Bayesian X-Ray Imaging in Astrophysics

Paul, Unik, Suchetana, Valentin (Group L2)

AI Lab, July 2023

Contents

1 Aim of the Experiment

High-energy X-rays that are released by numerous celestial objects and processes are referred to as "X-ray radiation from space." The observed X-ray radiation is mostly produced by hot gases in astronomical objects like the Intracluster Medium (ICM) in galaxy clusters, Supernova Remnants and Active Galaxy Nucleis. In this lab experiment our goal will be the reconstruction of the signal (flux) from the provided photon count data from the Chandra X-ray Observatory with the help of a generative forward model.
We will be working with Bayesian X-ray imaging methods, more specifically we will be implementing the two algorithms : the maximum a posteriori (MAP) estimation and the variational inference (VI) method. The code is going to be implemented using the NIFTy python package. After the reconstruction of the X-ray data, we will analyze the difference between the outputs of the two inference algorithms.

2 Introduction to the Physics Background

2.1 Bayesian Inference

In our work of reconstruction of the X-Ray signal, we will be using Bayesian Inference. Bayes' Theorem states

$$P(s|d) = \frac{P(d,s)}{P(d)} = \frac{P(d|s)P(s)}{P(d)}.$$

The signal knowledge after the measurement is the posterior $P(s|d)$ which can be obtained from the prior $P(s)$ and the likelihood $P(d|s)$. The likelihood describes the probability of observing the data d given the existence of the signal s and the prior contains our knowledge on the signal before measuring. The evidence $P(d) = \Sigma_s P(d,s)$ acts as a normalization factor, as:

$$\sum_s P(s|d) = \sum_s \frac{P(d,s)}{P(d)} = 1.$$

The evidence will not be computed, as we would have to marginalize over all possible states

2.1.1 Signal Prior and Sources

The prior $P(s)$ is a probability distribution function that encodes our knowledge on the signal. The signal in this instance consists of X-ray radiation produced by various physical processes that may be divided into spatially uncorrelated and spatially correlated structures. The difference between "signal" and "data" in this context is that signal is the incoming spectrum of emission from the source, and data is the discrete quantity recorded by our instruments.

As for the emission sources that we will work with, we can divide them into two categories: Uncorrelated or point sources, and correlated structures or diffusion emission. A point source is a signal source that is so far away in distance that its brightness can be modeled by a delta peak. A correlated source is spatially correlated structures like clouds of gas and they produce diffused emission. In order to describe them we may use a log normal distribution.

2.1.2 Likelihood

The likelihood that we have already mentioned, i.e $P(d|s)$ is used to describe the measurement process. We introduce a new variable in order to describe the likelihood better, which is called the instrument response $\mathcal{R}$. It describes how the signal is converted to data by the measuring instrument. The point spread function (PSF) and exposure operator can be concatenated to simulate the response $\mathcal{R}$ of the X-ray telescope. The probability distribution of the Poissoinian signal, where λ is the rate of the Poissonian distribution, is :

$$\mathcal{P}(d \mid \lambda) = \prod_{i=1}^{N} \frac{\left(\lambda^i\right)^{d^i} e^{-\lambda^i}}{d^i!}, \quad \lambda = \mathcal{R}\ s$$

2.2 Information Field Theory and NIFTy [3]

2.2.1 Introduction to IFT [2]

Information field theory (IFT) is information theory and logic under uncertainty applied to fields. Any quantity that is defined across a space can be a field. IFT explains how field attributes can be inferred using data and knowledge. The framework for signal processing and image reconstruction is entirely Bayesian. NIFTy is a software package that can carry out Bayesian field inference regardless of the underlying grid or resolution. Forward generative models, likelihoods, and inference may all be built using it, and in our experiment we will only be using NIFTy to build operators and execute tasks.
When the signal field s and the noise n of the data d are separate, zero-centered Gaussian processes with known covariances S and N, respectively, a free IFT occurs as :

$$\mathcal{P}(s, n) = \mathcal{G}(s, S)\mathcal{G}(n, N)$$

the measurement equation $d = \mathcal{R}s + N$ is linear in both signal and noise, where $\mathcal{R}$ is the response mapping the continuous signal field into the discrete data space. The information Hamiltonian is defined as :

$$\mathcal{H}(d, s) = -\log \mathcal{P}(d, s) = \frac{1}{2} s^\dagger S^{-1} s + \frac{1}{2}(d - Rs)^\dagger N^{-1}(d - Rs) + \text{ const}$$

In this case, the posterior is: $\mathcal{P}(s \mid d) = \mathcal{G}(s - m, D)$, with the posterior mean field being $m = Dj$, the posterior covariance operator as $D = \left(S^{-1} + R^\dagger N^{-1} R\right)^{-1}$ and $j = R^\dagger N^{-1} d$ as the information source.

2.2.2 Generative Modeling

In situations involving non-linear measurements or unknown covariances, we can infer the signal by using generative modeling. We can rewrite the above free theory as a generative model:

$$s = A\xi$$

with A being the amplitude operator such that it generates signal field realizations with the covariance $S = AA^\dagger$ when being applied to a white Gaussian field ξ with $\mathcal{P}(\xi) = \mathcal{G}(\xi, 1)$ The joint information Hamiltonian for the signal field ξ reads:

$$\mathcal{H}(d, \xi) = -\log \mathcal{P}(d, s) = \frac{1}{2}\xi^\dagger \xi + \frac{1}{2}(d - RA\xi)^\dagger N^{-1}(d - RA\xi) + const$$

2.2.3 Maximum a Posteriori

Maximum a Posteriori (MAP) is a field estimation method which requires minimizing the information Hamiltonian. The minimizing is done by gradient descent method which stops when

$$\frac{\partial \mathcal{H}(d, \xi)}{\partial\, \xi} = 0$$

.
To minimize the Hamiltonian, NIFTy derives the required gradient using a generative model of the signal and the data. In circumstances of deep hierarchical Bayesian networks, MAP frequently yields unsatisfactory outcomes. This is due to MAP's disregard for the parameter space volume factors, which shouldn't be ignored when determining whether a solution is feasible or not. These volume factors can vary by huge ratios in the high-dimensional environment of field inference. It may not be a good idea to use a MAP estimate since it only covers a small portion of the parameter space. This is also a reason why Bayesian statistics is hard to implement in higher dimensions.

2.2.4 Variational Inference

In Variational Inference (VI), the posterior $\mathcal{P}(\xi|d)$ is estimated by a Gaussian distribution distribution, $\mathcal{Q}(\xi) = \mathcal{G}(\xi - m, D)$. The parameters of $\mathcal{Q}$: the mean m and its covariance D are obtained by minimizing a suitable information distance between $\mathcal{Q}$ and $\mathcal{P}$. We will use the variational Kullback-Leibler (KL) divergence as:

$$\mathrm{KL}(m, D|d) = \mathcal{D}_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) = \int \mathcal{D}\xi \mathcal{Q}(\xi) \log\left(\frac{\mathcal{Q}(\xi)}{\mathcal{P}(\xi)}\right)$$

NIFTy allows us to use Metric Gaussian Variational Inference (MGVI) approximates the posterior precision matrix $\mathcal{D}^{-1}$ at the location of the current mean

m by the Bayesian Fisher information metric,

$$M \approx \left\langle \frac{\partial \mathcal{H}(d,\xi)}{\partial \xi} \frac{\partial \mathcal{H}(d,\xi)^{\dagger}}{\partial \xi} \right\rangle_{(d,\xi)}.$$

We only need to determine the approximate distribution's mean. The Hamiltonian of the real issue averaged throughout the approximation $KL(m|d) \hat{=} \langle \mathcal{H}(\xi,d) \rangle_{\mathcal{Q}(\xi)}$ is the only term within the KL-divergence that explicitly depends on it. Then, the KL-divergence and gradients are stochastically estimated using a set of samples taken from the approximative posterior distribution.

2.3 X Ray Telescopy [1]

X-Rays intercepted from the universe are electromagnetic radiation which are emitted from objects that are millions of degrees celsius : such as active galactic nuclei, pulsars, intracluster medium of galaxy clusters, and the accretion disk of black holes. X-ray telescopes measure X-ray amplitude from these extragalactic objects using specialized detector systems designed to capture and quantify the incoming X-ray photons. The XRT typically consists of a sensor called a charge-coupled device (CCD) or complementary metal-oxide-semiconductor (CMOS) detector.
The basic process involves the following steps:

1. **X-ray Interaction**: When X-ray photons from the celestial sources enter the XRT, they interact with the detector material. The X-rays can dislodge electrons from the atoms in the detector material.

2. **Electron Detection**: The dislodged electrons are then collected and converted into electrical signals by the CCD or CMOS detector. Each X-ray photon produces a specific amount of charge that corresponds to its energy.

3. **Signal Processing**: The electrical signals generated by the detector are amplified and processed by electronic circuits. These signals are then converted into digital data for further analysis.

4. **Data Representation**: The processed digital data is used to create images or other forms of data representation. The intensity of X-ray emission at different positions in the sky is mapped to create X-ray images, revealing the distribution and intensity of X-ray sources.

5. **Spectral Analysis**: In addition to imaging, X-ray telescopes can also perform spectroscopy. The detector measures the energy levels of the X-ray photons, allowing astronomers to analyze the X-ray spectrum of celestial objects. This spectrum provides valuable information about the chemical composition, temperature, and other physical properties of the X-ray-emitting sources.

One of the shortcomings of XRT is related to the point spread function (PSF) of the telescope. The PSF describes how a point source of X-rays appears on the detector after passing through the telescope's optics. Due to diffraction and other optical effects, the X-rays from a point source do not converge to a perfect point on the detector but instead form a blurred image with a finite extent. This blurring effect leads to decreased spatial resolution, making it difficult to precisely localize X-ray sources and potentially causing confusion between neighboring sources. Additionally, instrumental imperfections and calibration errors can further contribute to uncertainties in the reconstructed images.

3 Implementation and Analysis

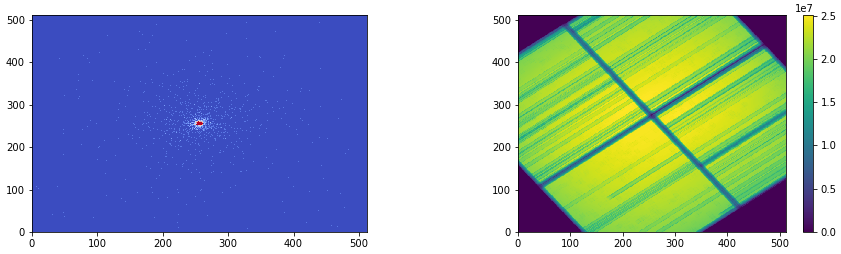
3.1 Description of the simulation chain

The simulation chain we used is described below.It provides a structured approach to analyze X-ray telescope data, accounting for exposure effects and PSF blurring.Then after that, by performing MAP approximation and variational inference, we gain valuable insights into the underlying astrophysical sources and their properties.

1. **Prior Data Model:** The simulation chain begins with a prior data model, which assumes that every pixel in the dataset is uncorrelated with each other. This initial assumption allows us to start with a baseline representation of the data before incorporating any specific characteristics.

2. **Exposure Masking:** To account for the effects of exposure, we create an exposure mask represented as a NIFTy field operator. This exposure mask accounts for variations in the observation time or sensitivity across the dataset. We then apply this exposure mask to our prior data, adjusting the pixel values accordingly.

3. **Convolution:** Next, we define the convolution operator for the NIFTy field. The convolution is performed between the prior data and the Point Spread Function (PSF) image. The PSF characterizes how a point source would be blurred due to the telescope's optical effects. By convolving the prior data with the PSF, we account for the blurring effect in the observation.



(a) Point Spread Function (PSF) Image (b) Exposure Image

Figure 1: PSF and Exposure Images

4. **Response Function Creation:** After applying exposure masking and convolution, we obtain the response function of the system. The response function represents the combined effects of exposure, masking, and PSF blurring on the prior data.

5. **Poissonian Likelihood:** With the response function in hand, we build a Poissonian likelihood model. The Poissonian likelihood accounts for the statistical nature of photon counting in X-ray observations. It allows us to estimate the likelihood of observing the given data, given the response function.

6. **MAP Approximate and Variational Inference:** Finally, we employ Maximum A Posteriori (MAP) approximation and variational inference techniques (as described in the report above(to estimate the parameters of interest in the model. These techniques enable us to find the most probable configuration of model parameters based on the observed data and the prior information.

3.2 Choosing a prior

We assume that the signal we are trying to model is mostly comprised of independently occurring point sources, since we do not know the power spectrum of the signal. Statistically we can model point sources by independent inverse gamma distributions, see [4] chapter 2.3.2. In terms of one variable the probability density function (pdf) of an inverse gamma distribution is

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-1-\alpha} \exp\left(-\frac{\beta}{x}\right).$$

There are two free parameters that need to be chosen α, the so called shape parameter, and β, the so called scale parameter. α determines how much of the density of the pdf is in its tail, if its value is smaller, more of its density is in its tail. For the mean of the distribution to be finite, we require $\alpha > 1$, which made

us choose a value of $\alpha = 1.0001$. The value of *beta* determines below which value the bulk of the density in the pdf is, as can be seen in figure 2, the orange line indicates the value of beta.

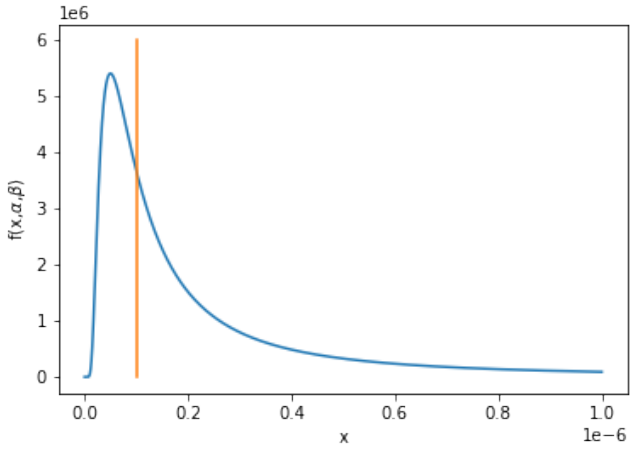


Figure 2: pdf of inverse gamma distribution and value of beta(orange)

In order to choose the value of beta we embrace the rule of thumb $|data| \approx |exposure||signal|$. In order to more accurately access the approximate value for the exposure, we apply the mask to the data and to the exposure, divide them and then apply the adjoint mask operator to be able to plot the result, see figure 3.

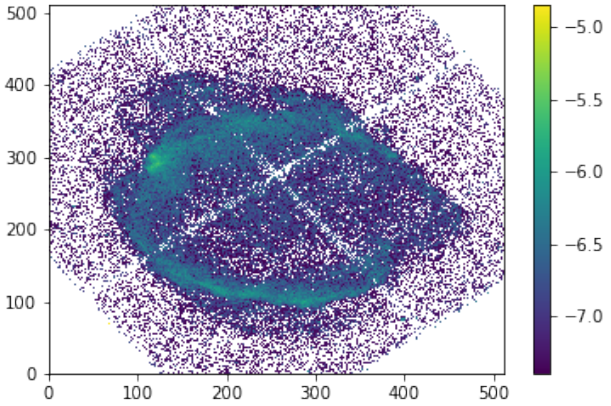


Figure 3: log10($|data|/|exposure|$)

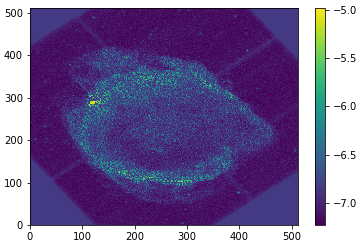
Since most of the values of $|data|/|exposure|$ are under the value 10^{-7}, with some exceptions of values up to 10^{-5}, we choose $\beta = 10^{-7}$, the resulting pdf is the one in figure 2.
As described in chapter 2.3.1 of [4], the signal also has a diffuse part, which can be described by an exponential of a correlated field. The power spectrum used has been given by the tutors. We will examine three different models, one with a prior for point sources, one for a diffuse prior and one that represents our believe about the true nature of the signal the best, a superimposed diffuse and point like signal.

3.3 Results

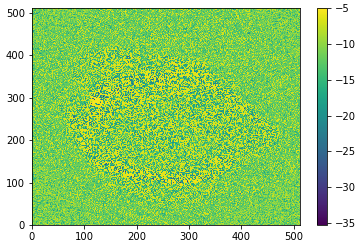
In this chapter we will briefly talk about the different result we obtained for the three prior models. The first three chapters will cover the results of the MAP approach and the last three the VI approach.

3.3.1 MAP Approach

The results of the point source prior estimate with a maximum a posteriori approach in Figure 4a, can be directly compared to the diffusion prior seen in Figure 4b. One can tell that the result of a diffuse prior perform way worse than that of the point source. The point source prior models the signal with no correlation between near points and one can see that it tries to reconstruct the different light sources point wise. The diffuse prior, in contrast, shows a high correlation of the reconstructed light sources. The corners of the point source prior show the similar behaviour as the assumed prior. A sample of the assumed prior can be seen in Figure 5.



(a) Point source prior



(b) Diffuse prior

Figure 4: Comparison of the mean of both priors using MAP

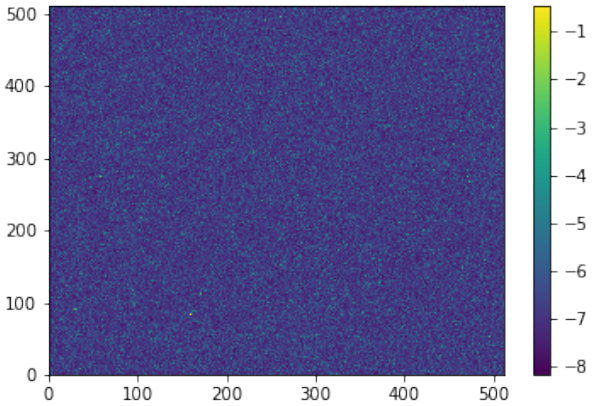


Figure 5: Prior sample

The combined prior results can be seen in Figure 6. The edges and the cross intersection of the detector structure are barely noticeable, which is as expected. One can clearly tell that the overall quality of the plot is lessened since there are plenty of null values now. This artifact could be caused by the general formulation of the MAP approach, where we only get a point estimate without considering the volume of the associated space.

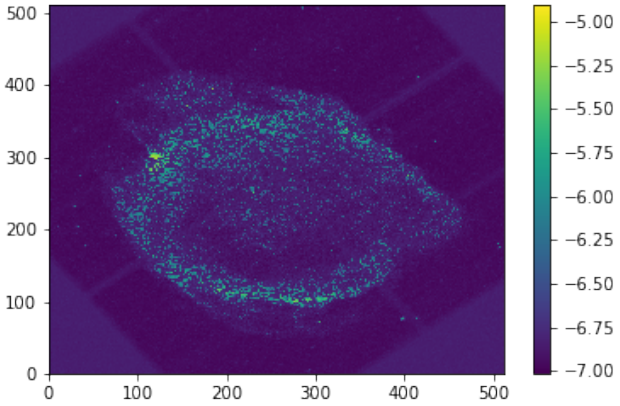
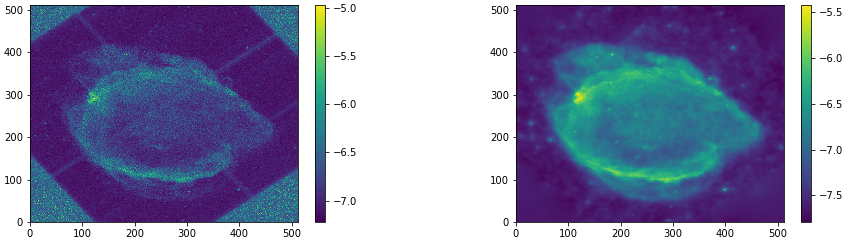


Figure 6: Point source and diffuse prior combined MAP

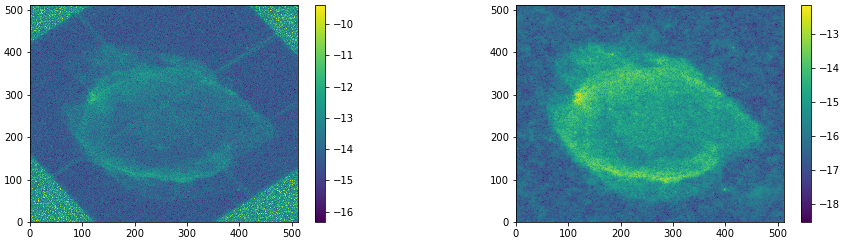
3.3.2 Variational Inference Approach

The results for the variational inference perform a lot better than that of the MAP approach, both for the point source and the diffuse prior. One can see the plots in Figure 7a and 7b respectively. The trend of the two different prior are the same but can be seen a lot clearer than in the MAP model, especially for the diffuse prior one can observe correlations of close by regions. A closer look at the correlation behavior can be seen in Figure 8a and 8b.



(a) Point source prior (b) Diffuse prior

Figure 7: Comparison of the mean of both priors using VI



(a) Point source prior (b) Diffuse prior

Figure 8: Comparison of the variance of both priors using VI

3.3.3 Point source and diffuse prior combined VI

One can tell that the combined prior for the variational inference model, which can be seen in Figure 9 performs a lot better than the MAP model. The combined model approach is expected to give the best results for this task because it specifically tries to separate both underlying priors.

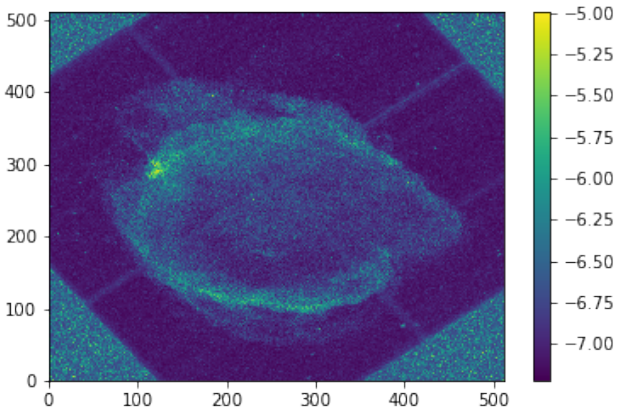


Figure 9: Point source and diffuse prior combined using VI

4 Who did what?

4.1 Report

Suchetana wrote 2.1, 2.2.2 and 2.2.3.
Unik wrote 2.2.1, 2.3 and 3.1.
Valentin wrote 1. and 3.3.
Paul wrote 3.2., 2.2.4.
We all read through the sections the others wrote and corrected some typos. Special thanks to Paul for typing section 4.

4.2 Code

Most of the group was written in a group or with two people working on one piece of code at a time, we also discussed a lot of parts with the whole group and the tutors, so the following list includes who mostly worked on which parts. There is also not a clear separation throughout the sections,
Valentin and Paul worked together on Section 1, Unik wrote the function for logarithmic plots in section 0. Everyone worked on Section 2. In Section 3, Valentin and Paul did a and c, Suchetana and Unik did part b. Some of the ways of how to create the Mask operator and how to estimate parameters we used in c was explicitly given by Vincent. Valentin and Paul did 4., 4.1, Section 5 was done by Suchetana and Unik. We all thought about the answers to the questions in 6 together. Paul and Valentin did Section 7 and added the required parts to the other sections (create combined and diffuse model,..). Unik added the config file. Paul wrote the code in inv_gamma_plot.ipynb.

References

[1] D. Grün. *Observing and Data Analysis Methods for Cosmological Surveys, SS 2023.*

[2] *IFT Script.* URL: https://wwwmpa.mpa-garching.mpg.de/~ensslin/lectures/Files/ScriptIT&IFT.pdf.

[3] *Nifty user guide.* URL: https://ift.pages.mpcdf.de/nifty/user/ift.html.

[4] Selig, Marco and Enßlin, Torsten A. "Denoising, deconvolving, and decomposing photon observations - Derivation of the D3PO algorithm". In: *A&A* 574 (2015), A74. DOI: 10.1051/0004-6361/201323006.