pixel at the given wavelength bin, and shown above each residual reconstruction image. The pure INLA results display generally the lowest residuals, however this technique’s extensive running time disqualifies it as a viable route to emulate MCRT codes. The PCA/NMF+INLA reconstructions have similar residuals, with ≤ 25% of the running time required for HPN reference (Table [2]). However, at certain wavelength bins, the PCA+INLA technique fails in the reconstruction of ~10% of spatial information, positioned at regions with the lowest flux density. Those regions with non-physical negative flux densities can be seen in Figures [10][12] and [13] as white pixels. The problem with non-physical reconstructions induced by PCA analysis can be avoided using NMF instead, since NMF enforces only positive elements in both original and decomposed matrices. Using NMF+INLA method the problematic regions are reconstructed, but with typically higher residuals compared to pure INLA.

Figure [14] shows the median of normalized residuals as a function of wavelength, for ‘face-on’ (left), ‘edge-on’ (middle) and ‘intermediate’ (right) cubes. Although similar to Figure [4] which shows normalized residuals for integrated SEDs, here we calculate residuals for each pixel at a given wavelength bin separately and then calculate their median value. Pixels with negative values are converted to zero, resulting in normalized residuals of 100% (equation [13]). Overall, all of the employed methods (pure INLA and PCA/NMF+INLA) results with significantly lower residuals compared to LPN input. Pure INLA (green) is able to recover spatial information over the whole range of wavelength bins, with the median of normalized residuals ≤ 20 % for ‘face-on’ cube, and ≤ 30 % for ‘edge-on’ and ‘intermediate’ cubes. On the other side, both PCA/NMF+INLA reconstructions have higher residuals at wavelengths ~ 20 – 150 μm. At these wavelengths both LPN input and HPN reference images lose dif-
Fig. 5: Single spaxel SEDs for HPN reference (black) and LPN input (red) ‘face-on’ cubes, together with pure INLA (green) and PCA/NMF+INLA (blue and cyan) reconstructions. Spatial positions of the spaxels represented here are shown in Fig. 1.

Fig. 6: Same as Fig. 5 for ‘edge-on’ cube.
fuse emission of low intensity that fills regions between spiral structure, otherwise present at other wavelength bins. PCA and NMF components carry information about this diffuse emission and will always try to reconstruct it, resulting in higher residuals at wavelengths where the emission is attenuated in the original images. However, statistics for these wavelength bins is governed by outskirt regions of the galaxy where PCA+INLA reconstructions fail, while NMF+INLA overestimates the flux density. When applied to a field cut that includes galaxy central region, both PCA/NMF+INLA method results in lower residuals compared to pure INLA technique (Figure 3). Thus, masking out the outskirt regions prior INLA would improve the reconstructions.

3.3. Summary and discussion

Table 2 summarises statistics for the spatial reconstructions of 'face-on', 'edge-on' and 'intermediate' cubes, using pure INLA and PCA/NMF+INLA implementations. We present results for two different LPN input images ($3 \times 10^8$ and $3 \times 10^9$ photon number), using randomly sampled 10% or 25% of the available spatial information. Both LPN input images and the reconstructions are compared to HPN reference. The median of the normalized residuals are shown for each cube, together with the total running time required for LPN SKIRT simulations and INLA reconstructions. Running times presented in Table 2 is the following:

$$t(\%) = \frac{t_{\text{LPN SKIRT}} + t_{\text{INLA}}}{t_{\text{HPN SKIRT}}} \times 100,$$

where $t_{\text{INLA}}$ refers either to pure INLA reconstructions (50 images at 50 wavelength bins) or to the total time for PCA/NMF analysis and PCA/NMF+INLA reconstructions.

For each realization, residuals for our reconstructions are significantly lower compared to LPN input residuals, proving that each of the employed methods is able to recover spatial structure using only 10% or 25% of the LPN input. Although pure INLA reconstructions result in the lowest residuals, this methodology does not provide a significant speedup. Using a sample size of 10% requires $\gtrsim 70\%$ of HPN reference SKIRT running time while using 25% of 'face-on' cube exceeds it. However, employing dimensionality reduction techniques successfully reduces the running times up to $\sim 10\%$ of HPN reference’s with similar residuals.

Additionally, the quality of our reconstructions scales with LPN input residuals, thus the 'face-on' cube has overall lower residuals compared to other tilt angles. Depending on the LPN input (3 $\times$ 10^8 or 3 $\times$ 10^9 photon number), and the sample size (10% or 25%) residuals of PCA/NMF+INLA reconstructions are in range $\sim 10$-20% for 'face-on' cube, and $\sim 15$-30% for 'edge-on' and 'intermediate' cubes. Naturally, the lowest residuals are achieved when using 25% of 3 $\times$ 10^9 photon number LPN input cube, which requires $\sim 25$-40% of the HPN reference running time. At the other end, by sampling 10% of the 3 $\times$ 10^8 photon number LPN input cube, the running times are reduced to $\sim 7$-20% of the HPN reference. The 'edge-on' cube typically has shorter running times due to a smaller size field (2401$\times$901$\times$50), compared to the 'face-on' cube (2151$\times$2151$\times$50).

When applied to a field cut, PCA+INLA method outperforms pure INLA, both by lower residuals and significantly shorter running times (Figure 3), since this cut does not include...
Fig. 8: Spatial distribution of LPN input cube and pure INLA, PCA/NMF+INLA reconstructions (upper panels) with the associated residuals (bottom panels) at wavelength bin 0.23 $\mu$m. Sample size: 25%, ‘face-on’ cube. Colour legends are given in logarithmic scale.

Fig. 9: Same as Fig. 8 at wavelength bin 7.88 $\mu$m.

galaxy outskirts. This proves that performing a careful prior segmentation of the galaxy or astrophysical source will aid best in our post-processing technique. Furthermore, PCA analysis has advantages over NMF: PCA analysis is faster and the number of PCA components does not need to be decided in advance. On the other hand, NMF provides always physically positive reconstructions, particularly useful at low flux regions like the outskirts of the image.

4. Conclusions
In this work, we use PCA and NMF dimensionality reduction techniques together with approximate Bayesian inference of
continuous Gaussian random fields with INLA for SKIRT simulations post-processing. We test our methodology using three images of Au-16 SKIRT Auriga galaxy, simulated with different tilt angles: ‘face-on’, ‘edge-on’, and ‘intermediate’. These HPN reference images (3 × 10^8 photon packages) served as the ‘ground truth’ to which we compared the performance of our method applied to LPN input images (3 × 10^7 or 3 × 10^9 photon packages).

Our results showed that spatially integrated SED closely follow the reference SED for each of the employed methods, with the median of the normalized residuals typically ≲ 0.3%. Spatial modelling suggests that the quality of our reconstructions changes with the position along the galaxy plane, with faint galaxy outskirts having the highest residuals.

Depending on the number of photon packages, the sample size and the desired quality of reconstruction, our method offers time-efficient reconstructions with spatial residuals ∼ 20 – 30%.
Fig. 12: Edge-on cube at wavelength bin 7.88 µm

Table 2: Median of the normalized residuals and the running times for each realization, compared to HPN reference.

| angle          | LPN input | pure INLA | PCA+INLA | NMF+INLA |
|----------------|-----------|-----------|----------|----------|
|                | photon    | median    | median   | median   |
|                | number (%)| sample (%)| median (%)| time (%) |
| face-on        | $3 \times 10^9$ | 17.68 | 9.97 | 10.68 | 11.35 |
|                |           | 25%      | 110    | 26     | 40    |
|                |           | 10%      | 11.61 | 12.21  | 12.79 |
|                |           |          | 12.15  | 29     | 35    |
|                | $3 \times 10^8$ | 60.97 | 15.57 | 15.46 | 17.65 |
|                |           | 25%      | 103    | 15     | 24    |
|                |           | 10%      | 19.22 | 18.20  | 20.73 |
|                |           |          | 20.00  | 29     | 19    |
| edge-on        | $3 \times 10^9$ | 30.23 | 13.36 | 20.62 | 15.85 |
|                |           | 25%      | 58     | 18     | 24    |
|                |           | 10%      | 19.22 | 22.54  | 17.5  |
|                |           |          | 22.63  | 43     | 20    |
|                | $3 \times 10^8$ | 96.88 | 23.42 | 31.18 | 24.92 |
|                |           | 25%      | 47     | 10     | 12    |
|                |           | 10%      | 29.38 | 35.28  | 19    |
|                |           |          | 27.47  | 28     | 8     |
| intermediate   | $3 \times 10^9$ | 24.91 | 12.10 | 12.63 | 12.73 |
|                |           | 25%      | 97     | 24     | 33    |
|                |           | 10%      | 14.57 | 14.41  | 21    |
|                |           |          | 14.51  | 76     | 26    |
|                | $3 \times 10^8$ | 86.13 | 20.03 | 19.27 | 23     |
|                |           | 25%      | 89     | 22     | 22    |
|                |           | 10%      | 25.10 | 22.63  | 13    |
|                |           |          | 25.89  | 66     | 16    |

requiring $\sim 7 - 20\%$ of the HPN reference running time. Higher quality reconstructions can be achieved by sampling 25% of $3 \times 10^9$ LPN input image, resulting in residuals $\sim 10 - 20\%$ within $\sim 20 - 40\%$ of the HPN reference running time.

In order to improve the quality of reconstructions, different sampling methods based on information on the spatial densities, and different pre-INLA data pre-processing might be worth exploring. Even though our results are not sensitive to the choice of sampling method (random or uniform), more complex and less homogeneous spatial maps might require non-uniform sampling.

Being able to efficiently perform large amounts of numerical simulations with varying physical characteristics is essential to compare to real observations in a quantitative way. Even if the accuracy is not immediately as good as that of a HPN simulation, such explorations with optimized LPN simulations can help narrowing down the parameters to then run a full HPN simulation. This work represents a crucial step in this direction.

Acknowledgements. We acknowledge the support by the programme of scientific and technological cooperation between the government of the Republic of Serbia and the government of the Republic of Portugal, (grant No. 337-00-00227/2019-09/53 and FCT 5581 DRI, Sérvia 2020/21). M.S. and M.S. acknowledge support by the Science Fund of the Republic of Serbia, PROMIS 6069/16, BOWIE and by the Ministry of Education, Science and Technological Development of the Republic of Serbia through the contract no. 451-03-9/2022-14/200002. J. R.-S. is funded by Fundação para a Ciência e a Tecnologia (PD/BD/150487/2019).
Fig. 13: ‘Intermediate’ cube at wavelength bin 7.88 µm

Fig. 14: Median of normalized residuals as a function of wavelength for ‘face-on’ (left), ‘edge-on’ (middle) and ‘intermediate’ (right) cubes.

via the International Doctorate Network in Particle Physics, Astrophysics and Cosmology. S.G.G and J.R.-S. acknowledge the support of FCT under Project CRISP PTDC/FIS-AST-31546/2017.

References
Baes, M. & Camps, P. 2015, Astronomy and Computing, 12, 33
Baes, M., Fritz, J., Gadotti, D. A., et al. 2010, A&A, 518, L39
Behrens, C., Pallottini, A., Ferrara, A., Gallerani, S., & Vallini, L. 2018, MNRAS, 477, 552
Boulas, A. et al. 2021, Astronomy & Astrophysics, 647, A105
Bruzual, G. &Charlot, S. 2003, MNRAS, 344, 1000
Camps, P. & Baes, M. 2015, Astronomy and Computing, 9, 20
Camps, P., Kapoor, A. U., Trcka, A., et al. 2022, MNRAS, 512, 2728
Collins, J. D., Hart, G. C., Haselman, T. K., & Kennedy, B. 1974, AIAA Journal, 12, 185
Crain, R. A., Schaye, J., Bower, R. G., et al. 2015, MNRAS, 450, 1937
De Geyter, G., Baes, M., De Looze, I., et al. 2015, MNRAS, 451, 1728
De Looze, I, Baes, M., Bendo, G. J., et al. 2012a, MNRAS, 427, 2797
De Looze, I., Baes, M., Fritz, J., & Verstappen, J. 2012b, MNRAS, 419, 895
De Lee, S., Drinkwater, M. J., Sweet, S. M., et al. 2021, MNRAS, 508, 895
Di Mascia, F., Gallerani, S., Ferrara, A., et al. 2021, MNRAS, 506, 3946
Dudzevičiūtė, U., Smail, I., Swinbank, A. M., et al. 2020, MNRAS, 494, 3828
Font, A. S., McCarthy, I. G., Poole-Mckenzie, R., et al. 2020, MNRAS, 498, 1765
Genel, S., Vogelsberger, M., Springel, V., et al. 2014, MNRAS, 445, 175
Gong, W., Reich, B. J., & Chang, H. H. 2021, Environmental Research Communications, 3, 101002
González-Gaitán, S., de Souza, R. S., Krone-Martins, A., et al. 2019, MNRAS, 482, 3880
Grand, R. J. J., Gómez, F. A., Marinacci, F., et al. 2017, MNRAS, 467, 179
Groves, B., Dopsa, M. A., Sutherland, R. S., et al. 2008, ApJS, 176, 438
Gómez-Rubio, V. 2021, Bayesian Inference with INLA (Chapman & Hall/CRC Press, Boca Raton, FL.)
Hotelling, H. 1933, Journal of Educational Psychology, 24, 417–441
Jaffé, R., Nunes, S., Filipe Dos Santos, J., et al. 2021, Environmental Research Letters, 16, 084034
Jolliffe, I. T. & Cadima, J. 2016, Philosophical Transactions of the Royal Society of London Series A, 374, 20150202

Article number, page 13 of 19
Appendix A: Spatial reconstructions of PCA/NMF components

Here, we further explore the reconstructions obtained using PCA and NMF methods. Figures A.1 and A.2 show examples of spatial maps of PCA and NMF coefficients attached to principal components before INLA (left panels) and after INLA (right panels). PCA sorts principal components by the amount of information they carry, which can be seen in Figure A.1. The first component reconstructs the highest flux density regions: galaxy centre and strong spiral structure. The second principal component reconstructs outer parts of the spiral arms and the galaxy outskirts. The remaining principal components reconstruct background emission and less prominent features in the spiral structure. Up to the sixth principal component, INLA reconstructions do not significantly differ from the input. However, for the remaining principal components INLA reconstructions predict less variability, which is especially evident for the sixth component.

On the other hand, NMF components are not sorted by the amount of variability. Figure A.2 shows that NMF components can be distinguished between those that reconstruct central parts, spiral structure, and a mixture of different features. Overall, the structure seen in all NMF components is preserved and slightly highlighted after INLA reconstructions.
Fig. A.1: Spatial maps of PCA coefficients attached to principal components before INLA (left panels) and after INLA (right panels). Sample size: 25%, ‘face-on’ cube.

Fig. A.2: Same as Fig. A.1 for NMF method.
Appendix B: Pure PCA/NMF reconstructions

In this section we compare PCA (Figure B.1) and NMF (Figure B.2) reconstructions before and after INLA. Figures B.1 and B.2 show pure PCA/NMF and PCA/NMF+INLA reconstructions (upper panels) and the associated residuals (lower panels) at different wavelength bins: 0.23 µm, 7.88 µm, 100.08 µm and 515.36 µm. Pure PCA/NMF reconstructions refer to results of PCA or NMF analyses performed using 100% of full LPN input cube and ten PCA/NMF components. The quality of these reconstructions is further enhanced with INLA using randomly sampled 25% of PCA/NMF coefficients. Results of both pure PCA/NMF and PCA/NMF+INLA reconstructions are compared to HPN reference images.

Figure B.1 shows that pure PCA method produces pixels with negative values. PCA/NMF+INLA reconstructions result in lower residuals compared to the pure PCA/NMF technique. Employing INLA reduces the amount of negative reconstructions leading to lower residuals and thus better quality of reconstructions. However, at wavelengths around ~100 µm the amount of negative values is still high after INLA, leading to higher residuals at the galaxy outskirts. On the other hand, NMF reconstructions (Figure B.2) do not suffer from nonphysically negative reconstructions, but at the same wavelength bins (~100 µm) NMF+INLA overestimate flux densities at the galaxy outskirts, leading to higher residuals compared to pure NMF reconstructions.
Fig. B.1: Pure PCA and PCA+INLA spatial reconstructions (upper panels) and the associated residuals (lower panels) at wavelength bins: 0.23 μm, 7.88 μm, 100.08 μm and 515.36 μm. Sample size: 25%, 'face-on' cube.
Fig. B.2: Same as Fig. B.1 for pure NMF and NMF+INLA reconstructions.