Sampling methods for statistical inference of non-linear inverse problems : spatial distribution of physico-chemical properties of the interstellar medium

Pierre Palud

Centrale Lille

Thèse

présentée en vue d'obtenir le grade de

Docteur

dans la spécialité

« Automatique, Génie informatique, Traitement du Signal et des Images »

par

Pierre PALUD

Doctorat délivré par Centrale Lille

Sampling methods for statistical inference of non-linear inverse problems : spatial distribution of physico-chemical properties of the interstellar medium

Méthodes d'échantillonnage pour l'inférence statistique de problèmes inverses non linéaires : distribution spatiale des propriétés physico-chimiques du milieu interstellaire

Soutenue le 7 décembre 2023 devant le jury d'examen :

Président	
Jérôme Idier	DR CNRS, LS2N (Nantes)
Rapporteurs	
Olivier Berné	DR CNRS, IRAP (Toulouse)
Jérôme Idier	DR CNRS, LS2N (Nantes)
Examinateurs	
Florence Forbes	Directrice de recherche, INRIA Grenoble Rhone-Alpes
Émilie Habart	Maîtresse de conférences, Univertité Paris-Saclay - IAS
Emeric Bron	Astronome adjoint, Observatoire de Paris
Pierre-Antoine Thouvenin	Maître de conférences, Centrale Lille Institut
Directeurs de Thèse	
Pierre Chainais	Professeur des universités, Centrale Lille Institut
Franck Le Petit	Astronome, Observatoire de Paris

Thèse préparée dans le **Laboratoire CRIStAL** et l'**Observatoire de Paris**
Ecole Doctorale MADIS 631

Centrale Lille
Cité Scientifique - CS20048 - 59651 Villeneuve d'Ascq Cedex - France
Tél. + 33 (0)3 20 33 53 53 - https://centralelille.fr

Contents

List of Figures

List of Tables

List of Algorithms

Notation

$\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)_{n=1}^N \in \mathbb{R}^{N \times D}$ set of N physical parameter vectors $\boldsymbol{\theta}_n$. This is the parameter to be inferred in an inverse problem. When $N = 1$, this set is a vector and the notation is simplified to $\boldsymbol{\theta} \in \mathbb{R}^D$. When $N = D = 1$, this set is a scalar and the notation is simplified to $\theta \in \mathbb{R}$. When computing gradients with respect to $\boldsymbol{\Theta}$, it is handled as a vector with conversion in lexicographic order, such that $\boldsymbol{\Theta} \in \mathbb{R}^{ND}$.

$\boldsymbol{\theta} \in \mathbb{R}^D$ physical parameter vector. A physical parameter $\theta_d \in \mathbb{R}$ can be for instance a thermal pressure, a visual extinction, or the intensity of a radiative field.

$\widehat{\boldsymbol{\Theta}} = (\boldsymbol{\theta}_n)_{n=1}^N \in \mathbb{R}^{N \times D}$ generic notation for an estimator of $\boldsymbol{\Theta}$. When $N = 1$, this set is a vector and the notation is simplified to $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^D$. When $N = D = 1$, this set is a scalar and the notation is simplified to $\widehat{\theta} \in \mathbb{R}$.

$\mathbf{Y} = (\mathbf{y}_n)_{n=1}^N \in \mathbb{R}^{N \times L}$ set of N individual observation vectors $\mathbf{y}_n$, from which $\boldsymbol{\Theta}$ is inferred. When $N = 1$, this set is a vector and the notation is simplified to $\mathbf{y} \in \mathbb{R}^L$. When $N = L = 1$, this set is a scalar and the notation is simplified to $y \in \mathbb{R}$.

$\mathbf{y} \in \mathbb{R}^L$ individual observation vector. An observation element $y_\ell \in \mathbb{R}$ can be e.g., the integrated intensity of an ionic, atomic or molecular emission line.

$\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{y}}_n)_{n=1}^N \in \mathbb{R}^{N \times L}$ set of N individual reproduced observation vectors $\widetilde{\mathbf{y}}_n$. This reproduced observation follows the posterior predictive distribution Eq. 2.46, and is used for posterior predictive assessment. Introduced in Chapter 2 (Section 2.3.2). Uses in interstellar medium studies are reviewed in Chapter 3 (Section 3.3.2). Our approach is presented in Chapter 5 (Section 5.3). When $N = 1$, this set is a vector and the notation is simplified to $\widetilde{\mathbf{y}} \in \mathbb{R}^L$. When $N = L = 1$, this set is a scalar and the notation is simplified to $\widetilde{y} \in \mathbb{R}$.

D dimension of an individual physical parameter vector $\boldsymbol{\theta}$.

L dimension of an individual observation vector $\mathbf{y}$.

N number of individual observations $\mathbf{y}_n$ in the full observation set $\mathbf{Y}$. Also number of individual physical parameter vector $\boldsymbol{\theta}_n$ in the full parameter set $\boldsymbol{\Theta}$.

$\ln x$ natural logarithm of a scalar $x \in \mathbb{R}$, i.e., $\ln x = \log_e x$. This notation is preferred to the usual $\log$ to avoid any confusion between logarithms in base e and 10.

$\pi(\cdot)$ probability density function (pdf). The distribution associated to a pdf is indicated by the variables in the parentheses. In this thesis, we use an abuse of notation to draw samples from a distribution: for instance, $\boldsymbol{\Theta}^{(1)} \sim \pi(\boldsymbol{\Theta})$ corresponds to a parameter $\boldsymbol{\Theta}^{(1)}$ drawn from the prior distribution.

$\pi(\mathbf{Y}|\boldsymbol{\Theta})$ likelihood function. The likelihood function is defined from an observation model $\mathcal{M}$. It is sometimes written $\pi(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M})$ to highlight this dependence. In this thesis, the negative log-likelihood is assumed to be twice differentiable, i.e., $\boldsymbol{\Theta} \mapsto -\ln \pi(\mathbf{Y}|\boldsymbol{\Theta}) \in \mathscr{C}^2$.

$\pi(\boldsymbol{\Theta})$ pdf of the prior distribution. In this thesis, the negative log-prior is assumed to be twice differentiable, i.e., $\boldsymbol{\Theta} \mapsto -\ln \pi(\boldsymbol{\Theta}) \in \mathscr{C}^2$.

$\pi(\boldsymbol{\Theta}|\mathbf{Y})$ pdf of the posterior distribution, defined in Eq. 2.1. Like the likelihood function, the posterior is defined from an observation model $\mathcal{M}$. It is sometimes written $\pi(\boldsymbol{\Theta}|\mathbf{Y}, \mathcal{M})$ to highlight this dependence. In this thesis, the negative log-posterior is assumed to be twice differentiable, i.e., $\boldsymbol{\Theta} \mapsto -\ln \pi(\boldsymbol{\Theta}|\mathbf{Y}) \in \mathscr{C}^2$.

$-\nabla \ln \pi(\boldsymbol{\Theta}|\mathbf{Y}) \in \mathbb{R}^{ND}$ gradient of the negative log-pdf of the posterior distribution with respect to $\boldsymbol{\Theta}$.

$-\nabla^2 \ln \pi(\boldsymbol{\Theta}|\mathbf{Y}) \in \mathbb{R}^{ND \times ND}$ Hessian matrix of the negative log-pdf of the posterior distribution with respect to $\boldsymbol{\Theta}$.

$\mathcal{L}$ generic loss function, to be minimized. In inverse problems, it is often defined as the negative log-likelihood or the negative log-pdf of the posterior distribution. For minimization algorithms, see Chapter 2 (Section 2.2.1). For possible definitions in machine learning, see Chapter 4. In this thesis, this function is assumed to be twice differentiable, i.e., $\mathcal{L} \in \mathscr{C}^2$.

$\mathcal{M}$ observation model. Includes both the forward model $\mathbf{f}$ and the noise model $\mathcal{A}$.

$\mathbf{f}$ true forward model. It is a vector function $\mathbf{f} : \boldsymbol{\theta} \in \mathbb{R}^D \mapsto \mathbf{f}(\boldsymbol{\theta}) = (f_\ell(\boldsymbol{\theta}))_{\ell=1}^L \in \mathbb{R}^L$. In this thesis, this function is assumed to be twice differentiable, i.e., $\mathbf{f} \in \mathscr{C}^2$.

$\widetilde{\mathbf{f}}$ approximation of the true forward model, also called emulator. Like the true forward model $\mathbf{f}$, this function is assumed to be twice differentiable, i.e., $\widetilde{\mathbf{f}} \in \mathscr{C}^2$. Using an emulator is common when the evaluation time of the true forward model is long. For a review of emulators in interstellar medium and cosmology, see Chapter 3 (Section 3.1.1). Defining this emulator is the core of Chapter 4.

$\boldsymbol{\psi}$ parameters of the approximation of the forward model.

$\mathcal{A}$ general noise model. See Chapter 3 (Section 3.1.2) for a review of noise models in interstellar medium studies. Chapter 3 (Section 3.4) introduces the noise model considered in the class of inverse problems addressed in this thesis.

$\boldsymbol{\varepsilon}^{(m)} = (\varepsilon_{n\ell}^{(m)}) \in \mathbb{R}^{N \times L}$ multiplicative noise.

$\boldsymbol{\varepsilon}^{(a)} = (\varepsilon_{n\ell}^{(a)}) \in \mathbb{R}^{N \times L}$ additive noise.

$\boldsymbol{\Sigma}^{(m)} \in \mathbb{R}^{NL \times NL}$ covariance matrix of multiplicative noise.

$\boldsymbol{\Sigma}^{(a)} \in \mathbb{R}^{NL \times NL}$ covariance matrix additive noise.

$\omega_{n\ell} \in \mathbb{R}$ censorship lower bound on observations $y_{n\ell}$.

$\mathcal{C} \subsetneq \mathbb{R}^D$ validity set for the physical parameter vector $\boldsymbol{\theta}$. It is usually defined as a product of validity intervals for each component θ_d. This set is used to define the uniform prior from Chapter 5 (Section 5.1.2).

$\boldsymbol{\tau} \in \mathbb{R}^D$ regularization weight for the spatial prior defined in Chapter 5 (Section 5.1.2).

$\eta > 0$ step size, used in gradient descent algorithms, e.g., in Eq. 2.14, or Langevin sampling algorithms, e.g., in Eq. 2.36.

$\mathbf{G} \in \mathbb{R}^{ND \times ND}$ preconditioning matrix, also called preconditioner. This matrix is usually a positive definite approximation of the Hessian matrix. Defined in Eq. 2.17. Our choice of preconditioner is presented in Chapter 5 (Section 5.2.1).

q proposal distribution in sampling algorithms.

$\rho \in [0, 1]$ acceptance probability of Metropolis-Hastings algorithm. Defined in Eq. 2.34.

$\widetilde{\rho} \in [0, 1]$ generalized acceptance probability of multiple-try Metropolis algorithm. Defined in Eq. 2.44.

$P_{\mathbf{th}} \geq 0$ thermal pressure, expressed in K cm^{-3}.

$G_0 \geq 0$ intensity of a UV radiative field. In this thesis, its is expressed in reference to the Habing field (Habing, 1968).

$A_V^{\mathbf{tot}} \geq 0$ visual extinction, expressed in magnitudes (mag).

Introduction

"*Every adventure requires a first step.*"

Cheshire Cat in Lewis Carroll's "Alice's Adventures in Wonderland"

Contents

Scientific context

This PhD project is a collaboration between statisticians and astrophysicists. It was co-supervised by Pierre Chainais, researcher in signal processing in the Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL), and Franck Le Petit, astronomer in the Observatory of Paris. It was also co-advised by Pierre-Antoine Thouvenin, researcher in signal processing in CRIStAL, and Emeric Bron, assistant astronomer at the Observatory of Paris. This project was funded by the Mission pour les Initiatives Transverses et Interdisciplinaires (MITI) of the Centre National de la Recherche Scientifique (CNRS), within the 80|Prime fund OrionStat.

At a larger scale, this project is part of the ORION-B consortium, led by Jérôme Pety and Maryvonne Gérin and founded in 2015. This consortium gathers experts in signal processing, machine learning, and astrophysics, both on the theoretical side and the observational side. Its members meet weekly for updates in ongoing research projects and twice a year for strategic and in-depth discussions on current and future projects. In particular, this consortium analyzes the observations of the IRAM-30m Large Program "Orion B", described in Pety et al. (2017). The observation program covered about 5 square degrees of the celestial sphere to map the Orion B giant molecular cloud (GMC). It produced a hyper-spectral image with 1 million pixels and $200\,000$ spectral channels in the radio frequency domain, allowing to map the emission of dozens of molecules over the whole cloud.

Figure 1 shows the resulting Orion B map in radio wavelengths along with an observation of the same region in the visible domain. Multiple statistical analyses were then led from this rich dataset, e.g., principal component analysis (PCA) on integrated line intensities (Gratier et al., 2017) and clustering of the intensities to segment the map (Bron et al., 2018b).

(a) In the visible domain. The molecular cloud is mostly not visible except for the Horsehead nebula (bottom right) in absorption.

(b) Emission of the cloud in radio domain, for the $J = 1-0$ transition of CO isotopologues. Blue: ^{12}CO, Green: ^{13}CO, Red: C^{18}O.

Figure 1: Orion B giant molecular cloud. On each image, the moon is shown for scale.

PhD project summary

This PhD project aims at developing statistical tools to study the physics of the interstellar medium (ISM). The ISM is a very diffuse medium that fills the extraordinarily large volume in a galaxy between celestial objects such as stars and black holes. It comprises a wide variety of environments. Most of the ISM volume is hot, ionized, and diffuse. Conversely, most of its mass is cold, neutral and dense enough for hydrogen to be in its molecular form, and constitutes relatively small regions called *molecular clouds*.

The study of the ISM carries fundamental questions such as star formation or the development of molecular complexity possibly leading to the formation of prebiotic molecules. This thesis focuses on star formation and their feedback on molecular clouds. Stars are born from the gravitational collapse of a part of these clouds. Newborn stars impact their parent cloud with their ultraviolet (UV) irradiation and stellar winds, as well as through supernovae explosions at the end of their lives (for the most massive ones). The overall impact of these feedback processes on the remains of the parent cloud is to this day only partially understood. The feedback of newborn stars might dissipate their parent cloud, which would prevent the formation of other stars. It might as well locally compress parts of the surrounding cloud, and thus favor the formation of other stars.

Observations and physical parameters – The observations considered in this work are hyperspectral maps of molecular clouds in the far infrared and millimeter wavelength domains. We focus on clouds that are illuminated and heated by nearby massive stars emitting UV photons. The surface layer of such clouds, where the UV irradiation heats and dissociates the molecular gas, is called a photodissociation region (PDR). The ions, atoms and molecules present in the cloud cool mostly through observable radiative emission associated with quantum transitions. These hyperspectral maps are reduced to multispectral maps that we denote $\mathbf{Y} = (\mathbf{y}_n) \in \mathbb{R}^{N\times L}$, where each pixel $\mathbf{y}_n$ contains the integrated intensities of $L \sim 5 - 30$ emission lines. We work on observation maps that contain from $N = 1$ to $N = \mathcal{O}(10^4)$ pixels. These maps are therefore much smaller than that of the Orion B cloud which contains other types of environments than PDRs such as dense cores.

These multispectral maps $\mathbf{Y}$ can be compared with predictions of a numerical model of the ISM such as the Meudon PDR code (Le Petit et al., 2006), that we denote $\mathbf{f}$. Such models can produce predictions of the line intensity maps for any provided map of the local physical parameters, denoted $\mathbf{\Theta} = (\boldsymbol{\theta}_n) \in \mathbb{R}^{N\times D}$, where each vector $\boldsymbol{\theta}_n$ contains $D \lesssim 10$ physical parameters such as the gas density, the thermal pressure, and the total thickness of the cloud along the

line of sight. The parameters $\boldsymbol{\Theta}$ thus live in high dimension, but show a simple pixel structure. Estimating a map of physical parameters $\boldsymbol{\Theta}$ from an observation map $\mathbf{Y}$ and the Meudon PDR code $\mathbf{f}$ is an instance of a general class of problems called *inverse problems*. Figure 2 illustrates the principle of an inference procedure.

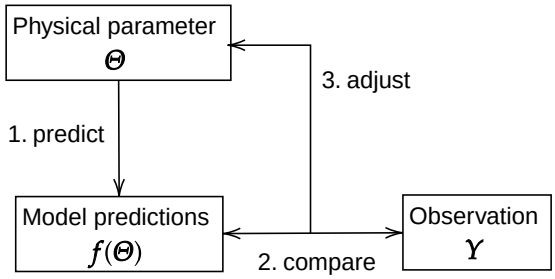


Figure 2: General principle of an inference procedure, repeated until a stopping criterion is satisfied.

Optimization-based inference and existence of multiple solutions – In astrophysics, the inverse problem is often formulated as an optimization problem with a *loss function* $\mathcal{L}$, which measures the difference between the model prediction $\mathbf{f}(\boldsymbol{\Theta})$ and the observation $\mathbf{Y}$. The estimated physical parameter $\widehat{\boldsymbol{\Theta}}$ is then the minimum of a loss function,

$$\widehat{\boldsymbol{\Theta}} \in \arg\min_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}; \mathbf{Y}\right), \tag{1}$$

where $\arg\min$ is the set of values of $\boldsymbol{\Theta}$ that minimize the loss function. For instance, the squared loss

$$\mathcal{L}\left(\boldsymbol{\Theta}; \mathbf{Y}\right) = \|\mathbf{f}(\boldsymbol{\Theta}) - \mathbf{Y}\|_2^2 \tag{2}$$

is a widespread loss function. It turns out that this inverse problem combines many challenges. One example of difficulty is the potential existence of multiple solutions.

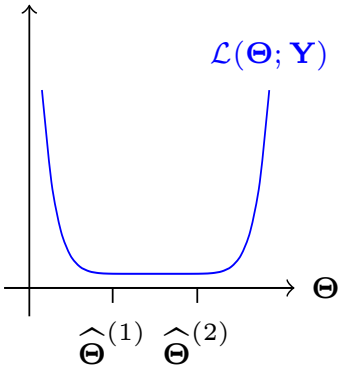
Figure 3 shows two cases where multiple solutions reconstruct equally well the observations, i.e., where the loss function has multiple local minima. The signal-to-noise ratio (SNR) can cover multiple decades in ISM observations. In low SNR regions, multiple physical parameter values $\boldsymbol{\Theta}$ fit the observations equally well. Figure 3a shows that such solution degeneracy is problematic for an optimization problem, as all the values in the wide valley are equally valid. Returning only one value hides this degeneracy. In high intensity regions, the SNR may be high enough for accurate estimations. However, ISM numerical models yield non-linear relations between the physical parameters and the observables so that the cost function is generally non-convex and may thus contain multiple local minima. Figure 3b shows a cost function with two equivalent local minima. Returning only one value hides the existence of the other local minimum.

Uncertainty quantification with Bayesian sampling-based inference – A Bayesian sampling approach quantifies the uncertainty associated with an inference, and can identify these two cases – degeneracy and multiple local minima. In such an approach, the unique estimator $\widehat{\boldsymbol{\Theta}}$ in the optimization problem from Eq. 1 is replaced by a random variable $\boldsymbol{\Theta}$. The cost function $\mathcal{L}(\boldsymbol{\Theta}; \mathbf{Y})$ is then replaced by a probability distribution $\pi(\boldsymbol{\Theta}|\mathbf{Y})$, called the *posterior distribution*.

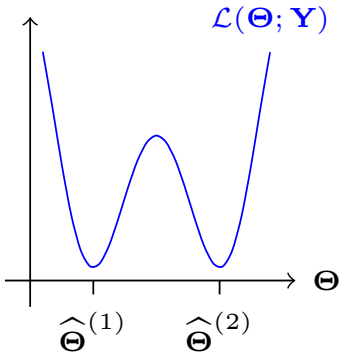
Figure 4 shows the two cost functions of Figure 3 replaced by posterior probability distributions. The posterior distribution contains the information on the physical parameters $\boldsymbol{\Theta}$ provided an observation $\mathbf{Y}$. In particular, it gives access to estimators, to high probability regions and credibility intervals, and permits to identify local minima, all at once.

Extracting information from the posterior distribution often involves integrals over the physical parameters $\boldsymbol{\Theta}$, e.g., for the posterior expectation

$$\mathbb{E}[\boldsymbol{\Theta}|\mathbf{Y}] = \int \boldsymbol{\Theta}\ \pi(\boldsymbol{\Theta}|\mathbf{Y})\ d\boldsymbol{\Theta}. \tag{3}$$

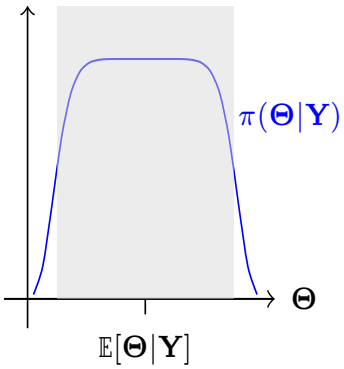


(a) Degeneracy of solutions.

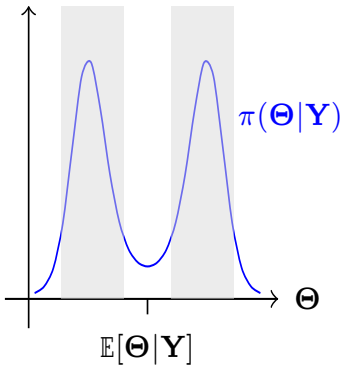


(b) Multiple local minima, here all global minima of the loss function.

Figure 3: Illustration of two types of solution multiplicity in an optimization problem. In both cases, an optimization procedure would return either $\widehat{\boldsymbol{\Theta}}^{(1)}$ or $\widehat{\boldsymbol{\Theta}}^{(2)}$, but would not detect the existence of the other.



(a) Unimodal distribution.



(b) Multimodal distribution.

Figure 4: Replacing a loss function by a posterior distribution. The gray areas represent high probability regions that define credibility intervals. They can be obtained with a Bayesian sampling approach.

In general, such integrals are intractable and require numerical evaluation. Using Riemannian integration is possible when the dimension of $\boldsymbol{\Theta}$ is low, i.e., when $ND \lesssim 10$. However, this method does not scale well to high dimensions as the number of necessary points grows exponentially with ND. In the considered astrophysical inverse problems, the number of pixels N ranges from $\mathcal{O}(1)$ to $\mathcal{O}(10^4)$, and the number of parameters per pixel ranges from 1 to 10. Integrals are thus approximated using Monte Carlo estimators (Robert and Casella, 2004, chapter 3), with T_{MC} samples of the distribution $\boldsymbol{\Theta}^{(t)} \sim \pi(\boldsymbol{\Theta}|\mathbf{Y})$. For instance, the expectation of the posterior distribution in Eq. 3, can be estimated with the empirical mean of samples, also called minimum mean square error (MMSE),

$$\widehat{\boldsymbol{\Theta}}_{\mathsf{MMSE}} = \frac{1}{T_{\mathsf{MC}}} \sum_{t=1}^{T_{\mathsf{MC}}} \boldsymbol{\Theta}^{(t)} \simeq \mathbb{E}[\boldsymbol{\Theta}|\mathbf{Y}]. \tag{4}$$

A general algorithm that efficiently draws independent and identically distributed (i.i.d.) samples from any probability distribution does not exist. For a large set of distributions, Markov chain Monte Carlo (MCMC) algorithms (Robert and Casella, 2004, chapters 6 and 7) generate correlated chains of samples $\boldsymbol{\Theta}^{(t)}$. Such algorithms are often called *samplers*.

Tasks addressed in this thesis – The aforementioned inverse problem is addressed within a Bayesian framework to address the absence of ground truth inherent to astrophysics. Uncertainty quantification, e.g., with credibility intervals, is derived along with point estimates. Solving this inverse problem involved a few main tasks. In each task, the uncertainty associated with proposed approximations is controlled.

- The Meudon PDR code requires a few hours per evaluation, which is prohibitively slow for inference that relies on many evaluations of the likelihood function. It is assumed to be a non-linear and twice differentiable function. Besides, it is a strictly positive function that covers multiple decades, which makes it non gradient Lipschitz continuous or with an extremely large Lipschitz constant. It will be approximated by a light, fast and accurate neural network-based emulator in Chapter 4.

- The full observation model described in Chapter 3 is complex as it involves two sources of noise – one additive and Gaussian, the other multiplicative and lognormal –, and censorship. Without neglecting any source, it leads to a likelihood function whose expression is challenging to address as is. The likelihood function will be approximated with a small and controlled error in Chapter 5.

- The signal-to-noise ratio (SNR) in observation maps greatly varies across lines and pixels. When the SNR is low, the physical parameters usually are poorly constrained. A spatial regularization prior will be included to improve the quality of estimations in these regions. It will enable low SNR pixels to access the information contained in the neighboring pixels.

- The resulting posterior distribution is difficult to sample from. In particular, it is non-log-concave and potentially multimodal because of the non-linearity of the forward PDR model. Besides, the log-posterior is twice differentiable but non gradient Lipschitz continuous. A dedicated MCMC algorithm will be proposed in Chapter 5 as a combination of two sampling kernels: one will efficiently explore the posterior distribution locally, the other will allow escaping from local minima.

- Due to the complexity of the ISM, the Meudon PDR code includes multiple microphysical processes and relies on simplifying assumptions. Its compatibility with the observation will be assessed thanks to a Bayesian hypothesis testing approach presented in Chapter 5.

- The resulting overall method will be applied to synthetic observations in Chapter 5 and to multiple real observations in Chapter 6.

An additional task was also addressed, that is the determination of the best emission lines to observe to obtain low uncertainties on physical parameter estimates. However, this work is still in progress. We currently work on it in collaboration with other members of the ORION-B consortium. Preliminary results are presented in Appendix A.

Addressing these tasks led to contributions in both the statistical and ISM communities.

Structure of the manuscript and reader's guide

This manuscript targets both the signal processing and astrophysics communities. Though the backgrounds of these two communities share some common references, they strongly differ in mindset and vocabulary. Key notions of both statistics and astrophysics are thus re-introduced and discussed. The "Notation" section gathers the main mathematical objects used throughout this thesis, along with the main hypotheses and properties. Along the manuscript, some boxes separated from the main text target one specific community to provide examples or a detailed explanation. In addition, sections whose title begins with "Illustration: ..." provide illustrative examples. These examples are detailed walkthroughs, applications on simplified cases or effect illustrations, but do not contain essential content. The reader may very well choose not to read these sections. Finally, this thesis is structured such that it can be read in two ways.

Option 1: full reading with the default structure – This manuscript is divided into six chapters and one appendix.

Chapter 1 provides some background on the ISM and the associated scientific questions. Then, it focuses on photodissociation regions (PDRs), a particular type of ISM environment. The model used throughout this work to simulate these regions, the Meudon PDR code, is also presented.

Chapter 2 provides some background in statistics. Necessary notions for modeling are introduced. Inference methods from optimization and Bayesian approaches, as well as existing alternatives, are then reviewed. Some Markov chain Monte Carlo (MCMC) algorithms that are necessary to introduce the proposed sampler of Chapter 5 are described.

Chapter 3 is a state-of-the-art that reviews applications of statistical models and inference algorithms to ISM studies.

Chapter 4 describes a model reduction step that is necessary for the proposed sampler of Chapter 5. An artificial neural network (ANN) is proposed to emulate the Meudon PDR code. Some specificities of the Meudon PDR code make traditional off-the-shelf ANN tools inefficient. We design and train ANNs that address these specificities. The corresponding paper, Palud et al. (2023c), was published in the international journal Astronomy & Astrophysics (A&A).

Chapter 5 details the statistical methods we developed to address the inverse problem and evaluate the results. The proposed model and sampler are described. This sampler was presented in a journal article published in the international journal Transactions on Signal Processing (TSP), Palud et al. (2023b). It was also described in two conference papers, one at the French Gretsi national conference (Palud et al., 2022a) and one at the international EUSIPCO conference (Palud et al., 2022b). A proposed model checking strategy enables validating the results of an inversion. This model checking approach is discussed in a conference Gretsi paper, Palud et al. (2023a).

In Chapter 6, the MCMC algorithm presented in Chapter 5 and the reduced model of Chapter 4 are applied to real observations. Inversion results are presented and analyzed. A journal article is currently in preparation, Palud et al. (in prep[a]).

Appendix A proposes a general method to determine which lines are most informative for inference using the differential entropy of a probability distribution. This method paves the way for a new variable selection method. An associated article is currently in preparation, Palud et al. (in prep[b]).

Option 2: reading according to topics – A second possibility is to read this document *by topic*. Table 1 lists the sections associated with each topic.

Table 1: Reading options per topic. "full" means that a given topic is covered in the whole chapter.

Topic	Ch. 1 ISM	Ch. 2 statistics	Ch. 3 interface ISM and statistics	Ch. 4 ANNs	Ch. 5 MCMC algorithm	Ch. 6 real data	App. A line selection
ISM	full	–	full	–	5.4.2	full	full
Statistical model	–	2.1	3.1, 3.4	full	5.1	full	full
forward model	1.2, 1.3	–	3.1.1	full	–	full	full
noise model	–	–	3.1.2, 3.4	–	–	full	full
prior	–	–	3.1.3	–	5.1.2	–	–
Inference	–	2.2, 2.A	3.2	–	5.2	full	full
local exploration	–	2.2,	3.2	–	5.2.1	–	–
escaping local modes	–	2.2, 2.A	3.2	–	5.2.2	–	–
Model relevance	–	2.3	3.3	–	5.3	full	–
model selection	–	2.3.1	3.3.1	–	–	–	–
model checking	–	2.3.2	3.3.2	–	5.3	full	–

List of publications

International journals

Palud, P., P.-A. Thouvenin, P. Chainais, E. Bron, and F. Le Petit (2023b). "Efficient sampling of non log-concave posterior distributions with mixture of noises". *IEEE Transactions on Signal Processing* 71, pp. 2491–2501.

Palud, P. et al. (2023c). "Neural network-based emulation of interstellar medium models". *A&A (in press)*.

International conferences

Palud, P. et al. (2022b). "Mixture of noises and sampling of non-log-concave posterior distributions". *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 2031–2035.

National conferences

Palud, P., P. Chainais, F. Le Petit, P.-A. Thouvenin, and E. Bron (2023a). "Problèmes inverses et test bayésien d'adéquation du modèle". *29° Colloque sur le traitement du signal et des images*. Grenoble: GRETSI - Groupe de Recherche en Traitement du Signal et des Images, p. 705–708.

Palud, P. et al. (2022a). "Mélange de bruits et échantillonnage de posterior non log-concave". *28° Colloque sur le traitement du signal et des images*. 001-0176. Nancy: GRETSI - Groupe de Recherche en Traitement du Signal et des Images, p. 705–708.

International journal articles in preparation

Palud, P., P.-A. Thouvenin, P. Chainais, E. Bron, and F. Le Petit (in prep[a]). "Bayesian inversion of large interstellar medium observation maps".

Palud, Pierre, Pierre-Antoine Thouvenin, Pierre Chainais, Emeric Bron, and Franck Le Petit (in prep[b]). "Entropy-based selection of most informative observables for inference from interstellar medium observations".

The above published conference articles were published in statistical conferences. I also participated to astrophysical international conferences:

- the IRAM[1] 2022 conference (https://iram2022nice.sciencesconf.org/),
- the PCMI[2] 2022 conference (https://pcmi2022paris.sciencesconf.org/),
- the EAS[3] 2023 conference (https://eas.unige.ch/EAS_meeting/index.jsp).

In the ISM community, conferences are based on abstracts and not on articles.

[1] Institut de radioastronomie millimétrique
[2] Physique et chimie du milieu interstellaire
[3] European astronomical society

Part I

Backgrounds: interstellar medium, statistical inference and interactions

Chapter 1

Background on the interstellar medium (ISM)

> " *The Milky Way is largely empty. Stars are separated by some 2 pc in the solar neighborhood. [...] If we take our Solar System as a measure, with a heliosphere radius of ≃235 AU, stars and their associated planetary systems fill about* $3 \times 10^{-8}\,\%$ *of the available space. This [work] deals with what is in between these stars: the interstellar medium (ISM).* "
>
> Tielens (2005, introduction of chapter 1)

Contents

This chapter provides a general overview of the ISM and of the associated scientific questions. It briefly introduces the physics of the ISM and the current state-of-the-art numerical models. The elements and notions covered in this chapter will be used throughout this thesis. Two fundamental notions necessary for the non astrophysicist are first introduced: distance units and scales.

Distance units – The solar system that we live in hosts one of the $\sim 10^{11}$ stars contained in our Galaxy, the Milky Way (MW). Then again, the Milky Way is one of the $\sim 10^{11}$ galaxies of the observable Universe. The study of the Universe at its largest scales falls into the domain of cosmology. In this work, we consider "smaller" astrophysical scales, between the scales of stars and of galaxies. Dedicated distance units are necessary to describe such scales. Three distance units are commonly used in astrophysics:

- The astronomical unit (au): 1 au corresponds to the distance between the Sun and the Earth, i.e., 1.5×10^6 km.
- The light-year (ly): 1 ly corresponds to the distance traveled in one year at the speed of light, i.e., 9.5×10^{12} km.
- The parsec (pc): 1 pc is defined as $\frac{648\,000}{\pi}$ au, i.e., about 3.26 ly. Figure 1.1 illustrates its definition. One parsec corresponds to the order of magnitude of the average distance between a star and the closest other star.

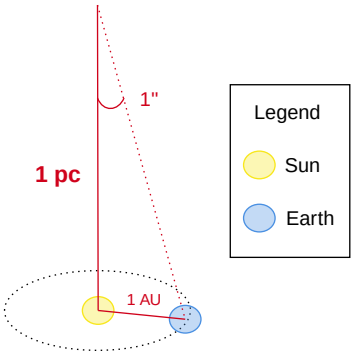


Figure 1.1: Illustration of the parsec (pc) definition. Distances are not at scale.

Scales – Figure 1.2 lists some typical astrophysical distances with the corresponding values in the three aforementioned units. It covers more than 16 decades. This work focuses on molecular clouds such as Orion-B. Molecular clouds correspond to one type of environment of the so-called ISM, which by definition lies between stars. They are extremely large: their sizes range from a few parsecs to a hundred parsecs. For instance, Orion-B is a "neighbor" molecular cloud, about 400 pc away from us. It is 10 pc large, which is about 10^{10} times larger than the diameter of the Earth. This size ratio corresponds to that of a person and a hydrogen atom.

In this chapter, we provide a general description of the ISM. Section 1.1 provides an overview of the ISM, of the associated physics and scientific questions. Section 1.2 describes how models of the ISM and observations can improve our understanding of star formation. We also list some existing ISM models. Finally, Section 1.3 focuses on the Meudon PDR code, which is used in the remainder of this work.

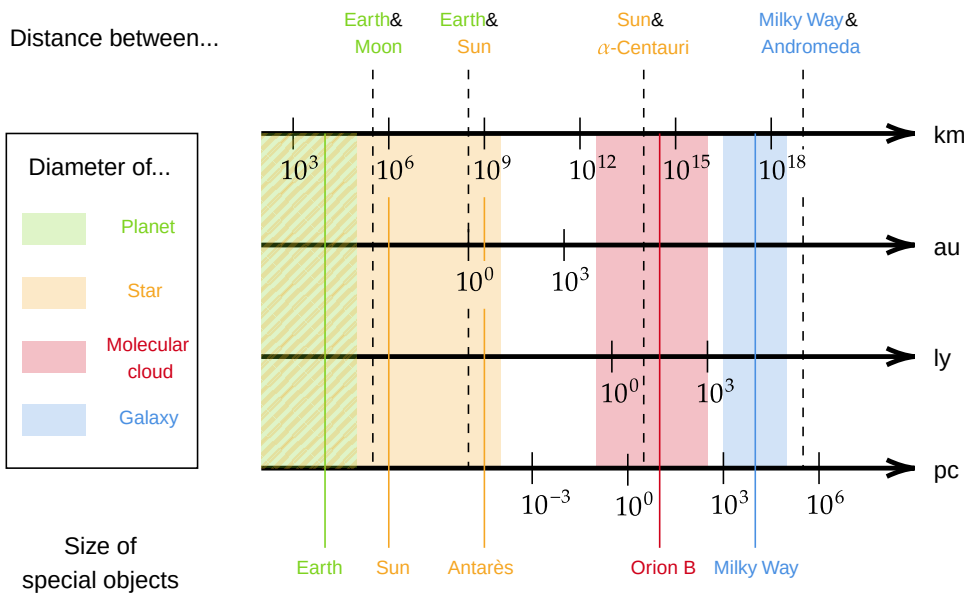


Figure 1.2: Scales in astrophysics. α-Centauri is the closest star to the Sun. The Andromeda galaxy is the closest galaxy to the Milky Way. Orion B is a giant molecular cloud (GMC) of the Milky Way.

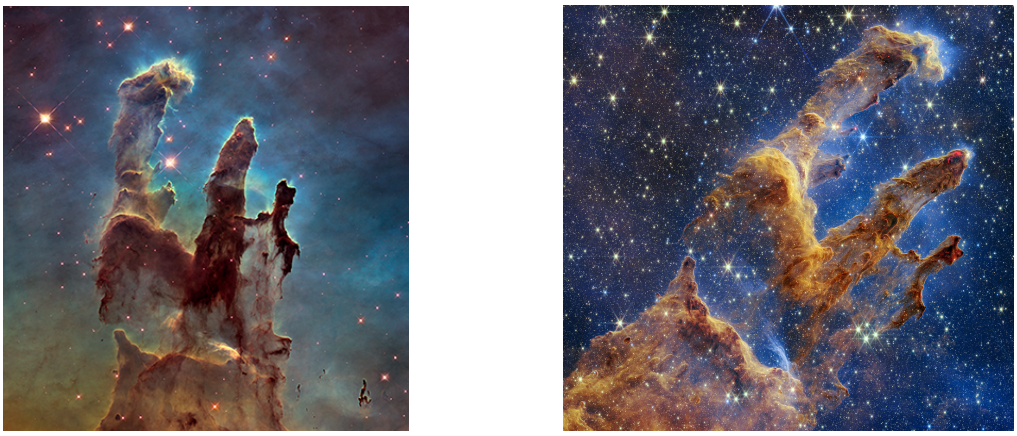
1.1 The interstellar medium (ISM)

In the Milky Way, stars and other celestial bodies such as black holes, planets, and comets represent about $90\,\%$ of the total baryonic mass, i.e., disregarding dark matter and dark energy. The remaining $10\,\%$ correspond to the extremely diffuse medium that lies in the volume between stars: the ISM. The study of the ISM addresses fundamental questions on the Universe including the formation of stars and planets, and the development of molecular complexity possibly leading to the formation of prebiotic molecules. In this section, we provide a general overview of the ISM with its composition, structure, and role in the cycle of matter in a galaxy. We also detail the aforementioned fundamental questions.

1.1.1 A short overview of the ISM

Composition – Hydrogen remaining from the Big Bang represents 90% in mass of the ISM. Helium corresponds to most of the remaining 10%. It comes in part from the Big Bang and in part from nuclear reactions inside stars. Other elements, from carbon to uranium, are also produced by nuclear reactions inside stars, and by supernovae for the most massive elements. They account for less than 1% of the total mass. These other elements, in particular carbon C and oxygen O, play a key role in the ISM physics and observations. The concentration of elements heavier than helium is called the *metallicity*. Table 1.1 lists estimated abundances of the most common elements in the ISM relatively to hydrogen in the solar neighborhood. These elements can exist in ionic, atomic, or molecular form, depending on the environment conditions.

In addition to the gas, the ISM contains small dust grains whose size typically ranges from 1 to 100 nm, and larger in the ISM densest regions and in protoplanetary disks. It is estimated that these dust grains represent about 1% of the ISM mass. They play a key role in the physics and chemistry of the ISM. They extinguish UV radiation and thus shield the gas from its ionizing and dissociating effects. Dust grains also convert a small fraction of the absorbed UV energy into gas heating through the photoelectric effect, which is one of the dominant heating processes in the ISM. Some species exist in both gas and dust phase, and are thus depleted in the gas phase – a phenomenon called the *interstellar depletion* (Draine, 2011, chapter 9). Heavy elements such as



(a) 2014 Hubble observation, in visible light.

(b) 2022 JWST observation, in NIR. Credits: NASA, ESA, CSA, STScI; Joseph DePasquale (STScI), Anton M. Koekemoer (STScI), Alyssa Pagan (STScI).

Figure 1.3: observation of the pillars of creation, in the Eagle Nebula, by the Hubble telescope and the James Webb spatial telescope (JWST). Unlike the observation in the visible light, the near infrared (NIR) observation shows the stars inside and behind the pillars.

silicon Si and iron Fe are extremely depleted, as their gas phase abundances correspond to only a few percent of the total expected abundance. These elements, along with lighter elements such as carbon C and oxygen O, are part of the components of the solid cores of grains. In addition, in the dense cold regions, water and CO molecules from the gas phase freeze out on the dust surface to form ice mantles. The depletion of C and O and other thus increases strongly in the dense cold regions where ice mantles can form. Many chemical reactions then occur at the surface of grains and within ice mantles.

Table 1.1: Solar abundance of elements in the gas phase of the interstellar medium relative to hydrogen. Adapted from Draine (2011, table 1.4).

Element	Abundance (relative to H)	
He	9.55×10^{-2}	$\pm 2.3\,\%$
C	2.95×10^{-4}	$\pm 12.2\,\%$
N	7.41×10^{-5}	$\pm 12.2\,\%$
O	5.37×10^{-4}	$\pm 12.2\,\%$
Si	3.55×10^{-5}	$\pm 9.6\,\%$
S	1.45×10^{-5}	$\pm 7.2\,\%$
Fe	3.47×10^{-5}	$\pm 9.6\,\%$

Phases in the ISM – The ISM is a very inhomogeneous medium. It goes from very diffuse hot ionized environments with a density of about 3×10^{-3} particles cm^{-3} that represent most of the volume of the ISM, to dense molecular clouds with 10^6 particles cm^{-3} that represent most of its mass. For comparison, at a standard pressure, the Earth atmosphere contains about 10^{19} particles cm^{-3} and the vacuum at the surface of the Moon about 10^5 particles cm^{-3} (Öpik, 1962).

Table 1.2, adapted from Galliano (2022), lists the main phases of the ISM along with their respective volume densities, average temperature and volume filling factor. The main distinctions between these phases are the state of hydrogen – ionized H^+, atomic H or molecular H_2 –, their temperature and their volume density, i.e., the average number of particles per cubic centimeter. The hot ionized medium (HIM), warm ionized medium (WIM), and HII regions correspond to hot plasmas. In particular, HII regions are located around a massive star where the gas is

ionized by the extreme ultraviolet (EUV) photons of the star. The EUV field contains photons with wavelengths smaller than $91.2\,\text{nm}$, i.e., $13.6\,\text{eV}$, the hydrogen ionization potential. This thesis focuses on the remaining four phases, where hydrogen is mostly neutral and where stars are formed. Hydrogen is mostly in its atomic form in warm neutral medium (WNM) and cold neutral medium (CNM), and in its molecular form in diffuse H_2 and dense H_2 regions. These regions form *molecular clouds* such as Orion-B, shown in Figure 1, or the pillars of creation, shown in Figure 1.3.

Table 1.2: Comparison of the main phases of the ISM. The sum of volume filling factors exceeds 100% because the associated uncertainties are large. Adapted from Galliano (2022).

Hydrogen state	Phase	Volume density n (in cm^{-3})	Temperature (in K)	Volume filling factor (in %)
ionized	HIM	$\simeq 3 \times 10^{-3}$	$\simeq 10^6$	$\simeq 50$
	WIM	$\simeq 0.1$	$\simeq 10^4$	$\simeq 25$
	HII regions	$\simeq 1 - 10^5$	$\simeq 10^4$	$\lesssim 1$
atomic	WNM	$\simeq 0.3$	$\simeq 10^4$	$\simeq 30$
	CNM	$\simeq 30$	$\simeq 100$	$\simeq 1$
molecular	Diffuse H_2	$\simeq 100$	$\simeq 50$	$\simeq 0.1$
	Dense H_2	$\simeq 10^3 - 10^6$	$\simeq 10$	$\simeq 0.01$

The cycle of matter in a galaxy – The ISM plays a key role in the evolution of matter in a galaxy. Figure 1.4 shows the cycle of baryonic matter. The ISM provides the building material for star formation. In average in the Milky Way, about $1.3\ M_\odot\ \text{yr}^{-1}$ of the ISM is converted into stars, with $1M_\odot = 2 \times 10^{30}$ kg corresponding to one solar mass. This process is detailed in the next section. This figure also shows that matter can go back from stars to the ISM in diverse ways, closing the cycle of matter in a galaxy. For instance, during its life, a star emits stellar winds that continuously feed the ISM in matter and in mechanical energy. Some stars also affect the ISM with violent deaths such as supernovae, feeding it both in energy and in elements heavier than helium.

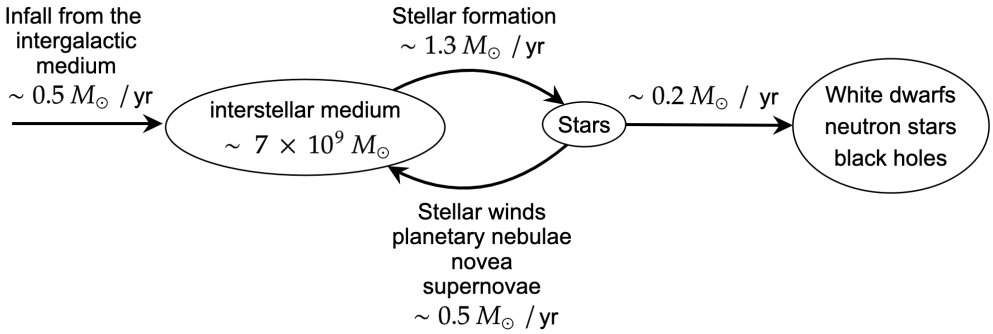


Figure 1.4: Observed flow of baryons in the Milky Way. Adapted from Draine (2011, chapter 1).

1.1.2 Star formation, stellar feedback and photodissociation regions (PDRs)

Star formation starts with the gravitational collapse of a cold and dense region of a molecular cloud. When this region collapses, the associated gas and dust form an infalling envelope and a rotating disk. Planetary systems form from this gas and dust. Figure 1.5 illustrates the star formation process. The first three steps typically last a few million years. During this dramatic compression process, the gas collapses from scales of ~ 1 pc to $\sim 10^{-7}$ pc, which corresponds to a density increase of a factor $\sim 10^{21}$ (Draine, 2011, chapter 41).

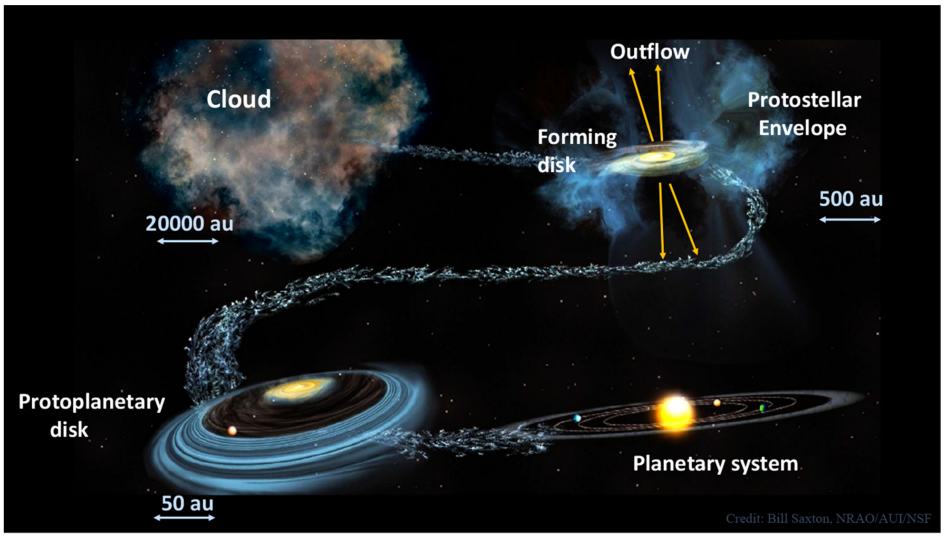


Figure 1.5: Key steps of the star formation process. Artist impression from Bill Saxton (NRAO/AUI/NSF), taken from Dishoeck and Bergin (2021).

The gravitational collapse starts when the self-gravity of a region of the cloud exceeds a certain threshold. Different definitions of this threshold such as the Jeans instability rely on different assumptions and geometries (Draine, 2011, chapter 41). Multiple physical phenomena resist the cloud self gravity such as the internal thermal pressure, the magnetic pressure, turbulence at the lowest scales, and the region angular moment. In this adversarial context, two main mechanisms contribute to attain the density threshold that triggers the collapse. First, large scale turbulence or other compressive events, such as Supernovae shock waves, or HII region expansion, can locally compress the cloud. Second, atomic and molecular radiative emission evacuates energy, and thus cools the cloud and helps reduce the thermal pressure of the cloud, allowing self-gravity to take over.

Despite good observational evidence and advanced models, the star formation process is to this day only partially understood. Observations lead to a star formation rate (SFR) averaged over the past $3\,\mathrm{Myr}$ in the Milky Way of about $1.3\ M_\odot\ \mathrm{yr}^{-1}$ – as shown in Figure 1.4. A simple theoretical estimate based on free-falling gas and using typical values of the Milky Way – total molecular gas $\sim 10^9\ M_\odot$ with density $\sim 50\ \mathrm{cm}^{-3}$ – leads to a maximum SFR of about $200\ M_\odot\ \mathrm{yr}^{-1}$, i.e., two orders of magnitude larger than the measured value (Draine, 2011, chapter 42). Therefore, the star formation process is highly inefficient.

A first explanation of this inefficiency is that only the densest regions of the ISM are expected to actually collapse. This hypothesis is confirmed by the Kennicutt-Schmidt empirical law, which relates the SFR to the surface density of gas with a power law with exponent around 1.4 (Draine, 2011, chapter 42). The exponent of this power law is yet to be physically explained.

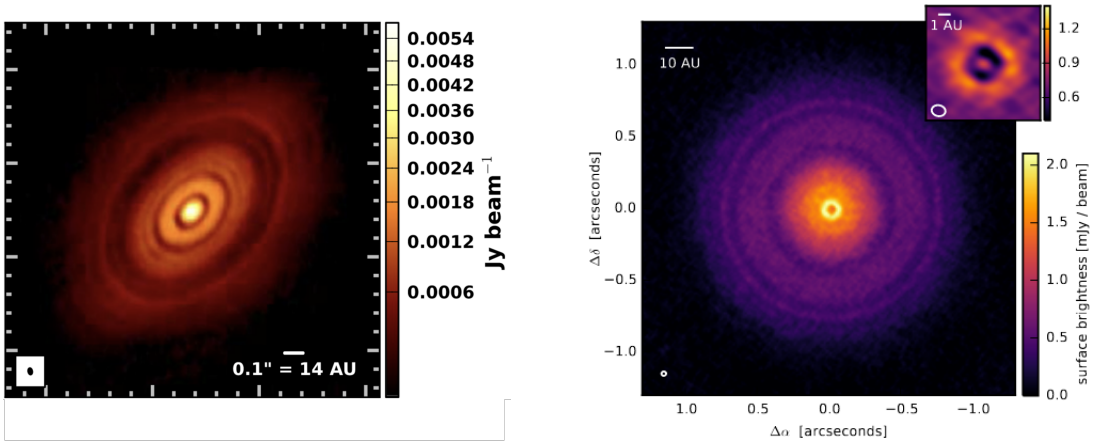
A second reason for this inefficiency is the *stellar feedback* that disrupts the parent cloud and thus decreases its star forming capability. We call stellar feedback all the processes by which a star injects energy into its parent cloud. It includes mechanical energy, such as stellar winds and supernovae, and radiative energy emitted in the UV domain. However, to this day, it is unclear whether stellar feedback favors or impedes star formation. Overall, stellar feedback tends to dissipate the parent cloud. For instance, the expansion of the HII bubble ionized by massive stars tends to disperse the surrounding molecular cloud. However, this expansion also causes the apparition of a compressed layer at the frontier between the molecular parent cloud and the HII region. This compressed layer at the edge of the molecular cloud and of the HII region is called a photodissociation region (PDR). This local compression of the gas might favor the formation of additional stars. Overall, the global parent cloud disruption and the local compression at the PDR are competing effects. Understanding the influence of stellar feedback requires studying the

structure of PDRs and the local physical conditions such as the thermal pressure, the volume density, and the cloud depth along the line of sight.

1.1.3 Chemical complexity and planet formation

In 2018, over 200 molecules had been detected in the ISM (McGuire, 2018), organic or not, including simple diatomic molecules such as H_2 and CO, more complex molecules such as water H_2O and methanol CH_3OH, and complex ions such as H_3O^+ and HCO^+. Molecules from the ISM can be observed in the optical/UV domain for electronic transitions, in the IR domain for vibrational transitions, and in the millimeter domain for rotational transitions. Dust grains play a key role in the formation of some complex molecules, acting as a catalyst. For instance, methanol mainly forms on the surface of grains in CO-rich ice. Water is more complex. It is the product of a variety of reactions, some occurring in the gas phase and some taking place on the surface of dust grains, like H_2 + O. It can freeze onto the grain, forming water ice mantles. This process occurs in the deep, cold, and dense regions of a molecular cloud. These ice mantles can serve as reservoirs for water and other simple molecules and allow for a rich chemistry that would be extremely slow in the gas phase in dense and cold conditions.

As already mentioned, stars and the associated planets form from the molecular gas and dust that gravitationally collapses. A significant part of the chemical composition of the planet-building material appears to be set in the cold pre- and protostellar stages and preserved to planet and comet construction sites. In the early phases of a star formation, the gas and dust spirals in a protoplanetary disk. Figure 1.6 shows two spatially well resolved protoplanetary disks observed by the ALMA telescope in the FIR and millimeter domains.



(a) Dust continuum observation at 1.3 mm of the disk surrounding the star HL Tau. Adapted from ALMA Partnership et al. (2015).

(b) Observations of the 870 µm continuum emission from the TW Hya disk. Taken from Andrews et al. (2016)

Figure 1.6: ALMA observations of two protoplanetary disks.

Protoplanetary disks models study stellar system formation from the initial molecular cloud collapse. Planets form with the aggregation of gas and dust grains. One of the key questions such models try to address is the origin of the chemical composition of planets, in particular for the necessary bricks for the apparition of life, such as water, simple sugars and peptide bonds. This composition might be inherited from the parent molecular cloud and the original dust grains and ice mantles, or might be reset with additional chemistry in the disk, either fully or partially. The study of comets and meteorites in the solar system indicates both a full reset of the chemistry and inheritance from the ISM (Dishoeck and Bergin, 2021). Therefore, a significant fraction of the water on Earth might come from the ISM.

1.1.4 PDRs: structure and physical parameters

Modeling and observing molecular clouds permits to understand the mechanisms of star formation and its feedback, and, within protoplanetary disks, of the formation of complex molecules. In this thesis, we focus on the process of star formation feedback in molecular clouds. More precisely, we focus on the surface layer of molecular clouds irradiated by stellar UV photons, i.e. photodissociation regions (PDRs). This section introduces the key parameters that describe the physical conditions of PDRs and thus quantify the impact of stellar feedback.

Figure 1.7 illustrates the structure of a plane-parallel PDR along with the associated main physical parameters. It also shows a PDR in the observation of the Carina nebula from the JWST. The gas around the star is impacted by its UV radiation. The ISM close to the star is ionized by the EUV part of the spectrum of a star. Such regions are called HII regions (in blue in the JWST observation). Once the ionizing photons are absorbed, the hydrogen can exist in neutral atomic form. Deeper in the cloud, hydrogen forms molecules. A PDR corresponds to the region with neutral hydrogen (red regions in the JWST observation, and shaded regions in the diagram). The physics and chemistry of a PDR are controlled by far ultraviolet (FUV) dissociating photons – with energy lower than the ionization potential of H, $13.6\,\mathrm{eV}$, i.e., with wavelengths greater than $91.2\,\mathrm{nm}$.

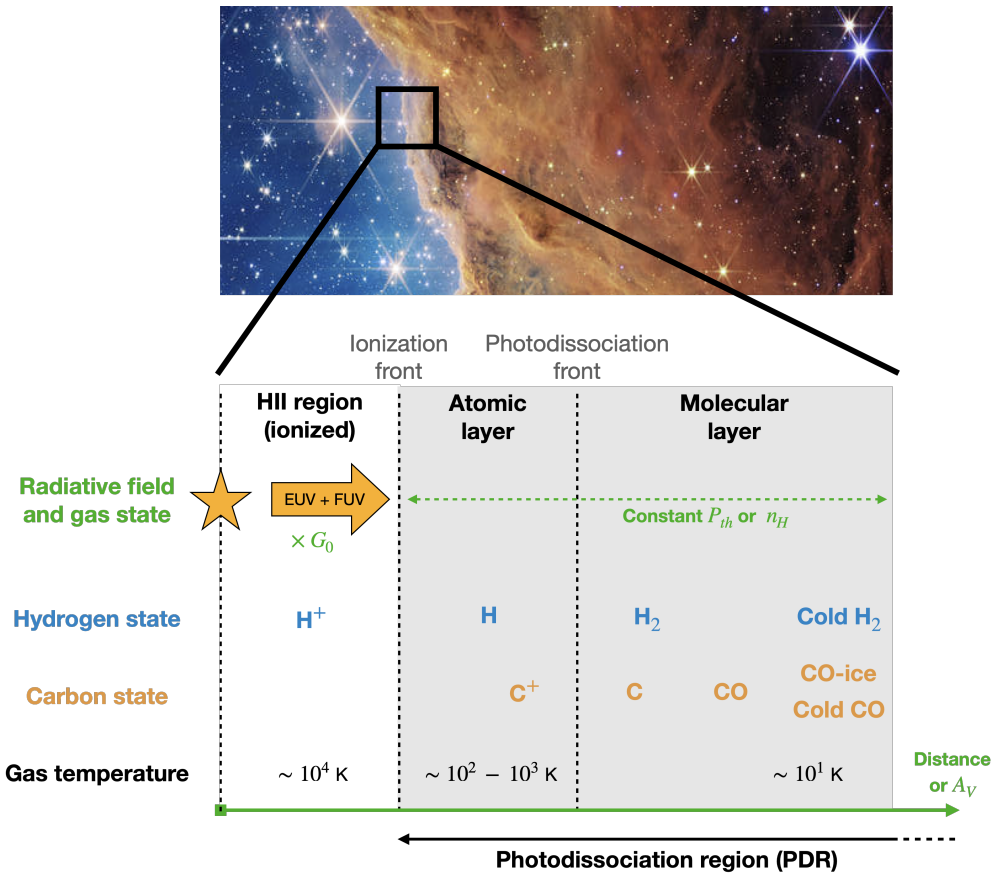


Figure 1.7: Structure of a PDR region. The image on top is extracted from the JWST Carina nebula observation. Credits to NASA, ESA, CSA, and STScI.

Temperature, volume density and thermal pressure – The physical state of the gas at any point in a PDR is usually described by its temperature T_{gas}, expressed in K, and its proton volume density $n_H = n(\mathrm{H}) + 2n(\mathrm{H_2}) + n(\mathrm{H^+})$, expressed in protons $\mathrm{cm^{-3}}$. Note that the proton volume density is not equal to the volume density of particles in the gas, which is $n_{\mathrm{tot}} = n(\mathrm{H}) + n(\mathrm{H_2}) + n(\mathrm{H^+}) + n(\mathrm{C}) + n(\mathrm{O}) + n(\mathrm{CO}) + \ldots$, expressed in particle $\mathrm{cm^{-3}}$. The gas temperature greatly varies in a PDR. In particular, it drops at the photodissociation front

from $\sim 10^2 - 10^3$ K to ~ 10 K. The proton volume density typically ranges from 10^2 to 10^8 cm^{-3}. The particle volume density and the temperature determine the thermal pressure of the gas $P_{\text{th}} = n_{\text{tot}} \times T_{\text{gas}}$, expressed in K cm^{-3}. In a PDR, the thermal pressure ranges from 10^5 to 10^9 K cm^{-3}.

Amount of matter: optical depth, total visual extinction and column density – Astrophysical observations are based on emission or absorption of photons in specific directions. However, one cannot locate photon sources along the line of sight. A natural quantity to describe the ISM – including PDRs – is thus the amount of matter along the line of sight.

Details for statisticians 1.1: The radiative transfer equation and optical depth

The radiative transfer equation relies on the specific intensity $I_\nu(x, \mathbf{k}, t)$, expressed in erg cm^{-2}s^{-1}Hz^{-1}sr^{-1}. $I_\nu(x, \mathbf{k}, t)$ quantifies the intensity of a radiation field at frequency ν, for a position x, a direction of propagation $\mathbf{k}$ and an instant t. Light and matter can interact in three ways: a photon can either be absorbed, scattered, or emitted. The radiative energy absorbed and scattered at position x are quantified by an absorption coefficient κ_ν and a scattering coefficient s_ν, respectively, both expressed in cm^{-1}. The energy emitted along the direction $\mathbf{k}$ is quantified by an emissivity term $\epsilon_\nu(\mathbf{k})$. The energy scattered from any direction $\mathbf{k}'$ back to $\mathbf{k}$ within a solid angle $d\Omega$ is quantified with a term $\widetilde{s}_\nu(x, \mathbf{k}', \mathbf{k})$. Neglecting time dependency terms, the radiative transfer equation reads

$$\frac{\partial I_\nu}{\partial x}(x, \mathbf{k}) = -I_\nu(x, \mathbf{k}) \left[\kappa_\nu(x) + s_\nu(x)\right] + \epsilon_\nu(x, \mathbf{k}) + \frac{1}{4\pi} \int I_\nu(x, \mathbf{k}') \widetilde{s}_\nu(x, \mathbf{k}', \mathbf{k}) d\Omega. \quad (1.1)$$

To simplify Eq. 1.1, the variable change $d\tau_\nu = (\kappa_\nu + s_\nu)dx$ is usually performed, where τ_ν is called the optical depth. Neglecting the integral term, Eq. 1.1 then becomes :

$$\frac{dI_\nu}{d\tau_\nu} = -I_\nu + S_\nu, \quad (1.2)$$

with $S_\nu = \epsilon_\nu/(\kappa_\nu + s_\nu)$ the source function. For some initial position x_0 – usually the back of the cloud – it reads (Draine, 2011, Eq. 7.14)

$$\tau_\nu(x) = \int_{x_0}^{x} d\tau_\nu(x')\, dx' = \int_{x_0}^{x} \left[\kappa_\nu(x') + s_\nu(x')\right]\, dx'. \quad (1.3)$$

The amount of matter along the line of sight can be quantified through radiative transfer and extinction. The optical depth quantifies the attenuation of the radiative field at frequency ν after passing through a cloud. It is positive except in case of optical pumping and MASER effect. When $\tau_\nu \gtrsim 1$, the cloud is said to be *optically thick*. A photon emitted inside the cloud is very likely to be re-absorbed – a phenomenon called *radiative trapping* (Draine, 2011, chapter 19). Therefore, an observer only detects photons emitted by the border of the cloud. Conversely, when $\tau_\nu \ll 1$, the cloud is said to be *optically thin*, and a photon is very likely to escape. In this case, an observer detects photons emitted in the entire cloud. The V ("visual") band is a reference frequency band that corresponds to a filter with a 551 nm effective wavelength and an 88 nm full width at half maximum. One then commonly uses the optical depth in the V band τ_V to measure the interstellar reddening, i.e., the extinction of visible light by interstellar dust. As astronomers commonly express intensities as magnitudes – which correspond to the decimal logarithm of physical fluxes – the extinction is defined in terms of a visual extinction

$$A_V = [2.5 \times \log_{10} e]\ \tau_V \simeq 1.086\, \tau_V. \quad (1.4)$$

Clouds with $A_V < 1$ are not deep enough to form molecules, as they get dissociated by the UV

photons. Dense cores can reach values of A_V of the order 10^2.

The proton column density, denoted N_H and expressed in cm^{-2}, counts the protons along the line of sight. For a cloud of depth x^{tot} (in cm),

$$N_H = \int_0^{x^{\text{tot}}} n_H \, dx. \tag{1.5}$$

In a cloud with constant volume density n_H, $N_H = x^{\text{tot}} \times n_H$. As dust is the dominant source of extinction in the V band and as one commonly assumes that dust is uniformly mixed with the gas, the column density is proportional to the visual extinction

$$N_H = \frac{R_V}{C_D} A_V \simeq 1.086 \frac{R_V}{C_D} \tau_V, \tag{1.6}$$

where $C_D = 5.8 \times 10^{21}$ cm^{-2} (Bohlin et al., 1978), and R_V is the so called total-to-selective extinction ratio and represents the slope of the extinction curve in the visible range. Typically, $R_V = 3.1$ in local diffuse clouds (Fitzpatrick and Massa, 1990).

Radiative field and G_0 – As already stated, PDRs are generally illuminated by a massive star or a cluster of stars. The shape and intensity of the UV radiative field reaching the surface of the PDR depends on the properties of the star and on the distance between the PDR and the star. The G_0 parameter quantifies the intensity of the FUV radiative field regardless of its shape. It is defined as the ratio between the integral of the spectrum of the radiative field over a selected wavelength range and of the same integral computed for a reference interstellar radiation field (ISRF). For instance, prescriptions of ISRF can be found in Habing (1968), Draine (1978), and Mathis et al. (1983). Conversions of G_0 exist between the different ISRF prescriptions. In this thesis, we measure G_0 in reference to the Habing field. Far from any star in our Galaxy, $G_0 \simeq 1$. In a moderately illuminated cloud such as the Horsehead nebula, $G_0 \simeq 10^2$. In a highly illuminated cloud such as the Orion bar, $G_0 \simeq 10^4$.

1.2 Linking physical conditions and observations: astrophysical numerical models

Astrophysical codes for ISM environments can model observed regions and link numerous observables such as line intensities to physical conditions such as the thermal pressure or the total visual extinction. There exist two main categories of codes: 3D (magneto)hydrodynamics simulations that model the dynamics at large scales, and detailed models of the local physico-chemistry of the gas and dust and of their interaction with the radiation field. In this thesis, we consider the latter category because they solve the microphysics of the cloud, compute chemical abundances and line intensities that can be compared to observations. In this section, we list different types of such numerical models of the ISM.

1.2.1 Radiative transfer models

FUV photons heat the gas by photoelectric effect on grains and excite atoms and molecules in electronic states. Molecules cascade in part in their rovibrational levels by emitting photons in FIR and radio domains. Radiative transfer thus carries two fundamental questions regarding radiative feedback in PDRs:

- What fraction of the stellar UV photons reaches a given position inside the cloud?
- What fraction of the photons emitted at a given position of the cloud escape and contribute to the observable emission and cooling, both for the dust continuum and in ionic, atomic, and molecular lines?

Answering these two questions requires accounting for absorption, scattering, and emission processes for the dust and the gas at each point of the cloud. Taking into account all these processes is challenging. First, the radiative transfer is coupled with quantum level populations of the emitting species for lines, and with the dust temperature for the continuum. Besides, the continuum and the lines are not independent. For instance, the dust continuum can pump the water molecules. Approximations are thus often used in radiative transfer models. A few remarkable models include RADEX, MOLPOP-CEP and Monte Carlo simulators.

RADEX[1] (van der Tak et al., 2007) is a simple code that considers the cloud as spatially uniform in terms of density, temperature, chemical abundances and level populations, and is thus called a 0D code. This code solves the populations in quantum states in non local thermal equilibrium (non-LTE). It takes into account collisional and radiative transitions, but neglects interspecies interactions or pumping by any other line or by the dust continuum. As it models the cloud as completely uniform, it has to rely on a simple approximation for radiative trapping: a photon emitted within the cloud either escapes from the cloud or is absorbed on-the-spot. The code proposes different escape probability approximations. As this code is easy to use, it is widespread in the ISM community despite the coarse approximations it relies on.

MOLPOP-CEP (Asensio Ramos and Elitzur, 2018) is a 1D code that smartly formulates the problem only in terms of level populations. This reformulation allows it to solve the problem of radiative transfer in the lines exactly, although neglecting overlaps between lines and pumping by the dust continuum.

Finally, Monte Carlo simulators approach the radiative transfer differently. Instead of solving approximately the radiative transfer equation, they simulate the stochastic propagation of many individual photons in the cloud, and compute statistics. Such codes include RATRAN[2] (Hogerheijde and van der Tak, 2000), LIME (Brinch and Hogerheijde, 2010), MCFOST (Pinte et al., 2022), or RADMC-3D (Dullemond et al., 2012). As they address more complex geometries, they usually simplify the physical processes solved at each point.

1.2.2 Astrochemical models

Astrochemistry codes are time-dependent 0D simulators that compute the evolution of the chemical abundance of up to hundreds of species accounting for thousands of reactions. A chemical network groups the list of photo-ionization, photodissociation, two-body or more rarely three-body reactions[3], either in gas phase or surface phase, i.e., on the surface of dust grains. The evolution of the volume density n_X of a species X is then controlled by a differential equation:

$$\frac{dn_X}{dt} = F_X - D_X \tag{1.7}$$

where F_X and D_X are formation rate and destruction rates, respectively, i.e., they sum the rates of all reactions in the network that contribute to the formation or destruction of species X. Combined with initial conditions, this set of equations permits to compute the volume density of all the species at any time t. The stationary state can be obtained by solving $F_X = D_X$.

Solving such a system of equations is numerically simple. The complexity lies in building the chemical network and modeling the reactions in ice and surface phases with the associated kinetic constants. Databases such as the KIDA database[4] (Wakelam et al., 2012) and the UMIST database[5] (McElroy et al., 2013) gather the kinetic constants of large chemical networks.

There exist many astrochemical codes based on these databases. NAHOON is associated with the KIDA database and solves networks for the gas phase. This model is not suited to dense

[1] https://home.strw.leidenuniv.nl/~moldata/radex.html
[2] https://home.strw.leidenuniv.nl/~michiel/ratran/
[3] Three-body reactions are rare in the gas phase of the ISM, as it is not dense enough.
[4] https://kida.astrochem-tools.org/
[5] http://udfa.ajmarkwick.net/

molecular regions, which require more processes involving dust grains. More advanced models such as AstroChem (Maret and Bergin, 2015) and UCLCHEM (Holdship et al., 2017) include e.g., freeze-out of species onto dust grains or non-thermal desorption of species from dust grains due to UV photons and cosmic rays. These two models can be used to model a wide variety of environments, including molecular gas and dense cores. More recent and sophisticated models include Nautilus[6] (Ruaud et al., 2016). The results of such time-dependent astrochemical models highly depend on the initial conditions that are poorly known. These codes can yield unexpected results such as oscillations, as demonstrated with the Chimes[7] code (Roueff and Le Bourlot, 2020).

The main limitation of such codes is their 0D geometry. This is problematic for photo-reactions as they do not solve radiative transfer. These models are suited to deep, cold and dark clouds that are protected from UV radiation. In such environments, the chemistry is driven by cosmic rays. The associated characteristic times are hundreds of million of years (Indriolo et al., 2007). Therefore, such environments cannot be considered in a stationary state.

1.2.3 Dust models

Dust grains play three central roles in the physics and chemistry of the PDRs: they absorb UV radiations and thus protect from photoionization and photodissociation, they heat the gas through the photoelectric effect, and they permit the formation of complex molecules on their surface. They also emit a continuum in the FIR domain. An observation of these emissions is called a spectral energy density (SED).

Understanding the many properties of dust grains is an active area of research. There exist different grain population models. Widespread models such as Themis[8] (Jones et al., 2013) or the Astrodust+PAH model (Hensley and Draine, 2023) involve grains mostly made of silicon, carbon, or a mixture. Their structure is either amorphous, crystalline, porous, or aggregated. Finally, the grain size is usually described with a distribution, either a power law, a lognormal distribution or a mixture thereof.

From a grain population model, dust emission models can compute the extinction and emissivity of grains and predict their SED. The most widespread dust emission model is DustEM[9] (Compiègne et al., 2011). Figure 1.8 shows a comparison of an observed SED (in gray) in the diffuse high galactic latitude medium with predictions from DustEM. The contribution to the predicted SED from different types of grain is represented, including polycyclic aromatic hydrocarbonss, small amorphous carbon grains, large amorphous carbon grains, and amorphous silicates.

1.2.4 Holistic models

The aforementioned numerical models focus on a specific subpart of the physics and chemistry of the ISM. Some other models adopt a more holistic approach and account for many physical phenomena at once as well as their couplings. For instance, Cloudy[10] (Ferland et al., 2017) simulates HII regions. The Paris-Durham code[11] (Godard et al., 2019) and the MAPPINGS code[12] (Sutherland et al., 2018) simulate shock-dominated regions, i.e., regions crossed by a shock wave caused e.g., by protostellar outflows, cloud collisions, supernovae, or galactic outflows. ProDimo[13] (Woitke et al., 2009) simulates protoplanetary disks. The Meudon PDR code (Le Petit et al., 2006) and KOSMA-τ[14] (Röllig and Ossenkopf-Okada, 2022) describe UV-irradiated

[6] https://kida.astrochem-tools.org/codes.html
[7] https://ism.obspm.fr/chimes.html
[8] https://www.ias.u-psud.fr/themis/THEMIS_model.html
[9] https://www.ias.u-psud.fr/DUSTEM/index.html
[10] https://trac.nublado.org/
[11] https://ism.obspm.fr/shock.html
[12] https://mappings.anu.edu.au/code/index.html
[13] https://prodimo.iwf.oeaw.ac.at/
[14] https://astro.uni-koeln.de/riechers/research/kosma-tau

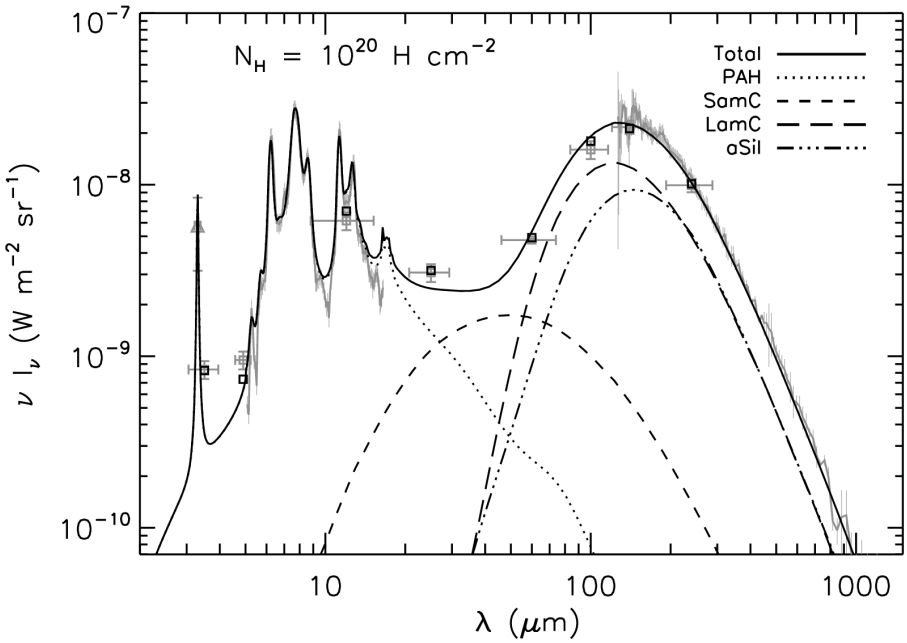


Figure 1.8: Illustration of a spectral energy density for dust. The contributions of dust grain subpopulations are represented, such as polycyclic aromatic hydrocarbonss (PAHs), small amorphous carbon grains (SamC), large amorphous carbon grains (LamC), and amorphous silicates (aSi). Taken from Compiègne et al. (2011).

medium at the edge of molecular clouds in star forming regions or diffuse gas, i.e., PDRs.

All these models compute the chemical and physical structure resulting from a coupled treatment of chemistry, radiative transfer, and thermal processes. To account for all these phenomena and their coupling while maintaining reasonable evaluation times, these models rely on approximations such as the stationary state or a 1D geometry – except for ProDimo that handles a 2D geometry, but resorts to other approximations. Despite these approximations, these models are usually slower and more computationally expensive than the aforementioned codes that address only a specific subpart of the physics and chemistry.

1.3 The Meudon PDR code

In this section, we give a short overview of the Meudon PDR code[15] (Le Petit et al., 2006). It is considered state-of-the-art PDR model as it includes many physical phenomena, which also causes each evaluation to last a few hours. This code is used in the remainder of this thesis to model each pixel individually in map inversion procedures. It can simulate a large variety of physical environments, such as diffuse clouds, dense PDRs, diffuse gas in the galactic center, nearby galaxies, damped Lyman alpha systems, circumstellar disks, etc. Indeed, part of the observable emissions of each of these environments comes from molecular gas irradiated by UV photons. In particular, this code permits to study effects such as the radiative feedback of a newborn star on its parent molecular cloud.

1.3.1 Physics and chemistry taken into account

The Meudon PDR code simulates the physics and chemistry of neutral interstellar gas illuminated with a FUV radiation field and computes its stationary state. The cloud is assumed to be a plane-parallel slab of neutral gas and dust with finite thickness, and is thus modeled with a

[15] https://ism.obspm.fr

one-dimensional geometry – the Meudon PDR code is called a 1D code. From a set of physical conditions of the cloud, the code iteratively solves large systems of coupled equations to compute the gas temperature, the chemical abundances and the excitation of species at each point of an adaptive grid. These equations include radiative transfer, thermal balance and chemistry. These intermediate and local results permit for example to compute integrated intensities comparable with observations. A full evaluation is computationally intensive and typically lasts a few hours. Figure 1.9 lists some physical phenomena taken into account for each physical aspect.

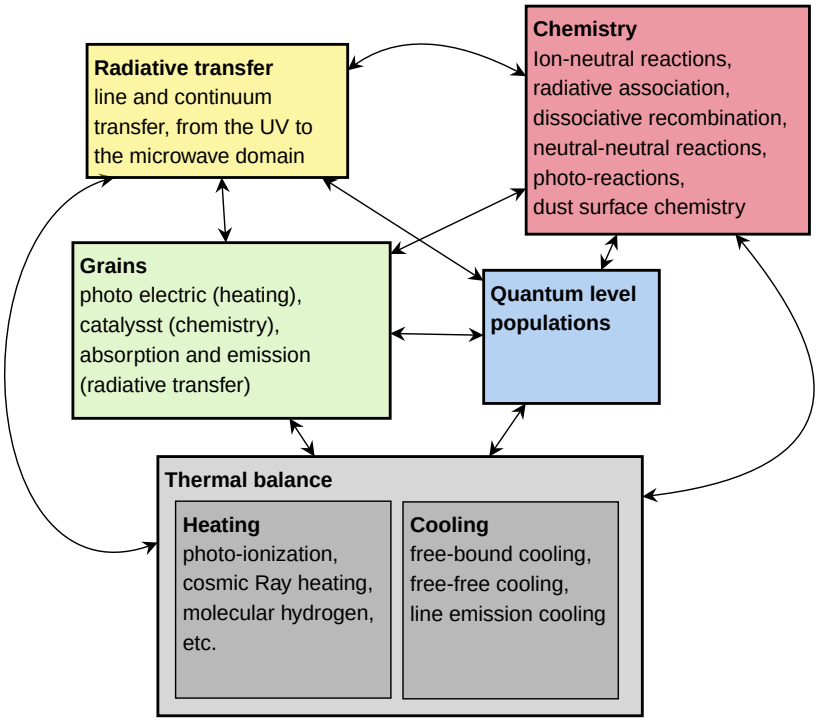


Figure 1.9: Summary of Meudon PDR code modeling components.

Radiative transfer equations are solved at each position on an adaptive spatial grid in a fully wavelength-dependent approach. In addition, the angular dependence is handled using a Legendre polynomial decomposition. Multiple phenomena are accounted for such as the redistribution across angles by dust scattering and the continuum absorption by dust and gas, including absorption in UV lines of H and H_2. For emission lines, the escape probability approximation of de Jong et al. (1980) is used. An emitted photon is therefore either absorbed on the spot or leaves the cloud.

The temperature of the gas at each position is solved at thermal balance,

$$\Gamma(T) = \Lambda(T), \tag{1.8}$$

where $\Gamma(T)$ and $\Lambda(T)$ are the heating rates of all heating and cooling processes, respectively. This equation is solved at each position of the grid to compute the gas temperatures from the specific intensity of the obtained radiation field. The heating rate $\Gamma(T)$ includes a large range of processes such as photoelectric heating, collisional de-excitation of UV-pumped H_2, photoionization and photodissociation heating, cosmic ray heating, and exothermic chemical reactions. The cooling rate $\Lambda(T)$ is estimated from the non local thermal equilibrium (non-LTE) excitation in the quantum levels of the main species by considering radiative and collisional processes as well as chemical formation and destruction. In total, the rovibrational states are considered for a few tens of species, and electronic states for some species such as H_2. Additional processes can either heat or cool the gas, such as H_2 heating or gas-grain collisions.

Finally, the chemical composition of the gas is computed at each position by solving for the chemical stationary state of a network of 200 species and $3\,000$ reactions. In the gas phase, the

code accounts for two body reactions, photo-reactions – with primary and secondary photons –, reactions induced by cosmic rays, and some three body reactions. On the surface of dust grains, adsorption and thermal and non-thermal desorption are included. Finally, in ices, the code includes two body reactions due to thermal agitation and tunnel effect diffusions. The chemical reaction network was built combining different sources including data from the KIDA database and the UMIST database as well as data from articles. For key photo-reactions, cross-sections are taken from Ewine van Dishoeck's photodissociation and photoionization database[16] (Heays et al., 2017).

The successive resolution of these three highly coupled aspects – radiative transfer, thermal balance, and chemistry equations – is iterated until a global stationary state is reached. After a run, the code provides density profiles of the chemical species and the temperature profiles of both the grains and the gas. It also outputs the line intensities emerging from the cloud that can be compared to observations. As of version 7, yet to be released, a total of $5\,409$ line intensities are predicted from 40 species such as H_2, HD, C^+, C, CO, ^{13}CO, $C^{18}O$, $^{13}C^{18}O$, SO, HCO^+, HCN, HNC, CH^+, CN or CS.

1.3.2 Input parameters and considered environments

For an input set of physical parameters, the Meudon PDR code computes the structure of the cloud, emitted line integrated intensities, and the dust continuum intensities. The predicted and observable intensities can be compared with actual observations to constrain the input physical parameters. The physical parameters to be estimated in this thesis characterize the radiative feedback. These parameters are the thermal pressure that describes local conditions, the visual extinction that quantifies the total amount of matter along the line of sight, and the intensity of the incident UV radiative field illuminating the surface of the PDR. Parameters associated with cosmic rays and dust grains are set to default values. Table 1.3 lists these default values.

Table 1.3: Secondary input parameters in the Meudon PDR code and their default values.

Parameter	Value	Unit	Note
cosmic rays ionization rate	10^{-16}	s^{-1}	Le Petit et al. (2004), Indriolo et al. (2007)
Dust extinction curve	Galaxy	–	Fitzpatrick and Massa (1990)
R_V	3.1	–	Fitzpatrick and Massa (1990)
$C_D = N_H/E(B-V)$	5.8×10^{21}	cm^{-2}	Bohlin et al. (1978)
Mass grain/Mass gas	0.01	–	–
Distribution on grain size a	$\propto a^{-3.5}$	–	Mathis et al. (1977)
Min grain radius $a_{\min}$	10^{-7}	cm	–
Max grain radius $a_{\max}$	3×10^{-5}	cm	–
Turbulent velocity	2.0	km s^{-1}	–

Constant density and constant pressure models – The Meudon PDR code handles constant density and constant pressure models. Constant density models result in large pressure gradients in the cloud, as the temperature strongly drops in the H / H_2 transition region. They are therefore not very consistent with the stationarity hypothesis. However, they are quite common in astrophysics as they are easier to interpret. Constant pressure models are more consistent with the stationarity hypothesis and generally considered more realistic (Marconi et al., 1998; Lemaire et al., 1999; Allers et al., 2005; Goicoechea et al., 2016; Joblin et al., 2018; Wu et al., 2018). We restrict ourselves to constant pressure models. However, the thermal pressure is harder to physically interpret than volume density, which plays a role both in chemistry and radiative

[16] https://home.strw.leidenuniv.nl/~ewine/photo/index.html

transfer – for instance with critical densities.

Total visual extinction – As already mentioned, the Meudon PDR code relies on an adaptive one-dimensional grid of a plane-parallel slab of gas. The grid of the Meudon PDR is a grid on the visual extinction A_V, which quantifies the amount of matter along the line of sight. The total depth of the cloud is thus defined with a total visual extinction A_V^{tot}.

Radiation field – The Meudon PDR code permits the use of any incident radiation field, including a stellar radiation field. For simplicity, in this thesis, we consider the radiation field to be the Mathis ISRF Mathis et al. (1983) scaled at the front of the cloud by one scalar parameter denoted radm. The conversion of radm to G_0 is done with

$$G_0 = \frac{1.2786}{2}\text{radm}, \tag{1.9}$$

where the 1.2786 factor comes from the conversion from the Mathis to Habing ISRF, and the $\frac{1}{2}$ factor from the fact that the front face of the cloud is illuminated from the outside of the cloud, and not from the inside.

1.3.3 Limitations

The geometry simplification can affect integrated intensities and the effective depth of penetration of UV photons. For instance, under an isotropic external radiation field, UV photons should penetrate deeper in a spherical cloud than in a plane-parallel cloud. Other PDR codes assume different cloud geometries, e.g., the KOSMA-τ numerical code (Röllig and Ossenkopf-Okada, 2022) assumes a 1D spherical geometry.

The stationarity assumption neglects gas dynamics and non-equilibrium effects. In other words, the timescales of all physical and chemical processes taken into account must be smaller than that of the changes in the external conditions, and of the gas dynamics. In a PDR, the physical processes are dominated by the photodissociation of H_2 by UV photons. The H_2 photodissociation rate at the surface of a cloud is $2.9 \times 10^{-11} G_0\,\mathrm{s}^{-1}$. For $G_0 = 10^4$, the characteristic time of H_2 photodissociations is about 40 days. As this duration is very short compared to gas dynamics timescales, the stationarity hypothesis is realistic. However, the UV photons do not penetrate much beyond the photodissociation front. In deeper clouds, the H_2 dissociation is dominated by cosmic rays. As indicated in Table 1.3, the cosmic ray ionization rate is $\simeq 10^{-16}\,\mathrm{s}^{-1}$ (Indriolo et al., 2007). The corresponding characteristic time is about $300\,\mathrm{Myr}$. Therefore, such clouds require time-dependent chemistry models.

Finally, the neutral gas assumption means that in the modeled region, hydrogen is either in its atomic form (H) or its molecular form (H_2) but never in its ionized form (H^+), even in the atomic layer, close to the ionization front. Other PDR codes that relax these hypotheses exist. For instance, the Hydra PDR code (Bron et al., 2018a) relaxes the stationarity and neutrality assumptions by including gas hydrodynamics and photo-evaporation. However, this code simplifies the modeling of other phenomena compared to the Meudon PDR code in order to maintain reasonable computational durations, i.e., a few hours per evaluation.

1.4 Conclusion

The ISM is a complex and inhomogeneous medium that fills the unimaginably large volume that lies between stars. It is made mostly of hydrogen and helium. It also contains traces of other elements and small dust grains that play a key role in its chemistry and physics. In this thesis, we focus on one particular type of environment called photodissociation region (PDR) that lies at the border between a massive star or cluster of stars and a molecular cloud. Studying PDRs permits to understand better the impact of radiative feedback from newborn stars on their parent cloud and its star forming capability.

Observations with high spatial resolution such as the Orion B map described in Pety et al. (2017), the OMC-1 map described in Goicoechea et al. (2019), and future observations from ALMA or the JWST contain much physical information regarding the radiative feedback. In this thesis, our goal is to extract this information. Our aim is to infer maps of the key physical parameters from such large multiline PDR observations. We model PDRs with the Meudon PDR code. As it iteratively solves large, coupled, and complex non-linear sets of equations, each evaluation of the Meudon PDR code requires from a few hours to a few days. In Chapter 4, we derive a fast and accurate neural-network based approximation of the Meudon PDR code to perform fast inference on large observation maps. In Chapter 5, we propose a new statistical inference algorithm that exploits this Meudon PDR code approximation, and validate it on synthetic data. In Chapter 6, this statistical inference algorithm is applied to multiple real observations.

Chapter 2

Background on Bayesian statistical modeling and inference

" The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which of times they are unable to account. "

Pierre-Simon Laplace, in Théorie Analytique Des Probabilités, 1814

Contents

In Chapter 1, we described the interstellar medium (ISM) and some numerical models that can be used to infer physical parameters from observations. This chapter describes notions and tools necessary to formulate and solve the inverse problem considered in this thesis – see Chapter 5. Some of these methods are already popular in ISM studies, as we will show in Chapter 3.

Section 2.1 describes the general Bayesian statistical modeling approach, the associated random variables, probability distributions and estimators. Section 2.2 presents two common approaches for estimator derivation, namely optimization methods and sampling methods. In particular, methods based on first and second order derivatives are covered to exploit the twice differentiable log-posterior and address its absence of gradient Lipschitz continuity. Block coordinate algorithms are also described to exploit the natural structure of the parameter $\boldsymbol{\Theta}$. The sampler proposed in Chapter 5 (Section 5.2) builds on the methods presented in this section. Finally, Section 2.3 depicts two methods that assess the compatibility of the model with the observations, namely model selection and model checking.

2.1 Bayesian statistical modeling

This section covers fundamental notions of Bayesian statistical modeling. First, the main variables and probability distributions involved in statistical modeling are reviewed, including the likelihood function and posterior distribution. Then, the main estimators are described.

2.1.1 Main random variables and probability distributions

The Bayesian framework is a statistical approach that encodes and describes uncertainty using random variables and probability distributions. In this context, a probability distribution indicates a of belief and not a limit of frequencies as in frequentist statistics. Bayesian statistics exploits random variables to model uncertainty in any quantity of interest such as observations, physical parameters, or model hyperparameters. For a review on random variables and probability distributions, see Le Gall (2022).

There are two main random variables of interest in Bayesian statistics: the physical parameters to infer $\boldsymbol{\Theta} \in \mathbb{R}^{N \times D}$ and the observation $\mathbf{Y} \in \mathbb{R}^{N \times L}$. Assuming an observation model $\mathcal{M}$, the Bayes theorem states that the conditional probability density functions (pdfs) of these variables verify

$$\pi\left(\boldsymbol{\Theta}|\mathbf{Y}, \mathcal{M}\right) = \frac{\pi\left(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M}\right) \pi\left(\boldsymbol{\Theta}\right)}{\pi\left(\mathbf{Y}|\mathcal{M}\right)}, \tag{2.1}$$

where $\pi(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M})$ is the *likelihood function*, $\pi(\boldsymbol{\Theta})$ is the pdf of the *prior distribution*, $\pi(\boldsymbol{\Theta}|\mathbf{Y}, \mathcal{M})$ is the pdf of the *posterior distribution*, and $\pi(\mathbf{Y}|\mathcal{M})$ is the *Bayesian evidence*.

The **likelihood function** $\pi(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M})$ evaluates how well the parameter $\boldsymbol{\Theta}$ and the model $\mathcal{M}$ reconstruct the observations $\mathbf{Y}$. It is the pdf of a distribution on the observation $\mathbf{Y}$ given a fixed value for the physical parameters $\boldsymbol{\Theta}$. In inference, the observation $\mathbf{Y}$ is fixed and the free variable is $\boldsymbol{\Theta}$. The denomination "likelihood function" emphasizes that $\boldsymbol{\Theta} \mapsto \pi(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M})$ is not a pdf. The likelihood function is derived from an *observation model* $\mathcal{M} = (\mathbf{f}, \mathcal{A})$ that groups a forward model $\mathbf{f}$, generally a numerical simulator in astrophysics, and a model noise $\mathcal{A}$ such that

$$\mathbf{Y} = \mathcal{A}\left(\mathbf{f}(\boldsymbol{\Theta})\right). \tag{2.2}$$

Details for astrophysicists 2.1: Example: Gaussian additive and uncorrelated noise

For a Gaussian additive and uncorrelated noise, the observation model in Eq. 2.2 reads

$$\forall n \in [\![1, N]\!], \quad \ell \in [\![1, L]\!], \quad y_{n\ell} = f_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)}, \quad \varepsilon_{n\ell}^{(a)} \sim \mathcal{N}(0, \sigma_{n\ell}^2). \tag{2.3}$$

Using the pdf of a Gaussian distribution, this observation model translates to the likelihood function

$$\pi(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M}) \propto \prod_{n=1}^{N} \prod_{\ell=1}^{L} \exp\left[-\frac{1}{2\sigma_{n\ell}^2} \left(f_\ell(\boldsymbol{\theta}_n) - y_{n\ell}\right)^2\right], \tag{2.4}$$

and to the negative log-likelihood function

$$-\ln \pi(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M}) = \sum_{n=1}^{N} \sum_{\ell=1}^{L} \frac{1}{2\sigma_{n\ell}^2} \left(f_\ell(\boldsymbol{\theta}_n) - y_{n\ell}\right)^2, \tag{2.5}$$

up to an additive constant. The χ^2 loss function, widespread in ISM studies, is equivalent to the negative log-likelihood function in Eq. 2.5 for this specific case of Gaussian additive and uncorrelated noise. The negative log transform turns the product to a sum and removes the exponential, which makes derivative computations easier.

Note that in this thesis, the natural logarithm of $x \in \mathbb{R}$ is denoted $\ln x$. This notation is preferred to the usual $\log$ to avoid any confusion between logarithms in base e and 10.

In this thesis, the forward model $\mathbf{f}$ is set to the Meudon PDR code, introduced in Chapter 1 (Section 1.3). This forward model maps physical parameter vectors $\boldsymbol{\theta} \in \mathbb{R}^D$ to observables $f_\ell(\boldsymbol{\theta}) \in \mathbb{R}^L$. Therefore, evaluating the Meudon PDR code for a map $\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)_{n=1}^N$ requires N evaluations: $\mathbf{f}(\boldsymbol{\Theta}) = (\mathbf{f}(\boldsymbol{\theta}_n))_{n=1}^N$. To accelerate the inference procedure, we proposed an approximation of the Meudon PDR code based on an artificial neural network with controlled error. The derivation of this approximation is the core of Chapter 4. For a review of how numerical models are handled in inference procedures in ISM studies, see Chapter 3 (Section 3.1.1). Similarly, Section 3.4 introduces the noise model considered in this thesis. For a review on noise models used in astrophysics, see Section 3.1.2.

The **prior distribution** $\pi(\boldsymbol{\Theta})$ encodes prior information on the physical parameters of interest $\boldsymbol{\Theta}$. It may be non-informative, such as a uniform distribution on a set of parameters, or more informative, e.g., to include the results of other trusted studies or to constrain acceptable solutions to be physically meaningful. Similarly to the likelihood function, using the negative log-prior $-\ln \pi(\boldsymbol{\Theta})$ simplifies computations. In the optimization framework, $-\ln \pi(\boldsymbol{\Theta})$ is often called a *regularization function*.

Chapter 3 (Section 3.1.3) reviews prior distributions used in ISM studies. Chapter 5 (Section 5.1.2) describes the prior distribution considered in this thesis.

The **posterior distribution** $\pi(\boldsymbol{\Theta}|\mathbf{Y}, \mathcal{M})$ combines the likelihood function and the prior distribution. Like the prior distribution, it is a distribution on $\boldsymbol{\Theta}$. The posterior describes the parameters of interest knowing the observations. It is the distribution from which estimators of the physical parameters $\boldsymbol{\Theta}$ are extracted, as well as the associated uncertainty. The posterior distribution associated with the inverse problem considered in this thesis is introduced in Chapter 5 (Section 5.1.3).

In some simple cases, the likelihood function and prior distribution enjoy a *conjugacy relation* (Robert and Casella, 2004, chapter 1). The posterior distribution is then an element of a

parametric family of distributions with closed-form formulae for its parameters. However, these simple cases are very limited and do not occur in astrophysical inference problems. For instance, there exists no conjugacy relation when the forward model is non-linear.

The **Bayesian evidence** $\pi(\mathbf{Y}|\mathcal{M})$ is a normalization constant that does not depend on the physical parameter $\boldsymbol{\Theta}$ and that is generally challenging to evaluate. It is also called *marginal likelihood*, as it marginalizes over the physical parameter $\boldsymbol{\Theta}$

$$\pi(\mathbf{Y}|\mathcal{M}) = \int \pi(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M})\, \pi(\boldsymbol{\Theta})\, d\boldsymbol{\Theta}. \tag{2.6}$$

Many approaches that solve inverse problems require the assumption of a model $\mathcal{M} = (\mathbf{f}, \mathcal{A})$[1]. The Bayesian evidence permits to assess the consistency of the model with the observations. It is a crucial quantity for Bayesian model selection, as we will discuss in Section 2.3.1. However, it is not always necessary for statistical inference. In inference, Eq. 2.1 is therefore often simplified to

$$\pi(\boldsymbol{\Theta}|\mathbf{Y}, \mathcal{M}) \propto \pi(\mathbf{Y}|\boldsymbol{\Theta}, \mathcal{M})\ \pi(\boldsymbol{\Theta}), \tag{2.7}$$

where $\propto$ means "proportional to", i.e., equal up to a multiplicative constant independent of $\boldsymbol{\Theta}$. The model $\mathcal{M}$ remains unchanged during statistical inference. Writing explicitly the dependence to the model $\mathcal{M}$ is useful when evaluating its relevance or comparing multiple models. Otherwise, this dependence is not explicitly written to avoid heavy notations. Therefore, Eq. 2.7 will often be simplified to

$$\pi(\boldsymbol{\Theta}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\boldsymbol{\Theta})\ \pi(\boldsymbol{\Theta}). \tag{2.8}$$

Hyperparameters can have a dramatic impact on the shape of the posterior distribution. They can appear in both the prior distribution and the likelihood function. For instance, a Gaussian prior on $\boldsymbol{\Theta}$ would be defined with a mean vector and a covariance matrix. Similarly, the likelihood function obtained with a Gaussian additive noise model is defined with a noise covariance matrix. These hyperparameters greatly influence the shape of the posterior distribution. Hierarchical models (Gelman et al., 2015, chapter 5) enable inferring simultaneously both $\boldsymbol{\Theta}$ and hyperparameters from the observations $\mathbf{Y}$. For now, we assume hyperparameters set to constant values, and restrict ourselves to non-hierarchical models.

2.1.2 Estimators

Two main types of estimators can be defined from the likelihood function and the posterior distribution: some are solutions of an optimization problem, while others are defined as an integral.

Estimators solutions of an optimization problem – The point that maximizes the likelihood function is by definition the point that best reproduces observations. It is called the maximum likelihood estimator (MLE), and denoted $\widehat{\boldsymbol{\Theta}}_{\text{MLE}}$. It is a natural estimator with good theoretical properties (Robert and Casella, 2004, chapter 1). Similarly, the posterior mode is the point that best balances observation reproduction and consistency with prior knowledge. It is called the maximum a posteriori (MAP), and denoted $\widehat{\boldsymbol{\Theta}}_{\text{MAP}}$. The MLE and the MAP are defined as solutions of a maximization problem, often converted to a minimization problem in log-scale:

$$\widehat{\boldsymbol{\Theta}}_{\text{MLE}} \in \underset{\boldsymbol{\Theta}\in\mathbb{R}^{N\times D}}{\arg\max}\, \pi(\mathbf{Y}|\boldsymbol{\Theta}) = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{N\times D}}{\arg\min} -\ln \pi(\mathbf{Y}|\boldsymbol{\Theta}), \tag{2.9}$$

and

$$\widehat{\boldsymbol{\Theta}}_{\text{MAP}} \in \underset{\boldsymbol{\Theta}\in\mathbb{R}^{N\times D}}{\arg\max}\, \pi(\boldsymbol{\Theta}|\mathbf{Y}) = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{N\times D}}{\arg\min} -\ln \pi(\boldsymbol{\Theta}|\mathbf{Y}). \tag{2.10}$$

[1] An example of approach that does not rely on the assumption of a model resorts to end-to-end neural networks – see e.g., conditional invertible neural networks (Ardizzone et al., 2019).

Details for astrophysicists 2.2: Well-posed and ill-posed inverse problems

Inverse problems are usually divided in two classes: *well-posed* problems and *ill-posed* problems.

In well-posed problems, the observations alone are sufficiently informative to yield physical parameter accurate estimations. In astrophysics, these problems are said to be *well constrained*. The MLE is very useful in such problems.

Conversely, in ill-posed problems, the observations may come in low SNR regime, or the observables may not be good tracers of the physical parameter of interest. In astrophysics, such problems are called *not constrained* or *badly constrained*. In such cases, the MLE leads to unstable results. By exploiting regularizing prior knowledge, the MAP can be more robust to noise and to non-linearities.

Estimators defined as an integral – Many estimators are defined as an integral, including the posterior mean and covariance. The posterior mean $\mathbb{E}[\boldsymbol{\Theta}|\mathbf{Y},\mathcal{M}]$ is the point estimate that minimizes the mean squared error $\mathbb{E}[\|\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\|_2^2 \mid \mathbf{Y},\mathcal{M}]$, i.e., the average squared distance with other points $\boldsymbol{\Theta}$ drawn from the posterior. This estimator is therefore usually called the minimum mean square error (MMSE), and denoted $\widehat{\boldsymbol{\Theta}}_{\text{MMSE}}$. The posterior covariance $\text{Cov}(\boldsymbol{\Theta}|\mathbf{Y},\mathcal{M})$ permits the detection of degeneracies between physical parameters[2]. In general, multiple estimators can be written as the integral of a function g on $\boldsymbol{\Theta}$ with respect to the posterior,

$$\mathbb{E}\left[g(\boldsymbol{\Theta})|\mathbf{Y}\right] \;=\; \int g(\boldsymbol{\Theta})\;\pi(\boldsymbol{\Theta}|\mathbf{Y})\;d\boldsymbol{\Theta}. \tag{2.11}$$

For instance, the posterior mean $\mathbb{E}\left[\boldsymbol{\Theta}|\mathbf{Y}\right]$ corresponds to $g:\boldsymbol{\Theta}\mapsto\boldsymbol{\Theta}$, i.e., the identity function. Similarly, the posterior covariance matrix corresponds to $g:\boldsymbol{\Theta}\mapsto(\boldsymbol{\Theta}-\mathbb{E}\left[\boldsymbol{\Theta}|\mathbf{Y}\right])(\boldsymbol{\Theta}-\mathbb{E}\left[\boldsymbol{\Theta}|\mathbf{Y}\right])^T$.

A more accurate uncertainty quantification on physical parameters than the posterior covariance can be obtained with a credible region (Pereyra, 2017). A credible region C_α with confidence level $1-\alpha$ verifies

$$\mathbb{P}\left[\boldsymbol{\Theta}\in C_\alpha\right] = \int \mathbb{1}_{C_\alpha}(\boldsymbol{\Theta})\;\pi(\boldsymbol{\Theta}|\mathbf{Y})d\boldsymbol{\Theta} = 1-\alpha. \tag{2.12}$$

A posterior distribution admits infinitely many credible regions C_α. The *highest posterior density region* C_α^* is the credible region with minimum volume. It is defined as

$$C_\alpha^* = \{\boldsymbol{\Theta}\mid -\ln\pi(\boldsymbol{\Theta}|\mathbf{Y})\leq c_\alpha\} \tag{2.13}$$

where $c_\alpha\in\mathbb{R}$ is set so that Eq. 2.12 is verified with C_α^*. For one parameter θ_{nd}, a credibility interval corresponds to the highest posterior density region $\pi(\theta_{nd}|\mathbf{Y})$ with all other parameters being marginalized. When the marginalized posterior $\pi(\theta_{nd}|\mathbf{Y})$ is unimodal, the $1-\alpha$ credibility interval is defined with the $\alpha/2$ and $1-\alpha/2$ percentiles. These percentiles which can be obtained with the inverse cumulative density function (cdf) of $\pi(\theta_{nd}|\mathbf{Y})$.

2.2 Statistical inference

This section covers fundamental methods of statistical inference that numerically compute the aforementioned estimators. Estimators defined as a minimum of a loss function are evaluated through optimization methods. Estimators defined with potentially high dimensional integrals are approximated with Monte Carlo (MC) estimators, which requires sampling from the posterior

[2] A degeneracy between physical parameters corresponds to an *ill-conditioned* Hessian matrix. The notion of condition number of a matrix will be introduced in Section 2.2.1.2.

distribution. Unlike the optimization approach, this sampling approach thus permits the use of one algorithm to evaluate multiple estimators. It can therefore extract more information from the posterior distribution, in particular regarding uncertainty. In this thesis, due to the absence of ground truth in ISM studies, we address inverse problems with a sampling approach.

The optimization approach, that we do not use for inference, is presented for three reasons. First, it is more intuitive than the sampling approach, and many links can be made between their respective methods. We believe that knowing optimization methods simplifies the understanding of sampling methods for the unfamiliar reader. Second, some proposed sampling methods are inspired from optimization methods that thus need to be introduced. Finally, as we will show in Chapter 3 (Section 3.2.1), optimization methods are widely exploited in ISM studies.

This section only covers methods that are used to derive the inference algorithm proposed in Chapter 5. In particular, variational Bayes inference and proximal methods are not reviewed. Variational Bayes inference is an intermediate approach between optimization and sampling. It consists in projecting the true posterior onto a class of simple distributions. This target class is generally chosen so that it is easy to extract information. For instance, for a Gaussian approximation, the mean vector and covariance matrix are adjusted to approximate the posterior distribution. Other insights such as credibility intervals are then very simple to derive from the approximating distribution. See e.g., Pereyra et al. (2016) for a review. These methods require prior knowledge of the shape of the posterior distribution to select a relevant class of approximating distribution. Because of the non-linear and non gradient Lipschitz continuous forward model, and of the complex noise model, we do not have such prior knowledge. Besides, the quality of the approximation greatly impacts the relevance of the provided uncertainty quantification. Therefore, we decided not to consider these methods, and do not introduce them formally in this thesis.

Proximal gradient descent and proximal sampling algorithms (Pereyra et al., 2016) are now widespread methods that permit to address cases where the posterior distribution is not differentiable. In the inverse problem considered in this thesis, the log-posterior is twice differentiable. Algorithms based on the proximal operator are therefore not covered in this thesis.

2.2.1 Evaluating estimators defined as solutions of optimization problems

Some estimators $\widehat{\boldsymbol{\Theta}}$ are defined as the minimum of a *loss function*, denoted $\mathcal{L}$, as described in the previous section. For instance, this loss function $\mathcal{L}$ corresponds to the negative log-likelihood for the MLE and to the negative log-posterior for the MAP. In most cases, such estimators cannot be written in closed-form. They are therefore numerically approximated. The most naive method is to look for the precomputed model that leads to the lowest loss value among a set of precomputed models. This method is usually called a *grid search*. It scales very poorly as the number of required precomputed models increases exponentially with the dimension. A wide variety of methods permit better and cheaper evaluation of these two point estimations. In the following, we describe the main alternative method, the gradient descent (GD) algorithm, and its preconditioned and block coordinate variants.

2.2.1.1 Gradient descent (GD) algorithm

In its simplest version, the gradient descent (GD) algorithm starts at some state $\boldsymbol{\Theta}^{(0)}$, and updates it with

$$\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}^{(t-1)} - \eta \nabla \mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right), \tag{2.14}$$

with $\eta > 0$ a step size hyperparameter.

The matrices $\boldsymbol{\Theta} \in \mathbb{R}^{N \times D}$ are converted to vectors of $\mathbb{R}^{ND}$ in lexicographic order to evaluate the gradient. In the remainder of this thesis, this vector form of the physical parameter $\boldsymbol{\Theta}$ is considered whenever a first or second order derivative is used. Iterations of the GD algorithm are

repeated until some stopping criterion is satisfied. This stopping criterion typically corresponds to a maximum number T of iterations or to a lower threshold on the gradient norm $\|\nabla\mathcal{L}\left(\boldsymbol{\Theta}^{(t)}\right)\|$ or, equivalently, on the distance of two successive iterates $\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^{(t-1)}\|$. Algorithm 2.1 summarizes the GD algorithm.

Algorithm 2.1: Gradient descent (GD) algorithm

Input: Starting point $\boldsymbol{\Theta}^{(0)}$, step size $\eta > 0$.

$t = 0$

while *Stopping criterion not satisfied* **do**

  Update $\boldsymbol{\Theta}^{(t)}$ with $\boldsymbol{\Theta}^{(t-1)}$ and $\nabla\mathcal{L}$ `// using Eq. 2.14, Eq. 2.17 or Eq. 2.22`

  $t = t + 1$

Output: Last iterate $\boldsymbol{\Theta}^{(t)}$

The gradient descent (GD) algorithm (Shalev-Shwartz and Ben-David, 2014, chapter 14) is generally preferred to grid search, as it generally yields better estimates $\widehat{\boldsymbol{\Theta}}$, i.e., estimates with lower loss value $\mathcal{L}\left(\widehat{\boldsymbol{\Theta}}\right)$, with less function evaluations. This advantageous comparison is due to two strengths of the GD algorithm. First, it exploits more information from the loss function $\mathcal{L}$, with its first and potentially second order derivative information. Second, its iterative nature exploits past evaluations of the loss functions to progressively converge to a critical point, potentially a local or global minimum.

Convergence properties – GD algorithms converge to critical points, i.e., points $\boldsymbol{\Theta}$ such that $\nabla\mathcal{L}(\boldsymbol{\Theta}) = 0$, under different sets of assumptions. Theoretical guarantee of convergence can be obtained for a loss function whose gradient $\nabla\mathcal{L}$ is a *Lipschitz continuous* function, i.e., if there exists a constant $\beta > 0$ such that

$$\forall\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \quad \|\nabla\mathcal{L}(\boldsymbol{\Theta}_1) - \nabla\mathcal{L}(\boldsymbol{\Theta}_2)\| \leq \beta\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|. \tag{2.15}$$

Such loss functions $\mathcal{L}$ are called *gradient Lipschitz continuous*, with *Lipschitz constant* β.

Details for astrophysicists 2.3: Two examples for gradient Lipschitz continuity

Example 1: gradient Lipschitz continuous – The gradient of the $x \in \mathbb{R} \mapsto x^2$ is $x \mapsto 2x$, which admits 2 as a Lipschitz constant. The function $x \in \mathbb{R} \mapsto x^2$ is therefore gradient Lipschitz continuous with constant $\beta = 2$.

Example 2: non gradient Lipschitz continuous – The gradient of $x \in \mathbb{R} \mapsto \exp(x^2)$ is $x \mapsto 2x\exp(x^2)$, which does not admit a Lipschitz constant. This second function is therefore not gradient Lipschitz continuous.

For non-convex and gradient Lipschitz continuous loss functions, convergence to a critical point is guaranteed when the step size η satisfies $\eta < 2/\beta$ (Beck, 2017, Theorem 10.15)[3]. Smaller step sizes lead to slower convergence, and the GD algorithm does not converge for values larger than the $2/\beta$ limit. In many imaging inverse problems, the log-posterior is gradient Lipschitz continuous with a known Lipschitz constant. In general, however, the log-likelihood and log-posterior are not gradient Lipschitz continuous, or the constant is either hard to compute or too large to be of any practical use.

[3] This theorem holds for proximal gradient descent. As the gradient descent algorithm is a special case of proximal gradient descent, the result also holds for gradient descent.

2.2.1.2 Preconditioned variants

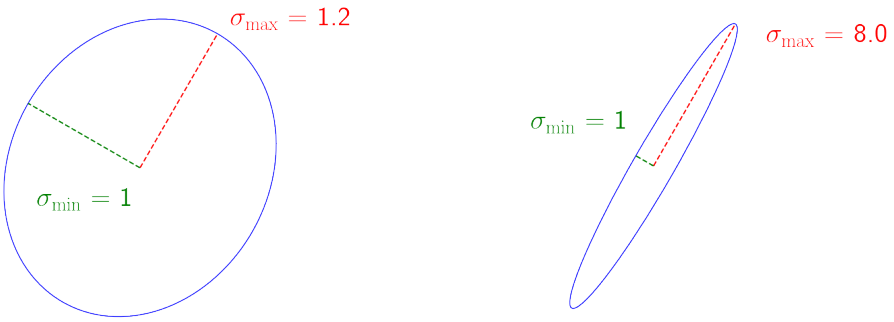
The update step from Eq. 2.14 can be prohibitively slow to converge, especially when $\boldsymbol{\Theta} \in \mathbb{R}^{N \times D}$ lives in high dimension or when the dynamics of its components θ_{nd} – $n \in [\![1, N]\!]$, $d \in [\![1, D]\!]$ – are very different. In this section, we introduce the condition number, that indicates how slowly the standard GD will converge. Then, we describe preconditioning techniques that accelerate convergence by locally reshaping the loss function.

Condition number and preconditioning – The condition number κ of the Hessian matrix $\nabla^2 \mathcal{L}(\boldsymbol{\Theta})$ quantifies the difference in dynamic ranges between components θ_{nd} or their combinations. It is defined as

$$\kappa\left(\nabla^2 \mathcal{L}(\boldsymbol{\Theta})\right) = \frac{\sigma_{\max}\left(\nabla^2 \mathcal{L}(\boldsymbol{\Theta})\right)}{\sigma_{\min}\left(\nabla^2 \mathcal{L}(\boldsymbol{\Theta})\right)} \geq 1, \tag{2.16}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ denote the maximal and minimal singular values of a matrix, respectively. These singular values are the absolute value of the eigenvalues. When $\kappa \simeq 1$, the Hessian matrix is said to be *well-conditioned*. It equals one, e.g., for the identity matrix. Conversely, when $\kappa \gg 1$, the Hessian matrix is said to be *ill-conditioned*.

Figure 2.1 shows two ellipses, each corresponding to a contour level of a quadratic function, i.e., with constant Hessian matrices. In Figure 2.1a, the condition number is close to 1. There is a slight anisotropy, but the dynamic ranges are roughly the same for both variables. Conversely, in Figure 2.1b, the condition number is 8. There is a strong anisotropy, and the dynamic ranges are not the same for the two variables. The GD algorithm performs best in the first case, and converges slowly in the second case – see e.g., Nocedal and Wright (2006, theorem 3.4).



(a) Low condition number (1.2). (b) High condition number (8).

Figure 2.1: Illustration of the condition number of the Hessian matrix for two two-dimensional quadratic functions. The ellipse corresponds to a contour line of the quadratic function.

The goal of preconditioning is to transform Figure 2.1b into Figure 2.1a. To do so, a preconditioning matrix $\mathbf{G}^{(t-1)} = \mathbf{G}(\boldsymbol{\Theta}^{(t-1)}) \in \mathbb{R}^{ND \times ND}$ added to Eq. 2.14

$$\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}^{(t-1)} - \eta\, \mathbf{G}^{(t-1)} \nabla \mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right). \tag{2.17}$$

In presence of a preconditioning matrix $\mathbf{G}^{(t-1)}$, the condition number is measured on the new Hessian $\mathbf{G}^{(t-1)} \nabla^2 \mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right)$. The goal when resorting to a preconditioner is thus to decrease as much as possible the condition number. In Guggenheimer et al. (1995), the authors provide a general upper bound on the condition number of an invertible matrix. They also include an analysis of situations where their bound is tight.

The inverse Hessian matrix $\mathbf{G}^{(t-1)} = \left[\nabla^2 \mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right)\right]^{-1}$ is the natural choice of preconditioner. It yields the Newton algorithm (Nocedal and Wright, 2006, chapter 10) applied to $\nabla \mathcal{L}$. By definition the Newton algorithm looks for zeros of a function. Applying it on $\nabla \mathcal{L}$ boils down

to finding critical points of the loss function $\mathcal{L}$. With this preconditioner, the condition number equals exactly 1. The modified Hessian matrix is thus perfectly conditioned. However, this preconditioner suffers from three main drawbacks. First, it can only be used when the Hessian matrix is invertible. Second, the Newton algorithm requires evaluating and inverting the full Hessian matrix, and then applying a matrix-vector product with the gradient. These three steps are very expensive, especially in high dimensions. A faster and more stable implementation directly evaluates the vector $\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}$ by considering it as the solution of a linear system:

$$\left[\nabla^2 \mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right)\right]\left(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\right) = -\nabla \mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right), \tag{2.18}$$

but remains expensive in high dimensions. Lastly, it can suffer from instabilities for loss functions that are not gradient Lipschitz continuous.

The Gauss-Newton and Levenberg-Marquardt algorithms (Nocedal and Wright, 2006, chapter 10) are preconditioned GD algorithms. As we will show in Chapter 3, Levenberg-Marquardt is quite common in ISM studies. These two algorithms resort to a cheaper and invertible approximation of the inverse Hessian that does not require the evaluation of second order derivatives. They enjoy similar convergence results and convergence rate as the Newton algorithm. However, they are specialized to non-linear least squares problems, i.e., to loss functions $\mathcal{L}$ of the form

$$\mathcal{L} : \boldsymbol{\Theta} \in \mathbb{R}^{ND} \mapsto \sum_{n=1}^{N} \sum_{\ell=1}^{L} \frac{1}{2\sigma_{n\ell}^2} (f_\ell(\boldsymbol{\theta}_n) - y_{n\ell})^2 \tag{2.19}$$

The Gauss-Newton algorithm uses $\mathbf{G}^{(t-1)} = \left[\nabla \mathcal{L}(\boldsymbol{\Theta}) \nabla \mathcal{L}(\boldsymbol{\Theta})^T\right]^{-1}$ to approximate the Hessian matrix. This approximation is symmetric by construction, and positive definite as long as the matrix $\nabla \mathcal{L}(\boldsymbol{\Theta}) \nabla \mathcal{L}(\boldsymbol{\Theta})^T \in \mathbb{R}^{ND \times ND}$ is full rank. This condition requires $L \geq D$, i.e., the number L of observables per pixel needs to be greater or equal to the dimension D of the parameter $\boldsymbol{\theta}$ per pixel. Though this condition is usually verified in inverse problems, it is not sufficient to guarantee this matrix is invertible. In some cases, the Gauss-Newton algorithm thus suffers from numerical instabilities, especially far from a local minimum. The Levenberg-Marquardt algorithm "damps" this approximation of the Hessian matrix: $\mathbf{G}^{(t-1)} = \left[\nabla \mathcal{L}(\boldsymbol{\Theta}) \nabla \mathcal{L}(\boldsymbol{\Theta})^T + \epsilon \mathbf{I}_{ND}\right]^{-1}$, with $\mathbf{I}_{ND}$ the $ND \times ND$ identity matrix and $\epsilon \geq 0$ a damping parameter. Using large enough values of ϵ ensures better numerical stability than Gauss-Newton. For large ϵ values, this preconditioner boils down to the identity matrix, which is equivalent to the standard GD in Eq. 2.14. For low values, one recovers the Gauss-Newton preconditioner. In practice, the damping parameter ϵ is often adjusted during the optimization procedure to benefit from the gradient descent stability in the first iterations and from the Gauss-Newton algorithm speed in the last iterations. Similarly to the Newton algorithm, the Gaussian-Newton and Levenberg-Marquardt algorithms involve either the inversion of a dense matrix and matrix-vector products or solving a linear system, which are both expensive in high dimensions.

Quasi-Newton optimization algorithms such as BFGS or limited memory BFGS (L-BFGS), propose efficient approximations of the inverse Hessian matrix for more general loss functions (Nocedal and Wright, 2006, chapter 6). In particular, the limited memory BFGS preconditioner does not require any matrix inversion and rely on vector-vector products only. Therefore, it can be used even in high dimensional settings.

Diagonal preconditioners for neural network training – Finally, neural networks training involves multiple dedicated diagonal preconditioners – see Chapter 4 for more details on neural networks. In the deep learning community, these preconditioners are known as "adaptive learning rates". They are dedicated to non-convex optimization in extremely high dimensions, i.e., $\mathcal{O}(10^6)$

and higher. Such preconditioners include the adaptive gradient algorithm (AdaGrad) (Duchi et al., 2011)

$$\mathbf{G}^{(t-1)} = \text{diag}\left(\frac{1}{\epsilon + \sqrt{\mathbf{v}^{(t-1)}}}\right), \quad \text{with} \quad \mathbf{v}^{(t-1)} = \sum_{\tau=1}^{t-1}\left[\nabla\mathcal{L}\left(\boldsymbol{\Theta}^{(\tau)}\right)\right]^2, \tag{2.20}$$

and the root mean squared propagation (RMSProp) (Tieleman and Hinton, 2012)

$$\mathbf{G}^{(t-1)} = \text{diag}\left(\frac{1}{\epsilon + \sqrt{\mathbf{v}^{(t-1)}}}\right), \quad \text{with} \quad \mathbf{v}^{(t-1)} = a\mathbf{v}^{(t-2)} + (1-a)\left[\nabla\mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right)\right]^2, \tag{2.21}$$

where all exponentiations are taken element-wise, $a \in]0,1[$ is an exponential decay rate, and $\epsilon > 0$ is a damping factor that prevents divisions by zero – typically, $\epsilon = 10^{-5}$. The AdaGrad preconditioner accumulates the squared gradients, causing the components of $\mathbf{v}$ to diverge. Conversely, RMSProp progressively forgets past gradients with exponential decay rate ϵ. As these two preconditioners are diagonal, they are simple to invert and only involve vector-vector products. In Dauphin et al. (2015), the authors propose another diagonal preconditioner that considerably decreases the upper bound on the condition number from Guggenheimer et al. (1995). They also show that RMSProp is very similar to their preconditioner, which partially explains the performance of RMSProp observed in practice.

As the inverse problem we consider in this thesis is high dimensional, the full rank preconditioners from the Newton, Gauss-Newton and Levenberg-Marquardt algorithms are too computationally expensive to be considered. L-BFGS and RMSProp scale better while enforcing low condition numbers. Section 2.2.2.5 describes how these two preconditioners can be exploited in sampling algorithms.

2.2.1.3 Block coordinate variant

In high dimensional settings, evaluating the full gradient at once may be quite costly, both in time and memory. In such cases, one can adopt a divide-and-conquer strategy by dividing the large physical parameter $\boldsymbol{\Theta}$ into J smaller blocks, e.g., $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_j)_{j=1}^J$ (Pereyra et al., 2016). This division replaces the computation of the full gradient vector $\nabla\mathcal{L}$ by the smaller gradient vector $\nabla_j\mathcal{L}$ with respect to part $\boldsymbol{\Theta}_j$. At step t, instead of being performed at once, the updated of iterate $\boldsymbol{\Theta}^{(t-1)}$ is divided into a succession of J updates of its parts, with

$$\boldsymbol{\Theta}_j^{(t)} = \boldsymbol{\Theta}_j^{(t-1)} - \eta\nabla_j\mathcal{L}\left(\boldsymbol{\Theta}_1^{(t)}, \boldsymbol{\Theta}_2^{(t)}, \ldots, \boldsymbol{\Theta}_{j-1}^{(t)}, \boldsymbol{\Theta}_j^{(t-1)}, \ldots, \boldsymbol{\Theta}_J^{(t-1)}\right). \tag{2.22}$$

Considering these faster and lighter individual and alternating updates typically speeds up individual updates and can lighten memory requirements. The updates for each part j can be combined with preconditioning for faster convergence. The update order of the blocks j can be defined either deterministically or stochastically. Besides, a random activation of blocks at each step can lead to a stochastic estimate of the full gradient, yielding a stochastic gradient descent algorithm (Pereyra et al., 2016). Such division of the parameter space typically comes at the price of slower convergence to a critical point in terms of number of iterations, especially in case of degeneracy between components $\boldsymbol{\Theta}_j$. See for instance Chouzenoux et al. (2016) for an application.

Limitations of optimization-based inference

Convergence to the global minimum for non-convex loss function – Until now, we have described the GD algorithm, its preconditioned variant and its block coordinate variant. At best, these algorithms are guaranteed to converge to a critical point – see e.g., Theorem 10.15 from Beck (2017). However, the MLE and MAP are defined as global minima of their twice differentiable loss functions $\mathcal{L}$. When $\mathcal{L}$ is in addition convex, any critical point is a global minimum. However,

the loss functions considered in this thesis are non-convex. Such functions have potentially many saddle points and local minima with higher loss values than the global minimum. In practice, some non-convex preconditioned variants such as RMSProp escape quickly from the neighborhood of saddle points and reach a local minimum (Dauphin et al., 2015). However, the reached local minimum may have a high loss value. GD algorithms usually fail to escape from such local minima.

A first solution to compensate for this effect is to launch the chosen minimization algorithm many times with different initial states. For high dimensional $\boldsymbol{\Theta}$, this solution requires numerous and potentially costly runs. This is also true in low dimensions when the loss function contains many local minima. Meta-heuristics form an alternative class of algorithms that addresses this multimodality issue. By proposing random candidates, they can escape from local minima with high loss value. In general, they converge to a local minimum with a loss value close to that of the global minimum. See Section 2.A.1 for more details on these methods.

Lack of uncertainty quantification – Estimators defined as solutions of an optimization problem do not naturally come with an uncertainty description. Some dedicated methods such as the Cramér-Rao bound propose approximate uncertainty quantification from a point estimate, but rely on strong hypotheses on the posterior, namely unimodality or Gaussianity.

With a point estimate $\widehat{\boldsymbol{\Theta}}$, one can derive uncertainties on predictions of reproduced observations $\widetilde{\mathbf{Y}}|\widehat{\boldsymbol{\Theta}}$. These uncertainties can be defined using the noise model from the likelihood in inverse problems. In machine learning, these uncertainties can also be defined without assuming any noise model e.g., with quantile regression (Koenker, 2005) or conformal prediction (Angelopoulos and Bates, 2022). We do not cover these distribution-free uncertainty quantification approaches in this thesis, as our goal is to provide uncertainties on the physical parameters $\boldsymbol{\Theta}$.

2.2.2 Approximating integrals with Monte Carlo estimators

As we showed in Section 2.1.2, some estimators can accurately quantify the uncertainty on physical parameters $\boldsymbol{\Theta}$. These estimators are defined as an integral of a function g on $\boldsymbol{\Theta}$ over the posterior distribution. In this section, we first introduce Monte Carlo (MC) estimators that approximate the integral using samples of the posterior distribution. Then, we present Markov chain Monte Carlo (MCMC) algorithms, including the widespread Metropolis-Hastings (MH) algorithm and Metropolis adjusted Langevin algorithm (MALA). Preconditioned and block coordinate variants are also presented. Finally, the multiple-try Metropolis (MTM), which extends the MH algorithm by generating multiple candidates at each iteration instead of one, is presented.

The sampler proposed in Chapter 5 (Section 5.2) builds on the algorithms presented in this section. It combines two sampling kernels. The first addresses the log-posterior gradient Lipschitz regularity issue. It builds on preconditioning to explore the parameter space more efficiently. The second addresses the posterior multimodality difficulty. It performs block coordinate updates and generates multiple candidates for each block to improve mixing properties.

Additional sampling algorithms dedicated to multimodal distributions are presented in Appendix 2.A. Three classes of algorithms are considered: adaptations of meta-heuristics to MCMC algorithms, exploitation of prior knowledge of the modes localization, and nested sampling. We do not use these algorithms to derive the proposed sampler. However, they are popular in interstellar medium studies, as we will show in Chapter 3 (Section 3.2.2). Besides, we will compare these algorithms to the proposed sampler in Chapter 5 (Section 5.4).

2.2.2.1 Monte Carlo (MC) estimators

In case of no conjugacy relation, two main approaches permit to numerically evaluate the integral in Eq. 2.11, reminded here,

$$\mathbb{E}\left[g(\boldsymbol{\Theta})|\mathbf{Y}\right] \;=\; \int g(\boldsymbol{\Theta})\,\pi(\boldsymbol{\Theta}|\mathbf{Y})\,d\boldsymbol{\Theta}. \tag{2.23}$$

The first approach relies on a deterministic quadrature of the parameter space, i.e., on a finite set of points with deterministic positions. The posterior pdf is then evaluated on each point. The integration is performed using Riemannian integration, or more accurate methods such as the trapezoidal rule or Simpson's rule. This approach is efficient in low dimensions (Robert and Casella, 2004, chapter 1). However, the number of points in the quadrature increases exponentially with the dimension. This method has already been applied in ISM studies, as detailed in Chapter 3 (Section 3.2.2).

The second approach relies on a stochastic quadrature of the parameter space. It first generates samples $\boldsymbol{\Theta}^{(t)}$ from the posterior distribution. Estimators are then evaluated as empirical means of the $g\left(\boldsymbol{\Theta}^{(t)}\right)$:

$$\mathbb{E}\left[g(\boldsymbol{\Theta})|\mathbf{Y}\right] \;\simeq\; \frac{1}{T_{\mathsf{MC}}}\sum_{t=1}^{T_{\mathsf{MC}}} g\left(\boldsymbol{\Theta}^{(t)}\right), \quad \boldsymbol{\Theta}^{(t)} \sim \pi\left(\boldsymbol{\Theta}|\mathbf{Y}\right), \quad t = 1,\ldots,T_{\mathsf{MC}}. \tag{2.24}$$

Such estimators are called Monte Carlo (MC) estimators (Robert and Casella, 2004, chapter 3) – hence the notation "MC" for T_{MC}.

For independent and identically distributed (i.i.d.) samples drawn from the posterior, the law of large numbers guarantees that

$$\lim_{T_{\mathsf{MC}}\to\infty} \frac{1}{T_{\mathsf{MC}}}\sum_{t=1}^{T_{\mathsf{MC}}} g\left(\boldsymbol{\Theta}^{(t)}\right) \;=\; \mathbb{E}\left[g(\boldsymbol{\Theta})|\mathbf{Y}\right], \tag{2.25}$$

and the central limit theorem (CLT) describes the distribution on the error of this empirical mean compared to the true one

$$\left(\frac{1}{T_{\mathsf{MC}}}\sum_{t=1}^{T_{\mathsf{MC}}} g\left(\boldsymbol{\Theta}^{(t)}\right) - \mathbb{E}[g(\boldsymbol{\Theta})|\mathbf{Y}]\right) \;\xrightarrow[T_{\mathsf{MC}}\to+\infty]{\mathcal{D}}\; \mathcal{N}\left(0,\; \frac{1}{T_{\mathsf{MC}}}\mathsf{Cov}\left[g(\boldsymbol{\Theta})|\mathbf{Y}\right]\right), \tag{2.26}$$

where the convergence is in distribution and $\mathsf{Cov}\left[g(\boldsymbol{\Theta})|\mathbf{Y}\right]$ is a fixed covariance matrix, generally unknown. Essentially, the CLT shows that an MC estimator built with i.i.d. samples converges with speed $1/\sqrt{T_{\mathsf{MC}}}$. The variances in $\mathsf{Cov}\left[g(\boldsymbol{\Theta})|\mathbf{Y}\right]$ scale polynomially with the dimension of the parameter space. The sampling approach therefore scales much better than deterministic integration, at the condition of efficient sampling of the posterior distribution.

2.2.2.2 From Monte Carlo to Markov chain Monte Carlo (MCMC) methods

There exists no efficient algorithm that draws independent samples from an arbitrary posterior distribution. Rejection sampling (Robert and Casella, 2004, chapter 2) can in principle draw independent samples from any posterior distribution. It generates candidates according to a proposal distribution and accepts or rejects them with a certain acceptance probability. In practice, the acceptance probabilities are prohibitively small in high dimensional settings with a complex model.

Some alternative methods generate correlated samples from which MC estimators can be evaluated. Despite the correlation between samples, MC estimators built from Markov chains enjoy similar convergence results as MC estimators built from independent samples. Such algorithms

include nested sampling, sequential MC (SMC) and Markov chain Monte Carlo (MCMC) methods. Nested sampling and SMC are widespread in astrophysics, but are not used directly in this thesis. They are presented in Appendix 2.A. They both build on or rely on MCMC methods.

An MCMC algorithm (Robert and Casella, 2004, chapters 6 and 7) generates a Markov chain whose stationary distribution is the posterior distribution. The goal is to generate a correlated set of samples from the posterior distribution to evaluate MC estimators (Eq. 2.24). To do so, each MCMC algorithm relies on a different transition kernel $\mathcal{K} : \mathbb{R}^{N\times D} \times \mathbb{R}^{N\times D} \to \mathbb{R}_+$. The MCMC class gathers a wide variety of algorithms that can tackle problems with diverse sets of properties.

Burn-in phase – A Markov chain generated with an MCMC algorithm contains two regimes. The chain is initialized with a starting point $\boldsymbol{\Theta}^{(0)}$. At first, it is in a transient regime, usually called the *burn-in phase*, where iterates are close to $\boldsymbol{\Theta}^{(0)}$ and evolve towards high probability regions. When it reaches a high probability region, the Markov chain enters a stationary regime, in which the elements $\boldsymbol{\Theta}^{(t)}$ of the chain can be considered as samples of the posterior. In particular, these samples are considered independent of $\boldsymbol{\Theta}^{(0)}$. The MC estimators are evaluated with these samples. The first $T_{\text{BI}} \geq 1$ iterates associated with the burn-in phase are removed to avoid biasing estimators. A good transition kernel $\mathcal{K}$ thus reaches quickly the posterior high probability region and explores it efficiently.

Convergence properties[4] – The convergence of the Markov chain to the posterior distribution and of the associated MC estimators to the true values are difficult to establish. Three properties are central. First, a Markov chain admits an *invariant distribution*. An invariant distribution π verifies

$$\begin{cases} \boldsymbol{\Theta}^{(1)} \sim \pi \\ \boldsymbol{\Theta}^{(2)} \sim \mathcal{K}(\boldsymbol{\Theta}^{(1)}, \cdot) \end{cases} \implies \boldsymbol{\Theta}^{(2)} \sim \pi. \tag{2.27}$$

The *detailed balance* is a sufficient condition to establish that the posterior distribution is the invariant distribution of a Markov chain. A transition kernel $\mathcal{K}$ verifies the detailed balance property with the posterior if and only if

$$\forall \boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)} \in \mathbb{R}^{N\times D}, \quad \mathcal{K}\left(\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}\right) \pi(\boldsymbol{\Theta}^{(1)}|\mathbf{Y}) = \mathcal{K}\left(\boldsymbol{\Theta}^{(2)}, \boldsymbol{\Theta}^{(1)}\right) \pi(\boldsymbol{\Theta}^{(2)}|\mathbf{Y}) \tag{2.28}$$

A Markov chain with such a transition kernel is reversible and admits the posterior as invariant density (Robert and Casella, 2004, Theorem 6.46).

Second, the *ergodicity* of a Markov chain guarantees the asymptotic convergence of the Markov chain to its invariant with respect to a distance on probability distributions, the total variation. In particular, this result holds for any starting point $\boldsymbol{\Theta}^{(0)}$. It also guarantees that MC estimators asymptotically converge to their true values, i.e., that the law of large numbers (Eq. 2.25) applies. The ergodicity theorem (Robert and Casella, 2004, Theorem 6.63) states that a sufficient condition for ergodicity is for the chain to be Harris recurrent. We now provide an intuition of the definition of Harris recurrence – see Robert and Casella (2004, Definition 6.32) for the exact definition. The *irreducibility* of a Markov chain means that its transition kernel $\mathcal{K}$ allows for a transition from any $\boldsymbol{\Theta}^{(1)} \in \mathbb{R}^{N\times D}$ to any $\boldsymbol{\Theta}^{(2)} \in \mathbb{R}^{N\times D}$, potentially with multiple steps. Besides, for any subset of $\mathbb{R}^{N\times D}$ of non-zero measure, if the transition kernel $\mathcal{K}$ permits to remain in the subset with non-zero probability, the chain is said to be *aperiodic*. The *recurrence* of a Markov chain means that the expected number of visits of any subset of $\mathbb{R}^{N\times D}$ of non-zero measure is infinite. From Robert and Casella (2004, Theorem 6.30), an irreducible chain is either recurrent or transient. The *Harris recurrence* is a stronger recurrence property for an irreducible chain in which the probability of visiting any subset of $\mathbb{R}^{N\times D}$ of non-zero measure an infinite number of times is exactly 1.

[4] This paragraph contains mathematical details that are not mandatory for the remainder of the thesis.

Third, the speed of convergence of a Markov chain to its invariant distribution permits the derivation of stopping rules for required precision levels. There are two main characterizations of the speed of convergence for Markov chains. The first is *geometric ergodicity* (Robert and Casella, 2004, Definition 6.54). In essence, it ensures that an extended total variation distance between the Markov chain and its invariant distribution decreases at least geometrically. The second, *uniform ergodicity* (Robert and Casella, 2004, Definition 6.58), enforces a stronger condition on the rate of geometric convergence. Both conditions can be associated with a CLT (Eq. 2.26) - see Robert and Casella (2004, Theorems 6.67 and 6.77). A third condition permits to derive a CLT requires the Markov chain to be aperiodic, irreducible and reversible with the posterior as invariant distribution.

CLT and effective sample size (ESS) – In the context of Markov chains, the three previous CLTs yield, for a scalar component i of function g,

$$\left(\frac{1}{T_{\text{MC}}} \sum_{t=1}^{T_{\text{MC}}} g_i\left(\boldsymbol{\Theta}^{(t)}\right) - \mathbb{E}[g_i(\boldsymbol{\Theta})|\mathbf{Y}] \right) \rightarrow \mathcal{N}\left(0, \frac{\gamma_{g_i}^2}{T_{\text{MC}}}\right), \tag{2.29}$$

with

$$\frac{\gamma_{g_i}^2}{T_{\text{MC}}} = \frac{1}{T_{\text{MC}}} \left(\text{Var}\left[g_i(\boldsymbol{\Theta})|\mathbf{Y}\right] + 2 \sum_{\tau=1}^{\infty} \text{Cov}\left[g_i(\boldsymbol{\Theta}^{(t)}), g_i(\boldsymbol{\Theta}^{(t+\tau)})|\mathbf{Y}\right] \right). \tag{2.30}$$

The first element in the definition of $\gamma_{g_i}^2$ corresponds to the variance in the CLT for i.i.d. samples (Eq. 2.26). The second term sums the autocovariances in the Markov chain with lag $\tau \geq 1$. These autocovariances can be converted into more interpretable autocorrelations,

$$\frac{\gamma_{g_i}^2}{T_{\text{MC}}} = \frac{1}{T_{\text{MC}}} \text{Var}\left[g_i(\boldsymbol{\Theta})|\mathbf{Y}\right] \left(1 + 2 \sum_{\tau=1}^{\infty} r_i^{(\tau)} \right), \tag{2.31}$$

where $r_i^{(\tau)}$ is the autocorrelation of the chain with time lag $\tau \geq 0$ for $g_i(\boldsymbol{\Theta})$. Finally, one can rewrite Eq. 2.31

$$\frac{\gamma_{g_i}^2}{T_{\text{MC}}} = \frac{\text{Var}\left[g_i(\boldsymbol{\Theta})|\mathbf{Y}\right]}{\text{ESS}^{(T_{\text{MC}})}}, \tag{2.32}$$

where the ESS stands for the effective sample size (ESS) (Robert and Casella, 2004, Section 12.3.5). It permits to rewrite the CLT in Eq. 2.26 in the context of Markov with a size $\text{ESS}^{(T_{\text{MC}})}$, instead of T_{MC}. Therefore, the ESS measures the number of independent samples with the same estimation power as a correlated chain. Ideally, the correlation between iterates $\boldsymbol{\Theta}^{(t)}$ should be as small as possible. In the extreme case where the correlation in the chain is negligible, $\text{ESS}^{(T_{\text{MC}})} \simeq T_{\text{MC}}$, i.e., the Markov chain can be considered as a set of independent samples. Conversely, highly correlated chains typically "mix" poorly, as they need a large amount of iterations to explore the posterior.

In the general definition above, the ESS depends on the function g. In practice, for simplicity, it is often reduced to the case $g : \boldsymbol{\Theta} \mapsto \boldsymbol{\Theta}$, i.e.,

$$\text{ESS}_{nd}^{(T_{\text{MC}})} = \frac{T_{\text{MC}}}{1 + 2\sum_{\tau=1}^{\infty} r_{nd}^{(\tau)}}, \tag{2.33}$$

with $r_{nd}^{(\tau)}$ the autocorrelation of the chain with time lag $\tau \geq 0$ for the parameter θ_{nd}. As the above ESS is defined for a scalar parameter, for a ND-dimensional parameter space, one obtains ND estimations of the ESS. In this thesis, we implement the ESS estimation presented in Gelman et al. (2015). This definition accounts for results of potentially $M \geq 1$ Markov chains. It can therefore be used when an MCMC algorithm is run M times.

2.2.2.3 Metropolis-Hastings (MH) algorithm

The Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) is historically the first proposed MCMC algorithm, and remains arguably the most famous and widespread to this day. It is very general and versatile, and enjoys strong theoretical properties. The two kernels proposed in Chapter 5 (Section 5.2) build on this algorithm.

The algorithm – The MH algorithm can start from any point $\boldsymbol{\Theta}^{(0)}$. Then, at each step t, a candidate $\boldsymbol{\Theta}_c^{(t)}$ is sampled from a proposal distribution $q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}^{(t-1)})$. Unlike rejection sampling, the proposal distribution can depend on the current iterate $\boldsymbol{\Theta}^{(t-1)}$. This candidate is accepted with probability

$$\rho^{(t)} = \min\left(1, \frac{\pi\left(\boldsymbol{\Theta}_c^{(t)}|\mathbf{Y}\right)}{\pi\left(\boldsymbol{\Theta}^{(t-1)}|\mathbf{Y}\right)} \frac{q\left(\boldsymbol{\Theta}^{(t-1)}|\boldsymbol{\Theta}_c^{(t)}\right)}{q\left(\boldsymbol{\Theta}_c^{(t)}|\boldsymbol{\Theta}^{(t-1)}\right)}\right). \tag{2.34}$$

When the candidate is not accepted, the new iterate $\boldsymbol{\Theta}^{(t)}$ is set to $\boldsymbol{\Theta}^{(t-1)}$. Therefore, the transition kernel $\mathcal{K}$ of the MH algorithm is defined as a combination of the proposal q and the accept-reject step. Algorithm 2.2 summarizes the MH algorithm.

Algorithm 2.2: Metropolis-Hastings (MH) algorithm

Input: Starting point $\boldsymbol{\Theta}^{(0)}$, proposal q, numbers of samples T_{BI} and T_{MC}

for $t = 1, \ldots, T_{MC}$ **do**

- $\boldsymbol{\Theta}_c^{(t)} \sim q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}^{(t-1)})$ `// generate candidate`
- $\rho^{(t)}$ `// compute acceptance probability using Eq. 2.34`
- $\zeta \sim \text{Unif}(0,1)$ `// update iterate`
- $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}_c^{(t)}$ if $\zeta \leq \rho^{(t)}$ else $\boldsymbol{\Theta}^{(t-1)}$

Output: Chain of samples $\{\boldsymbol{\Theta}^{(t)}\}_{t=T_{\text{BI}}+1}^{T_{\text{MC}}}$ `// The first` T_{BI} `are rejected`

Theoretical convergence properties – The theoretical convergence properties of the MH algorithm are simple. The following is a short summary of Robert and Casella (2004, Sections 7.3.1 and 7.3.2). First, the accept-reject step allows $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}^{(t-1)}$. This condition is sufficient for the Markov chain to be aperiodic. Then, the main (sufficient) condition for convergence relies on the positivity of the proposal q on $\mathbb{R}^{N\times D}$, i.e.,

$$\forall \boldsymbol{\Theta}_c, \boldsymbol{\Theta}^{(t-1)} \in \mathbb{R}^{N\times D}, \quad q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}^{(t-1)}) > 0. \tag{2.35}$$

This condition is verified e.g., for a Gaussian proposal with any mean vector and covariance matrix. When it is verified, the transition kernel satisfies the detailed balance property (Eq. 2.28). Therefore, the Markov chain generated by the MH algorithm admits the posterior distribution as an invariant probability distribution (Robert and Casella, 2004, Theorem 7.2). Besides, it is also a sufficient condition for the chain $(\boldsymbol{\Theta}^{(t)})$ to be irreducible. Therefore, with such a proposal q, the law of large numbers (Eq. 2.25) applies to the MH Markov chain (Robert and Casella, 2004, Theorem 7.4). Note that this condition is only sufficient. Similar results hold for less restrictive conditions (Robert and Casella, 2004, Corollary 7.7).

Common proposal distributions q – As the theoretical results show, the choice of the proposal distribution q is crucial to efficiently explore the posterior distribution.

Choosing a proposal q that is independent of current iterate $\boldsymbol{\Theta}^{(t-1)}$, i.e., $q(\cdot|\boldsymbol{\Theta}^{(t-1)}) = q(\cdot)$, seems a simple and natural choice. Such a proposal yields the so-called independent Metropolis-Hastings (I-MH). As stated in Robert and Casella (2004, Theorem 7.8), this algorithm produces a uniformly ergodic chain if there exists a constant $M \geq 1$ such that $Mq(\boldsymbol{\Theta}) \geq \pi(\boldsymbol{\Theta}|\mathbf{Y})$ for all

$\boldsymbol{\Theta} \in \mathbb{R}^{N \times D}$. In this case, the expected probability acceptance is at least $1/M$ when the chain is stationary. As for rejection sampling, expected probability acceptance may thus be extremely low in practice for posteriors whose high probability region are concentrated in a region with much smaller volume than that of the proposal. This is typically the case in high dimensions, due to the curse of dimensionality.

Random walk Metropolis-Hastings (RWMH) is another very simple proposal that is more widely spread. It generates candidates $\boldsymbol{\Theta}_c^{(t)} = \boldsymbol{\Theta}^{(t-1)} + Z^{(t)}$, with generally $Z^{(t)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for a covariance matrix $\boldsymbol{\Sigma}$. In this case, $q(\cdot|\boldsymbol{\Theta}^{(t-1)}) = \mathcal{N}(\boldsymbol{\Theta}^{(t-1)}, \boldsymbol{\Sigma})$. As this proposal q is symmetric, i.e., $q(\boldsymbol{\Theta}_c^{(t)}|\boldsymbol{\Theta}^{(t-1)}) = q(\boldsymbol{\Theta}^{(t-1)}|\boldsymbol{\Theta}_c^{(t)})$, the acceptance probability in Eq. 2.34 simplifies to the ratio of posterior pdfs of the candidate and of the current iterate. RWMH is never uniformly ergodic. It can be shown to be geometrically ergodic for log-concave distributions (Robert and Casella, 2004, Theorem 7.15). However, as the Markov chain is reversible, the CLT still applies. In practice, the covariance matrix $\boldsymbol{\Sigma}$ plays a crucial role in RWMH: a matrix $\boldsymbol{\Sigma}$ closer to the covariance of the posterior will lead to larger steps with higher acceptance probabilities. However, even with an adapted covariance matrix, RWMH does not scale up well due to its blind nature (Pereyra et al., 2016).

As RWMH performs local steps, its optimal expected acceptance probability is not 1. Indeed, a covariance matrix $\boldsymbol{\Sigma}$ with small variances will generate candidates very close to the current iterate. In case of smooth log-posterior, these candidates will have an expected acceptance probability close to 1, but will lead to slow exploration of the parameter space. Conversely, if the variances are large, the candidates may fall far from the current iterate, potentially out of the posterior high probability region. In high dimensions, such candidates have an expected acceptance probability close to 0. However, each accepted candidate corresponds to a potentially large step, which improves the exploration. In general, the optimal mixing for RWMH is achieved with a trade-off between large steps and large acceptance probability. Optimal mixing for RWMH is achieved with an expected acceptance probability of 20% - 25% (Pereyra et al., 2016).

2.2.2.4 Using the gradient with Metropolis adjusted Langevin algorithm (MALA)

Metropolis adjusted Langevin algorithm (MALA) (Roberts and Stramer, 2002) and Hamiltonian Monte Carlo (HMC) (Neal, 2011) are both MH algorithms that exploit gradient information in the mean of the proposal q. Thanks to the gradient information, they both scale better than RWMH (Pereyra et al., 2016).

HMC relies on the introduction of a momentum auxiliary variable and exploits Hamiltonian dynamics to explore the posterior. MALA is similar to a gradient descent algorithm on the negative log-posterior. HMC generally mixes better than MALA for a fixed number of samples. However, HMC requires tuning more parameters and has a higher computational cost per iteration. Besides, MALA can be easily associated with advanced optimization methods (Pereyra et al., 2016). It is therefore simpler to transfer advanced optimization techniques to MALA than to HMC. For these reasons, we favor MALA over HMC in this thesis.

MALA can be defined from a particular Langevin diffusion process. This continuous diffusion process admits the posterior as an invariant distribution and is geometrically ergodic. First, the diffusion process is discretized, usually following the Euler-Maruyama scheme. The resulting random process is called unadjusted Langevin algorithm (ULA), and reads

$$\boldsymbol{\Theta}_c^{(t)} = \boldsymbol{\Theta}^{(t-1)} + \eta \nabla \ln \pi \left(\boldsymbol{\Theta}^{(t-1)} | \mathbf{Y} \right) + \sqrt{2\eta}\, Z^{(t)}, \quad Z^{(t)} \sim \mathcal{N}\left(0_{ND},\, \mathbf{I}_{ND}\right). \tag{2.36}$$

ULA is a GD algorithm perturbed with an additive Gaussian noise. It therefore enjoys some similar theoretical properties. The discretization of the diffusion process introduces an error in ULA. Due to this error causes, the posterior is generally not the invariant distribution of ULA. In Durmus and Moulines (2017), the authors propose three upper bounds on the distance between an ULA Markov chain and the posterior distribution. Each upper bound is associated with a

class of posterior distribution: super-exponential outside a ball around the mode – i.e., posteriors with heavy tails –, log-concave, and strongly log-concave. In all three cases, the log-posterior is assumed to be gradient Lipschitz continuous with Lipschitz constant β. The three upper bounds depend on the Lipschitz constant β, either directly or indirectly through the step size $\eta \in]0, 1/\beta[$. Similarly to GD in optimization, step sizes close to $1/\beta$ lead to faster convergence to the invariant distribution.

MALA corrects the discretization error in ULA thanks to an accept-reject step. This accept-reject steps makes MALA an MH algorithm. Its proposal distribution can thus be written

$$q(\cdot|\boldsymbol{\Theta}^{(t-1)}) = \mathcal{N}\left(\boldsymbol{\Theta}^{(t-1)} + \eta \nabla \ln \pi\left(\boldsymbol{\Theta}^{(t-1)}|\mathbf{Y}\right),\ 2\eta\, \mathbf{I}_{ND}\right), \tag{2.37}$$

which is an equivalent form of Eq. 2.36. Similarly to the covariance matrix in RWMH, the step size η defines a trade-off between large steps and high expected acceptance probability. As it exploits gradient information, MALA produces better candidates than RWMH. In case of a gradient Lipschitz continuous log-posterior of Lipschitz constant β, a step size η close to $1/\beta$ leads to good mixing properties. Overall, the MALA optimal expected acceptance probability is 50% - 60% (Pereyra et al., 2016), i.e., two to three times larger than RWMH.

Discretization error – As already mentioned, the Euler-Maruyama scheme introduces a discretization error. In MALA, this error is corrected with the accept-reject step. However, in case of no discretization error, all candidates would be accepted, which would accelerate the posterior exploration. Therefore, a lower discretization error leads to more relevant candidates. For instance, in Durmus et al. (2017), the authors exploit an Ozaki discretization scheme, which is more accurate than the usual Euler-Maruyama. They obtain considerable improvement of the mixing performance of MALA, and derive an optimal expected acceptance probability around 70%. In general, using a better discretization scheme reduces this discretization error and therefore generates better candidates.

2.2.2.5 Preconditioning with MALA to improve mixing properties

As stated in Section 2.2.1 for optimization methods, preconditioning permits to reduce the condition number of the Hessian matrix. This reduction permits a more efficient exploration of the parameter space. As ULA is defined as a GD algorithm perturbed with Gaussian noise, it seems natural to extend it with preconditioning.

Preconditioning and Riemannian manifold – In Girolami and Calderhead (2011), the authors associate preconditioning with sampling on a Riemannian manifold. The preconditioner is then closely related to the metric of the manifold. The authors extended MALA and HMC to Riemannian manifolds, yielding the Manifold Metropolis adjusted Langevin algorithm (MMALA) and Riemannian manifold HMC (RMHMC) algorithms. In Xifara et al. (2014), a minor correction was applied to MMALA, leading to a renamed preconditioned Metropolis adjusted Langevin algorithm (PMALA). We use PMALA to derive one of the two sampling kernels of the MCMC algorithm proposed in Chapter 5 (Section 5.2). The associated proposal distribution q reads

$$q\left(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}^{(t-1)}\right) = \mathcal{N}\left(\boldsymbol{\Theta}_c|\boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}^{(t)}\right) \tag{2.38}$$

with

$$\begin{cases} \boldsymbol{\mu}^{(t)} = \boldsymbol{\Theta}^{(t-1)} - \eta \mathbf{G}^{(t-1)} \nabla \mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right) + 2\eta\, \boldsymbol{\gamma}^{(t-1)}, \\ \boldsymbol{\Lambda}^{(t)} = 2\eta\, \mathbf{G}^{(t-1)}, \end{cases} \tag{2.39}$$

where η is a step size and $\boldsymbol{\gamma}^{(t-1)}$ is the additional drift term due to the position-dependent

preconditioner (Xifara et al., 2014). In full generality, for all $i \in [\![1, ND]\!]$,

$$\gamma_i^{(t-1)} = \frac{1}{2} \sum_{j=1}^{ND} \frac{\partial G_{ij}^{(t-1)}}{\partial \theta_j^{(t-1)}}. \tag{2.40}$$

Depending on the complexity of the preconditioner $\mathbf{G}$, this additional drift term can be quite complicated to evaluate. For instance, for a $\mathbf{G}$ defined with second order derivative information, evaluating this additional drift term will require accessing third order derivatives.

Stochastic gradient MCMC (SG-MCMC) and preconditioner choice – SG-MCMC (Welling and Teh, 2011) is a special class of MCMC algorithms that makes an extensive use of preconditioning and of the results from Girolami and Calderhead (2011). They extend MALA and HMC for stochastic estimates of the gradient. As in optimization, the stochasticity is generally included to avoid a costly evaluation of the full gradient. A first example of SGD algorithm was presented in Section 2.2.1.3. A second example of application for these methods is to sample a posterior distribution on the parameters of a neural network – see Chapter 4 for more details on neural networks. Such posterior distributions are generally not log-concave and very high-dimensional - $\mathcal{O}(10^6)$. Like in optimization methods for neural network training, preconditioning is necessary in this context.

We consider three main preconditioners proposed in SG-MCMC for a stochastic gradient MALA: the expected Fisher information matrix (Patterson and Teh, 2013), L-BFGS (Simsekli et al., 2016) and RMSProp (Li et al., 2016).

Using the full expected Fisher information matrix leads in general to a dense covariance matrix, i.e., with mostly non-zero terms. In high dimensions such as $\mathcal{O}(10^3)$ or $\mathcal{O}(10^4)$, inverting such a matrix is prohibitively expensive.

The L-BFGS preconditioner and its inverse are defined with sequences of vectors and inner products. They actually never require to compute a matrix or a matrix-vector product. This preconditioner thus scales very well. In addition, in Simsekli et al. (2016), the authors slightly modify the definition of L-BFGS to ensure $\boldsymbol{\gamma}^{(t-1)} = 0$ for all $t \geq 1$. However, this preconditioner was designed for log-concave posterior distribution. A positive regularization parameter is proposed to ensure positive definiteness of the preconditioner, but this parameter can become extremely large for highly non log-concave distributions. In the limit case of a very high regularization parameter, the preconditioner boils down to the identity matrix. Therefore, although this preconditioner is able to capture correlations between parameters, it is not applicable in the inverse problem considered in this thesis.

Finally, RMSProp is a diagonal preconditioner and thus is trivial to invert. As we mentioned in Section 2.2.1.2, this preconditioner shows good performance in practice. This performance in practice was partially explained in Dauphin et al. (2015). For this reason, in Chapter 5 (Section 5.2.1), we propose a PMALA sampling kernel equipped with the RMSProp preconditioner. Note that RMSProp being diagonal, it is not efficient in case of degeneracies, i.e., of high correlation between two parameters. Other preconditioners with particular structure such as a band matrix or block diagonal may still be used to address a limited number of degeneracies. However, as we will show in Chapter 5, RMSProp yields a drift term $\boldsymbol{\gamma}$ that can be evaluated with limited additional cost.

2.2.2.6 Gibbs sampling: a block coordinate MH variant

One of the current main challenges in MCMC is sampling from high dimensional posterior distributions, e.g., distributions on images. In Section 2.2.1.3, we showed that block coordinate optimization divides a high dimensional problem into smaller ones. In images, each block j can correspond to one pixel or to a group of pixels. Similarly, Gibbs sampling (Geman and Geman, 1984) is an MCMC algorithm that performs component-wise updates.

Gibbs and Metropolis-within-Gibbs algorithms – In Gibbs sampling, at each step t, the components $j = 1, \dots, J$ are updated sequentially. Each component j is updated by sampling from the conditional posterior

$$\boldsymbol{\Theta}_j^{(t)} \sim \pi\left(\boldsymbol{\Theta}_j | \mathbf{Y}, \boldsymbol{\Theta}_{\setminus j}^{\left(t-1+\frac{j-1}{J}\right)}\right), \tag{2.41}$$

with $\boldsymbol{\Theta}_{\setminus j}^{\left(t-1+\frac{j-1}{J}\right)} = \left(\boldsymbol{\Theta}_1^{(t)}, \dots, \boldsymbol{\Theta}_{j-1}^{(t)}, \boldsymbol{\Theta}_{j+1}^{(t-1)}, \dots, \boldsymbol{\Theta}_J^{(t-1)}\right)$. The $\boldsymbol{\Theta}_{\setminus j}$ notation corresponds to $\boldsymbol{\Theta}$ with the j^{th} block removed. As shown in Geman and Geman (1984), the Gibbs sampler leads to ergodic Markov chains for mild assumptions on the conditional posterior.

As we already discussed, sampling directly from the conditional posterior may be out of reach. The so-called *Metropolis-within-Gibbs* algorithm (Gelman et al., 2015, chapter 11) resorts to MH steps for each of the J components, i.e., using proposals q_j and performing an accept-reject step. Algorithm 2.3 summarizes the Metropolis-within-Gibbs algorithm. Both Gibbs sampling and Metropolis-within-Gibbs can produce geometric and uniform ergodic chains (Johnson et al., 2013). At each step t, Gibbs sampling and Metropolis-within-Gibbs generate J samples from conditional posterior distributions with lower dimension, instead of generating one sample from the full posterior distribution. Such a divide-and-conquer approach can be very useful for distributions that show a particular structure, such as images or time series.

Algorithm 2.3: Metropolis-within-Gibbs sampling

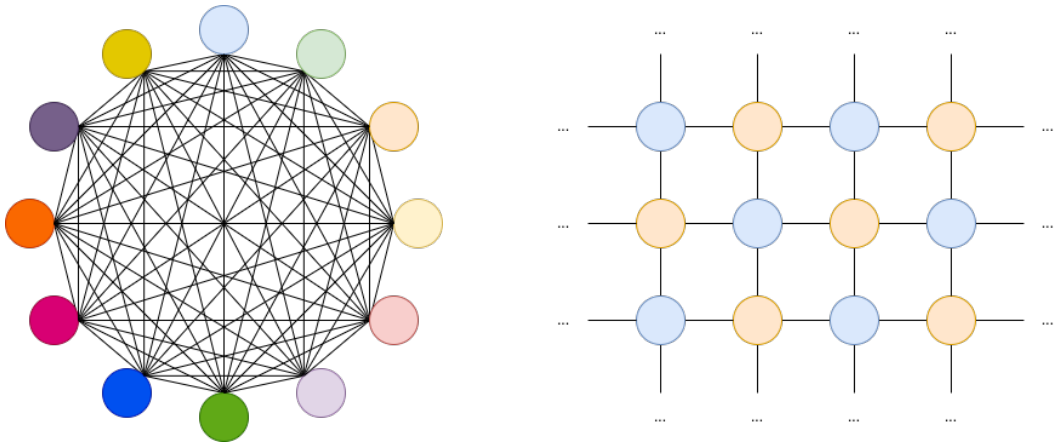
Input: Starting point $\boldsymbol{\Theta}^{(0)}$, number of samples T_{MC}

for $t = 1, \dots, T_{MC}$ **do**
  for $j = 1, \dots, J$ **do**
    $\boldsymbol{\Theta}_j^{(c)} \sim q_j\left(\boldsymbol{\Theta}_j | \mathbf{Y}, \boldsymbol{\Theta}_{\setminus n}^{\left(t-1+\frac{j}{J}\right)}\right)$ `// Generate candidate for component` j
    $\rho_j^{(t)}$ `// Compute acceptance probability using Eq. 2.34`
    $\zeta_j \sim \text{Unif}(0, 1)$ `// update iterate`
    $\boldsymbol{\Theta}_j^{(t)} = \boldsymbol{\Theta}_j^{(c)}$ if $\zeta_j \leq \rho_j^{(t)}$ else $\boldsymbol{\Theta}_j^{(t-1)}$

Output: Chain of samples $\{\boldsymbol{\Theta}^{(t)}\}_{t=1}^{T_{\text{MC}}}$

Chromatic Gibbs sampling – One constraint of the Gibbs and Metropolis-within-Gibbs algorithms is the sequential sampling of each component n. This sequential sampling is necessary to take into account the latest value of components n_0 and n_1 to update some other $n_2 \neq n_1$. However, in case of conditional independence between $\boldsymbol{\theta}_{n_1}$ and $\boldsymbol{\theta}_{n_2}$ when $\boldsymbol{\theta}_{n_0}$ is given, then $\boldsymbol{\theta}_{n_1}$ and $\boldsymbol{\theta}_{n_2}$ can be sampled in parallel. In general, the set of conditional dependencies between individual components $\boldsymbol{\theta}_n$ is represented as a graph. This graph can be colored such that two neighbor nodes never share the same color. Each color then corresponds to a set of components n that can be sampled in parallel using the chromatic Gibbs algorithm (Gonzalez et al., 2011). This parallelization, combined with vectorized computations, can considerably accelerate the sampling. Besides, this algorithm also produces ergodic Markov chains.

Figure 2.2 shows two examples of graphs of conditional dependence. The graph on the left shows the extreme case of a fully connected graph. Such a graph requires one color per node. In general, finding the minimum number of colors for a given graph is a NP-hard problem. However, in many practical use cases, one can easily find the smallest number of colors. The graph on the right shows a graph of pixels in an image. In this graph, a pixel only interacts with its direct neighbors. In this case, it is easy to see that the graph can be partitioned into two colors. Therefore, in this case, at each step t, only two steps are required, whatever the size of the image. In Chapter 5 (Section 5.1.2), we set a spatial regularization prior on maps of physical parameters that yields this exact graph. The sampling kernel that we propose in Chapter 5 (Section 5.2.2) thus exploits chromatic Gibbs sampling.



(a) Fully connected case: one color per node (b) An image as a Markov random field: two colors

Figure 2.2: Graphs of conditional dependence and associated optimal coloring.

2.2.2.7 Generating multiple candidates

As mentioned in Sections 2.2.2.3 and 2.2.2.4, the acceptance rate that leads to the best mixing is roughly 20%-25% for RWMH and 50%-60% for MALA. A proposal distribution that both achieves such acceptance rates and takes relatively large steps can be out of reach, especially in high dimensions. In such cases, very long chains are necessary to fully explore the posterior. An alternative to taking long chains is to generate multiple candidates at each step, which leads to an increase of the acceptance probability. This idea is at the core of multiple-try Metropolis (MTM) algorithms (Liu et al., 2000).

At each step t, MTM algorithms first generate $K \geq 1$ candidates $\boldsymbol{\Theta}_c^{(k)}$ instead of 1 in MH algorithms. Then, using an importance weight function

$$w(\boldsymbol{\Theta}_c^{(k)}) = \frac{\pi(\boldsymbol{\Theta}_c^{(k)}|\mathbf{Y}, \boldsymbol{\Theta}^{(t-1)})}{q(\boldsymbol{\Theta}_c^{(k)}|\boldsymbol{\Theta}^{(t-1)})}, \tag{2.42}$$

one candidate i is selected according to the set of selection probabilities $(w_k)_{k=1}^K$ with a multinomial distribution

$$w_k = \frac{w(\boldsymbol{\Theta}_c^{(k)})}{\sum_{j=1}^K w(\boldsymbol{\Theta}_c^{(j)})}. \tag{2.43}$$

The accept-reject step is then performed with the selected candidate i and the generalized acceptance probability (Liu et al., 2021; Martino, 2018)

$$\widetilde{\rho}^{(t)} = \min\left(1, \frac{w(\boldsymbol{\Theta}_c^{(i)}) + \sum_{j=1, j\neq i}^K w(\boldsymbol{\Theta}_c^{(j)})}{w(\boldsymbol{\Theta}^{(t-1)}) + \sum_{j=1, j\neq i}^K w(\boldsymbol{\Theta}_c^{(j)})}\right). \tag{2.44}$$

Algorithm 2.4 summarizes the MTM sampler. Note that for $K = 1$, MTM becomes equivalent to MH. Like MH, MTM verifies the detailed balance property and thus admits the posterior as invariant distribution. In addition, like MH, MTM produces ergodic Markov chains. Finally, like I-MH, proposal distributions q that are independent of the current iterate $\boldsymbol{\Theta}^{(t-1)}$ – independent multiple-try Metropolis (I-MTM) – leads to uniformly ergodic Markov chains if there exists a finite constant M such that $Mq(\boldsymbol{\Theta}) \geq \pi(\boldsymbol{\Theta}|\mathbf{Y})$ for all $\boldsymbol{\Theta}$. In particular, such proposal distributions do not get trapped in local modes. See Martino (2018) for a review on MTM algorithms.

Algorithm 2.4: Multiple-Try Metropolis (MTM) algorithm

Input: Starting point $\boldsymbol{\Theta}^{(0)}$, proposal q, number of candidates K, number of samples T_{MC}

for $t = 1, \ldots, T_{MC}$ **do**

`// Propose` K `candidates, select one`

$\boldsymbol{\Theta}_c^{(k)} \sim q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)})$ for $k = 1, \ldots, K$

$w(\boldsymbol{\Theta}_c^{(k)})$ for $k = 1, \ldots, K$ `// using Eq. 2.42`

w_k for $k = 1, \ldots, K$ `// using Eq. 2.43`

i `// select one candidate following probabilities` w_k

`// Accept or reject`

$\widetilde{\rho}^{(t)}$ `// using Eq. 2.44`

Draw $\zeta \sim \mathsf{Unif}\,(0, 1)$

$\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}_c^{(i)}$ **if** $\zeta \leq \widetilde{\rho}^{(t)}$ **else** $\boldsymbol{\Theta}^{(t-1)}$

Output: Chain of samples $\{\boldsymbol{\Theta}^{(t)}\}_{t=T_{\mathsf{BI}}+1}^{T_{\mathsf{MC}}}$

Advantages of MTM compared to MH – Using the MTM algorithm instead of MH might not seem advantageous at first, as it requires more computations per iteration. It may also seem that one step of MTM would be equivalent to K steps of MH, but with loss of the intermediate iterates. However, MTM is roughly to MH what chromatic Gibbs is to Gibbs sampling. The generation of the K candidates and the evaluation of their posterior pdf can be vectorized or performed in parallel. One step of MTM is therefore much faster than K iterations of MH, and about as fast as 1.

Besides, resorting to an MTM sampling kernel shortens the burn-in phase. In Chapter 5 (Section 5.2), we propose an MCMC algorithm that combines two kernels. Having at least one of the kernels based on the MTM algorithm permits to reach the posterior high probability region with few iterations. The other proposed kernel is based on RMSProp preconditioning (Eq. 2.21), which keeps in memory the history of the log-posterior gradient. Reaching the high probability region faster enables this second kernel to focus on gradient properties within the high probability region.

2.3 Comparing and checking observation models

As stated in Gelman et al. (2015, chapter 6): "Once we have accomplished the first two steps of a Bayesian analysis — constructing a probability model and computing the posterior distribution of all [physical parameters] — we should not ignore the relatively easy step of assessing the fit of the model to the data and to our substantive knowledge."

Solving an inverse problem relies on the assumption that "the forward model accurately simulates the observed phenomena and that the uncertainty model accurately describes the uncertainty sources that affect observations". For the inverse problem considered in this thesis, one might question the relevance of the Meudon PDR code to model a given region, the validity of the values of the secondary parameters listed in Table 1.3, or the chosen uncertainty model. These choices for the observation model have a great impact on the estimates. Unrealistic choices can greatly impact the estimation relevance for an end user. One thus needs to check whether the aforementioned hypothesis is satisfied or not, or compare models to select the one that is most compatible with the observations.

In this section, after briefly mentioning model selection and evaluation, we describe a well established Bayesian model assessment approach. Applications of both model evaluation and assessment in the ISM are described in Chapter 3 (Section 3.3). In this thesis, as considering more than one model is expensive, we focus on model assessment. The associated Bayesian hypothesis testing method is extended to be more robust in Chapter 5 (Section 5.3).

2.3.1 Model selection and evaluation

As already stated in Section 2.1, we group the forward model $\mathbf{f}$ and the noise model $\mathcal{A}$ in a single observation model $\mathcal{M}$. Model selection consists in comparing the results obtained with different models $\mathcal{M}_i$ and choosing the best one with respect to a quantitative criterion. There are two dominant approaches in Bayesian statistics. The first relies on the Bayesian evidence $\pi\left(\mathbf{Y}|\mathcal{M}_i\right)$ (Eq. 2.1), the second on the expected log-predictive density (elpd) and information criteria. We succinctly describe the two approaches.

Bayesian evidence approach – Having access to the evidence, also called marginal likelihood, for each considered model $\mathcal{M}_i$ permits to perform Bayesian model selection. Indeed, in the case of two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$, the Bayes theorem yields the relative model posterior probability

$$\frac{\pi\left(\mathcal{M}_1|\mathbf{Y}\right)}{\pi\left(\mathcal{M}_2|\mathbf{Y}\right)} = \frac{\pi\left(\mathbf{Y}|\mathcal{M}_1\right)}{\pi\left(\mathbf{Y}|\mathcal{M}_2\right)} \frac{\pi\left(\mathcal{M}_1\right)}{\pi\left(\mathcal{M}_2\right)}, \tag{2.45}$$

where $\pi\left(\mathcal{M}_1\right)$ and $\pi\left(\mathcal{M}_2\right)$ encode prior model preferences and sum to 1.

Computing the evidence is generally hard. Some methods exploit samples from the posterior – see e.g., McEwen et al. (2022), or Chib and Jeliazkov (2001) for MH samplers. Nested sampling (Skilling, 2004; Skilling, 2006) focuses on computing the evidence. Sequential MC (Del Moral et al., 2006), computes it as a byproduct. Nested sampling and SMC are described in Appendix 2.A. For a review on evidence estimation, see e.g., Friel and Wyse (2012).

Information criteria approach – A second option is to measure its so-called *predictive accuracy*, which relies on predicted observations $\widetilde{\mathbf{Y}}$. The predicted observations, also called reproduced observations, are distributed according to the *posterior predictive distribution*, given by

$$\pi\left(\widetilde{\mathbf{Y}}|\mathbf{Y},\mathcal{M}\right) = \int \pi\left(\widetilde{\mathbf{Y}}|\boldsymbol{\Theta},\mathcal{M}\right) \pi\left(\boldsymbol{\Theta}|\mathbf{Y},\mathcal{M}\right) d\boldsymbol{\Theta}. \tag{2.46}$$

In the context of atomic or molecular emission line observations, these predicted observations $\widetilde{\mathbf{Y}}$ can correspond to lines not used during the inversion. They can also simply correspond to synthetic reproduced observations of the same environment, i.e., with the same underlying physical conditions $\boldsymbol{\Theta}$, but with different noise realizations. In the following, we consider the latter case.

The predictive accuracy is usually measured with the expected log-predictive density (elpd) (Gelman et al., 2015, chapter 7),

$$\text{elpd} = \sum_{n=1}^{N}\sum_{\ell=1}^{L} \int \ln \pi\left(\widetilde{y}_{n\ell}|\mathbf{Y},\mathcal{M}\right)\ \pi_t(\widetilde{y}_{n\ell})\, d\widetilde{y}_{n\ell} \tag{2.47}$$

$$= \sum_{n=1}^{N}\sum_{\ell=1}^{L} \int \ln \left[\int \pi\left(\widetilde{y}_{n\ell}|\boldsymbol{\Theta},\mathcal{M}\right) \pi\left(\boldsymbol{\Theta}|\mathbf{Y},\mathcal{M}\right) d\boldsymbol{\Theta}\right]\ \pi_t(\widetilde{y}_{n\ell})\, d\widetilde{y}_{n\ell}, \tag{2.48}$$

with $\pi_t(\widetilde{y}_{n\ell})$ the pdf of the true generating process of new observations $\widetilde{y}_{n\ell}$, which is unknown. The elpd therefore cannot be evaluated directly, but can be approximated with cross-validation or information criteria. Cross-validation is computationally demanding, as it requires solving multiple inverse problems instead of one, but is accurate. Information criteria do not require solving additional inverse problems and are quite cheap to evaluate. They rely on a simplified estimator of the elpd, the log-predictive density (lpd)

$$\text{lpd} = \sum_{n=1}^{N}\sum_{\ell=1}^{L} \ln \pi(y_{n\ell}|\mathbf{Y},\mathcal{M}) \tag{2.49}$$

$$= \sum_{n=1}^{N}\sum_{\ell=1}^{L} \ln \left[\int \pi\left(y_{n\ell}|\boldsymbol{\Theta},\mathcal{M}\right) \pi\left(\boldsymbol{\Theta}|\mathbf{Y}\mathcal{M}\right)\right] d\boldsymbol{\Theta}, \tag{2.50}$$

which exploits observations $\mathbf{Y}$ used in the inversion to define the posterior distribution. Using T_{MC} samples $\boldsymbol{\Theta}^{(t)}$ of the posterior distribution, e.g., obtained with an MCMC algorithm, the lpd can be estimated with

$$\widehat{\text{lpd}} = \sum_{n=1}^{N} \sum_{\ell=1}^{L} \ln \left[\frac{1}{T_{\text{MC}}} \sum_{t=1}^{T_{\text{MC}}} \pi \left(y_{n\ell} | \boldsymbol{\Theta}^{(t)}, \mathcal{M} \right) \right]. \tag{2.51}$$

The lpd uses the observations $\mathbf{Y}$ in both the posterior distribution $\pi(\boldsymbol{\Theta}|\mathbf{Y})$ and in the predictive distribution $\pi(y_{n\ell}|\boldsymbol{\Theta}^{(t)}, \mathcal{M})$. Therefore, the lpd introduces an optimistic bias in the accuracy measure. In particular, the lpd might be largely overestimated in case of overfitting.

In the information criterion approach, this optimistic bias in the lpd is compensated with a positive correction term that is subtracted from the lpd. For any information criterion IC, the elpd thus is estimated with

$$\widehat{\text{elpd}}_{\text{IC}} = \widehat{\text{lpd}} - \widehat{p}_{\text{IC}} \tag{2.52}$$

where $\widehat{p}_{\text{IC}}$ is called an *effective number of parameters*. Its definition depends on the choice of the information criterion. For instance, the deviance information criterion (DIC) (Spiegelhalter et al., 2002), the widely applicable information criterion (WAIC) (Watanabe, 2010), or the WBIC (Watanabe, 2012) rely on different definitions of this effective number of parameters. We now present two examples of common information criteria and the associated effective number of parameters.

Example 1: Akaike information criterion (AIC) – Although we favor probabilistic approaches in this thesis, the information criterion approach can be used with a point estimate $\widehat{\boldsymbol{\Theta}}$, such as the maximum likelihood estimator or the maximum a posteriori. The posterior $\pi\left(\boldsymbol{\Theta}|\mathbf{Y}, \mathcal{M}\right)$ is then replaced by a Dirac mass located at the estimate $\widehat{\boldsymbol{\Theta}}$. The posterior predictive distribution then simplifies to the likelihood function, and the lpd to $\ln \pi(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \mathcal{M})$. Besides, directly using the number of inferred parameters – in this thesis the dimension of $\boldsymbol{\Theta}$, ND – is the simplest possible definition for the effective number of parameters. By combining a point estimate and this definition for the effective number of parameters, one obtains the classical Akaike information criterion (AIC) (Gelman et al., 2015, chapter 7)

$$\widehat{\text{elpd}}_{\text{AIC}} \;=\; \widehat{\text{lpd}} - \widehat{p}_{\text{AIC}} \;=\; \ln \pi \left(\mathbf{Y} | \widehat{\boldsymbol{\Theta}}, \mathcal{M} \right) - ND \tag{2.53}$$

which is usually defined with a factor -2, such that

$$\text{AIC} = -2 \ln \pi \left(\mathbf{Y} | \widehat{\boldsymbol{\Theta}}, \mathcal{M} \right) + 2ND \tag{2.54}$$

Example 2: the Bayesian information criterion (BIC) – Limited to the maximum likelihood estimator $\widehat{\boldsymbol{\Theta}}_{\text{MLE}}$, the BIC (Schwarz, 1978) uses $\widehat{p}_{\text{BIC}} = ND \ln(NL)/2$. The BIC is not exactly a Bayesian information criterion in the sense that it considers a point estimate $\widehat{\boldsymbol{\Theta}}_{\text{MLE}}$ and not a posterior distribution $\pi(\boldsymbol{\Theta}|\mathbf{Y}, \mathcal{M})$. Its name "Bayesian" comes from the fact that it is an approximation of the log-evidence based on Laplace's method (Konishi and Kitagawa, 2008, chapter 9).

For more information on elpd, cross validation and information criteria based on the posterior predictive distribution, see e.g., Gelman et al. (2015, chapter 7) or Vehtari et al. (2017). For a general review on information criteria, see, e.g., Konishi and Kitagawa (2008).

In general, an isolated value of Bayesian evidence lacks interpretability. This quantity is thus mostly used relatively, to compare multiple models. This is also true for information criteria. In particular, there exists no simple and general rule to decide whether a model $\mathcal{M}$ accurately reproduces observations.

2.3.2 Model checking using Bayesian hypothesis testing

When only one model $\mathcal{M}$ is considered, one might want to check whether it accurately reproduces the observations $\mathbf{Y}$, using hypothesis testing. One key advantage of this hypothesis testing compared to model selection or evaluation is that its metric, the p-value, is easily interpretable. In Bayesian statistics, such a hypothesis test is called *posterior predictive assessment*, or *checking*.

This subsection introduces posterior predictive assessment key notions, such as the underlying null hypothesis, the associated p-value, its Monte Carlo estimator, and the discrepancy measure T. Chapter 3 (Section 3.3.2) details three applications in interstellar medium studies. In Chapter 5 (Section 5.3), we extend the test to make it more robust to wrong decisions due to the Monte Carlo estimator error.

Posterior predictive checking was introduced in Guttman (1967) and Rubin (1984). Gelman et al. (1996) extended the definition to *discrepancy measures* T, more general than test statistics. It is based on the following null hypothesis: "The observation model $\mathcal{M}$ can reproduce the observations $\mathbf{Y}$". Like model evaluation methods based on information criteria, this test relies on reproduced observations $\widetilde{\mathbf{Y}}$ and on the posterior predictive distribution $\pi(\widetilde{\mathbf{Y}}|\mathbf{Y})$. The goal is to compare predicted observations $\widetilde{\mathbf{Y}}$ and the true observations $\mathbf{Y}$ with respect to a discrepancy measure $T : (\mathbf{Y}, \boldsymbol{\Theta}) \mapsto T(\mathbf{Y}, \boldsymbol{\Theta}) \in \mathbb{R}$. If the null hypothesis is true, then predictions $\widetilde{\mathbf{Y}}$ should be as likely as the true observations $\mathbf{Y}$ with respect to the measure T. Common discrepancy measures include $T(\widetilde{y}, \boldsymbol{\Theta}) = \widetilde{y}$ for scalar observations (Gelman et al., 2015), the negative log-likelihood, and the L_2-norm

$$T(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta}) = \sum_{n=1}^{N} \sum_{\ell=1}^{L} \frac{(\widetilde{y}_{n\ell} - f_\ell(\boldsymbol{\theta}_n))^2}{\sigma_{n\ell^2}}, \tag{2.55}$$

that is also called χ^2 loss in astrophysics articles.

The Bayesian p-value associated with this test corresponds to the probability of obtaining values $T(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta})$ at least as unlikely as $T(\mathbf{Y}, \boldsymbol{\Theta})$ under the null hypothesis.

$$p \;=\; \mathbb{P}_{(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta})} \left[T\left(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta}\right) \geq T\left(\mathbf{Y}, \boldsymbol{\Theta}\right) \middle| \mathbf{Y}, \mathcal{M} \right]. \tag{2.56}$$

Equivalently, it is the measure of the set $I = \{(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta}) \mid T(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta}) \geq T(\mathbf{Y}, \boldsymbol{\Theta})\}$ when using the model $\mathcal{M}$

$$p \;=\; \int \mathbb{1}_I \left(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta}\right) \; \pi\left(\widetilde{\mathbf{Y}} | \boldsymbol{\Theta}, \mathcal{M}\right) \; \pi\left(\boldsymbol{\Theta} | \mathbf{Y}, \mathcal{M}\right) \; d\boldsymbol{\Theta} d\widetilde{\mathbf{Y}}. \tag{2.57}$$

For classical choices of T, if the p-value is below a threshold α chosen prior to the analysis (typically 0.05 or 0.01), then the null hypothesis can be rejected with confidence $1 - \alpha$. For some other choices of T (as in Gelman et al. (2015, chapter 6)), the null hypothesis can be rejected when the p-value is below $\alpha/2$ or above $(1 - \alpha)/2$.

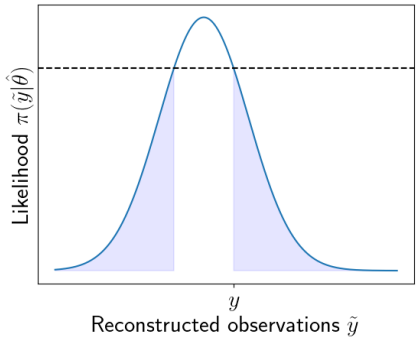
Analytical computation of the Bayesian p-value can be performed in some simple cases (Meng, 1994). For instance, for a point estimate $\widehat{\boldsymbol{\Theta}}$, i.e., a posterior distribution reduced to a Dirac on $\widehat{\boldsymbol{\Theta}}$, the statistic in Eq. 2.55 follows a χ^2_{NL} distribution. The associated p-value,

$$p = \mathbb{P}_{T(\widetilde{\mathbf{Y}}, \widehat{\boldsymbol{\Theta}}) \sim \chi^2_{NL}} \left[T(\widetilde{\mathbf{Y}}, \widehat{\boldsymbol{\Theta}}) \geq T(\mathbf{Y}, \widehat{\boldsymbol{\Theta}}) | \mathbf{Y}, \mathcal{M} \right], \tag{2.58}$$

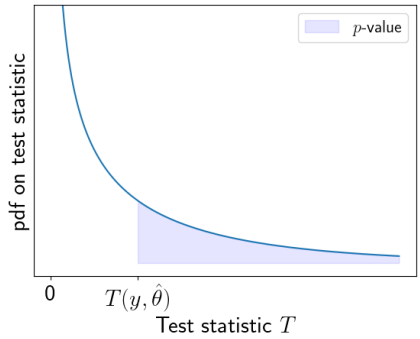
can thus be easily computed with the cdf of the χ^2 distribution.

Evaluation of the p-value – Figure 2.3 illustrates the presented p-value on two examples with a scalar observation $\mathbf{Y} = y \in \mathbb{R}$. In both cases, the statistic T is set to the negative log-likelihood. A point estimate $\widehat{\boldsymbol{\Theta}}$ is considered for simplicity and visualization. The first row shows a Gaussian additive noise model. The test statistic thus boils down to the χ^2 statistic (Eq. 2.55).

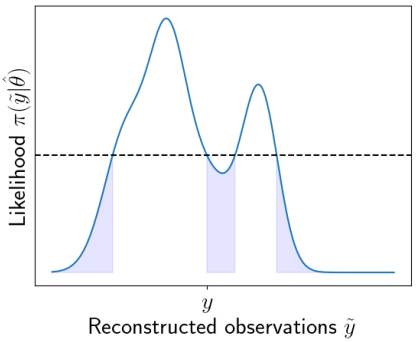
The second row corresponds to a more complicated fictitious observation model, defined as a Gaussian mixture model (GMM) with 3 components. The left column shows for both cases the likelihood distribution pdf on y with the parameter $\boldsymbol{\Theta}$ fixed to the estimator $\widehat{\boldsymbol{\Theta}}$. The black dashed line shows the attained likelihood pdf value for y, $\pi(y|\widehat{\boldsymbol{\Theta}})$. The blue area corresponds to the area to be integrated for the p-value computation (Eq. 2.57), i.e., the area for reproduced observations $\widetilde{y}$ such that $T(\widetilde{y}, \boldsymbol{\Theta}) \geq T(y, \boldsymbol{\Theta})$. The right column shows the pdf on the test statistic T. In Figure 2.3b, as $T \sim \chi^2_1$, the cdf can be evaluated efficiently. In Figure 2.3d, the considered test statistic T does not follow a simple distribution. Evaluating the cdf on the test statistic T exactly requires integrating with respect to $\widetilde{y}$. When the posterior distribution $\boldsymbol{\Theta}$ is not reduced to a dirac, on and , which is unrealistic in high dimensions.



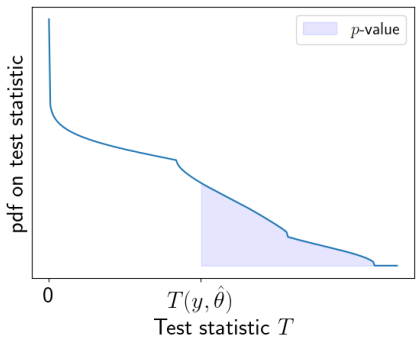
(a) Gaussian noise model



(b) The p-value can be evaluated efficiently



(c) Non-Gaussian noise model



(d) The p-value can not be evaluated efficiently

Figure 2.3: Application of the Bayesian p-value (Eq. 2.57) for a point estimate $\widehat{\boldsymbol{\Theta}}$.

For these general cases, the p-value is approximated with a Monte Carlo estimator (Gelman et al., 1996). As in model selection methods based on information criteria, this Monte Carlo estimator exploits T_{MC} samples $\boldsymbol{\Theta}^{(t)}$ from the posterior, typically obtained with an MCMC algorithm. In addition, it requires sampling observation reproductions $\widetilde{\mathbf{Y}}^{(t)}$ from the observation model $\mathcal{M}$ and the sample $\boldsymbol{\Theta}^{(t)}$. The p-value counts the number of iterations t where $T(\widetilde{\mathbf{Y}}^{(t)}, \boldsymbol{\Theta}^{(t)}) \geq T(\mathbf{Y}, \boldsymbol{\Theta}^{(t)})$, i.e., it is an empirical frequency

$$\widehat{p} = \frac{1}{T_{\text{MC}}} \sum_{t=1}^{T_{\text{MC}}} \mathbb{1}_I \left(\widetilde{\mathbf{Y}}^{(t)}, \boldsymbol{\Theta}^{(t)} \right), \quad \boldsymbol{\Theta}^{(t)} \sim \pi \left(\boldsymbol{\Theta} | \mathbf{Y}, \mathcal{M} \right), \quad \widetilde{\mathbf{Y}}^{(t)} \sim \pi \left(\widetilde{\mathbf{Y}} | \boldsymbol{\Theta}^{(t)}, \mathcal{M} \right), \tag{2.59}$$

where $\boldsymbol{\Theta}^{(t)} \sim \pi(\boldsymbol{\Theta}|\mathbf{Y})$ indicates that $\boldsymbol{\Theta}^{(t)}$ is drawn from the distribution of pdf $\pi(\boldsymbol{\Theta}|\mathbf{Y})$. In an MCMC algorithm, the iterates from burn-in phase should be disregarded for the p-value evaluation, as they are not considered to be samples from the posterior distribution. However, their influence on the estimated p-value decreases as the size of the Markov chain increases. Note that unlike model selection methods based on information criteria or the three-case χ^2 rule widespread

in astrophysics, this hypothesis testing framework does not penalize the number of free parameters. Overfitting cases thus cannot be detected, regardless of the test statistic $T(\mathbf{Y}, \boldsymbol{\Theta})$. Indeed, in such a case, by definition of overfitting, the null hypothesis is verified.

Marginal predictive checks – As mentioned in Gelman et al. (2015, chapter 6), a more informative hypothesis testing approach can be applied on individual observations $\widetilde{\mathbf{y}}_n$ instead of the full observation set $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{y}}_n)_{n=1}^N$. Such tests are called *marginal predictive checks* as they focus on the marginal predictive distribution $\pi(\widetilde{\mathbf{y}}_n|\mathbf{Y}, \mathcal{M})$ instead of the joint predictive distribution $\pi(\widetilde{\mathbf{Y}}|\mathbf{Y}, \mathcal{M})$. These tests can be used e.g., to find outliers among the observations $\mathbf{y}_n$ or to simplify the identification of issues in a problematic model $\mathcal{M}$.

Interpretability discussion – One interpretability limitation of this Bayesian p-value is that its null distribution, i.e., the distribution of p (Eq. 2.57) when the null hypothesis is true, is not uniform on $[0, 1]$ Bayarri and Berger (2000) and makes rejection rarer. This statistical behavior is illustrated in two examples in Gelman (2013) along with a discussion on the impact for the practical interest of the described p-value. Some alternative predictive posterior checks were calibrated to admit the uniform distribution asymptotically (Robins et al., 2000). These alternative checks are generally expensive to compute, while computing the presented p-value is very cheap. In Gelman et al. (2015, chapter 6), the authors emphasize that the described p-value is by construction a valid probability and should be used as such. In the following, we still use the presented p-value while acknowledging this limitation for interpretation.

In Chapter 3 (Section 3.3.2), we show applications of this model assessment approach in interstellar medium studies. Though the estimator $\widehat{p}$ (Eq. 2.59) converges to the theoretical p-value (Eq. 2.57) with the law of large numbers, using an approximation instead of the theoretical value introduces an error. In Chapter 5 (Section 5.3), we extend the test by taking this error into account. The test is therefore made more robust to wrong decisions.

2.4 Conclusion

In this chapter, we covered the three main steps involved in solving an inverse problem: statistical modeling, statistical inference and model assessment.

Bayesian statistical modeling consists in setting the inverse problem to solve. The distribution of interest is called the posterior distribution. It is defined with the Bayes' theorem and proportional to the product of the likelihood function and the pdf of the prior distribution. The normalization constant, called the Bayesian evidence, is not considered in this thesis. The likelihood function encodes the observation model, which includes the forward model and the noise model. The prior distribution encodes prior knowledge on the physical parameters to infer.

In Chapter 3 (Section 3.1), we review statistical modeling in interstellar medium studies. In Chapter 5 (Sections 5.1.1, 5.1.2, 5.1.3), we set the likelihood, prior and posterior of the general inverse problem addressed in this thesis, respectively.

For statistical inference, we described the two main approaches: optimization-based and sampling-based. In both cases, the most common methods were recalled. The optimization-based approach only provides point estimates. Conversely, the sampling-based approach, e.g., MCMC algorithms, naturally yields uncertainty quantification on the physical parameters $\boldsymbol{\Theta}$. Due to the absence of ground truth in ISM studies, we favor sampling methods to solve the inverse problem considered in this thesis. Some advanced MCMC methods such as preconditioned Metropolis adjusted Langevin algorithm (PMALA), Gibbs sampling, and multiple-try Metropolis (MTM) were covered.

Chapter 3 (Section 3.2) reviews applications of statistical inference in ISM studies. The sampler proposed in Chapter 5, combines a PMALA kernel that performs efficient local explorations

and a second kernel that combines chromatic Gibbs and MTM. By exploiting the natural structure of the parameter space, this second kernel can escape from local modes.

For model assessment, we covered Bayesian model selection and Bayesian model checking. The two main approaches of model selection, based on Bayesian evidence and information criteria, were explained. For model checking, we presented the Bayesian p-value from Gelman et al. (1996). Applications of both approaches in ISM studies are reviewed in Chapter 3 (Section 3.3). In ISM studies, often only one model is available, as generating forward models is expensive. In Chapter 5 (Section 5.3), we extend the model checking strategy by including uncertainties on the associated p-value.

Appendix 2.A Samplers widespread in astrophysics dedicated to multimodal distributions

In astrophysics, the considered forward models are almost exclusively non-linear, even in very simple cases – see, e.g., applications in Chapter 3. This non-linearity causes the negative log-posterior to be non-convex with possibly multiple modes. Multimodal distributions are known to be difficult to sample, especially when modes are isolated, i.e., when they are separated by large low-probability regions. In practice, samplers such as MH, MALA, or HMC fail to explore the full distribution. Indeed, these samplers tend to get stuck in one local minimum, and would require an unrealistically large amount of samples to visit all modes.

This appendix presents some existing sampling algorithms dedicated to multimodal distributions. Some of these algorithms are already popular in the ISM community, as we outline in Chapter 3 (Section 3.2.2). We do not use the algorithms listed in this appendix to solve the considered inverse problem. In Chapter 5 (Section 5.2.2), we propose a sampling kernel that addresses multimodality with a combination of chromatic Gibbs and MTM. However, we will compare them with the proposed sampler in Chapter 5 (Section 5.4).

2.A.1 Adapting meta-heuristics to MCMC

In optimization, meta-heuristics are common methods to address multimodality. After introducing meta-heuristics and particularly tempering and population methods, we present some samplers adapt these methods to the MCMC context. In particular, sequential MC (SMC) and the so-called affine invariant sampler, two popular methods in ISM studies – se Chapter 3 –, are presented.

Meta-heuristics algorithms were initially designed for discrete non-convex optimization problems such as the traveling salesman problem. They are stochastic in two ways: at each step t, they randomly generate candidates $\boldsymbol{\Theta}_c$, and then keep them with a certain probability that depends on the loss value $\mathcal{L}(\boldsymbol{\Theta}_c)$. The most popular meta-heuristics are simulated annealing (van Laarhoven and Aarts, 1987), genetic algorithms (Chelouah and Siarry, 2000), particle swarm (Eberhart and Kennedy, 1995; Kennedy and Eberhart, 1995), and some combinations of these algorithms (Kao and Zahara, 2008).

Genetic algorithm and particle swarm are both *population algorithms*. They first initialize a population of points, and then make the population evolve towards local minima with low loss values by combining current points or applying random mutations.

Simulated annealing resorts to *tempering*. Tempering relies on an inverse temperature parameter $\phi^{(t)} \in [0, 1]$. It first flattens the loss function by considering a tempered loss function $\boldsymbol{\Theta} \mapsto \phi^{(t)}\mathcal{L}(\boldsymbol{\Theta})$, with $\phi^{(t)}$ close to 0. This allows simple transitions between modes and permits to consider points with high loss to escape from local minima. During the optimization procedure, the inverse temperature $\phi^{(t)}$ progressively increases to 1 and transitions become more conservative. Simulated annealing enjoys some strong theoretical properties, though only valid for extremely slow temperature evolutions. For a complete introduction to meta-heuristics for both discrete and continuous optimization problems, see e.g., Dreo et al. (2006).

Tempering-based samplers without population – The equi-energy sampler (Kou et al., 2006) and the adaptive parallel tempering algorithm (Miasojedow et al., 2013) exploit tempering, as simulated annealing. These two algorithms run parallel interacting Markov chains at different temperatures. High temperature chains can navigate between modes and only one chain at low temperature is actually used for estimations.

Population-based samplers without tempering – Some other samplers adapt evolutionary methods such as the genetic algorithms to the MCMC context. For instance, Evolutionary MC (EMC) (Liang and Wong, 2000), distributed genetic MC (DGMC) (Hu and Tsui, 2010) do not

include any tempering, but run parallel interacting chains to generate relevant candidates. Besides, the so-called affine-invariant MCMC sampler (Goodman and Weare, 2010) is a special case of EMC that only applies the *snooker crossover* (Liang and Wong, 2001). To the best of our knowledge, this link between EMC and the affine-invariant sampler was never identified before.

The affine-invariant sampler runs $K > 1$ interacting Markov chains in parallel. At each step t, it updates an iterate $\mathbf{\Theta}^{(k,t-1)}$ by making it interact with another iterate $\mathbf{\Theta}^{(j,t-1)}$ with $j \neq k$. First, it draws a scaling from the distribution of pdf

$$\pi(z) \propto \frac{1}{\sqrt{z}} \mathbb{1}_{\left[\frac{1}{2},2\right]}(z).. \tag{2.60}$$

Second, it proposes a candidate $\mathbf{\Theta}_c^{(k,t)}$ using an affine combination of $\mathbf{\Theta}^{(k,t-1)}$ and $\mathbf{\Theta}^{(j,t-1)}$

$$\mathbf{\Theta}_c^{(k,t)} = \mathbf{\Theta}^{(j,t-1)} + z\left(\mathbf{\Theta}^{(k,t-1)} - \mathbf{\Theta}^{(j,t-1)}\right). \tag{2.61}$$

The candidate is then accepted with probability

$$\rho_{\text{emcee}}^{(t)} = \min\left(1,\ z^{ND-1}\frac{\pi(\mathbf{\Theta}_c^{(k,t)}|\mathbf{Y})}{\pi(\mathbf{\Theta}^{(k,t-1)}|\mathbf{Y})}\right). \tag{2.62}$$

Algorithm 2.5 summarizes the affine-invariant sampler.

Algorithm 2.5: Affine-invariant sampler

Input: Number of parallel chains K, starting points $(\mathbf{\Theta}^{(k,0)})_{k=1}^K$, number of iterations T_{BI} et T_{MC}

for $t = 1, \ldots, T_{MC}$ **do**
  for $k = 1, \ldots, K$ **do**
    `// generate candidate`
    $j \sim \text{Unif}(\{1, \ldots, k-1, k+1, \ldots, K\})$
    $z \sim \pi(z)$ `// using Eq. 2.60`
    $\mathbf{\Theta}_c^{(k,t)} = \mathbf{\Theta}^{(j,t-1)} + z\left(\mathbf{\Theta}^{(k,t-1)} - \mathbf{\Theta}^{(j,t-1)}\right)$ `// snooker crossover (Eq. 2.61)`
    `// Accept-reject step`
    $\rho_{\text{emcee}}^{(t)}$ `// using Eq. 2.62`
    $\zeta \sim \text{Unif}(0, 1)$
    $\mathbf{\Theta}^{(k,t)} = \mathbf{\Theta}_c^{(k,t)}$ **if** $\zeta \leq \rho_{\text{emcee}}^{(t)}$ **else** $\mathbf{\Theta}^{(k,t-1)}$

Output: Set of samples $\{\mathbf{\Theta}^{(k,t)},\ t = T_{\text{BI}}+1, \ldots, T_{\text{MC}},\ k = 1, \ldots, K\}$

This affine-invariant MCMC sampler – and the associated EMCEE package (Foreman-Mackey et al., 2013) – is a common sampler in astrophysics, as we state in Chapter 3 (Section 3.2.2). In principle, the snooker crossover enables to jump between modes. In Foreman-Mackey et al. (2013), the authors state the affine-invariant sampler is in practice not suited to multimodal cases. They recommend initializing the chains in a small ball centered around a point expected to be close to the MAP to avoid getting trapped in a low probability local mode.

Sequential MC (SMC): combining tempering and population methods – SMC (Del Moral et al., 2006) combines tempering with a population approach. It was originally defined as a generalization of Kalman filters to non-linear or non-Gaussian state spaces. In this context, SMC, also known as Particle filter (Del Moral, 2004), draws samples from a succession of distributions. It was adapted to multimodal sampling by defining a succession of K tempered posterior distributions

$$\pi_k(\mathbf{\Theta}|\mathbf{Y}) \propto [\pi(\mathbf{Y}|\mathbf{\Theta})]^{k/K}\ \pi(\mathbf{\Theta}). \tag{2.63}$$

It first generates T_{MC} samples from the prior distribution. These samples are then modified as the likelihood is progressively included in the tempered posterior by increasing an inverse temperature parameter $k/K \in [0,1]$[5]. Other sequences of inverse temperature can also be used.

At each step k, in a first correction step, the weights $w^{(k,t)}$ of the T_{MC} samples $\boldsymbol{\Theta}^{(k-1,t)}$ are evaluated. Then, in a selection step, the population is resampled according to the weights $\{w^{(k,t)}\}_{t=1}^{T_{\text{MC}}}$. Finally, in a mutation step, each member of the resampled population goes through a few steps of the MH algorithm with a kernel $\mathcal{K}_k$ that admits the tempered posterior $\pi_k(\cdot|\mathbf{Y})$ as invariant distribution. This last step can be performed in parallel. Algorithm 2.6 summarizes the SMC algorithm. As a by-product, SMC can also yield the Bayesian evidence, which is useful for model selection – see Section 2.3.

Algorithm 2.6: Sequential Monte Carlo (SMC)

Input: number of intermediate steps K, number of samples T_{MC}, number of MH steps N_{MH}, transition kernels $\mathcal{K}_k$

Initialization: Starting points $\boldsymbol{\Theta}^{(0,t)} \sim \pi(\boldsymbol{\Theta})$ for $t = 1, \dots, T_{\text{MC}}$

for $k = 1, \dots, K$ **do**

`// Correction: compute weights`

$\widetilde{w}^{(k,t)} = \frac{\pi_k(\mathbf{Y}|\boldsymbol{\Theta}^{(k-1,t)})}{\pi_{k-1}(\mathbf{Y}|\boldsymbol{\Theta}^{(k-1,t)})}$ for $t = 1, \dots, T_{\text{MC}}$

$w^{(k,t)} = \frac{\widetilde{w}^{(k,t)}}{\sum_{t=1}^{T_{\text{MC}}} \widetilde{w}^{(k,t)}}$ for $t = 1, \dots, T_{\text{MC}}$ `// normalize the weights`

`// Selection: resample the population`

sample $\widetilde{\boldsymbol{\Theta}}^{(k,t)}$ from $\{\boldsymbol{\Theta}^{(k-1,t)}\}_{t=1}^{T_{\text{MC}}}$ with selection probabilities $\{w^{(k,t)}\}_{t=1}^{T_{\text{MC}}}$

`// Mutation: apply` N_{MH} `steps of MH`

$\boldsymbol{\Theta}^{(k,t)} \sim \mathcal{K}_k^{N_{\text{MH}}}\left(\widetilde{\boldsymbol{\Theta}}^{(k-1,t)}, \cdot\right)$ for $t = 1, \dots, T_{\text{MC}}$

Output: Set of samples $\{\boldsymbol{\Theta}^{(K,t)}\}_{t=1}^{T_{\text{MC}}}$

2.A.2 Prior or parallel identification of the modes

Some methods resort to an augmented distribution with a latent mode index. The posterior distribution is approximated by a mixture model on $\boldsymbol{\Theta}$, where each model corresponds to a mode. Such methods sample from the posterior using two kernels: a local kernel that explores around a mode, and a jump kernel that permits jumps between already identified modes. Such methods include darting MC (DMC) (Andricioaei et al., 2001), jumping adaptative multimodal sampler (JAMS) (Pompe et al., 2020), regeneration darting MC (RDMC) (Ahn et al., 2013) and wormhole Hamiltonian Monte Carlo (WHMC) (Lan et al., 2014). WHMC is a particular case of the Riemannian manifold HMC algorithm introduced in Section 2.2.2.5. The metric of the corresponding manifold combines the standard Euclidean distance and a wormhole metric that shortens the distances between already identified modes, which simplifies transitions from one to another.

DMC and JAMS require a prior identification of the distribution modes by some optimization methods. RDMC and WHMC allow running optimization methods in parallel to the sampler and update the distribution and the sampler parameters at so-called random *regeneration times* (Gilks et al., 1998). However, in high-dimensional settings and with a non-linear forward model, the posterior has potentially many modes with only a few of significant weight in the mixture. The

[5] The presented tempering approach is called *likelihood tempering*. Other types of tempering exist, such as *data tempering* or *model tempering* (Mlikota and Schorfheide, 2022).

identification of the relevant modes with standard optimization methods remains difficult in this context.

2.A.3 Nested sampling

Nested sampling (Skilling, 2004; Skilling, 2006) is a class of algorithms whose primary goal is to compute the Bayesian evidence and that produces posterior samples as a by-product. Nested sampling is therefore not an MCMC algorithm. To evaluate the evidence $\pi(\mathbf{Y}|\mathcal{M})$, here noted Z, it considers the prior mass associated with a likelihood value greater than $v \geq 0$

$$x(v) = \int_{\pi(\mathbf{Y}|\mathbf{\Theta})>v} \pi(\mathbf{\Theta})\, d\mathbf{\Theta}. \tag{2.64}$$

By construction, the function x decreases on $[0, \pi(\mathbf{Y}|\widehat{\mathbf{\Theta}}_{\text{MLE}})]$. It verifies $x(0) = 1$ and $x(\pi(\mathbf{Y}|\widehat{\mathbf{\Theta}}_{\text{MLE}})) = 0$. Besides, a range $[x, x + dx]$ corresponds to the prior weight associated with likelihood values in $[v, v + dv]$. The evidence Z can then be rewritten as the uni-dimensional integral

$$Z = \int_0^1 v(x)dx, \tag{2.65}$$

where $v : x \mapsto v(x)$ is the inverse function of $x : v \mapsto x(v)$. As x is decreasing on $[0, \pi(\mathbf{Y}|\widehat{\mathbf{\Theta}}_{\text{MLE}})]$, so is v on $[0, 1]$.

In its first version Skilling (2004); Skilling (2006), nested sampling considers a population of K points $\mathbf{\Theta}^{(k,t)}$. At each iteration t, the point $\mathbf{\Theta}^{(i,t-1)}$ with lowest likelihood v_t is exploited to update the Bayesian evidence estimator $\widehat{Z}^{(t)}$. This estimator implements the rectangle approximation of an integral:

$$\begin{aligned} \widehat{Z}^{(t)} &= \sum_{\tau=1}^{t} v_\tau \left[\exp\left(-\frac{\tau-1}{T_{\text{MC}}}\right) - \exp\left(-\frac{\tau}{T_{\text{MC}}}\right) \right] \\ &= \widehat{Z}^{(t-1)} + v_t \left[\exp\left(-\frac{t-1}{T_{\text{MC}}}\right) - \exp\left(-\frac{t}{T_{\text{MC}}}\right) \right]. \end{aligned} \tag{2.66}$$

Then, the point $\mathbf{\Theta}^{(i,t-1)}$ is updated to $\mathbf{\Theta}^{(i,t)}$ with MCMC steps. The other points in the population are passed as is to the next iteration, i.e., $\mathbf{\Theta}^{(k,t)} = \mathbf{\Theta}^{(k,t-1)}$, for all $k \neq i$. In these steps, the transition kernel $\mathcal{K}$ samples from the prior distribution with the constraint on the likelihood value $\pi(\mathbf{Y}|\mathbf{\Theta}^{(k)}) > v_t$. Therefore, in nested sampling, the main difficulty lies in sampling from the prior with a hard lower-level constraint on the likelihood values. Algorithm 2.7 summarizes the procedure of nested sampling. Figure 2.4 illustrates the principle of this original version of nested sampling in a two-dimensional case. In particular, note that higher population sizes K lead to horizontally finer rectangles, and that higher number of iterations T_{MC} permit to explore high likelihood regions.

In more recent papers, better sampling strategies were proposed including ellipsoidal rejection sampling (Feroz and Hobson, 2008) - and the associated code MultiNest (Feroz et al., 2009), diffusive sampling (Brewer et al., 2011) and slice sampling (Handley et al., 2015) - and the associated code PolyChord. For a general review on nested sampling, see e.g., Buchner (2023).

Nested sampling is applicable for distributions with dimension up to $\mathcal{O}(10^3)$ for the latest methods. One proximal nested sampling algorithm was proposed in Cai et al. (2022) to be applied to distributions on very high dimensional spaces – up to $\mathcal{O}(10^6)$ dimensions. However, this algorithm was built for imaging analysis and thus is only valid for log-concave posterior distributions. This property is seldom verified in astrophysical studies, due to the non-linearity of the forward model.

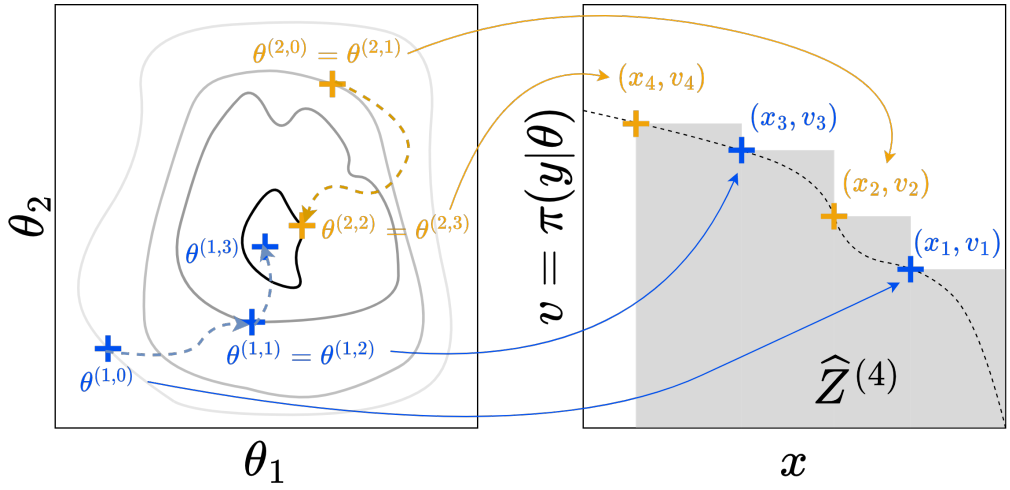


Figure 2.4: Illustration of nested sampling principle in a 2-dimensional parameter space with $T_{\text{MC}} = 4$ steps and a population of size $K = 2$. (Left) Parameter space with a population of $K = 2$ points $\boldsymbol{\theta}^{(k,t)} \in \mathbb{R}^2$. The gray lines space represent the associated likelihood isocontours in the parameter space, with darker lines corresponding to higher likelihood. At each step t, one of the two points is selected to update the evidence estimator $\widehat{Z}^{(t)}$. The selected point is then updated: the dashed arrows show their trajectories. (Right) illustration of the integral in Eq. 2.65, with x_t the prior mass associated with a likelihood greater than v_t, as defined in Eq. 2.64. The true evidence is the area under the dashed curve. The evidence estimator $\widehat{Z}^{(t)}$ from Eq. 2.66 is the grayed area.

Algorithm 2.7: Nested sampling

Input: population size K, number of iterations T_{MC}, sampling kernel $\mathcal{K}$, number of MCMC iterations N_{MCMC}

Initialization: Sample K points from the prior $\boldsymbol{\Theta}^{(k,0)} \sim \pi(\boldsymbol{\Theta})$, set of samples $\mathcal{S}^{(0)} = \emptyset$, Bayesian evidence estimator $\widehat{Z}^{(0)} = 0$

for $t = 1, \ldots, T_{MC}$ **do**

- $i = \arg\min_k \pi(\mathbf{Y}|\boldsymbol{\Theta}^{(k,t-1)})$ `// select lowest likelihood`
- $v_t = \pi(\mathbf{Y}|\boldsymbol{\Theta}^{(i,t-1)})$
- $x_t = \exp(-t/T_{\text{MC}})$
- $w_t = x_{t-1} - x_t$ `// weight of sample` $\boldsymbol{\Theta}^{(k)}$ `in the set of samples`
- $\widehat{Z}^{(t)} = \widehat{Z}^{(t-1)} + v_t w_t$ `// update evidence estimator (Eq. 2.66)`
- $\mathcal{S}^{(t)} = \mathcal{S}^{(t-1)} \cup \{(\boldsymbol{\Theta}^{(i,t-1)}, w_t)\}$ `// update set of samples`
- $\boldsymbol{\Theta}^{(i,t)} \sim \mathcal{K}^{N_{\text{MCMC}}}(\boldsymbol{\Theta}^{(i,t-1)}, \cdot)$ `// update` $\boldsymbol{\Theta}^{(k)}$ `to increase its likelihood`
- $\boldsymbol{\Theta}^{(k,t)} = \boldsymbol{\Theta}^{(k,t-1)}$ for all $k \neq i$ `// the other points are not updated`

Output: Bayesian evidence estimator $\widehat{Z}^{(T_{\text{MC}})}$, set of samples $\mathcal{S}^{(T_{\text{MC}})}$

Chapter 3

Review of statistical inference in ISM studies and position of our problem

"No computer is ever going to ask a new, reasonable question. It takes trained people to do that."

Grace Hopper

Contents

In Chapter 2, we described some statistical modeling and inference methods, including model selection and model assessment. This chapter reviews the main statistical approaches adopted in ISM studies. As we will show, ISM studies involving integrated intensities of atomic or molecular emission lines are mostly based on grid search. Few sampling-based studies have been proposed, limited to low-dimensional settings. Conversely, cosmology was among the first astrophysics fields to widely adopt sampling methods and to use advanced models. For instance, many works published in the early $2\,000$'s infer cosmological parameters from observations of the cosmological microwave background, sometimes combined with other observations. Therefore, we expand the review to stellar physics and cosmology, where the degree of adoption of Bayesian statistical modeling and inference is higher.

The goal of this chapter is to detail existing applications in ISM studies at each step of the modeling or inference to be able to motivate our contributions. It isolates the different steps and components of the Bayesian methodology. Section 3.1 describes the statistical modeling tendencies, including how numerical forward models are handled and the choice of noise model and of prior distribution. Section 3.2 lists popular methods for both optimization-based and sampling-based inference. Section 3.3 reviews applications of model selection and model checking. Finally, Section 3.6, briefly compares our contributions with the literature.

3.1 Statistical modeling

As presented in Chapter 2 (Section 2.1), statistical modeling aims at defining a posterior distribution $\pi(\boldsymbol{\Theta}|\mathbf{Y})$. The posterior involves a prior distribution $\pi(\boldsymbol{\Theta})$ and a likelihood function $\pi(\mathbf{Y}|\boldsymbol{\Theta})$. In turn, the likelihood function is defined from a forward model – usually a numerical model in astrophysics – and a noise model. In this section, we discuss choices and standards in the ISM community and some related communities on the forward model, the noise model and the prior distribution.

3.1.1 Forward model: handling a numerical model

For any numerical model, small changes in the physical parameters can lead to very different predicted observables. Adopting a realistic forward model $\mathbf{f}$ that predicts observables from physical parameters is critical for inference. Codes that model the observed environment more realistically lead to more meaningful estimations. However, the complexity of the physics considered in a code directly impacts its evaluation time and hence its usability. This is especially true for inference, which requires many evaluations of the forward model.

3.1.1.1 Direct evaluations or using precomputed models

There are two main approaches to handle a numerical forward model in inference, depending on their speed. The first approach consists in running the code during the inversion procedure. This is only possible for fast numerical codes. The second approach targets slower codes. It consists in generating a grid of precomputed models prior to running the inference procedure.

Running the code during the inversion – On the one hand, the lighter codes can run in a few seconds and are often used directly in optimization or Bayesian inference. In ISM studies, many inverse problems rely on a fast model, such as those listed in Chapter 1 (Section 1.2). Combinations of these codes can also be used. For instance, Behrens et al. (2022) and Keil et al. (2022) rely on a combination of RADEX and ULCCHEM – see Chapter 1 (Sections 1.2.1 and 1.2.2). Even for the fastest codes, the evaluation of the numerical code is usually the slowest step in the inversion procedure. In astrochemistry, many efforts were spent on reducing the execution time. Some large chemical networks were simplified to smaller ones (Holdship et al., 2018). Alternatively, the original numerical model is used directly in the inversion, but the number of

evaluations is limited by progressively reusing past evaluations (Keil et al., 2022).

Generating a dataset of precomputed models prior to the analysis – On the other hand, a more comprehensive model such as the Meudon PDR code Le Petit et al. (2006) – introduced in Chapter 1 – that handles multiple physical processes on a 1D spatial grid typically requires several hours of computations. These durations are prohibitively long for inferring physical parameters on large observation maps. These heavier numerical codes are frequently evaluated on a grid $\{\boldsymbol{\theta}_i\}$ on the parameter space prior to any analysis. Following the ISM community standard, we call *precomputed model* one evaluation $\mathbf{f}(\boldsymbol{\theta}_i)$ of a numerical model $\mathbf{f}$ prior to any analysis. Besides, we call a set of $(\boldsymbol{\theta}, \mathbf{f}(\boldsymbol{\theta}))$ pairs a *dataset of precomputed models* or *grid of precomputed models* interchangeably, whether the dataset has a regular grid structure or not.

Datasets of precomputed models can be used for grid search in optimization approach or to approximate integrals with Bayesian-like approaches – see Section 3.2. They can also be used to derive faster emulators of the numerical model.

We now discuss the choice of the structure of the dataset, which can impact both uses, and strategies to derive emulators.

3.1.1.2 Datasets structure

A dataset of precomputed models can either be built at once, prior to any analysis, or iteratively until an accuracy criterion is satisfied. In Chapter 4 (Section 4.3), we resort to the first option. We build a dataset of precomputed models at once using a lattice structure.

Generation of the dataset prior to any analysis – The first approach is widespread in ISM studies. Most datasets of precomputed models are structured as lattices, i.e., as uniform regular grids, or as near-uniform grids with few and light differences in spacings. For instance, Sheffer et al. (2011) resorts to a lattice structure to perform a grid search in an optimization approach. Similarly, Pérez-Montero (2014) and Blanc et al. (2015) use this structure to approximate integrals in a Bayesian approach. Finally, Ramambason et al. (2022) and Smirnov-Pinchukov et al. (2022) exploit this structure to train emulators of their numerical models based on a nearest-neighbor interpolation and a k-nearest-neighbor regression, respectively.

Although very common, uniform or near-uniform grids are not systematically used. For instance, to derive an emulator of the Meudon PDR code, Wu et al. (2018) resorts to a non-uniform grid of models with more points in the region of interest in the parameter space. This non-uniformity ensures good accuracy of the emulator in this region. Besides, Bron et al. (2021) generates two datasets using independent and identically distributed (i.i.d.) draws from the uniform distribution on a cube defined with lower and upper values in the physical parameter space.

In cosmology, datasets are often generated using the latin hypercube sampling (LHS) algorithm – to be introduced in Chapter 4. For instance, see Spurio Mancini et al. (2022); Mootoovaloo et al. (2022); Agarwal et al. (2012). Some of its variants are also popular Heitmann et al. (2009); Ramachandra et al. (2021).

Iterative construction of the dataset – There are only two cases in ISM studies, to the best of our knowledge:

- In Vale Asari et al. (2016), the authors start from a coarse lattice for a first approximation, and refine it with an octree. An octree is a tree structure that represents the first grid as its first "layer". Regions of the parameter space in which the emulator is not accurate enough are divided into smaller regions and explored with finer grids. See Saftly et al. (2013) for an introduction to octrees with an application in a Monte Carlo radiative transfer model.

- In Galliano (2018), the authors start from a coarse grid and progressively refine it using a mid-point strategy until they satisfy an accuracy criterion.

For a given size of the dataset, the iterative approach usually takes much more time compared to the first approach, but can lead to better performance by focusing on challenging regions of the parameter space. In other words, a non-iterative approach can lead to oversampled parameter spaces, i.e., to more evaluations than needed to obtain a good approximation of the original model. For large grids, this oversampling can result in high storage and computational costs. In particular, the grid might be too large to be loaded in memory.

Our proposal – In Chapter 4 (Section 4.3), we present the results obtained with a lattice structured dataset of precomputed models of the Meudon PDR code. This dataset was generated to derive an emulator of the Meudon PDR code.

3.1.1.3 Learning an emulator

When the numerical model is too slow to be used directly in an inversion procedure, it is common to derive faster emulators. In this thesis, an *emulator* of a numerical code is an approximation of the relation between its inputs $\boldsymbol{\theta}$ and its outputs $\mathbf{f}(\boldsymbol{\theta})$, derived from a dataset of precomputed models. Emulating the numerical model, e.g., with interpolation methods, permits to predict observables for new points $\boldsymbol{\theta}$ with lower evaluation time. In particular, an emulator does not perform the same intermediate computations as the original code.

Emulating the full likelihood function, which includes both the numerical model and the observation uncertainty model, is possible but quite rare. Indeed, it requires specifying a noise model, and is therefore either observation-specific or very generic. For instance, the SkyNet artificial neural network (ANN) (Graff et al., 2014) emulates the full likelihood function by approximating a numerical model and by assuming a noise model. The noise model is considered Gaussian and uncorrelated with fixed variance for continuous variable inference. The likelihood function is set to the cross-entropy (Bishop, 2006, chapter 4) for classification tasks. For a short introduction to ANNs, see Chapter 4 (Section 4.2.1). In this thesis, we focus on numerical model emulation to handle separately the forward model and the noise model.

In **ISM studies**, many numerical model emulations have already been proposed, with different choices of class of functions to define the emulator:

- Interpolation methods are well spread in the ISM community. For instance, linear interpolation was used in Galliano (2018) to emulate a combination of simple dust codes. Linear interpolation was also used in Ramambason et al. (2022), combined with nearest-neighbor interpolation to emulate Cloudy, a HII region model. Similarly, radial basis function interpolation are used in Wu et al. (2018) to emulate the Meudon PDR code. Spline interpolation methods are sometimes used to compute integrals in Bayesian approaches (Blanc et al., 2015), but not to emulate numerical models in the ISM community. However, they are used for emulator derivation in other communities. For instance, Bailer-Jones (2011) emulates a simulator using a thin-plate spline to infer a set of stellar parameters.
- Classic machine learning regression approaches are punctually used. For instance, in Smirnov-Pinchukov et al. (2022), a k-nearest-neighbor regression emulates a protoplanetary disks model. In Bron et al. (2021), a random forest emulates a chemistry model.
- ANNs form a versatile class of functions that is often preferred to address the complexity of comprehensive ISM models. For instance, ANN emulators of astrochemical models are derived in de Mijolla et al. (2019); Holdship et al. (2021); Grassi et al. (2022). In addition, in Grassi et al. (2011), the authors derive a new simulation code and an associated ANN emulator.

In **cosmology**, deriving emulators of numerical codes such as Camb that compute power spectra has received a lot of attention. Polynomial regression was used at first (Jimenez et al.,

2004; Fendt and Wandelt, 2007). Since 2007, Gaussian process regression and ANNs are the two main approaches. Gaussian processes are used, e.g., in Heitmann et al. (2009); Ramachandra et al. (2021); Mootoovaloo et al. (2022), and ANNs e.g., in Auld et al. (2007); Agarwal et al. (2012); Manrique-Yus and Sellentin (2019). The current state-of-the-art emulators, CosmoPower (Spurio Mancini et al., 2022), are fast ANNs that achieve an estimated mean error of 0.4% on power spectra. These state-of-the-art emulators were derived to be used as fast surrogate models in MCMC algorithms for Bayesian parameter inference.

Our proposal – We resort to ANNs to approximate the Meudon PDR code. We chose this family of functions as 1) it enjoys strong theoretical properties, 2) it proved state-of-the-art in a variety of regression problems, including in ISM studies, and 3) it allows to easily compute first and second order derivatives using auto-differentiation. Chapter 4 (Section 4.6) presents the design and training strategies we proposed to address specificities of complex ISM numerical models.

3.1.2 Noise model and forward model misspecification

Once the forward model is selected, an uncertainty model on the observations needs to be specified. Indeed, it is highly unlikely for a forward model to reproduce exactly the observations, unless over-parameterized. The noise model defines the likelihood function, which quantifies how distant a prediction $\mathbf{f}(\boldsymbol{\theta})$ is from the corresponding observation $\mathbf{y}$. In this subsection, we describe existing noise models used in ISM studies.

The noise model of the inverse problem considered in this thesis will be fully introduced in Section 3.4. It involves an additive thermal Gaussian noise and a multiplicative lognormal noise. The multiplicative uncertainty source encodes both calibration errors and forward model misspecification. In addition to noise, observations may contain censored observations due to the sensitivity limits of the telescopes.

3.1.2.1 Notation

For this subsection on noise models, we extend the notation, in particular $\mathbf{Y}$ and $\boldsymbol{\Theta}$. In this subsection, the L observables are not restricted to the integrated intensity of emission lines, and may correspond to the flux in spectral channels. Similarly, the N components are not restricted to pixels of a map, but correspond to the more general notion of *observation beams*, which does not assume a map structure. Maps of N pixels and with integrated intensities of L ionic, atomic or molecular emission lines are thus a special case of this extended notation. In general, the observation model can be written

$$\forall n \in [\![1, N]\!], \quad \forall \ell \in [\![1, L]\!], \quad y_{n\ell} = \mathcal{A}(I_{n\ell}), \tag{3.1}$$

where $I_{n\ell}$ is the original true and unaltered value for observable ℓ and beam n and $\mathcal{A}$ is a general observation operator. The observation operator $\mathcal{A}$ can include measurement noise, calibration error, modeling error from the choice of forward model, or censorship, i.e., upper bounds on observables. The number of beams ranges from $N = 1$, e.g., in Pérez-Montero (2014); Keil et al. (2022), while some others consider multiple beams, e.g., $N = 176$ pixels in Wu et al. (2018) and $N = 798$ galaxies in Galliano et al. (2021). Similarly, depending on the observed signal, the number L of observables per beam can be $\mathcal{O}(10^2)$ to $\mathcal{O}(10^3)$ for non spectrally integrated signals and $\mathcal{O}(10^0)$ to $\mathcal{O}(10^1)$ for line integrated intensities.

Inverse problem usually rely on the assumption that the true signals $I_{n\ell}$ can be reproduced exactly and simultaneously for all n and ℓ by a forward model $\mathbf{f}$ for a physical parameter $\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)_{n=1}^N$. In other words, it is assumed that there exists a physical parameter $\boldsymbol{\Theta}$ such that for all n and ℓ, $I_{n\ell} = f_\ell(\boldsymbol{\theta}_n)$. The observation model (Eq. 3.1) is therefore rewritten

$$\forall n \in [\![1, N]\!], \quad \forall \ell \in [\![1, L]\!], \quad y_{n\ell} = \mathcal{A}(f_\ell(\boldsymbol{\theta}_n)). \tag{3.2}$$

In the following, we list some observation operators $\mathcal{A}$ used in ISM studies, including noise models and censorship.

3.1.2.2 Purely Gaussian noise models

Gaussian and additive uncertainty models are widespread in ISM studies:

$$y_{n\ell} = f_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)}, \quad \boldsymbol{\varepsilon}^{(a)} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^{(a)}), \tag{3.3}$$

where $\boldsymbol{\varepsilon}^{(a)}$ is a measurement noise with zero mean and covariance matrix $\boldsymbol{\Sigma}^{(a)}$. Such a model is often a good approximation for noise models thanks to the central limit theorem. Besides, it is extremely simple to manipulate. For these two reasons, Gaussian distributions are a by-default choice in statistical modeling in ISM inverse problems (see e.g., Galliano et al. (2003); Chevallard et al. (2013); Pérez-Montero (2014); Chevance et al. (2016); Wu et al. (2018); Lee et al. (2019); Roueff et al. (2021); Keil et al. (2022); Behrens et al. (2022)). In all these works, the covariance matrix $\boldsymbol{\Sigma}^{(a)}$ is diagonal, i.e., all the noise components are assumed independent. In Galliano et al. (2003); Chevance et al. (2016); Lee et al. (2019), this observation model underlies the choice of the χ^2 as a loss function:

$$\mathcal{L}(\boldsymbol{\Theta}) = \chi^2_\nu(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \sum_{\ell=1}^{L} \frac{\left(f_\ell(\boldsymbol{\theta}_n) - y_{n\ell}\right)^2}{\sigma_{n\ell}^2}, \tag{3.4}$$

where ν is a so-called *degree of freedom*. This choice of notation comes from the fact that a sum of ν i.i.d. squared Gaussian random variables follows a χ^2_ν distribution. The degree of freedom is often distinct from the dimension of $\boldsymbol{\Theta}$, even for linear models $\mathbf{f}$ (Andrae et al., 2010). In other words, in general, $\nu \neq NL$. In some papers (e.g., Vale Asari et al. (2016)), the Gaussian noise is considered on the log of the observations $\log y_{n\ell}$, which is equivalent to a multiplicative noise $\boldsymbol{\varepsilon}^{(m)}$ following a lognormal distribution. A non-diagonal covariance matrix $\boldsymbol{\Sigma}^{(a)}$ permits to account for correlation in Gaussian noise models. This assumption appears, e.g., in star property inference (Bailer-Jones, 2011) where $\boldsymbol{\Sigma}^{(a)}$ is known.

3.1.2.3 Combining additive and multiplicative noises

It is common in ISM studies to combine two sources of noise to account for thermal noise, calibration noise, or model misspecification. Recall that in the inverse problem considered in this thesis, we consider a Gaussian additive thermal noise and a multiplicative lognormal noise. This latter noise includes both calibration error and model misspecification.

Gaussian additive and Gaussian multiplicative distributions – Some ISM studies combine a Gaussian additive noise $\boldsymbol{\varepsilon}^{(a)}$ and a Gaussian multiplicative noise $\boldsymbol{\varepsilon}^{(m)}$. This multiplicative error can represent calibration noise (Gordon et al., 2014; Ciurlo et al., 2016; Galliano, 2018; Galliano et al., 2021) or model misspecification (Blanc et al., 2015; Vale Asari et al., 2016; Jóhannesson et al., 2016). The observation model then reads

$$y_{n\ell} = \varepsilon_{n\ell}^{(m)} f_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)}, \quad \boldsymbol{\varepsilon}^{(m)} \sim \mathcal{N}(1, \boldsymbol{\Sigma}^{(m)}), \quad \boldsymbol{\varepsilon}^{(a)} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^{(a)}). \tag{3.5}$$

The covariance matrices $\boldsymbol{\Sigma}^{(a)}$ are assumed known and diagonal in all these papers. In Vale Asari et al. (2016) and Jóhannesson et al. (2016), the multiplicative error $\varepsilon_{n\ell}^{(m)}$ is a single scalar value equal for all observations, and handled as a nuisance parameter, i.e., inferred and marginalized. In Blanc et al. (2015), the multiplicative uncertainty source $\varepsilon_{n\ell}^{(m)}$ is also a unique scalar with fixed variance. When the observation $\mathbf{Y}$ is a multispectral map, a hyperspectral map, or a set of multichannel observations, the covariance matrix $\boldsymbol{\Sigma}^{(m)}$ is assumed known. In such cases, it needs to be specified both spatially and spectrally. Ciurlo et al. (2016) assumes no correlation,

i.e., independent realizations for all n and ℓ. Gordon et al. (2014) considers a partial spectral correlation and no spatial correlation. Galliano (2018); Galliano et al. (2021) consider a partial spectral correlation and a spatial correlation of exactly 1.

Using a Gaussian model for multiplicative noise simplifies computation, as the overall uncertainty model is Gaussian. However, depending on the standard deviation of the multiplicative error, a Gaussian model may not accurately account for multiplicative uncertainty. A typical model for multiplicative noise is the lognormal distribution, i.e., a Gaussian distribution for the multiplying factor on the log scale. The lognormal distribution is symmetric in log scale, i.e., the chances of multiplying or dividing by a given factor are equal, which is not the case with a Gaussian model.

Figure 3.1 illustrates this property. It shows that a Gaussian approximation is relevant for a low mean calibration error (left), but inappropriate for larger multiplicative errors (right). In particular, in the latter case, a Gaussian distribution allows negative multiplicative factors, while they cannot be accommodated by the lognormal distribution.

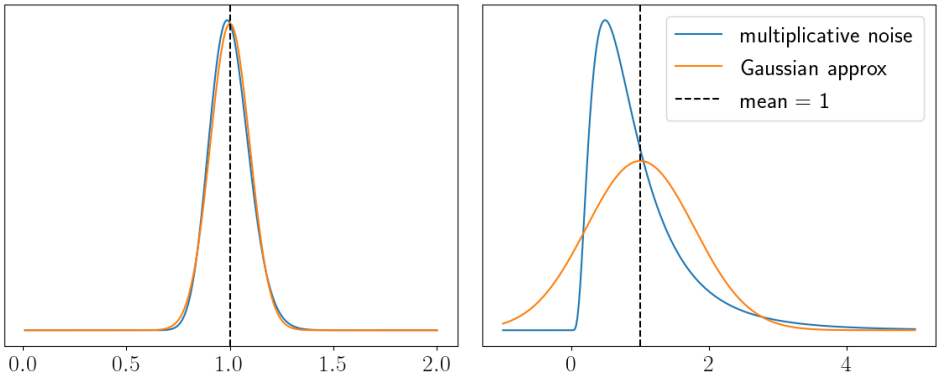


Figure 3.1: Quality of Gaussian approximation of a lognormal distribution for low and high variance. Left: the multiplicative noise corresponds to an error of 10% in average. The associated lognormal distribution is quite symmetric and is thus well approximated by a Gaussian distribution. Right: the multiplicative noise corresponds to an error of a factor 2 in average. The associated lognormal distribution is significantly asymmetric and is thus poorly approximated by a Gaussian distribution.

Student's t additive and Student's t multiplicative distributions – Using a non-Gaussian uncertainty model usually leads to a complex likelihood function with no closed-form expression. For instance, in Kelly et al. (2012), the observation model is similar to Eq. 3.5, but with different noise models: the additive noise $\boldsymbol{\varepsilon}^{(a)}$ and the log of the multiplicative noise $\log \boldsymbol{\varepsilon}^{(m)}$ are assumed to be independent and to follow a Student's t distribution, whose pdf is

$$\forall n, \ell, \quad \pi\left(\varepsilon_{n\ell}^{(a)}\right) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{\left(\varepsilon_{n\ell}^{(a)}\right)^2}{v}\right)^{-\frac{v+1}{2}}, \tag{3.6}$$

with v the degree of freedom of the distribution, set to $v = 8$ in Kelly et al. (2012). The Student's t distribution boils down to the Cauchy distribution for $v = 1$ and to the Gaussian distribution for $v = \infty$. In the article, the covariance matrix $\boldsymbol{\Sigma}^{(a)}$ of the additive noise is known and diagonal, and the calibration errors $\varepsilon_{n\ell}^{(m)}$ are assumed independent spectrally and with spatial correlation of exactly 1.

Setting a Student's t distribution on the additive noise $\varepsilon_{n\ell}^{(a)}$ and on the log of the multiplicative noise $\log \varepsilon_{n\ell}^{(m)}$ is equivalent to assuming a Gaussian distribution where the standard deviations are unknown but described with an inverse gamma prior. As the estimation of the standard deviations

is performed prior to the inverse problem, using a Student's t distribution is a statistically relevant way to encode uncertainty on the noise standard deviations.

With these two Student's t noise sources, the likelihood cannot be written in closed-form. The authors obtained a closed-form likelihood by using a hierarchical model, i.e., by including $\log \boldsymbol{\varepsilon}^{(m)}$ in the set of parameters to infer. This auxiliary variable was thus sampled, and ignored when deriving estimators.

Other combinations of noises – Mixtures of noise models (Robert and Casella, 2004, chapter 1) are sometimes exploited to set one model for standard observations and one for outliers. In astrochemistry, Gratier et al. (2016) uses a Gaussian mixture model (GMM) with two components

$$y_{n\ell}|\boldsymbol{\theta}_n \sim (1-p_o)\,\mathcal{N}(\mu_s,\sigma_s^2) + p_o\,\mathcal{N}(\mu_o,\sigma_o^2), \quad p_o \in [0,1], \tag{3.7}$$

where $\mathcal{N}(\mu_s,\sigma_s^2)$ is the noise model for standard observations, $\mathcal{N}(\mu_o,\sigma_o^2)$ is a model for outliers, typically with $\sigma_o \gg \sigma_s$, and p_o is the probability for the observation to be an outlier. Similarly, to evaluate distances to local molecular clouds, Zucker et al. (2019) mixes a Gaussian model and a uniform one. In both cases, the mixture parameter p_o is also inferred, which makes the model hierarchical. Gratier et al. (2016) goes farther and also infers most of the noise model parameters from the data.

3.1.2.4 Including censorship in the likelihood function

Upper bounds on observations are quite common in ISM studies. The corresponding observables are sometimes discarded in the inversion process. Omitting observables leads to a loss of information that could damage the inference results. Including these upper bounds on observations in the likelihood function permits to account for all constraints provided by the observations. In statistics, censorship permits to include such upper limits $\omega \geq 0$. In case of Gaussian noise, the observation model becomes

$$y_{n\ell} = \max\left\{\omega,\ f_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)}\right\} = \begin{cases} \omega & \text{if } f_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)} \leq \omega \\ f_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)} & \text{otherwise} \end{cases} \tag{3.8}$$

In (Ramambason et al., 2022), censorship is modeled with a half-normal distribution $\mathcal{N}_-(\omega,\sigma^2)$. Up to an additive constant, the associated likelihood function is

$$-\ln \pi(y_{n\ell}|\boldsymbol{\theta}_n) = \frac{1}{2\sigma^2}\left(f_\ell(\boldsymbol{\theta}_n) - \omega\right)^2\ \iota_{[-\infty,\omega]}\left(f_\ell(\boldsymbol{\theta}_n)\right), \tag{3.9}$$

where $\iota_{[-\infty,\omega]}(f_\ell(\boldsymbol{\theta}_n)) = 0$ if $f_\ell(\boldsymbol{\theta}_n) \leq \omega$ and $+\infty$ otherwise. This approach suffers from two main drawbacks. First, it prohibits values that are above the censoring threshold ω, though values slightly above ω should not be unrealistic. Second, it biases $f_\ell(\boldsymbol{\theta}_n)$ towards values close to the threshold ω, while values of $f_\ell(\boldsymbol{\theta}_n) \ll \omega$ are also compatible with the upper bound constraint.

In statistics, the standard approach to encode censorship in the likelihood is (Robert and Casella, 2004, chapter 1)

$$-\ln \pi(\omega|\boldsymbol{\theta}_n) = -\ln \int_{-\infty}^{\omega} \pi(y_{n\ell}|\boldsymbol{\theta}_n)\, dy_{n\ell} \tag{3.10}$$

The integral covers all the integrated intensities $y_{n\ell} \leq \omega$, and sums the associated uncensored likelihoods $\pi(y_{n\ell}|\boldsymbol{\theta}_n)$.

Figure 3.2 illustrates this censorship modeling for a Gaussian additive noise model $\varepsilon_{n\ell}^{(a)} \sim \mathcal{N}(0,\sigma^2)$. In this case, the integral in Eq. 3.10 can be evaluated efficiently for Gaussian noise using the Gaussian cumulative density function (cdf) $\Phi(\omega|f_\ell(\boldsymbol{\theta}),\sigma^2)$. Unlike the half Gaussian mentioned above, it does not forbid all physical parameters $\boldsymbol{\theta}$ such that $f_\ell(\boldsymbol{\theta}) > \omega$. Instead, it

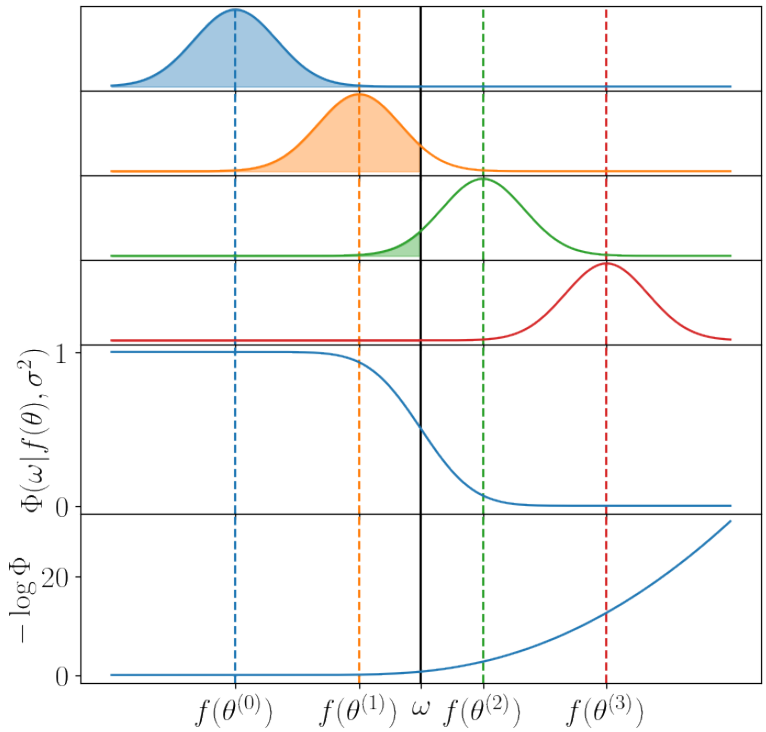


Figure 3.2: Illustration of censorship in likelihood for Gaussian additive noise models $\varepsilon_{n\ell}^{(a)} \sim \mathcal{N}(0,\sigma^2)$. The first four panels show four values of $f(\boldsymbol{\theta}^{(i)})$, $i = 0,\ldots,3$ and the associated Gaussian additive pdf. The censored portion, below the censorship threshold ω, is highlighted. The fifth panel shows the likelihood function taking censorship into account. The sixth panel shows the corresponding negative log-likelihood.

smoothly and increasingly penalizes them as $f_\ell(\boldsymbol{\theta})$ gets farther to the threshold limit, i.e., as the probability of $y_{n\ell} \leq \omega$ becomes smaller.

This censorship modeling is already used in some ISM inversion codes (Blanc et al., 2015; Holdship et al., 2018; Thomas et al., 2018).

3.1.2.5 Avoiding the manual specification of a noise model

Finally, two types of approaches do not require specifying a noise model to perform the inference. The first one abandons the noise model specification, while the second learns it from observation without any signal.

Likelihood-free approach – The most famous approaches that avoid the noise specification are called *likelihood-free*, with for instance the approximate Bayes computing (ABC) methods, introduced in Beaumont et al. (2002) for statistical inference in genetics. See, e.g., Cranmer et al. (2020) for a review on likelihood-free methods and Beaumont (2010) for a review on ABC methods. Put in simple words, ABC methods replace the likelihood function by a distance between summary statistics of true and simulated observations. They were first used in cosmology in Cameron and Pettitt (2012). Then, they became quite popular, e.g., with Weyant et al. (2013), COSMOABC (Ishida et al., 2015), ABCPMC (Akeret et al., 2015)[1], and ASTROABC (Jennings and Madigan, 2017). ABC sampling is much rarer in ISM studies. It is used e.g., in Robin et al. (2014) to study the thick disk in the Milky Way.

Learning the noise from observations – Abandoning all hope of specifying an accurate noise model and replacing it with a distance on some statistics seems suboptimal. A recent alternative consists in learning a correct noise model from observations of noise realizations without any signal. For instance, in Hahn et al. (2019), the authors combine GMMs and independent component analysis (ICA) (see, e.g., Hastie et al. (n.d., chapter 14)) to better approximate the likelihood from datasets usually used for covariance analysis. In Legin et al. (2023), the noise distribution of some observations of the Hubble space telescope and of the James Webb spatial telescope (JWST) are learned with a score-based diffusion generative model (see, e.g., Song and Ermon (2020)). This type of approach fully learns the noise distribution. Hierarchical models can also learn from observations to estimate noise parameters. However, these parameters depend on a noise model set a priori, which is more restrictive. To the best of our knowledge, no published work in the ISM community tries to fully learn a noise distribution from observations.

Our proposal

In the general inverse problem considered in this thesis, we consider a Gaussian additive noise $\boldsymbol{\varepsilon}^{(a)} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^{(a)})$, a lognormal multiplicative error $\boldsymbol{\varepsilon}^{(m)} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^{(m)})$ and a censorship level $\omega \geq 0$. The multiplicative noise includes calibration error and, depending on the observation, a model misspecification error. The two covariance matrices $\boldsymbol{\Sigma}^{(a)}$ and $\boldsymbol{\Sigma}^{(m)}$ are assumed to be known and diagonal. Censorship is modeled as in Eq. 3.10. As the multiplicative noise is not Gaussian, the overall likelihood function has no closed-form expression. In Section 3.4, we detail the uncertainty model, its origin and the corresponding assumptions. Chapter 5 (Section 5.1.1) proposes an approximation with controlled error of the associated intractable likelihood function.

3.1.3 Prior distributions and regularization functions

A prior distribution, or equivalently a regularization function, can encode any physical knowledge on the parameters to infer $\boldsymbol{\Theta}$. There are two main philosophies regarding inference. The first considers that the statistical model should only exploit information contained in the observed data, and thus sets the prior as non-informative as possible, if any. The second prefers to add physical

[1] https://abcpmc.readthedocs.io/en/latest/

prior knowledge, e.g., to favor desirable properties in the posterior or to rule out non-physical solutions.

Non-informative and weakly-informative priors are extremely widespread in ISM studies. The most common is a uniform prior on $\boldsymbol{\Theta}$, with validity intervals on its components. This prior is either stated explicitly (e.g., in Behrens et al. (2022); Blanc et al. (2015); Thomas et al. (2018); Holdship et al. (2018)) or implicitly, when working on a lattice dataset (e.g., in Joblin et al. (2018); Sheffer et al. (2011); Sheffer and Wolfire (2013)). This uniform prior can be set on the linear or log scale, depending on the physical parameter and its dynamic range. For this prior to be weakly informative, the lower and upper bounds are set to obtain wide intervals. These bounds typically correspond to the orders of magnitude spanned by the physical parameter – see, e.g., Wu et al. (2018); Joblin et al. (2018).

Informative priors are punctually exploited to encode additional knowledge. There are two main classes of informative priors in the considered astrophysics communities:

- The first class exploits the structure of the physical parameter $\boldsymbol{\Theta}$. For instance, Paumard et al. (2014); Ciurlo et al. (2016); Paumard et al. (2022) fit Gaussian line profiles on hyperspectral observations. Exploiting the map structure of the physical parameter, they enforce spatial smoothness in the maps by adding an L_1-L_2 spatial regularization. The ROHSA fitting algorithm (Marchal et al., 2019) also fits Gaussian line profiles on hyperspectral maps using a spatial regularization. However, ROHSA relies on the L_2 norm of the map Laplacian as a regularization. Spatial regularization improves reconstructions particularly in the low signal-to-noise ratio regions, where the observations mostly contain noise.

- The second class directly encodes physical information. For instance, Wu et al. (2018) rules out non-physical solutions with a physically informed prior on predicted observations $\mathbf{f}(\boldsymbol{\theta})$. This prior constrained the reconstructed integrated intensities to be decreasing for the CO lines $J = 11 - 10$, $J = 12 - 11$, and $J = 13 - 12$, as this condition seems always satisfied in observations. In stellar properties studies, more advanced physics-informed priors are sometimes used, such as the Hertzsprung-Russell Diagram prior on population of stars in Bailer-Jones (2011). This prior ensures that inference results are physically consistent.

In the case of informative priors, one may know a relevant family of distributions for a considered inverse problem, but face uncertainty on the values of its parameters, e.g., its mean or covariance. This hyperparameter tuning problem is crucial for inference, as different values yield different priors and thus different trade-offs. For instance, Ciurlo et al. (2016) presents a tedious manual setting of the six hyperparameters of its L_1-L_2 spatial regularization. Hierarchical Bayesian models, also called multi-level models, represent an intermediate between weakly-informative priors and wrong over-informative priors. They offer a way to decompose the prior knowledge into several elementary components, while accounting for uncertainties over each included item of knowledge. The parameters of the prior distribution can then be inferred from the data along with the physical parameters of interest $\boldsymbol{\Theta}$.

In the ISM community, Kelly et al. (2012) popularized hierarchical models for dust studies. For instance, Galliano (2018) – which introduces the dust Bayesian inversion code HERBIE – and Galliano et al. (2021) rely on a multivariate Student's t distribution prior on all physical parameters. Using a hierarchical statistical model, the mean and covariance parameters of this prior distribution are inferred along with the physical parameters. To simplify the inference of the covariance matrix, they exploit the separation method (Barnard et al., 2000) to divide covariances into correlations and standard deviations. Hierarchical models have been also used in other dust studies e.g., in Juvela et al. (2013); Veneziani et al. (2013). To the best of our knowledge, this approach still remains to be applied in other ISM fields.

Our proposal – The proposed approach will exploit the spatial regularity of the parameters to be inferred through the L_2 norm of the Laplacian operator, while restraining their values to physically consistent ranges.

3.2 Statistical inference

In this section, we review the inference methods widespread in ISM studies to evaluate estimators defined either as minimum of a loss function or with an integral.

3.2.1 Optimization-based inference

The MLE and MAP are widespread estimators in ISM studies. The methods listed below are introduced in Chapter 2 (Section 2.2.1).

Grid search – Searching for the point in the dataset that best reproduces the observations is very common in the inversion of integrated intensities of ionic, atomic or molecular lines e.g., in Sheffer et al. (2011); Sheffer and Wolfire (2013); Joblin et al. (2018); Lee et al. (2019). This grid search approach only works in low dimensional settings, as the number of required precomputed models grows exponentially with the number of physical parameters. Besides, it requires very fine grids to return accurate estimations, which means many evaluations of the forward model.

Gradient descent (GD) – GD algorithms are also quite common in ISM studies. They lead to more accurate results and work in higher dimensional settings. One of the most common GD method is the Levenberg-Marquardt algorithm, used e.g., in Schilke et al. (2010) to fit Herschel/HIFI spectra of an ortho-H_2O^+ line with a radiative transfer code, and in Galliano et al. (2003) in a dust study. Other GD algorithms are sometimes used in ISM studies, such as conjugate gradient descent in Paumard et al. (2022) and preconditioned gradient descent – using the limited memory BFGS (L-BFGS) preconditioner – in Wu et al. (2018). These gradient descent algorithms perform well with unimodal posteriors. However, unlike grid search methods, they can get trapped in local minima in case of a multimodal posterior.

Meta-heuristics – In multimodal cases, meta-heuristics such as simulated annealing and genetic algorithms are sometimes preferred to escape from local minima, e.g., in the MAGIX inversion code (Möller et al., 2013). These algorithms are also used in cosmology (Hannestad, 1999) or in the inference of star properties (Sarro et al., 2018).

The main limitation of optimization-based methods is that they only provide point estimates, with no information about the associated uncertainties. This is highly problematic, in particular with non-informative or weakly-informative priors and ill-posed problems, i.e., badly constrained problems. For instance, low SNR observations or tracers with limited sensitivity with respect to physical parameters can lead to non-physical results. Additional studies are necessary to quantify uncertainty. For instance, Roueff et al. (2021) quantifies the uncertainty associated with the MLE with the Cramér-Rao bound. As noted by Panter et al. (2003), this approach is only relevant when the posterior is well approximated by a Gaussian at its mode, which is generally not the case in astrophysics. In contrast, sampling-based approaches provide credibility intervals on the physical parameter $\boldsymbol{\Theta}$ along with point estimates for general posterior distributions.

3.2.2 Sampling-based inference

A sampling-based approach gives access to estimators defined with an integral – see Chapter 2 (Section 2.1.2). Several methods were already used in the ISM community.

Riemann integration with a grid of pdf evaluations – A simple method to evaluate an integral performs a Riemann integration with a grid in the physical parameter space. The negative log-posterior is evaluated at each point of the grid. This method was applied in some dust physics studies (Da Cunha et al., 2008; Pacifici et al., 2012). The Bayesian-like HII-CHI-MISTRY code (Pérez-Montero, 2014) uses a grid of CLOUDY models to study HII regions. From a grid of models, it returns a mean estimator computed with a weighted χ^2. This grid approach was implemented for other ISM inverse problems with extremely low-dimensional settings ($N = 1$, $D \leq 10$), e.g., in the codes IZI (Blanc et al., 2015), BOND (Vale Asari et al., 2016), NEBULABAYES (Thomas et al., 2018) or in Villa-Vélez et al. (2021). As the required number of precomputed models grows exponentially with the dimension, this approach does not scale to higher dimensions.

Markov chain Monte Carlo (MCMC) algorithms, introduced in Chapter 2 (Section 2.2.2.2), scale better to high dimensions. MCMC algorithms were popularized in astronomy through cosmology in Christensen et al. (2001). The first public MCMC code, COSMOMC was published in Lewis and Bridle (2002). These articles used random walk Metropolis-Hastings (RWMH) to generate posterior samples, arguably the most widespread sampling kernel. COSMOMC then became common in cosmological parameter inference, see, e.g., Tegmark et al. (2004), and was also applied in stellar formation (Acquaviva et al., 2011), dust studies in nearby galaxies (Chevallard et al., 2013) and fundamental properties of galaxies (Serra et al., 2011). All these applications had low dimensionality – up to 11 in Lewis and Bridle (2002). RWMH was also applied in ISM studies e.g., in dark interstellar clouds (Makrymallis and Viti, 2014), in dust (Paradis et al., 2010) or in astrochemistry (Makrymallis and Viti, 2014).

Some more advanced Bayesian methods, introduced in Chapter 2 (Section 2.A), are also already popular in the ISM community:

- The affine-invariant MCMC sampler and the associated EMCEE package is sometimes considered as the most popular MCMC algorithm in astronomy (Thrane and Talbot, 2019). It was applied in astrochemistry (Gratier et al., 2016; Holdship et al., 2018; Keil et al., 2022), extragalactic molecular gas (Yang et al., 2017). It is also popular out of the ISM community, e.g., in stellar physics (Jackiewicz, 2020; Kashyap et al., 2021). This sampler can exclusively address low dimensional problems, and requires to be initialized close to a mode (Foreman-Mackey et al., 2013).

- Nested sampling is also a very popular Bayesian framework in astrophysics and ISM studies. Popular algorithms include MULTINEST (Feroz et al., 2009), DYNESTY (Speagle, 2020) and ULTRANEST (Buchner, 2016b; Buchner, 2021), that were developed by and for astrophysicists. MULTINEST is applied e.g., to study extragalactic molecular clouds (Kamenetzky et al., 2014; Chevallard and Charlot, 2016) and cosmic ray propagation (Jóhannesson et al., 2016). DYNESTY is applied e.g., to evaluate distances between the Earth and nearby molecular clouds (Zucker et al., 2018; Zucker et al., 2019) or to reconstruct their 3D structure (Zucker et al., 2021). ULTRANEST is applied e.g., on extragalactic SEDs (Behrens et al., 2022). In particular, these codes can handle multimodal distribution, but remain limited to low dimensional distributions, e.g., up to 30 in a toy case for MULTINEST Feroz et al. (2009) and in the inverse problem considered in Jóhannesson et al. (2016).

- Sequential MC (SMC) is quite rare in ISM but is nonetheless applied in Ramambason et al. (2022) in an inversion of integrated intensities of ionic, atomic and molecular emission lines – the problem we are interested in in this thesis. In that work, it is chosen for its ability to sample from multimodal distributions in low dimension settings - up to 28 in the paper, for three sectors - and to compute the Bayesian evidence. SMC is also used when the likelihood is not specified with an ABC approach, especially in cosmology (see, e.g., Cameron and Pettitt (2012); Weyant et al. (2013); Jennings and Madigan (2017)).

- One sampling algorithm based on a combination of MH, Gibbs sampling and ancillarity-sufficiency interweaving strategy (ASIS) (Yu and Meng, 2011) is widespread in dust studies that involve a hierarchical model. ASIS is a sampling strategy that relies on complementary data augmentation schemes to better explore the posterior, including along strongly correlated directions. This sampler was first used in a dust study in Kelly et al. (2012), with implementation details provided in Kelly (2011). It was then coupled with a Gibbs block coordinate approach to exploit the structure in the physical parameter $\boldsymbol{\Theta}$ in very high dimensional applications. For instance, Galliano (2018) involve a few thousand dimensions, and (Galliano et al., 2021) about $10\,000$.

Our proposal – Recall that in the inverse problem considered in this thesis, the posterior distribution is non-log-concave and thus potentially multimodal. The negative log-posterior is also assumed to be twice differentiable and non gradient Lipschitz continuous. Besides, the physical parameters we aim at inferring live in high dimensions – up to $\mathcal{O}(10^4)$. The RWMH or the affine-invariant sampler do not scale well with dimensionality and thus are not able to sample from the considered posterior. Similarly, nested sampling and sequential MC could bypass the multimodality and non gradient Lipschitz issues, but do not scale well to high dimensions. Finally, ASIS is based on two complementary data augmentation schemes. There exists other data augmentation schemes in the signal processing community dedicated to accelerating MCMC algorithms, e.g., (Vono et al., 2019). In this thesis, we do not resort to data augmentation strategies. In Chapter 5 (Section 5.2), we define a dedicated sampler. Section 5.4 demonstrates that it yields state-of-the art performance on two multimodal distributions. Future work may include a comparison of the proposed sampler with MCMC algorithms involving data augmentation.

3.3 Comparing and checking observation models

As discussed in Chapter 2 (Section 2.3), once the inference is performed, one may want to evaluate the quality of the fit and compare it to other fits. Model selection is a quite common approach in astrophysics. However, it only compares models, as the values of the associated criteria are not interpretable. Conversely, model checking assesses a model individually, and yields interpretable diagnoses.

In this section, we first list applications of model comparison with Bayesian evidence and information criteria. Then, applications of model checking methods are reviewed.

3.3.1 Model comparison

Model comparison is an active research topic in astrophysics and cosmology that motivated the development of multiple statistical methods. As discussed in Chapter 2, model comparison can be performed by estimating the expected log-predictive density (elpd) or the marginal likelihood, also called Bayesian evidence.

The **elpd and information criteria** are rarely used in ISM studies, except for some simple methods. For instance, in a study of star formation history in other galaxies, Acquaviva et al. (2011) compares models $\mathcal{M}_i$ using the χ^2 values of their respective best-fit. As it uses an additive uncorrelated Gaussian noise model for a point estimate $\widehat{\boldsymbol{\Theta}}$, this criterion is equivalent to the log-predictive density (lpd), introduced in Chapter 2 (Section 2.3.1). The extragalactic SED analysis from Villa-Vélez et al. (2021) and the study of the thick disc in the Milky Way in Robin et al. (2014) both use Bayesian approaches and approximate the full posterior distribution. Both evaluate their results with the Bayesian information criterion (BIC) (Section 2.3.1) using the maximum a posteriori $\widehat{\boldsymbol{\Theta}}_{\text{MAP}}$. The ABC-based approach from Robin et al. (2014) resorts to a surrogate likelihood function to evaluate the BIC. Finally, Lebouteiller and Ramambason (2022) evaluates the widely applicable information criterion (WAIC) and leave-one-out cross-validation

approximation from Vehtari et al. (2017), in addition to the Bayesian evidence.

The **Bayesian evidence** is a more widespread model comparison criterion in astrophysics. The need for methods that address multimodal distributions in high dimension led to many dedicated statistical developments. For instance, new statistical methods were proposed by astrophysicists to evaluate the marginal likelihood from posterior samples (McEwen et al., 2022) or with nested sampling, with e.g., development of MultiNest (Feroz and Hobson, 2008), UltraNest (Buchner, 2016b; Buchner, 2021) and Dynesty (Speagle, 2020). Evaluating the Bayesian evidence quickly became popular in cosmology to compare models – see Trotta (2008) for a review. It is also widespread in astrophysical fields such as stellar physics, e.g., in Hatta et al. (2022). In ISM studies, the evidence is sometimes evaluated:

- Ramambason et al. (2022) computes the evidence directly with the SMC sampler in a study of extragalactic HII regions.
- The inversion code Beagle (Chevallard and Charlot, 2016), which analyzes the SEDs of galaxies, evaluates the evidence using MultiNest.
- Zucker et al. (2021) evaluates the evidence using Dynesty to study the 3D structure of nearby molecular clouds.
- Kamenetzky et al. (2014), which studies molecular clouds properties in nearby galaxies, resorts to model selection to compare local modes of one posterior distribution, using "local evidences". This evaluation of local evidence of each mode is performed with MultiNest.

3.3.2 Model checking with posterior predictive assessment

In ISM studies, the obtained loss function value in a non-linear least squares problem, often noted χ^2, is used in a three-case interpretation (Ivezić et al., 2020, chapter 4). When $\chi^2 > 1$, the estimated parameters are judged not able to reproduce observations. A $\chi^2 \ll 1$ suggests an overfit or an overestimation of uncertainties on observations. Finally, $\chi^2 \simeq 1$ is the ideal case, indicating physical parameters that reasonably reproduce the observations. This interpretation was applied in ISM studies e.g., in Chevance et al. (2016); Joblin et al. (2018); Villa-Vélez et al. (2021).

This interpretation comes with some limitations and was already criticized in the astrophysics community (Andrae et al., 2010). In particular, the degree of freedom is challenging to estimate for non-linear forward models and when a prior distribution is considered. Besides, this χ^2 rule only applies to Gaussian noise models. Finally, when $\boldsymbol{\Theta}$ is described by a posterior distribution and not by a point estimate $\widehat{\boldsymbol{\Theta}}$, the obtained χ^2 value is approximated with a Monte Carlo (MC) estimator. Replacing the true χ^2 with an Monte Carlo (MC) estimator introduces an error that is seldom accounted for in astrophysics.

In Chapter 2 (Section 2.3), we introduced the so-called Bayesian p-value (Eq. 2.57). This p-value permits to check the ability of the forward model to reproduce observations. The aforementioned χ^2 rule can be transformed into a Bayesian p-value approach for a point estimate $\widehat{\boldsymbol{\Theta}}$ using the L_2 test statistic from Eq. 2.55, reminded below

$$T(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta}) = \sum_{n=1}^{N} \sum_{\ell=1}^{L} \frac{(y_{n\ell} - f_\ell(\boldsymbol{\theta}_n))^2}{\sigma_{n\ell}^2}. \tag{3.11}$$

The inversion code Beagle (Chevallard and Charlot, 2016) aims at modeling and interpreting spectral energy densities (SEDs) of galaxies. In Beagle, y_ℓ is the observed flux in band ℓ, $f_\ell(\boldsymbol{\Theta})$ the flux predicted by the model in band ℓ, and σ_ℓ the standard deviation for band ℓ. This code uses the χ^2 both as a likelihood function and as a test statistic T for the Bayesian

p-value. As the authors consider a posterior distribution and not a point estimate $\widehat{\boldsymbol{\Theta}}$, this p-value is intractable. The authors approximate it using the MC estimator presented in Gelman et al. (1996) and in Eq. 2.59.

Galliano et al. (2021) infers dust properties from the SED of nearby galaxies using the HERBIE inversion code (Galliano, 2018). In this article, y_ℓ is the observed flux in band ℓ, $f_\ell(\boldsymbol{\Theta})$ the flux predicted by the model in band ℓ, and σ_ℓ the standard deviation for band ℓ. Lebouteiller and Ramambason (2022) fits physical parameters to integrated intensity of atomic and molecular emission lines using CLOUDY. In this work, y_ℓ and $f_\ell(\boldsymbol{\Theta})$ are integrated intensities of an atomic or molecular emission line ℓ, and σ_ℓ the corresponding observation standard deviation. In both these articles, the authors use a likelihood more complex than uncorrelated Gaussian noise. For instance, both include censored observations, i.e., upper limits on some low intensity observations. With a censored likelihood, the posterior enforces observation reproductions $\widetilde{y}_\ell$ to be below an upper limit. Both these articles also use the χ^2 as a test statistic T. With this statistic T, reproductions that are far below this upper limit are heavily penalized by the test statistic, while relevant for the likelihood function. A test statistic that favors the same behaviors as the likelihood should be preferred. Finally, both high and low p-values are rejected in Galliano et al. (2021). While rejection for high p-values is indicated in Gelman et al. (2015) for a simple test statistic, p-values close to 1 do not indicate a problematic fit with the χ^2 statistic. On the contrary, with this discrepancy measure T, p-values close to 1 indicate that the model can reproduce observations with smaller errors than the uncertainties in the noise model.

Our proposal – In Chapter 5 (Section 5.3), we resort to hypothesis testing to assess the compatibility of a forward model with the observations. We evaluate the Bayesian p-value from Gelman et al. (1996), with three specificities:

1. We do not set the test statistic to the χ^2. To ensure that the p-value is consistent with the likelihood function in general cases, we set the test statistic to the negative log-likelihood $T(\mathbf{Y}, \boldsymbol{\Theta}) = -\ln \pi(\mathbf{Y}|\boldsymbol{\Theta})$. This choice generalizes the case of BEAGLE to arbitrary observation models.

2. We compute maps of p-values instead of one global p-value. This approach helps to identify regions that are badly modeled by the Meudon PDR code in analyses.

3. We account for uncertainties on the p-value that come from the MC evaluation. This avoids the rejections that are due to chance and caused by insufficient number of samples.

3.4 Our observation model

In this thesis, we consider hyperspectral observations such as the large maps IRAM-30m "ORION-B" Large Program. Such maps are usually transformed into maps of integrated intensities of $L \sim 5 - 30$ ionic, atomic or molecular emission lines. In this section, we describe the reduction process considered in this thesis and the associated uncertainty model.

3.4.1 Noise model on the original hyperspectral cube

We start with an observed hyperspectral cube $\mathbf{O} = (o_{nk}) \in \mathbb{R}^{N \times K}$, where N is the number of pixels (observed spatial positions) and K the number of spectral channels (observed frequencies). For instance, the Orion-B observation described in Pety et al. (2017) contains $N = 10^6$ pixels and $K = 200\,000$ spectral channels. Usually, observation maps contain between $N = \mathcal{O}(1)$ and $\mathcal{O}(10^4)$ pixels.

Figure 3.3 shows examples of spectra of molecular emission lines from the Orion-B observations presented in Pety et al. (2017). Each emission line is limited to an effective small range of spectral channels.

As in Kelly et al. (2012); Galliano (2018); Galliano et al. (2021), we assume a mixture of multiplicative and additive noises on the voxels, i.e., the elements of the cube:

$$\forall n \in [\![1, N]\!], \quad \forall k \in [\![1, K]\!], \quad o_{nk} = \widetilde{\varepsilon}^{(c)}_{nk} I_{nk} + \widetilde{\varepsilon}^{(a)}_{nk}, \tag{3.12}$$

with n a pixel index, k a spectral channel, and I_{nk} the unaltered signal of interest. The term $\widetilde{\boldsymbol{\varepsilon}}^{(c)} = (\widetilde{\varepsilon}^{(c)}_{nk})_{nk}$ is a multiplicative noise due to calibration error, and $\widetilde{\boldsymbol{\varepsilon}}^{(a)} = (\widetilde{\varepsilon}^{(a)}_{nk})_{nk}$ an additive measurement noise.

The measurement noise $\widetilde{\boldsymbol{\varepsilon}}^{(a)}$ combines two sources of uncertainty. A common observation procedure defines o_{nk} as the difference between two measurements, one "off" (observation of a dark portion of the sky), and one "on" (observation of the position of interest n). Each measurement counts photons on a small area during a fixed duration. As photons are considered independent, the number of observed photons is modeled as independent realizations of a Poisson distribution. When many photons are observed, the Poisson distribution can be well approximated by a Gaussian distribution with known variance. Finally, the difference of the two measurements yields a difference of two independent Gaussian distributions, which is itself Gaussian. Besides, thermal noise also affects the observations. This thermal noise can be accurately modeled by an additive Gaussian noise as well.

As a sum of these two Gaussian noises, $\widetilde{\varepsilon}^{(a)}_{nk}$ is modeled by a Gaussian distribution. In this thesis, its covariance matrix $\widetilde{\boldsymbol{\Sigma}}^{(a)}$ is assumed diagonal, i.e., $\widetilde{\boldsymbol{\Sigma}}^{(a)} = \mathrm{diag}(\widetilde{\sigma}^2_{a,n\ell})$, for simplicity. Realizations on different voxels (n, k) are thus assumed independent. This assumption is generally not verified on real observations. For instance, Einig et al. (2023) the structure of the noise covariance in the IRAM-30m Orion-B observations (Pety et al., 2017).

The calibration noise $\widetilde{\boldsymbol{\varepsilon}}^{(c)}$ represents a relative error with respect to the true signal I_{nk} of 10% or lower (Einig et al., 2023). As described in Chapter 3 (Section 3.1.2.3), an additive Gaussian noise cannot capture accurately such relative uncertainty, contrary to a multiplicative noise. As in Kelly et al. (2012), we assume a lognormal model, i.e., a Gaussian distribution on $\ln \widetilde{\boldsymbol{\varepsilon}}^{(c)}$. The spectral correlation is partly taken into account. We assume that the multiplicative noise realizations $\widetilde{\varepsilon}^{(c)}_{nk}$ in all the spectral channels associated with one emission line ℓ are equal. The noise realization value is denoted $\varepsilon^{(c)}_{n\ell}$. Conversely, we neglect the correlation between the spectral channels of distinct lines. As in the additive noise, the noise spatial correlation is neglected. Overall, for a line ℓ and a pixel n, we assume $\varepsilon^{(c)}_{n\ell} \sim \text{Lognormal}(-\sigma_c^2/2, \sigma_c^2)$, where the non-zero mean on the log scale ensures $\mathbb{E}[\varepsilon^{(c)}_{n\ell}] = 1$.

Intensities I_{nk} can cover multiple decades within one hyperspectral cube – see, e.g., Pety et al. (2017). The dominant noise for a voxel o_{nk} thus depends on the intensity I_{nk}. Indeed, if $I_{nk} \gg \widetilde{\sigma}_{a,n\ell}$, the additive noise becomes negligible and the multiplicative dominates. Conversely, if $I_{nk} \ll \widetilde{\sigma}_{a,n\ell}$, the additive noise dominates. Therefore, one cannot neglect one of the two sources of uncertainty.

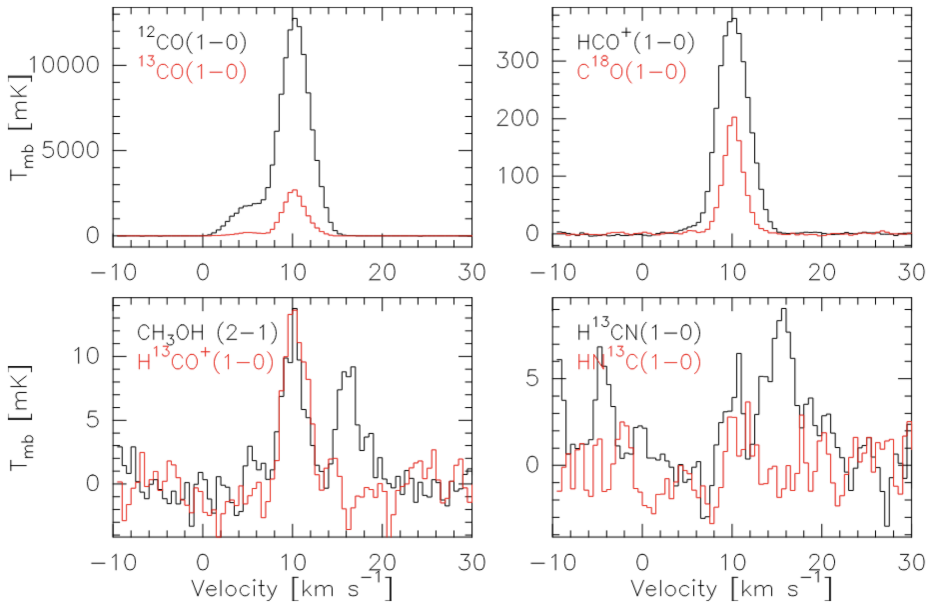


Figure 3.3: Examples of molecular emission line spectra for different molecules in the Orion-B cloud. The line profiles displayed here are averaged over all pixels in the observed map. The velocities in the horizontal axes are translated for each line ℓ such that the 0 corresponds to its characteristic frequency ν_ℓ. Adapted from Pety et al. (2017).

Details for statisticians 3.1: Units for specific and integrated intensities

In Figure 3.3, the spectral channels are called velocity channels and expressed in km s^{-1}, and the intensities are expressed as temperatures. Integrated intensities are thus expressed in K km s^{-1}. These conventions are widespread in radio observations. In physics, the integrated intensities are expressed in erg cm^{-2} s^{-1} sr^{-1}, using the centimetre-gram-second (cgs) unit system. The conversion is detailed here.

The Doppler effect provides a simple bijective relation between frequency ν and velocity v

$$d\nu\ [\text{Hz}] = \frac{\nu}{c}\, dv\ [\text{m s}^{-1}] = \left(10^3\, \frac{\nu}{c}\right)\, dv\ [\text{km s}^{-1}]. \tag{3.13}$$

The measured specific intensity I_ν in a spectral channel of frequency ν is converted to a temperature T using a black body model and the Rayleigh-Jeans (RJ) approximation:

$$I_\nu\ [\text{W m}^{-2}\ \text{Hz}^{-1}\ \text{sr}^{-1}] = \frac{2h\nu^3}{c^2}\frac{1}{\exp\left\{\frac{h\nu}{k_B T\ [\text{K}]}\right\} - 1} \overset{\text{RJ}}{\simeq} \left(\frac{2\nu^2}{c^2} k_B\right) T\ [\text{K}], \tag{3.14}$$

with k_B the Boltzmann constant and h the Planck constant. Combining Eq. 3.13 and Eq. 3.14, the integrated intensity I_ℓ reads

$$\begin{aligned} I_\ell\ [\text{erg cm}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}] &= 10^3\ I_\ell\ [\text{W m}^{-2}\ \text{sr}^{-1}] && (3.15)\\ &= 10^3 \int I_\nu\ [\text{W m}^{-2}\ \text{Hz}^{-1}\ \text{sr}^{-1}] \times d\nu\ [\text{Hz}] && (3.16)\\ &= \left(2 \times 10^6\, \frac{\nu^3}{c^3} k_B\right) \int T\ [\text{K}] \times dv\ [\text{km s}^{-1}]. && (3.17) \end{aligned}$$

3.4.2 From the hyperspectral cube to maps of integrated intensities

There are multiple methods to evaluate the integrated intensity of an emission line ℓ at a given pixel n from the hyperspectral cube. A widespread approach is to assume a line profile shape, usually Voigt or Gaussian, and to fit it to the observations.

Details for statisticians 3.2: Evaluating integrated intensity with a line profile

For a static cloud, an emission line ℓ that corresponds to a quantum de-excitation is characterized by a central frequency ν_ℓ and a line profile ϕ such that $\int \phi(\nu)d\nu = 1$. The integrated intensity I_ℓ associated with an emission line ℓ is therefore obtained with

$$I_\ell = \int I_\nu \phi(\nu - \nu_\ell) d\nu. \tag{3.18}$$

In astrophysics, the line profile ϕ is usually a Voigt profile, i.e., the convolution of a Gaussian profile and of a Lorentzian profile. The Gaussian profile accounts for the speed distribution of the gas particles and for microturbulence. The Lorentzian profile, also called Cauchy profile, models a quantum effect that causes uncertainty on the energy level values. The Voigt profile accounts for the three aforementioned sources of uncertainty. It is commonly used for absorption observations, e.g., for H or the H_2 absorption lines in the UV domain. For emission lines, it is often simplified to a Gaussian profile, as the Lorentzian wings get drowned in thermal noise. See Draine (2011, chapter 6) for more information on standard line profiles.

In real observations, a cloud is generally not static, which causes perturbations in the characteristic frequency ν_ℓ. Codes such as CubeFit (Paumard et al., 2022) perform a Gaussian profile fit to each spectrum to account for this perturbation. Other codes such as Rohsa (Paumard et al., 2022) perform a Gaussian mixture fit to account for the existence of multiple components in the cloud, each with a specific speed.

Figure 3.3 shows spectra of line emissions. In the high SNR regime, fitting a profile can provide an accurate fit and yield an accurate estimate of the integrated intensity $y_{n\ell}$. For instance, fitting a Gaussian profile seems relevant HCO^+ $J = 1 - 0$ and $C^{18}O$ $J = 1 - 0$ emission lines (top right). Similarly, fitting a Gaussian mixture on the ^{12}CO $J = 1 - 0$ profile (top left) seems relevant. However, it is unclear how to define an uncertainty model on the resulting integrated intensities. Besides, in the low SNR regime (bottom left) or in the absence of signal (bottom right), fitting a Gaussian profile may either provide a non-physical result or fail. In the latter case, the pixel is usually considered censored, but it is unclear with which upper bound – see Chapter 3 (Section 3.1.2.4) for an introduction to likelihood model in presence of observation censoring. See e.g., Gaudel et al. (2023) for another application to the Orion-B cloud to study the gas kinematics.

A second approach consists in summing the spectral channels associated with a given line ℓ. Some approaches such as Einig et al. (2023) may denoise the spectrum beforehand. Such denoising relies on assumptions on the spectrum regularity (e.g., continuity or differentiability) and on the noise (e.g., zero mean), but does not assume any shape for the profile. We do not resort to denoising to avoid distorting the signal of interest and to exploit a relevant noise model. For a given emission line ℓ at pixel n, the integrated intensity $y_{n\ell}$ is defined as the sum of a set of spectral channels $[\![k_{\min}^{(\ell)}, k_{\max}^{(\ell)}]\!]$:

$$y_{n\ell} = \sum_{k=k_{\min}^{(\ell)}}^{k_{\max}^{(\ell)}} o_{nk} = \sum_{k=k_{\min}^{(\ell)}}^{k_{\max}^{(\ell)}} \left(\varepsilon_{n\ell}^{(c)} I_{nk} + \widetilde{\varepsilon}_{nk}^{(a)} \right) = \varepsilon_{n\ell}^{(c)} \sum_{k=k_{\min}^{(\ell)}}^{k_{\max}^{(\ell)}} I_{nk} + \sum_{k=k_{\min}^{(\ell)}}^{k_{\max}^{(\ell)}} \widetilde{\varepsilon}_{nk}^{(a)}. \tag{3.19}$$

The first sum defines the integrated intensity I_{nk} before any alteration by noise. The second sum is a sum of independent Gaussian distributions with known standard deviations $\widetilde{\sigma}_{a,nk}$, and is therefore itself a Gaussian distribution $\varepsilon_{n\ell}^{(a)} \sim \mathcal{N}(0, \sigma_{a,n\ell}^2)$, with $\sigma_{a,n\ell}^2 = \sum_{k=k_{\min}^{(\ell)}}^{k_{\max}^{(\ell)}} \widetilde{\sigma}_{a,nk}^2$. We rewrite Eq. 3.19 as

$$y_{n\ell} = \varepsilon_{n\ell}^{(c)} I_{n\ell} + \varepsilon_{n\ell}^{(a)}. \tag{3.20}$$

The set of spectral channels $[\![k_{\min}^{(\ell)}, k_{\max}^{(\ell)}]\!]$ may be fixed for all pixels n or adaptive. An adaptive set may result in reduced variance on additive noise, but include a bias on the integrated intensities (Pety et al., 2017). In particular, an adaptive approach may cut the wings of the line profile and underestimate the integrated intensity $I_{n\ell}$. We use a fixed interval $[\![k_{\min}^{(\ell)}, k_{\max}^{(\ell)}]\!]$ for all pixels n, at the cost of potentially higher noise variance. The intervals were set conservatively, reasonably large to include the whole line profile, even in case of Doppler shifting.

3.4.3 Forward model, observational effects and censorship

The Meudon PDR code as a forward model and model misspecification – A usual hypothesis in inversion procedures is that the forward model $\mathbf{f}$ can predict exactly this integrated intensity, i.e., that there exists a physical parameter $\boldsymbol{\theta}_n$ such that $I_{n\ell} = f_\ell(\boldsymbol{\theta}_n)$. In this thesis, the forward model is set to the Meudon PDR code, introduced in Chapter 1 (Section 1.3). As already mentioned, the Meudon PDR code might not be able to exactly reconstruct the observations. For instance, the reaction constants, collision rates and photo-reaction cross-sections used in the code come from laboratory measurements or from theoretical computations. The uncertainty on these values induces uncertainty on the code predictions. Besides, the encoded physics are based on physical assumptions and simplifications, such as the parallel slab geometry. Finally, in inference procedures the Meudon PDR code $\mathbf{f}$ is replaced by the neural network emulator $\widetilde{\mathbf{f}}$ proposed in Chapter 4, which introduces another approximation error.

We introduce a second multiplicative error source $\varepsilon_{n\ell}^{(v)}$ to encode uncertainty on the model validity such that $I_{n\ell} = \varepsilon_{n\ell}^{(v)} \widetilde{f}_\ell(\boldsymbol{\theta}_n)$. For simplicity, we assume $\varepsilon_{n\ell}^{(v)} \sim \text{Lognormal}(-\sigma_{\text{mod}}^2/2, \sigma_{\text{mod}}^2)$ with independent and identically distributed (i.i.d.) realizations for different pixels n or lines ℓ. In this case, the full multiplicative uncertainty is $\varepsilon_{n\ell}^{(m)} = \varepsilon_{n\ell}^{(v)} \varepsilon_{n\ell}^{(c)}$. As a product of lognormal distributions, this multiplicative uncertainty is itself a lognormal distribution $\varepsilon_{n\ell}^{(m)} \sim \text{Lognormal}(-\sigma_m^2/2, \sigma_m^2)$, with $\sigma_m^2 = \sigma_c^2 + \sigma_{\text{mod}}^2$. The observation model in Eq. 3.20 thus becomes

$$y_{n\ell} = \varepsilon_{n\ell}^{(c)} \left(\varepsilon_{n\ell}^{(v)} \widetilde{f}_\ell(\boldsymbol{\theta}_n) \right) + \varepsilon_{n\ell}^{(a)} = \varepsilon_{n\ell}^{(m)} \widetilde{f}_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)}. \tag{3.21}$$

Observational geometry effects – To account for beam dilution and for signal amplification or attenuation effects due to the angle with the surface of the cloud, we introduce a last multiplicative parameter $\boldsymbol{\kappa} = (\kappa_n)_{n=1}^N$. Unlike the other uncertainty parameters, this parameter is to be inferred. It scales all lines identically so that Eq. 3.21 is rewritten

$$y_{n\ell} = \varepsilon_{n\ell}^{(m)} \kappa_n \widetilde{f}_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)}. \tag{3.22}$$

Such a parameter was also introduced e.g., in Sheffer and Wolfire (2013), where κ_n is defined as a product of 4 terms. Two of these terms are smaller than 1, one of which being the beam dilution factor. The two remaining are greater than 1. Similar multiplicative parameters were also considered in Joblin et al. (2018). We assume that for all n, $\log_{10} \kappa_n \in [-1, 1]$. To simplify notation, this parameter κ_n is added to the parameter vector $\boldsymbol{\theta}_n$ such that $\widetilde{\mathbf{f}}(\boldsymbol{\theta}_n)$. Adding this parameter does not affect the twice differentiable property of the forward model $\widetilde{\mathbf{f}}$, as it is simple to compute $\frac{\partial \widetilde{\mathbf{f}}}{\partial \kappa_n}(\boldsymbol{\theta}_n)$ and higher order derivatives.

Censorship – For real data applications as in Chapter 6, we only have access to reduced multiline integrated intensity maps instead of the raw hyperspectral map. In these cases, a simpler observation model may be imposed. Upper values $\omega_{n\ell} \geq 0$ on observations are sometimes provided when $I_{n\ell} \ll \sigma_{a,n\ell}$. These upper values $\omega_{n\ell}$ on observations correspond to lower sensitivity bounds of the telescope. In these cases, we consider censorship as described in Chapter 3 (Eq. 3.10). Overall, the general observation model considered in this thesis is

$$y_{n\ell} = \max\left\{\omega_{n\ell},\ \varepsilon_{n\ell}^{(m)}\widetilde{f}_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)}\right\} = \begin{cases} \omega_{n\ell} & \text{if } \varepsilon_{n\ell}^{(m)}\widetilde{f}_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)} \leq \omega_{n\ell} \\ \varepsilon_{n\ell}^{(m)}\widetilde{f}_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)} & \text{otherwise} \end{cases}. \tag{3.23}$$

The absence of censorship corresponds to the case where $\omega_{n\ell} = -\infty$.

3.5 Modeling, inference and model assessment choices

This section summarizes the choices we made in this thesis regarding statistical modeling, inference and model assessment. These choices will be detailed in Part II.

Statistical modeling:

1. The forward model is set to the Meudon PDR code, as we assume that it simulates photodissociation regions accurately. Unlike lighter codes, this numerical simulator requires a few hours per evaluation. Evaluating it during the inference would result in an extremely slow inversion procedure. To accelerate inference and be able to compute gradients efficiently, we derive an accurate, fast, and light ANN emulator of the Meudon PDR code. This emulator is trained from a training dataset with a regular grid structure on the parameter space. The derivation of this emulator is the core of Chapter 4.

2. The uncertainty model from Eq. 3.23 combines two sources of noise: one additive and Gaussian, associated with thermal noise, and one multiplicative that includes both calibration and model misspecification errors. The additive Gaussian noise dominates in low SNR observations and becomes negligible in the high SNR regime. Conversely, the multiplicative lognormal noise dominates at high SNR observations and is negligible in the low SNR regime. Therefore, one cannot neglect one of the two noise sources. The resulting likelihood function has an expression challenging to address as is. In Chapter 5 (Section 5.1.1), we propose a closed-form and non-hierarchical approximation of the likelihood function with controlled error.

3. The proposed prior distribution combines a weakly-informative smooth uniform prior on a cube and an informative spatial regularization. The former encodes validity intervals for the components of $\boldsymbol{\Theta}$. The latter exploits the map structure in $\boldsymbol{\Theta}$. It is based on the L_2 norm of the maps Laplacian, as in Marchal et al. (2019). The resulting negative log-prior is also twice differentiable. The proposed prior is fully introduced in Chapter 5 (Section 5.1).

Statistical inference – Chapter 5 (Section 5.2) presents a new MCMC algorithm dedicated to map-structured data and high-dimensional multimodal distributions. The proposed sampler combines two kernels:

1. The first kernel addresses the regularity issue of the posterior. It performs efficient local exploration of the posterior by exploiting its local geometry, encoded by a RMSProp preconditioner.

2. The second kernel addresses the multimodality of the posterior. It permits escapes from the local minima by exploiting the map structure.

Model assessment – We use Bayesian hypothesis testing as Chevallard and Charlot (2016); Galliano et al. (2021); Lebouteiller and Ramambason (2022). In Chapter 5 (Section 5.3), we extend the test with three specificities:

1. The discrepancy measure T is set to the negative log-likelihood for generality. The resulting p-value thus remains relevant for non-Gaussian noise models.

2. We compute one p-value per pixel to help identify potential regions where the Meudon PDR code does not simulate the physics accurately.

3. Uncertainties associated with the Monte Carlo estimators of the p-values are accounted for. The test is therefore made more robust to wrong decisions caused by the error inherent to MC estimation.

3.6 Conclusion

In this chapter, we reviewed statistical modeling and inference approaches adopted in the ISM community. We showed that the choices of model and inference method heavily depend on the community. Cosmology seems to be a pioneer in using state-of-the-art statistical approaches, e.g., with early uses of ANNs to emulate heavy numerical models and of advanced MCMC algorithms to perform inference. Dust studies in the ISM community are also quite advanced, with e.g., the use of sophisticated uncertainty models and the inference of prior hyperparameters with hierarchical models. In particular, works like Galliano (2018) or Galliano et al. (2021) infer physical parameters in very high dimensions – $\mathcal{O}(10^4)$ – and with advanced noise models. To the best of our knowledge, these works are currently the only sampling-based applications with such dimensionality in the ISM community. The only other mentioned works relying on such dimensions are Paumard et al. (2014); Ciurlo et al. (2016); Paumard et al. (2022), which infer maps of parameters by exploiting spatial regularization within an optimization algorithm.

In comparison, inverse problems based on maps of integrated intensities of ionic, atomic and molecular lines have mainly been addressed for maps with a limited number of pixels. For instance, Sheffer et al. (2011); Sheffer and Wolfire (2013); Joblin et al. (2018); Ramambason et al. (2022) work on single-pixel observations. In Wu et al. (2018), the 176-pixel observation is handled with a pixel-by-pixel approach. In the coming years, with the upcoming JWST data and large observation surveys such as the IRAM-30m "ORION-B" Large Program, large observation maps with varying SNR are expected to be more common. The inverse problems we are interested in are therefore expected to increase in dimensionality.

In this thesis, we aim at inferring maps of physical parameter $\boldsymbol{\Theta}$ with up to $\mathcal{O}(10^4)$ pixels. We perform this inference using a state-of-the-art numerical model, the Meudon PDR code, and state-of-the-art statistical modeling and inference methods. The resulting likelihood function is highly non-linear, which causes the posterior distribution to be potentially multimodal. Usual sampling algorithms such as RWMH or the affine-invariant MCMC usually fail to explore such distributions and remain stuck in one local minimum. In addition, the code Meudon PDR is assumed to be twice differentiable, but not gradient Lipschitz continuous. This latter property makes the local exploration of the posterior complicated for classic gradient-based sampling algorithms such as MALA or HMC.

Chapter 4 derives a fast, light and accurate approximation of the Meudon PDR code. Chapter 5 (Section 5.1) details a likelihood function approximation, proposes a spatial regularization prior and the resulting posterior distribution. Chapter 5 (Section 5.2) presents the proposed MCMC algorithm used for inference. Chapter 5 (Section 5.4.2) presents a first very high dimensional synthetic case. Chapter 6 presents multiple applications of the proposed approach to real observations.

Part II

Solving inverse problems on ISM multiline maps

Chapter 4

Fast simulations of photodissociation region models

> " *Learning with neural networks was proposed in the mid-20th century. It yields an effective learning paradigm and has recently been shown to achieve cutting-edge performance on several learning tasks.* "
>
> Shalev-Shwartz and Ben-David (2014, chapter 20)

Contents

In Chapter 3 (Section 3.1.1), we reviewed how numerical forward models are handled in studies of the interstellar medium (ISM). In case of a simple and fast astrophysical model, the numerical simulation is sometimes directly used within the inversion, e.g., in Holdship et al. (2018). The inverse problem studied in this thesis – to be fully introduced in Chapter 5 – involves large observation maps. Each pixel of these maps is modeled using the Meudon PDR code. As the Meudon PDR code requires a few hours for each evaluation, evaluating the associated likelihood during the inversion procedure would be prohibitively slow. We farther described some simulation-based inference approaches, including approximate Bayes computing methods and likelihood emulation.

This chapter is focused on emulating the Meudon PDR code from a set of evaluations instead of the full likelihood function. This approach avoids re-training an emulator whenever the nature or level of the noise changes. The resulting surrogate model will be included in the observation model detailed in Chapter 5. A similar strategy was adopted e.g., in Wu et al. (2018) and Ramambason et al. (2022). Using an emulator instead of the original numerical code induces approximation errors that should be minimized.

In this chapter, we propose a neural network-based emulator, and compare it with standard interpolation methods in terms of speed, memory requirements and accuracy. We eventually obtain a fast, light and accurate artificial neural network (ANN) emulator. This emulator is used as a forward model in Chapter 5.

Section 4.1 provides a description of the considered interpolation and regression methods. Section 4.2, introduces ANNs and how to fit them to a grid of precomputed models. Section 4.3 presents some methods to generate datasets of precomputed models with a good coverage of the parameter space. The illustrative Section 4.4 compares all considered methods and dataset structures on a simple case, the log Rosenbrock function. Section 4.5 describes the dataset of precomputed models and introduces the metrics used to compare surrogate models. In Section 4.6, we design ANNs that address the specificities of ISM numerical codes. Section 4.7 compares these ANNs with classic interpolation methods in terms of speed, memory requirements and accuracy.

This chapter is based on the journal article Palud et al. (2023c) and the associated Gretsi conference article Einig et al. (n.d.). This article is the product of an equal collaboration with Lucas Einig, another PhD candidate of the ORION-B consortium.

4.1 Deriving emulators with interpolation or regression

Deriving an emulator consists in estimating the function $\widetilde{\mathbf{f}} : \mathbb{R}^D \to \mathbb{R}^L$ that maps input vectors $\boldsymbol{\theta}$ to output vectors $\mathbf{y} = \mathbf{f}(\boldsymbol{\theta})$ as closely as possible. This function $\widetilde{\mathbf{f}}$ is *learned*, or equivalently, *fitted*, from a dataset of precomputed models $\mathcal{D} = \{(\boldsymbol{\theta}_n, \mathbf{y}_n) \in \mathbb{R}^D \times \mathbb{R}^L, n = 1, \dots, N\}$. We remind that in this thesis, the input vector $\boldsymbol{\theta}$ corresponds to a vector of physical parameters, e.g., temperature, thermal pressure, volume density, and the output vector $\mathbf{y}$ to observables computed by a numerical code, e.g., integrated intensities of specific emission lines.

In this Section, we first summarize four interpolation methods we will use in our comparison. Then, we present ANNs and the associated regression methods. For a more detailed introduction to ANNs, see Shalev-Shwartz and Ben-David (2014, chapter 20).

4.1.1 Interpolation methods

Interpolation methods have become a common approach to build surrogates of comprehensive ISM models over the last years thanks to their conceptual and implementation simplicity. Four families of interpolation methods are usually considered:

- Nearest-neighbor interpolation assigns to a new point the value of the closest point in the dataset. It is fast but generally performs poorly in terms of accuracy. It was used e.g., in Ramambason et al. (2022).

- Piecewise linear interpolation generally performs better, while remaining quite fast. It first triangulates the dataset, so that a new point is associated with a cell of the triangulation. Then, it returns a weighted average of the cell points values. It was used e.g., in Thomas et al. (2018).
- Spline interpolation (Bojanov et al., 1993) methods are based on piecewise polynomials, yielding an even more accurate and still fast surrogate model. It was used e.g., in Blanc et al. (2015).
- Radial basis function (RBF) interpolation (Fasshauer, 2007, chapter 6) uses the full dataset for each evaluation. For a new point, it returns a weighted sum of the values of all the dataset points, where the weights depend on the distance to this new point. Surrogate models defined with RBF interpolation are generally very accurate but slow. It was used e.g., in Wu et al. (2018).

Interpolation methods suffer from some drawbacks. By definition, a surrogate model defined with an interpolation method passes exactly through the points of the dataset. This constraint does not guarantee good accuracy on points not used during the fit. Besides, evaluating a surrogate model defined with an interpolation method requires loading the whole dataset, which can induce a very heavy memory cost when it contains many precomputed models or many quantities associated with each model. Finally, although they are generally faster than the original numerical codes, interpolation methods handle outputs (i.e., observables) independently. They are thus quite slow when the number of outputs is large.

Accuracy can be improved by relaxing the constraint of passing through the points of the dataset. The memory and speed drawbacks can be addressed using an approach that allows predicting all outputs at once.

4.1.2 Regression methods

Relaxing the constraint of passing through the points of the dataset turns the interpolation problem into a regression problem. The first step towards emulating a numerical model $\mathbf{f}$ is to restrict the search of $\widetilde{\mathbf{f}}$ to a class of functions. Most classes of functions are parametrized with vectors $\boldsymbol{\psi}$. In the following, potential surrogate functions are sometimes denoted $\widetilde{\mathbf{f}}_{\boldsymbol{\psi}}$ to emphasize this point. For instance, in linear regression, an affine function $\widetilde{\mathbf{f}}_{\boldsymbol{\psi}} : \boldsymbol{\theta} \mapsto \mathbf{W}\boldsymbol{\theta} + \mathbf{b}$ is uniquely described by $\boldsymbol{\psi} = (\mathbf{W}, \mathbf{b})$. Given the complexity of ISM numerical models, the class of affine functions is too restrictive to produce accurate surrogate models, and richer classes are required.

Multiple classes of functions and the associated regression algorithms enable the emulation of complex non-linear functions from data of precomputed models, such as polynomial functions, k-nearest-neighbor regression (Shalev-Shwartz and Ben-David, 2014, chapter 19) – used e.g., in Smirnov-Pinchukov et al. (2022) –, Gaussian process regression (Rasmussen and Williams, 2006), decision trees and the associated ensemble methods such as random forest (RF) (Shalev-Shwartz and Ben-David, 2014, chapter 18) – used e.g., in Bron et al. (2021) – or XGBoost (Chen and Guestrin, 2016), ANNs (Shalev-Shwartz and Ben-David, 2014, chapter 20). All methods based on decision trees or nearest-neighbors yield piecewise functions, which prevents from enforcing desirable regularity property in the surrogate model such as continuity or differentiability. Besides, all the listed algorithms, except ANNs and nearest-neighbor interpolation, handle multiple outputs independently, which slows prediction when the number of outputs is high. Conversely, ANNs predict all outputs at once from a common sequence of intermediate computations, which is considerably faster. In addition, ANNs are known to yield very accurate surrogate models both in theory and in practice (Shalev-Shwartz and Ben-David, 2014, chapter 20). Finally, an ANN comes with the ability to automatically and efficiently compute the first or second order derivative of its outputs with respect to its inputs $\boldsymbol{\theta}$, using automatic differentiation (Paszke et al., 2017). In other words, for any output ℓ of an ANN, one can efficiently access the gradient $\nabla \widetilde{f}_\ell$ or the Hessian matrix $\nabla^2 \widetilde{f}_\ell$. This property is desirable to efficiently explore high dimensional parameter

space in inverse problems, as detailed in Chapter 3 (Section 3.1.1). We consequently adapt the rich and versatile class of ANNs to address the complexity of ISM numerical models, to exploit prior knowledge on the regularity of the function to approximate, and to efficiently predict all outputs at once.

4.2 Neural networks for regression

The class of ANNs is first introduced. We then explain how to fit ANNs to a dataset.

4.2.1 Generalities on neural networks

Artificial neural networks (ANNs) form a class of mathematical models inspired from biological neural systems. The first ANN was proposed in McCulloch and Pitts (1943) to perform logical operations. Since then, multiple hardware and algorithmic developments such as GPU computing and back-propagation (Rumelhart et al., 1986) made them capable of learning more complex patterns and relationships in data. They enjoy strong theoretical results. For different sets of assumptions on the architecture, universal approximation theorems establish that ANNs can approximate almost any continuous function (Hornik et al., 1989; Leshno et al., 1993). They gained a widespread popularity after the 2012 ImageNet Challenge, an image classification competition in which an ANN significantly outperformed competing methods. They are nowadays state-of-the-art for a variety of tasks in vector, image, sound or text processing in multiple scientific and industrial fields. As we showed in Chapter 3 (Section 3.1.1), ANNs are already used in astrophysics, including ISM studies, to emulate numerical models (Grassi et al., 2011; de Mijolla et al., 2019; Holdship et al., 2021; Grassi et al., 2022).

Formal description of an ANN – Throughout this thesis, an ANN is considered as a function $\widetilde{\mathbf{f}} : \boldsymbol{\theta} \in \mathbb{R}^D \mapsto \widetilde{\mathbf{f}}(\boldsymbol{\theta}) \in \mathbb{R}^L$, where D and L are input and output dimensions, respectively. For a numerical model, D is the number of considered physical parameters, e.g., thermal pressure or visual extinction, and L is the number of predicted observables, e.g., line intensities. An ANN is made of $h+1$ intermediate functions $\widetilde{\mathbf{f}}^{(j)}$, called *layers*. Intermediate layers $1 \leq j \leq h$ are called the *hidden layers* and the final layer is the *output layer*. The j^{th} layer takes an intermediate vector $\boldsymbol{\theta}^{(j)} \in \mathbb{R}^{i_j}$ as input and computes an intermediate output $\mathbf{y}^{(j)} \in \mathbb{R}^{o_j}$. Intermediate dimensions i_j and o_j can be chosen arbitrarily, except for $i_1 = D$ and $o_{h+1} = L$. In a feedforward neural network, layers interact following an acyclic graph. The output of a layer j feeds one or more of the next layers $j' > j$, hence the notion of direction in a feedforward neural network.

Figure 4.1 shows the structure of a simple ANN that contains $h = 2$ hidden layers and one output layer. This ANN takes in input $D = 2$ physical parameters and predicts $L = 10$ observables. It is an instance of a feedforward neural network as each layer only feeds layers closer to the output layer, as shown on the left-hand side of the figure. More precisely, its layer graph is linear: the output of one of its layers j is the input of the next layer $j+1$, as the $\boldsymbol{\theta}^{(j+1)} = \mathbf{y}^{(j)}$ and $i_{j+1} = o_j$ equalities show. Alternative feedforward architectures with non-linear layer graph exist, such as residual networks (He et al., 2016) and dense networks (Huang et al., 2017). These architectures include skip connections between layers that bypass the activation function to preserve original input information and intermediate computations. However, linear layer graphs remain the simplest and most widespread multi-layer architectures for vector classification or regression tasks. In the remainder of this chapter, ANNs are considered with such an architecture unless mentioned otherwise.

A **hidden layer** combines an affine transformation and a non-linear scalar function $g^{(j)}$ applied element-wise:

$$\widetilde{\mathbf{f}}^{(j)} : \boldsymbol{\theta}^{(j)} \mapsto \mathbf{y}^{(j)} = g^{(j)}(\mathbf{W}^{(j)}\boldsymbol{\theta}^{(j)} + \mathbf{b}^{(j)}), \tag{4.1}$$

with $\mathbf{W}^{(j)} \in \mathbb{R}^{o_j \times i_j}$ and $\mathbf{b}^{(j)} \in \mathbb{R}^{o_j}$ the weight matrix and bias vector of the affine transformation, respectively. The non-linear scalar function $g^{(j)}$ is called an *activation function*. Common

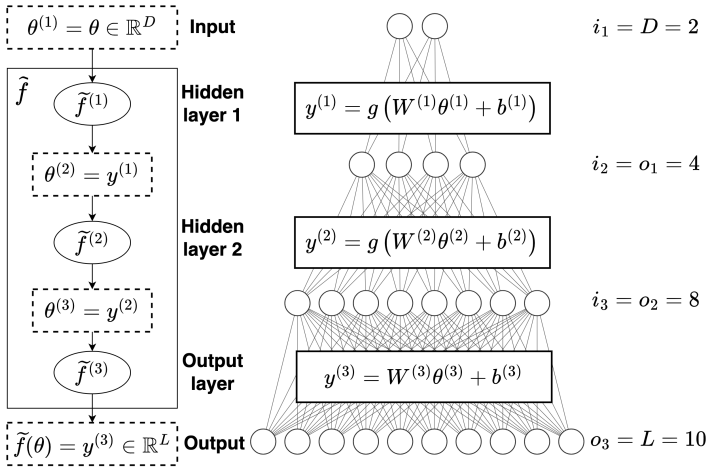


Figure 4.1: Structure of a simple feedforward neural network with $h = 2$ hidden layers and a linear layer graph, shown on the left.

activation functions include the sigmoid, hyperbolic tangent, rectified error linear unit (ReLU), and multiple variants – see Nwankpa et al. (2021) for a review. Choosing a specific activation function $g^{(j)}$ for each of the h hidden layers might lead to better performance but would require training many ANNs. A unique g is therefore generally set for all hidden layers. As an ANN is a composition of affine transformations and activation functions, the regularity properties of the activation function g apply to the full ANN. For instance, if g is differentiable, so is the full ANN $\widetilde{\mathbf{f}}$. The same implication holds for infinitely continuously differentiable functions.

The **output layer** transforms the outputs of one or more hidden layers into the desired prediction with an affine transformation and a second activation function. This second activation function depends on the considered problem. The sigmoid and softmax functions are usually employed to return probabilities in binary and multi-class classification, respectively. In regression tasks, the identity function is generally used.

Architecture and learnable parameters – Overall, in a regression context, an ANN is uniquely defined by its layer graph, an activation function g, a number of hidden layers $h \geq 0$, a sequence of sizes of its layers $(i_j, o_j)_{j=1}^{h+1}$, and the sequence of weight matrices and bias vectors $\boldsymbol{\psi} = (\mathbf{W}^{(j)}, \mathbf{b}^{(j)})_{j=1}^{h+1}$. All but $\boldsymbol{\psi}$ are manually set by the user prior to the fitting procedure. They define the so-called *architecture* of the ANN. The only learnable parameters are the weight matrices and bias vectors $\boldsymbol{\psi} = (\mathbf{W}^{(j)}, \mathbf{b}^{(j)})_{j=1}^{h+1}$. The fitting procedure adjusts them so that $\widetilde{\mathbf{f}}_{\boldsymbol{\psi}} \simeq \mathbf{f}$.

4.2.2 Fitting a neural network to a dataset

In regression, once the class of function is set, the associated parameter $\boldsymbol{\psi}$ is adjusted so that $\widetilde{\mathbf{f}}_{\boldsymbol{\psi}}$ fits the dataset $\mathcal{D}$ of precomputed models. In our case, the class of functions is set by selecting an ANN architecture, and the associated parameter $\boldsymbol{\psi}$ contains the weights and biases of the network's layers. A loss function $\mathcal{L}(\widetilde{\mathbf{f}}\,; \mathcal{D})$ quantifies the distance between predictions $\widetilde{f}_\ell(\boldsymbol{\theta})$ and the corresponding true values y_ℓ. It is based on an error function, e.g., the absolute error (AE)

$$\mathrm{AE}\left(\widetilde{\mathbf{f}}\,; (\boldsymbol{\theta}, y_\ell)\right) = \left|\widetilde{f}_\ell(\boldsymbol{\theta}) - y_\ell\right|, \tag{4.2}$$

or the squared error (SE)

$$\mathrm{SE}\left(\widetilde{\mathbf{f}}\,; (\boldsymbol{\theta}, y_\ell)\right) = \left(\widetilde{f}_\ell(\boldsymbol{\theta}) - y_\ell\right)^2. \tag{4.3}$$

The loss function summarizes the set of $N \times L$ errors obtained on the dataset $\mathcal{D}$. The mean is often used for computational efficiency of evaluation and differentiation, yielding e.g., the mean

squared error (MSE) or the mean absolute error (MAE). Obtaining the best function $\widetilde{\mathbf{f}}$ boils down to minimizing the loss function with respect to the parameter $\boldsymbol{\psi}$

$$\widetilde{\mathbf{f}} \in \underset{\boldsymbol{\psi}}{\arg\min}\, \mathcal{L}\left(\widetilde{\mathbf{f}}_{\boldsymbol{\psi}}\,;\mathcal{D}\right). \tag{4.4}$$

In general, problems of the form of Eq. 4.4 do not admit a closed-form solution. Furthermore, with ANNs, the loss function $\mathcal{L}\left(\widetilde{\mathbf{f}}_{\boldsymbol{\psi}}\,;\mathcal{D}\right)$ is generally not convex, with multiple saddle points and local minima (Shalev-Shwartz and Ben-David, 2014, chapter 20). When $\boldsymbol{\psi}$ is low dimensional, such problems can be solved approximately using a meta-heuristic such as a genetic algorithm or simulated annealing – see Section 2.A.1. As ANNs typically contain at least hundreds of parameters to tune, meta-heuristic methods are prohibitively slow.

In contrast, gradient descent (GD) methods are computationally very efficient. They rely on auto-differentiation to efficiently evaluate the gradient of the loss function $\nabla_{\boldsymbol{\psi}}\mathcal{L}$ and on back-propagation (Rumelhart et al., 1986) to efficiently update $\boldsymbol{\psi}$. Stochastic gradient descent (SGD) Shalev-Shwartz and Ben-David, 2014, chapter 14 accelerates the search by using *batches* instead of the full dataset in gradient evaluations. Preconditioned variants such as RMSProp (Tieleman and Hinton, 2012) or Adam (Kingma and Ba, 2017) exploit the local geometry of the loss function to escape from saddle points and farther accelerate convergence to a good local minimum. This optimization procedure is often called *training* or *learning phase* with ANNs, because the network progressively learns from the data as the loss function decreases.

4.3 Dataset structure

Like the choice of the class of function, the structure of the training dataset $\mathcal{D}$ in the $\boldsymbol{\theta}$-space can have a dramatic impact on the accuracy of the obtained emulator. To approximate the Meudon PDR code, we propose to resort to a lattice structure on a cube of dimension D generated prior to any fit. This approach is very common in ISM studies (Joblin et al., 2018; Wu et al., 2018; Ramambason et al., 2022) – see Chapter 3.

Lattice structured datasets – A regular grid structure has many advantages. First, if the parameter space is a D-dimensional cube and the number of points N is such that $N = K^D$ for some $K \geq 1$, the lattice structure maximizes the minimum distance between two points of the dataset. Therefore, this structure enforces good coverage of the parameter space. Second, it is often more convenient to manually inspect a dataset with such a structure, especially for $D \geq 3$, when the full dataset cannot be visualized at once. Besides, it allows the use of efficient interpolation methods such as splines, for which regular grid structures are mandatory. Also, the regularity of the grid can be exploited to accelerate nearest-neighbor and linear interpolations. Indeed, it facilitates the localization of the new point within the dataset and thus the identification of the points to use for prediction. However, this structure is not necessary for RBF interpolation methods or in regression approaches.

Despite these advantages, the lattice structure may not yield optimal emulators, and other structures might yield better accuracy. For instance, for coarse grids, a lattice structure creates wide regions in the parameter space without any point in the dataset $\mathcal{D}$. An emulator fitted in this context could struggle to reproduce accurately brutal changes occurring in these regions. Besides, it creates a strong anisotropy in the cube, which can reduce the accuracy of an emulator. For instance, large gradient variations occurring between two slices of the lattice would be difficult to emulate accurately.

Alternative methods to generate datasets with good coverage – Many alternative methods enable sampling a D-dimensional cube. The simplest would be to sample independent and identically distributed (i.i.d.) points from a uniform distribution on the cube. However, this method would leave some regions uncovered, and some pairs of points could be arbitrarily close,

which limits the information brought by each. Better methods enforce repulsion between points to ensure a good coverage of the cube. Some of these methods are deterministic, with low discrepancy sequences such as the Halton sequences used in quasi-Monte Carlo (QMC) methods (Asmussen and Glynn, 2007, chapter 9). Some others are partially random, such as latin hypercube sampling (LHS) (McKay et al., 1979), its variants orthogonal LHS (OLHS) (Tang, 1993) and symmetric LHS (SLHS) (Ye et al., 2000), or stratified Monte Carlo (Haber, 1966). These methods are simple to implement. To generate N points in the unit cube $[O, 1]^D$, LHS handles each dimension independently. For each dimension, it divides the $[0, 1]$ segment in N smaller segments of equal size, draws one sample per segment with a uniform distribution, and randomly permutes the draws. Stratified Monte Carlo divides the cube in N smaller regions of equal volume and draws one sample from a uniform distribution on each smaller region. Orthogonal LHS combines both LHS and stratified MC to ensure better coverage. Symmetric LHS is a generalization of OLHS that enforces symmetry properties to farther improve the coverage of the cube. Other LHS algorithms that optimize criteria such as the entropy or the minimum distance between two points also exist - see Joseph (2016) for a review.

Figure 4.2 illustrates how each method would cover the unit square, i.e., for $D = 2$. The i.i.d. draws with uniform distribution on square yield the visually worst datasets. The lattice structure best covers corners and edges, but leaves large gaps between points. LHS and stratified MC lead to datasets with better coverage of the square. By combining them, orthogonal LHS yields even better datasets.

As illustrated in the next section, these methods may yield better accuracy than the lattice structure. However, the associated accuracy has a high variance, as it depends on the presence or absence of points in regions with high gradients. Using the lattice structure is therefore a good heuristic.

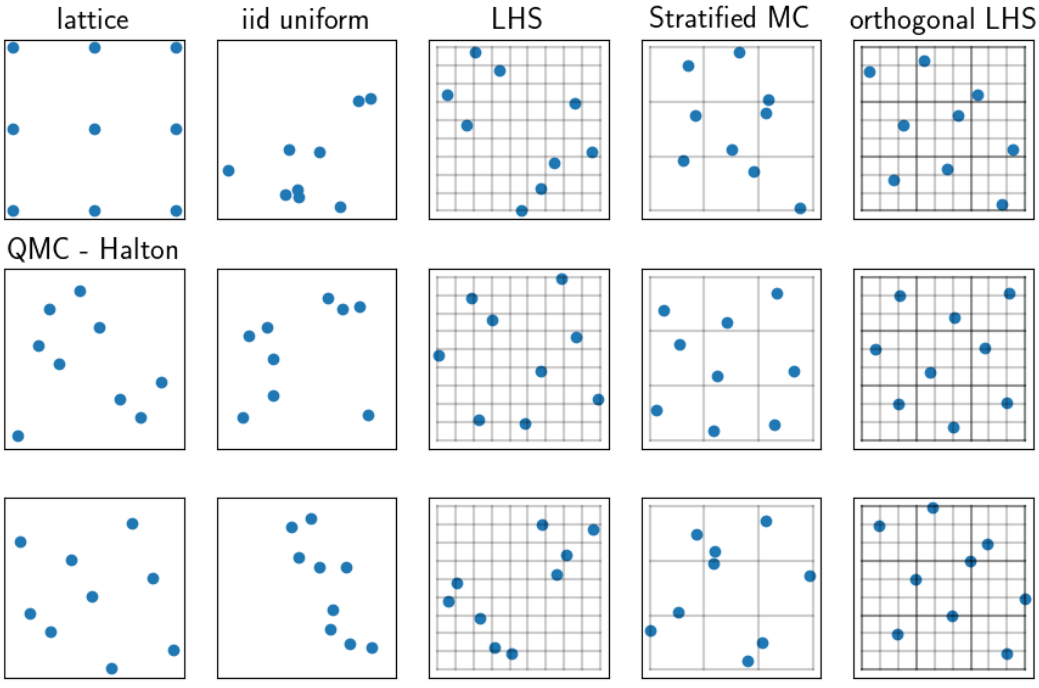


Figure 4.2: Illustration of methods to generate non-lattice datasets of $N = 9$ points. The sampling principle of LHS, stratified MC and orthogonal LHS are indicated with the grids.

4.4 Illustration: the log Rosenbrock function

In this Section, we illustrate the four interpolation methods mentioned in Section 4.1.1 and two regression methods, namely random forest (RF) and ANNs, on the log Rosenbrock func-

tion (Rosenbrock, 1960):

$$\ln R : \boldsymbol{\theta} \in \mathbb{R}^2 \mapsto \ln \left[1 + (1 - \theta_1)^2 + 100 \left(\theta_2 - \theta_1^2 \right)^2 \right], \tag{4.5}$$

which is positive and admits a minimum at $(1, 1)$ such that $\ln R(1, 1) = 0$. We restrict ourselves to the $[-1, 1] \times [-0.5, 1.5]$ square.

The SciPy[1] (Virtanen et al., 2020) implementation is used for the interpolation methods. For RBF interpolation, the cubic kernel is used as it yields the best results. For random forest and ANNs, the scikit-learn[2] (Pedregosa et al., 2011) implementations are used. Note that the PyTorch implementation is used in the remainder of this thesis. The scikit-learn implementation is used in this example for simplicity. The ANN architecture is set to a perceptron with $H = 2$ hidden layers of 32 neurons each. The Rosenbrock function is smooth, i.e., infinitely continuously differentiable. The tanh activation function is used to enforce this smoothness property in the emulator.

Figure 4.3 shows the values taken by this function on this subspace. Its banana shape being relatively thin, it is challenging to reproduce accurately with a limited amount of points. The results are compared over six dataset structures, namely lattice, Halton sequence, i.i.d. uniform, LHS, stratified MC and orthogonal LHS. Each training dataset contains $N = 49$ points. The accuracies of emulators are evaluated on a 151×151 grid on the $[-1, 1] \times [-0.5, 1.5]$ square, using the maximum absolute error, the MAE and the MSE.

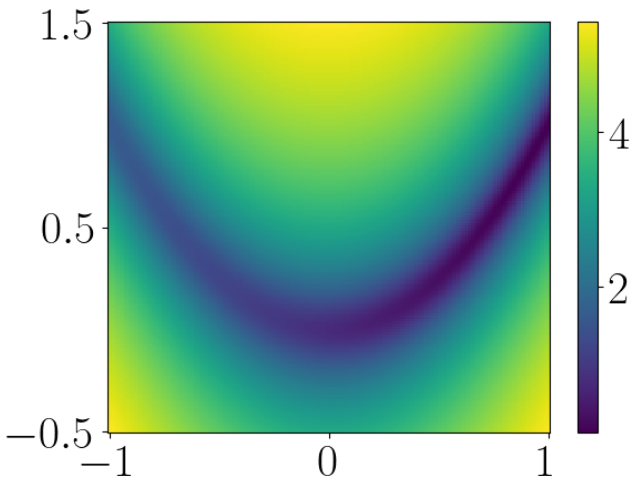


Figure 4.3: log Rosenbrock function $\ln R$ (Eq. 4.5).

Figure 4.4 shows the emulator obtained for each method and the associated errors. All methods struggle to reproduce accurately the banana shape of the Rosenbrock function, as there are few points in the training set. In ISM numerical models, such strong and fast variations can correspond to a change of regime and are thus of crucial importance. The overall performance of an emulator highly relies on the presence of a few points in such regions. The presence or absence of these key points in datasets generated by random methods induces a high variance in the accuracy of emulators built from these datasets.

Table 4.1 quantifies the results of this comparison with three metrics: the maximum absolute error, the MAE and the MSE. In all scenarios, ANNs and RBF interpolation yield the most accurate emulators. Besides, ANNs provide the best emulator with respect to each of the three metrics. However, the best results are not always obtained with a lattice structure.

Overall, this illustrative example shows three important aspects of the emulator derivation of the Meudon PDR code:

[1] https://scipy.org/
[2] https://scikit-learn.org/stable/index.html

1. ANNs often yield much more accurate emulators than interpolation methods.

2. The lattice structure might not always lead to optimal emulators. However, as it covers the square very well, it still leads to good emulators. Besides, as the other considered methods are random by nature and as only a few training points have a great impact on the emulator accuracy, two datasets generated with the same method can lead to emulators with very different accuracies. To limit uncertainties on causes of the quality of the results, we prefer to exploit a lattice dataset in our emulation of the Meudon PDR code.

3. The result of this comparison thus depends on the error function. Choosing a relevant accuracy metric is therefore crucial to derive a good emulator. In Section 4.5.2, we introduce a new accuracy metric, the Error Factor, tailored to the Meudon PDR code.

Table 4.1: Comparison of popular interpolation and regression methods on the log Rosenbrock function with datasets of $N = 49$ points. The values correspond to the datasets and emulators displayed in Figure 4.4. Bold values indicate the best emulator class for each training dataset structure, and blue values indicate the best couple (structure, emulator class) for each accuracy metric. As cubic splines can only be defined on a lattice, there is no value for the other dataset structures. As linear interpolation is only defined inside the convex envelope of the training dataset, its error can be computed on a fraction of the test set for all structures except the lattice. For this reason, its values are indicated in parentheses.

Structure	Interpolation				Regression	
	nearest neighbor	linear	cubic spline	RBF	random forest	neural network
Max absolute error						
Lattice	2.974	1.880	1.611	1.568	2.754	**0.787**
i.i.d. uniform	4.404	(3.120)	-	2.742	4.001	**2.104**
QMC	2.585	(1.909)	-	**1.766**	2.850	2.742
LHS	3.740	(3.259)	-	1.860	3.371	**1.098**
Stratified MC	3.148	(3.555)	-	**1.853**	3.613	3.106
Orthogonal LHS	3.173	(3.378)	-	2.127	3.225	**1.736**
mean absolute error (MAE)						
Lattice	0.419	0.185	0.151	0.131	0.399	**0.082**
i.i.d. uniform	0.516	(0.332)	-	0.225	0.653	**0.163**
QMC	0.391	(0.167)	-	0.138	0.528	**0.133**
LHS	0.431	(0.302)	-	0.139	0.544	**0.072**
Stratified MC	0.406	(0.162)	-	**0.138**	0.566	0.177
Orthogonal LHS	0.407	(0.190)	-	0.136	0.536	**0.083**
mean squared error (MSE)						
Lattice	0.324	0.094	0.060	0.052	0.280	**0.022**
i.i.d. uniform	0.547	(0.343)	-	0.177	0.788	**0.092**
QMC	0.276	(0.079)	-	**0.062**	0.480	0.098
LHS	0.351	(0.276)	-	0.080	0.549	**0.025**
Stratified MC	0.313	(0.074)	-	**0.065**	0.529	0.152
Orthogonal LHS	0.307	(0.122)	-	0.069	0.526	**0.034**

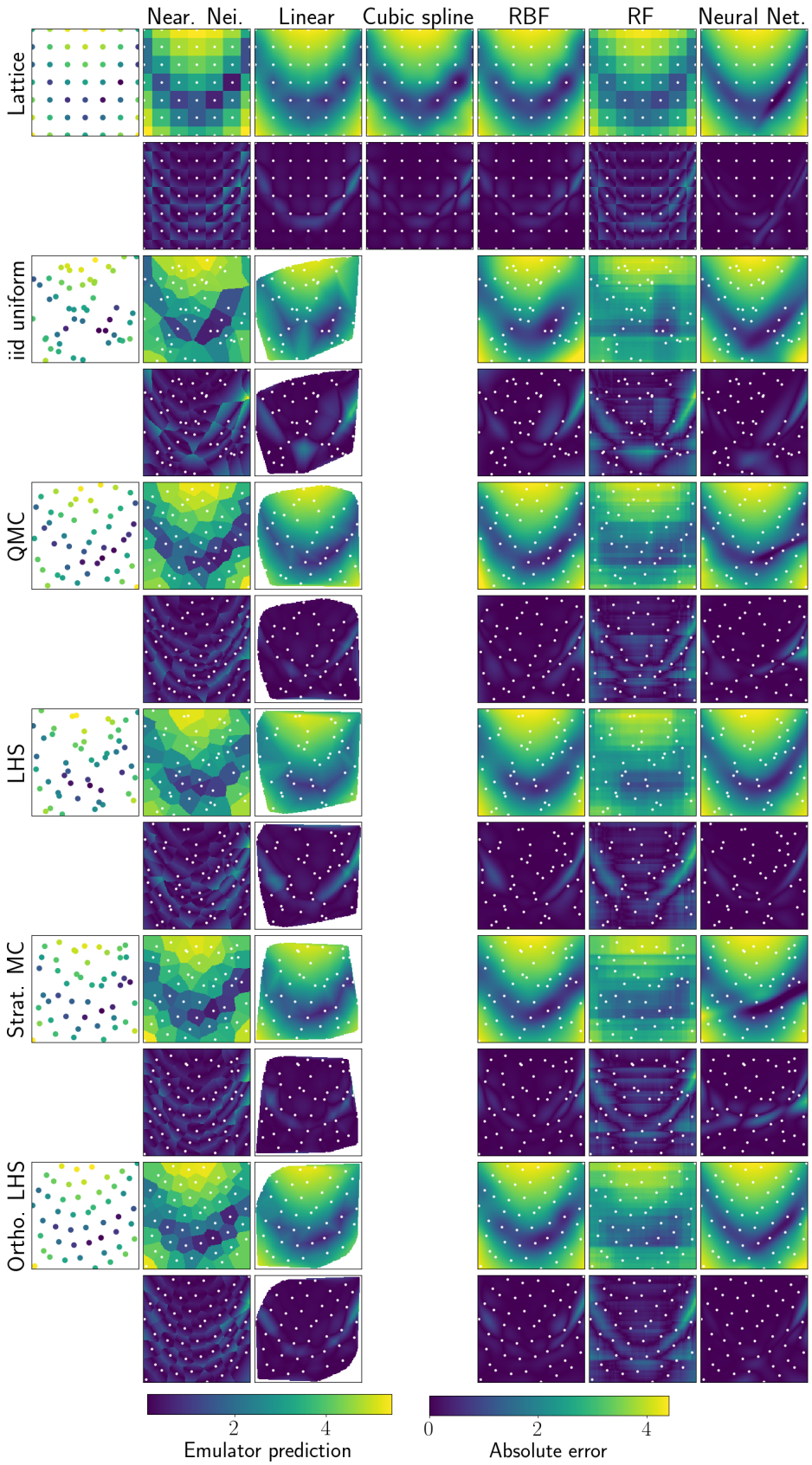


Figure 4.4: Comparison of four popular interpolation methods and two regression methods, namely random forest (RF) and ANN, on the log Rosenbrock function $\ln R$ (Eq. 4.5). For each dataset structure, the first row shows the emulator obtained with the considered algorithm, and the second row shows the absolute error. Cubic splines are only defined on a lattice. For linear interpolation, white values indicate that the emulator is not defined for the corresponding value of $\boldsymbol{\theta}$.

4.5 Towards an emulator of the Meudon PDR code

In this Section, we describe the setup used to define and fit emulators of the Meudon PDR code (Le Petit et al., 2006), introduced in Chapter 1. We first present the train and test sets, and then introduce the metrics used to compare emulators.

4.5.1 Datasets

Physical parameters – We approximate the Meudon PDR code with respect to the $D = 4$ input parameters that are most relevant for inference (Wu et al., 2018). The three main ones are the thermal pressure P_{th}, the scaling factor G_0 of the interstellar standard radiation field and the size of the slab of gas measured in total visual extinction A_V^{tot}. As in Wu et al. (2018), we consider a wide variety of environments with $P_{\text{th}} \in [10^5, 10^9]$ K cm^{-3}, $G_0 \in [1, 10^5]$ and $A_V^{\text{tot}} \in [1, 40]$ mag. The Meudon PDR code computes line intensities for multiple angles φ between the cloud surface and the line of sight. In the Meudon PDR code, this angle φ can cover a $[0, 60]$ deg interval. A face-on geometry corresponds to $\varphi = 0$ deg, and $\varphi = 60$ deg is the closest to an edge-on geometry. To enable analyses of PDRs with known edge-on geometry such as the Orion Bar (Joblin et al., 2018), this angle is added to the considered physical parameters. Table 4.2 details the considered ranges of the main input parameters. The secondary parameters of the code, listed in Section 1.3, are set to their default values – see Table 1.3.

Table 4.2: Input parameters in the Meudon PDR code and structure of training dataset.

Parameter	Range of values	Unit	Training set lattice structure	
			Number of points	Grid
Thermal pressure P_{th}	$[10^5, 10^9]$	K cm^{-3}	14	log spacing
Radiative intensity G_0	$[1, 10^5]$	Mathis	14	log spacing
Total visual extinction A_V^{tot}	$[1, 40]$	mag	14	log spacing
Line-of-sight angle φ	$[0, 60]$	deg	7	linear spacing

Datasets and first preprocessing – Two datasets of Meudon PDR code evaluations are defined: a training set and a test set[3].

The **training set** is used to fit all surrogate models. It contains $N_{\text{train}} = 19\,208$ points, structured as a $14 \times 14 \times 14 \times 7$ uniform regular grid on $(\log_{10} P_{\text{th}}, \log_{10} G_0, \log_{10} A_V^{\text{tot}}, \varphi)$. Table 4.2 shows its lattice structure. This uniform grid structure is chosen to include spline interpolation in the comparison and to simplify the outlier identification procedure presented in Section 4.6.1. The Meudon PDR code predicts line intensities that are strictly positive and span multiple decades, from 10^{-50} to 10^{-2} erg cm^{-2} sr^{-1} s^{-1}. To avoid giving more weight to high line intensities in the regression and disregarding the lowest ones, we consider the intensities in log scale. In other words, we derive emulators to reproduce the log-intensities $\ln \mathbf{y} \in \mathbb{R}^L$, i.e., we derive emulators of $\ln \mathbf{f}$.

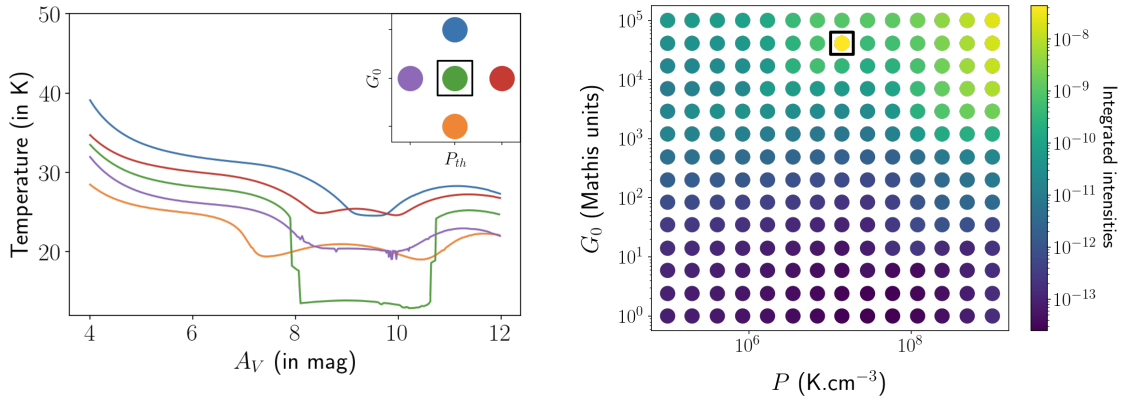
Similarly, P_{th}, G_0, and A_V^{tot} are considered in log scale. Even in log scale, the parameters of interest cover intervals with quite different sizes. For instance, $\log_{10} G_0 \in [0, 5]$ while $\log_{10} A_V^{\text{tot}} \in [0, 1.602]$, i.e., A_V^{tot} covers an interval more than 3 times smaller than G_0. Both interpolation methods and neural network-based regression typically suffer from this difference. All D parameters are thus standardized to have a zero mean and a unit standard deviation. This simple transformation generally improves accuracy for both interpolation methods and ANNs (Shalev-Shwartz and Ben-David, 2014, chapter 25).

The **test set** is used to assess the accuracy of surrogate models on data not used in the training step. It contains $N_{\text{test}} = 3\,192$ points. These points are generated with 456 independent

[3]Both datasets can be found in https://ism.obspm.fr/files/ArticleData/2023_Palud_Einig/2023_Palud_Einig_data.zip.

random draws from a uniform distribution on the $(\log_{10} P_{\text{th}}, \log_{10} G_0, \log_{10} A_V^{\text{tot}})$ cube and with a regular grid of 7 values on φ. To ensure consistent preprocessing between the two sets, both the input values $\boldsymbol{\theta}$ and output values $\mathbf{y}$ of the test set undergo the same transformations as the training set. In particular, the standardization applied to its input values $\boldsymbol{\theta}$ relies on the means and standard deviations obtained on the training set, and its output values $\mathbf{y}$ are considered in $\log_{10}$-scale.

Outliers – Numerical codes may yield numerical instabilities. Such behavior was reported, e.g., in Auld et al. (2007) about the WMAP 3-yr likelihood code. In its domain of validity, the Meudon PDR code produces few of them. However, the considered complex non-linear physics can also lead to physical bistabilities or multistabilities. For example, the H_2 heating process can produce bistable solutions (Burton et al., 1990; Röllig and Ossenkopf-Okada, 2022). In such a case, spatial profiles computed by the code, e.g., of a species density or of the gas temperature, can oscillate between the possible solutions at each position in the modeled cloud. The line integrated intensities computed from these profiles can contain errors of up to a factor of 100 and thus are highly unreliable. Figure 4.5 shows one example of bistability and of its impact on one of the integrated intensities. This case can be easily detected as the bistability affects the temperature profile of the cloud, which is used to compute the integrated intensity of many emission lines. More often, only a handful out of the L lines show such pathological behavior.



(a) Temperature profiles in five simulated clouds, including the one with clear bistability.

(b) Two-dimensional slice of the training set for one line ℓ. The point indicated by a black square corresponds to the bistable profile on the left.

Figure 4.5: Illustration of an invalid integrated intensity due to a bistability in temperature profile. The corresponding point should be withdrawn from the training set.

The code being deterministic, an input vector $\boldsymbol{\theta}$ consistently leads to a unique output vector $\mathbf{y}$. However, in the regions of the parameter space with such multistabilities, variations of intensities can be very chaotic and challenging for a surrogate model to reproduce. Such chaotic values thus deteriorate the accuracy of any surrogate model, interpolation or ANN, and thus should not be used. Unfortunately, as of today there exists no simple or complete procedure to check the physical validity of a precomputed model of the Meudon PDR code. With a first scan of the datasets, we remove a few lines that are particularly affected. The total number of considered lines is therefore reduced from $5\,409$ to $L = 5\,375$. This simple filter leaves other outliers in the training and test datasets. Though we observe that these outliers are rare – we expect less than 1%, we do not have any specific a priori knowledge on their location nor on their exact proportion. Manually checking the validity of each value is unrealistic given the sizes of the two datasets. The most informative hypothesis we can make on outliers is that if one line in a precomputed model is identified as an outlier, then it is likely for this precomputed model to contain other outliers,

especially in lines of the same species or of isotopologues. This hypothesis is exploited in the more thorough outlier detection method exploiting an ANN, introduced in Section 4.6.1.

Summary of the setup – The Meudon PDR code version to emulate is a function $\mathbf{f} : \boldsymbol{\theta} \in \mathbb{R}^D \mapsto \mathbf{y} \in \mathbb{R}^L$, with $D = 4$ and $L = 5\,375$. As the integrated intensities predicted by the Meudon PDR code cover many decades, we train emulators of $\ln \mathbf{f}$, denoted $\ln \widetilde{\mathbf{f}}$, to reproduce log-intensities $\ln \mathbf{y}$. We assume the log-Meudon PDR code, $\ln \mathbf{f}$, to be twice continuously differentiable, except in the case of outliers that should be disregarded. This hypothesis $\ln \mathbf{f} \in \mathscr{C}^2$ is a reasonable physical assumption exploited in Chapter 5 (Section 5.2.1). In Section 4.6, we build emulators $\ln \widetilde{\mathbf{f}}$ so that they satisfy this regularity property.

4.5.2 Comparison metrics

Interpolation methods and ANNs are compared on evaluation speed, memory requirements and approximation accuracy. We describe here the metrics used for the comparison, regardless of how the surrogate models are defined or trained.

The **evaluation speed** is measured on the full set of L lines for $1\,000$ random points. The measurements are performed on a personal laptop equipped with a 11th Gen Intel(R) Core(TM) i7-1185G7, with 8 logical cores running at $3.00\,\text{GHz}$. ANNs and interpolation methods are run on CPU to obtain a meaningful comparison. Running ANNs on a GPU could farther reduce their evaluation times. The implementations of interpolation methods are from the SciPy Python package, popular in ISM studies (Wu et al., 2018). Nearest-neighbor, linear and RBF interpolation implementations allow for the evaluation of a vector function at once. Conversely, the spline interpolation implementation requires an explicit loop over the L lines, which is slow in Python. To avoid an unfair comparison with the other methods, the spline interpolation speeds are not evaluated.

The **memory requirements** are quantified with the number of parameters necessary to fully describe the surrogate model. Interpolation methods, for instance, require storing the full training set. It corresponds to $N_{\text{train}}(D + L) \simeq 1.03 \times 10^8$ parameters. In Python, these parameters are stored using 64-bits floating-point numbers. Storing the full grid requires about $1.65\,\text{GB}$.

The **accuracies** of surrogate models are evaluated on the test set, composed of points not used during training. To quantify accuracies, we define a new metric called the error factor (EF). This metric is a symmetric version of relative errors. Unlike the AE or the SE on $\ln \widetilde{\mathbf{f}}$ and $\ln \mathbf{y}$, this metric does not favor predictions lower than the true value. Indeed, as line intensities are considered in log-scale, the absolute error (Eq. 4.2) corresponds to the log-ratio of the predicted and true line intensities. The error factor is this ratio transformed back to linear scale. For a surrogate model $\widetilde{\mathbf{f}}$ on a given tuple $(\boldsymbol{\theta}, \ln \mathbf{y})$ and line ℓ, it reads

$$\text{EF}\left(\ln \widetilde{\mathbf{f}}\,; (\boldsymbol{\theta}, \ln y_\ell)\right) = \exp\left\{\left|\ln \widetilde{f}_\ell(\boldsymbol{\theta}) - \ln y_\ell\right|\right\} = \max\left\{\frac{\widetilde{f}_\ell(\boldsymbol{\theta})}{y_\ell}, \frac{y_\ell}{\widetilde{f}_\ell(\boldsymbol{\theta})}\right\}. \tag{4.6}$$

As the absolute value ensures positivity in log scale, an error factor is always superior or equal to 1. It can be expressed in percent using a $100 \times (\text{EF} - 1)$ transformation. For readability, the error factor is displayed in percent when $\text{EF} < 2$, i.e., 100%. An error factor that is not in percent is indicated by the multiplication sign, e.g., "$\times 3$" corresponds to $\text{EF} = 3$.

Figure 4.6 compares the relative error and the error factor. The error factor is a symmetrized relative error, as the absolute value also ensures symmetry in log scale. For small errors, i.e., $\text{EF} \simeq 1$, it is similar to the standard relative error. However, for larger errors, the error factor is more relevant in our case. A standard relative error would return 100% for a factor of 2 too high and 50% for a factor of 2 too low, while in both cases $\text{EF} = 2$. In the worst case, a relative

error of 100% corresponds to a factor of 2 too high or a prediction of exactly 0, while EF $= 2$ in the former case and EF $= +\infty$ in the latter. Minimizing a standard relative error would therefore lead to an under-estimation tendency, which is not the case for the proposed error factor.

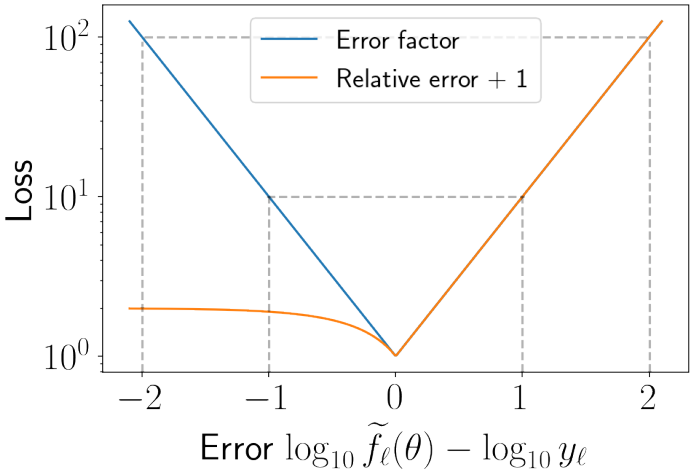


Figure 4.6: Comparison of the error factor and the relative error. To simplify the visualization, the errors are plotted in decimal log scale, while the surrogate models train on the natural log scale. We add 1 to the relative error to simplify the comparison and to avoid a divergence of the log-relative error towards $-\infty$ when $\widetilde{f}_\ell(\boldsymbol{\theta}) \to y_\ell$. The relative error $+ 1$ tends to 2 for large negative errors, i.e., the relative error tends to and is upper bounded by 100%. This effect favors underestimations compared to overestimations. Using the error factor cancels this undesirable effect.

When applied to the full test set, the error factor yields a distribution of errors. This distribution is summarized by its mean, its 99th percentile and its maximum. The mean provides an estimation of the average error to expect. The 99th percentile and maximum provide upper bounds on the error. The maximum is very sensitive to outliers while the 99th percentile is more robust. Figure 4.7 illustrates this distribution summary on a fictional dataset of $20\,000$ points including 0.5% of outliers. The maximum is affected by the outliers, which induces a pessimistic bias for the corresponding error upper bound estimation. The 99th percentile is not significantly affected by the outliers, and provides a more relevant estimator of the actual upper bound of the error factor for this fictional dataset. This example shows that the choice of percentile is a trade-off based on the expected proportion of outliers. Lower percentiles such as 90th or 95th underestimate the upper bound on the error factor, and percentiles higher than 99.5th would in turn be sensitive to outliers like the max. The training and test sets generated with the Meudon PDR case are expected to contain less than 1% of outliers. The 99th percentile is therefore expected to be a relevant estimator of the error upper bound, robust to outliers.

In current IRAM-30m observations, the relative day-to-day calibration accuracy ranges from 3% to 10% – see e.g., Einig et al. (2023). The absolute flux calibration accuracy for ground based observations is more difficult to estimate but cannot be better than the relative calibration accuracy. For a surrogate model to be relevant for observations analysis and physical parameter inference, we set the constraint that satisfactory surrogate models must have a mean error factor below 10%.

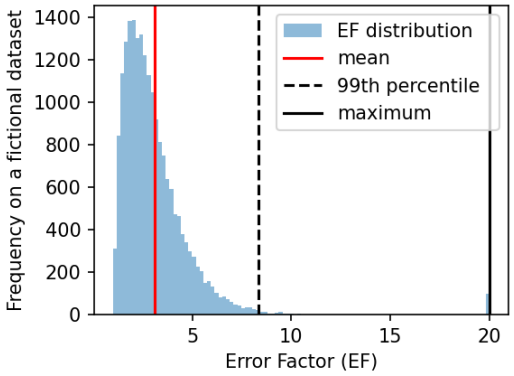


Figure 4.7: Illustration of the impact of outliers on the mean, the 99th percentile and the max estimators of the error factor. Note that the dataset used for this illustration is fictitious.

4.6 Designing and training adapted ANNs

The choice of architecture and training approach of ANNs are now discussed. In the following, ANNs are trained with the MSE loss function. We set the activation function g to the Gaussian error linear unit (GELU) (Hendrycks and Gimpel, 2023) to ensure that the resulting ANN is twice continuously differentiable. Unless explicitly mentioned, our ANNs have $H = 3$ hidden layers of equal size. This choice may not be optimal. A hyperparameter optimization step could improve the network performance, but would require a validation dataset and the training of many networks. As the results of Section 4.7 will show, this step is not necessary to obtain satisfactory results.

ISM models such as the Meudon PDR code present specificities, namely the presence of outliers and the unusual dimensions of the problem – very few inputs to predict many outputs. To address these specificities, dedicated strategies are needed and are described in the subsections below:

1. we apply an outlier removal procedure;
2. with a clustering method, we derive homogeneous line groups simpler to emulate with separate networks;
3. to select an adequate size for hidden layers, we resort to a dimension reduction method;
4. we apply a polynomial transform to augment the input data and thus ease the learning of nonlinearities;
5. we replace the standard ANN architecture by a dense architecture that exploits values in intermediate layers to re-use intermediate computations.

4.6.1 Removing outliers from the training set

Outliers that come from either numerical instabilities or physical bistabilities or multistabilities can be found in both the training and test sets, as described in Section 4.5.1. With a loss function such as the MSE, outliers in the training set greatly deteriorate the quality of a fitted ANN. A similar problem was reported in Auld et al. (2007) about the WMAP 3-yr likelihood code, and addressed by removing them from the dataset. In Auld et al. (2007), the outliers could be identified with a simple binary test. Such a test is not available in our case. More sophisticated methods are therefore required.

Performing regression in presence of outliers is a crucial topic in machine learning. Multiple methods exist for non-linear regression (Rousseeuw and Leroy, 1987). We resort to the method proposed in Motulsky and Brown (2006), which consists of 3 steps:

1. A statistical model is fitted to the training set with a strategy robust to outliers.
2. The training points with largest errors are reviewed.
3. Identified outliers are removed and a new statistical model is fitted to the cleaned training set.

To avoid any risk of biasing the analysis towards optimistic results, we do not remove any value from the test set. We now detail these three steps.

Step 1: first fit robust to outliers – For this first fit, we resort to an ANN designed as described at the introduction of Section 4.6. The size of hidden layers is fixed with the dimension reduction strategy that will be described in Section 4.6.2. We also include the polynomial transform of the input, to be described presented in Section 4.6.3. For specific outlier removal step, this fit is performed using the Cauchy loss function (CL). For a surrogate function $\widetilde{\mathbf{f}}$ and a pair $(\boldsymbol{\theta}, \ln \mathbf{y})$, it reads

$$\mathrm{CL}\left(\ln \widetilde{\mathbf{f}}\,; (\boldsymbol{\theta}, \ln y_\ell)\right) = \ln\left[1 + \left(\ln \widetilde{f}_\ell(\boldsymbol{\theta}) - \ln y_\ell\right)^2\right]. \tag{4.7}$$

Figure 4.8 shows how the squared error (Eq. 4.3), the absolute error (Eq. 4.2), the error factor (Eq. 4.6), and the Cauchy loss function penalize errors. The Cauchy function gives less weight to very large errors, which makes it more robust to outliers.

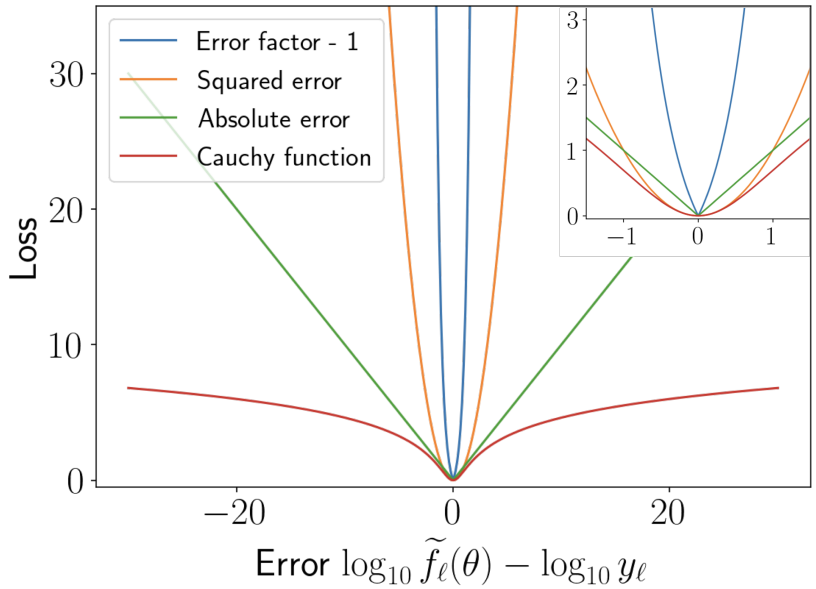


Figure 4.8: Evolution of errors penalization by the considered loss functions. To simplify the visualization, the errors are plotted in decimal log scale, while the surrogate models train on the natural log scale. An error of 30 thus corresponds to a factor of 10^{30}. Since some line intensities range from 10^{-50} to 10^{-2}, very high error can occur early in the training phase.

Step 2: review of training points with largest errors – The review of training points with high errors is performed with a manual procedure. An instability in a given model of the grid does not affect all lines, as all lines are not emitted in the same spatial regions of the model. Therefore, we only remove affected lines instead of the full model. To accelerate this procedure, we exploit similarities between lines. For instance, when one water line intensity is identified as an outlier, it is highly likely that most of the water line intensities of the corresponding precomputed model are outliers. We emphasize that outliers are associated with instabilities or multistabilities. Physically consistent intensities that are challenging to reproduce, e.g., due to fast variations in

a change of regimes, are not considered as outliers and maintained in the training set. In total, $71\,239$ values were identified as outliers, i.e., 0.069% of the training set. Note that this outlier identification step is very informative, as it reveals regions of the parameter space that lead to multistabilities. However, studying these regions is beyond the scope of this thesis.

Step 3: fit on the cleaned dataset – Once outliers are identified, they are removed from the dataset, through a binary mask matrix $\mathbf{M} = (m_{n\ell})_{n\ell}$. This mask permits to disregard only the identified outliers instead of removing all L lines of precomputed models with at least one outlier. In this binary mask, $m_{n\ell} = 1$ indicates that $y_{n\ell}$ is an outlier and should not be taken into account, and $m_{n\ell} = 0$ indicates that $y_{n\ell}$ is not an outlier. Elements of the training set $(\boldsymbol{\theta}_n, \mathbf{y}_n) \in \mathbb{R}^D \times \mathbb{R}^L$ are augmented with corresponding binary mask vectors $\mathbf{m}_n \in \{0,1\}^L$. An ANN, on the one hand, can easily take this mask into account for training by computing the loss function and its gradient on non-masked values only. In the following, a masked version of the MSE relying on the binary mask $\mathbf{M}$ is used when this outlier removal step is taken into account.

Existing implementations of interpolation methods, on the other hand, lack flexibility to handle such a mask during the fit. As some points of the grid are removed for some lines, the spline interpolation can not be applied on the masked training set. Nearest-neighbor, piecewise linear and RBF interpolation methods can be applied but would require line by line fits and predictions, as outliers don't occur for the same training points $\boldsymbol{\theta}$ for all lines. Such line by line manipulation would be extremely slow with a PYTHON implementation. To present a somewhat meaningful comparison between ANNs and interpolation methods on the masked dataset, the masked values are imputed. This imputation relies on a line by line fit of an RBF interpolator with linear kernel. Masked values are replaced by interpolations computed from available data points. Interpolation methods are then fitted with this imputed training set.

Note also that it would be possible to avoid the manual review by training a network and automatically identifying outliers at once. Appendix 4.A outlines a possible approach as well as the reasons why we chose not to apply it in this particular case.

4.6.2 Exploiting correlations between line intensities

Line intensities computed by the Meudon PDR code come from the radiative de-excitation of energy levels. While non-local effects are accounted for in the radiative transfer, the excitation of many lines is affected to a large extent by local variables such as the gas temperature or density. As a result, high correlations between some lines are expected. Figure 4.9 shows the $L \times L$ matrix of absolute Pearson correlation coefficients, with lines grouped by molecule. We indeed find some strong correlations. In particular, lines from the same species are often highly correlated, especially for water isotopologues and molecular hydrogen. However, some species produce lines that are not correlated. For instance, high energy lines of SO have a very small correlation with low energy lines, as the corresponding sub-matrix has a diagonal shape. Finally, some lines from different species are highly correlated, e.g., OH^+, SH^+ and H_2. Handling the L lines independently, as in interpolation methods, ignores these correlations in the line intensities. We exploit these correlations with two strategies: line clustering and dimension reduction.

Line clustering to divide and conquer

Figure 4.9 highlights some clusters of highly correlated lines. These clusters are not simply related to the line carrier. We derive clusters of lines automatically from the correlation matrix using the spectral clustering algorithm (Shalev-Shwartz and Ben-David, 2014, chapter 22). Spectral clustering defines clusters such that lines from a same cluster are as similar as possible and such that lines from different clusters are as different as possible. It relies on a similarity measure such as the Pearson correlation, while most clustering algorithms are distance-based. We set the number of clusters to the value that maximizes the ratio of intra to inter-cluster mean correlations.

Figure 4.10 presents the 4 clusters we obtain. The mean intra and inter-cluster correlations

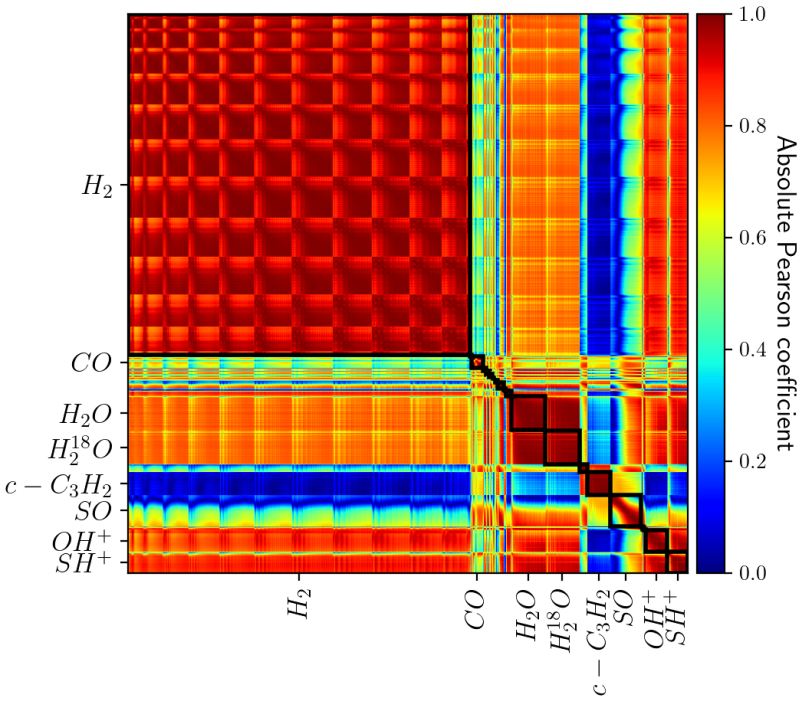


Figure 4.9: Meudon PDR code correlations among the $L = 5\,375$ predicted lines from 27 chemical species, shown with the $L \times L$ matrix of absolute Pearson correlation coefficients. A value of exactly 1 for two lines means that there exists an exact affine relationship between their log-intensities. The black squares on the diagonal group lines from a common chemical species. For readability reasons, only the names of species with more than 100 lines predicted by the Meudon PDR code are displayed.

are 0.895 et 0.462 respectively, while the mean correlation among all lines is 0.73. The obtained clusters contain $3\,712$, $1\,272$, 241 and 150 lines, respectively. This imbalance between clusters comes from the imbalance between molecules. For instance, H_2 corresponds to $3\,282$ lines, i.e., 61% of the lines computed by the Meudon PDR code. Figure 4.9 shows that all these lines are highly correlated. Appendix 4.B provides a more complete description of the content of these 4 clusters. With this approach, an ANN is trained for each cluster.

Using PCA to set the size of the last hidden layer

A second and complementary approach to exploit these correlations is to assume that a vector $\mathbf{y}$ with the L line intensities can be compressed to a vector of size $\widetilde{L} < L$ with limited loss of information. Formally, we assume that the line intensities $\mathbf{y}$ live in a subspace of dimension $\widetilde{L} < L$ where $\widetilde{L}$ can be estimated using a dimension reduction algorithm. We resort to a principal component analysis (PCA) (Shalev-Shwartz and Ben-David, 2014, chapter 23) on the training set, which performs compressions using only affine transformations. We obtain that compressing all $L = 5\,375$ lines with only $\widetilde{L} \simeq 1\,000$ principal components and then decompressing leads to a mean error factor below 0.1% on the training set, which confirms our hypothesis.

Figure 4.1 shows that in an ANN such that $D \ll L$, most parameters belong to the last hidden layer. The size of this layer is thus critical to obtain a good accuracy. Too large a layer might lead to overfitting; too small a layer could not capture the nonlinearities of the dataset. Using PCA, we determine the dimension $\widetilde{L} < L$ of a subspace in which the code outputs can be described with limited accuracy loss. In regression, this last hidden layer also applies an affine transformation. We therefore set the size of the last hidden layer to the estimated dimension $\widetilde{L}$. To predict the $\widetilde{L}$ intermediate values of the last hidden layer, which are then used to predict all L line intensities, the first two hidden layers are set with the same size.

Link with principal component regression – This approach shares similarities with principal component regression, used e.g., in Spurio Mancini et al. (2022), but is not equivalent. Using PCA, one obtains the optimal affine transformation to a subspace of dimension $\widetilde{L} \ll L$. The proposed approach uses PCA to estimate $\widetilde{L}$ to set the size for the hidden layers to a relevant value. Principal component regression (PCR) goes farther by fixing the output layer of the network to the learned transformation. This approach might seem very advantageous in the considered case. Since the output layer is the largest layer in our networks, it would reduce the number of parameters to learn. In this work, we tried both approaches. We chose to leave the output layer as a learnable parameter as PCR leads to worsse results than those presented in the manuscript.

Application to the four clusters of lines – For the networks trained on the four clusters of lines obtained in Section 4.6.2, the size of the last hidden layer is also set to the minimum number of principal components that ensures a decompression with mean error factor below 0.1% on the training set. The obtained sizes $\widetilde{L}$ are approximately 500 (about 13% of the $L = 3\,712$ lines of the cluster), 350 (about 28%), 100 (about 41%) and 75 (50%), respectively. As the bigger clusters are the most homogeneous, they have the smallest ratio $\widetilde{L}/L$ of subspace dimension $\widetilde{L}$ with the total number of lines L. The number of parameters necessary to describe four small specialized ANNs is thus greatly reduced in comparison to a single larger general network.

4.6.3 A polynomial transform to learn nonlinearities

The nonlinearities in the Meudon PDR code make the approximation task challenging. In an ANN, nonlinearities come from the activation function g. However, learning meaningful and diversified nonlinearities is difficult with few hidden layers. Conversely, an ANN with numerous layers can lead to overfitting and requires more time for evaluations and memory for storage. Preprocessing the physical parameters $\boldsymbol{\theta}$ with a variety of predefined non-linear functions eases this learning task while maintaining a small network architecture. We choose to apply a polynomial transform P_p

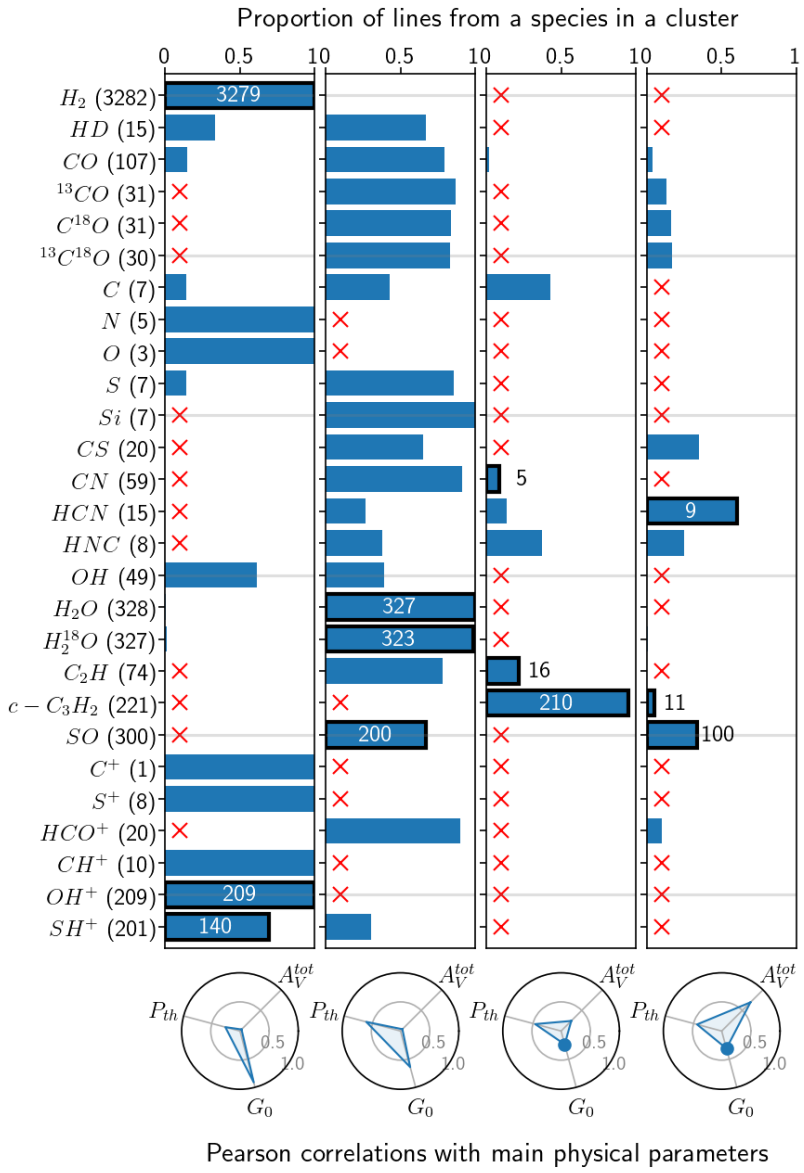


Figure 4.10: On both top and bottom plots: each column corresponds to one of the four line clusters obtained in Section 4.6.2.
Top row: composition of each cluster. Each bar indicates the proportion of lines of a species in a cluster. The red crosses correspond to exactly zero line. For each cluster, the 3 species with the most lines are highlighted.
Bottom row: Pearson correlation of the most representative line of each cluster with the 3 main physical parameters. The most representative line of a cluster is defined as the line with the highest average correlation with the other lines. A round marker at a vertex indicates a negative correlation.

which replaces the input vector $\boldsymbol{\theta}$ of dimension D with an input vector containing all monomials computed from the D entries of degree up to p. For instance, for $D = 3$ and $p = 2$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ is replaced with $P_2(\boldsymbol{\theta}) = (\theta_1,\ \theta_2,\ \theta_3,\ \theta_1^2,\ \theta_2^2,\ \theta_3^2,\ \theta_1\theta_2,\ \theta_1\theta_3,\ \theta_2\theta_3) \in \mathbb{R}^9$. For $D = 4$ and $p = 3$, $P_3(\boldsymbol{\theta}) \in \mathbb{R}^{34}$. This approach is classic in regression (Ostertagová, 2012), but less common to train ANNs.

It is well known in polynomial regression that a high maximum degree p can lead to overfitting (Shalev-Shwartz and Ben-David, 2014, chapter 11). The analysis of the physical processes indicates that the gas structure and emission properties depend on control quantities combining G_0, n_H (or P_{th}) and A_V^{tot}. For instance, G_0/n_H is known to play an important role in PDRs (Sternberg et al., 2014). It is therefore important to consider monomials combining these 3 physical parameters. In contrast, the angle φ is assumed to have a simpler role in the model. To avoid overfitting, we choose the minimum value that combines the three parameters, $p = 3$, and thus consider the polynomial transforms P_3. This transformation is applied to the input variables after the preprocessing step described in Section 4.5.1 – log scale for P_{th}, G_0 and A_V^{tot}, and standardization of the $D = 4$ parameters. It is implemented as an additional first fixed hidden layer. The gradient $\nabla_{\boldsymbol{\theta}}\widetilde{\mathbf{f}}$ can thus be efficiently evaluated with auto-differentiation methods.

4.6.4 Dense networks to reuse intermediate computations

Introduction to the dense architecture – The fully connected ANNs architecture considered so far, shown in Figure 4.1, is widely used in the deep learning community. However, this architecture struggles to maintain input information in hidden layers, as it is transformed with non-linear activation functions. It might therefore fail to reproduce very simple relationships. For instance, the intensity of UV-pumped lines of H_2 is highly correlated with G_0. Using G_0 directly to predict intensities of such lines thus might be more relevant than passing it through non-linear transformations. This architecture also struggles to pass gradient information all the way back to the first hidden layers. This phenomenon, called *gradient vanishing*, might lead to largely suboptimal trained networks. The recent residual (He et al., 2016) and dense architectures (Huang et al., 2017) address these two issues. We use the dense architecture for our regression problem.

A dense architecture is a special type of feedforward architecture where the input of a layer $j+1$ is the concatenation of the input and output vectors of the previous layer j: $\boldsymbol{\theta}^{(j+1)} = [\![\boldsymbol{\theta}^{(j)}, \mathbf{y}^{(j)}]\!]$. This architecture focuses on reusing intermediate values in hidden layers and can thus reduce the number of parameters to train.

Figure 4.11 illustrates this dense architecture for a simple ANN with $H = 2$ hidden layers and the same sequence of layer input sizes $(i_j)_{j=1}^{H+1}$ used to illustrate the standard feedforward architecture in Figure 4.1. The output sizes o_j of hidden layers are much smaller with the dense architecture, as the input of layer j concatenates the input and output of layer $j-1$. The weight matrices $\mathbf{W}^{(j)}$ of hidden layers are thus much smaller as well, which reduces the total number of parameters to train. By lowering the number of parameters to learn while providing the same number of inputs to the output layer, this architecture limits overfitting risks.

Considered dense architecture – As the number of parameters per layer is reduced, we define ANNs with $H = 9$ hidden layers, i.e., 6 more layers than for the proposed networks with the standard architecture, and yet with a similar total number of parameters. By definition, the size of the hidden layers in a dense architecture is strictly increasing, as the size i_{j+1} of a layer input is the sum $i_j + o_j$ of the input and output sizes of the previous layer. The network is set so that the input i_{j+1} of a layer $j+1$ is 50% larger than the input of the previous layer i_j. With this geometric progression and the polynomial transform P_3, the input of the output layer contains $1\,296$ neurons, which is 29.6% larger than the recommendation from PCA obtained in Section 4.6.2. However, out of these $1\,296$ neurons, 34 correspond to the input values, 17 to the output of the first hidden layer, 25 to the output of the second hidden layer, etc. In other words, though the input of the output layer contains more neurons for the considered dense neural network than the PCA recommendation, a majority of these neurons are the result of fewer

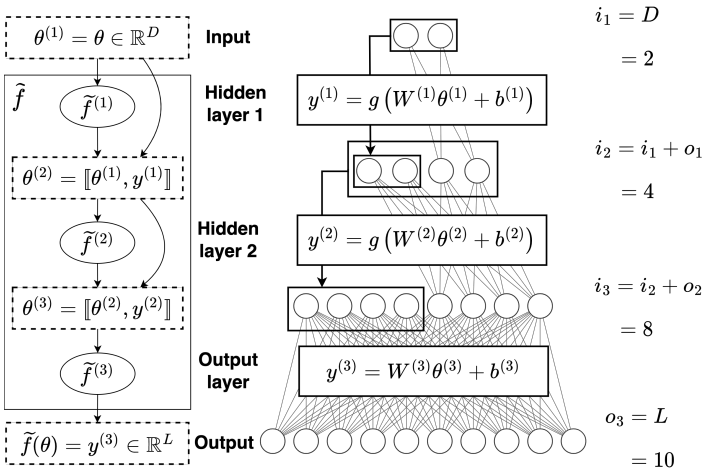


Figure 4.11: Structure of a dense neural network with $H = 2$ hidden layers and the same sequence of layer input sizes $(i_j)_{j=1}^{H+1}$ used to illustrate the feedforward architecture in Figure 4.1.

transformations.

When using this dense architecture strategy with the clustering approach, four dense networks with $H = 9$ hidden layers are designed. The size of the last hidden layer is also set to a slightly larger value than the PCA recommendation. The geometric progressions of these 4 networks are set to 35%, 30%, 15% and 10%, respectively.

4.7 Experiments: application to the Meudon PDR code

The ANNs designed and trained with the proposed strategies are compared with interpolation methods with respect to accuracy, memory and speed. Table 4.3 shows the results of the comparison. It is divided in two halves. The first presents models trained on the raw training set, while the second contains models trained on the cleaned training set. The cleaned dataset is based on the mask defined with the outlier detection procedure of Section 4.6.1. In each half, the results of interpolation methods are first listed, followed by ANNs combining one or more of the presented strategies.

The code and data produced for this experiment are public. The training set, the test sets, and the mask on the train set can be found at https://ism.obspm.fr/files/ArticleData/2023_Palud_Einig/2023_Palud_Einig_trained_ANN.zip. The code can be found at https://github.com/einigl/ism-model-nn-approximation. All proposed ANNs were implemented using the PyTorch[4] Python library (Paszke et al., 2019). The most accurate ANN is available at https://ism.obspm.fr/files/ArticleData/2023_Palud_Einig/2023_Palud_Einig_data.zip.

4.7.1 Performance analysis

The proposed ANNs outperform all interpolation methods on all aspects by a large margin: they are between 100 and $1\,000$ times faster than reasonably accurate interpolation methods, and between 14 and 38 times lighter in terms of memory. Interpolation methods handle the prediction of L lines as L independent operations, while ANNs handle the L lines at once, which is much faster. Interpolation methods require storing the full training set that contains 103 million 64-bit floating point numbers, i.e., 1.65 GB in size. In contrast, ANNs use shared intermediate values in hidden layers to predict all lines, which limits redundant computations and effectively compresses the dataset. They can thus be fully described with between 2.7 and 7.8 million parameters, i.e.,

[4] https://pytorch.org/

Table 4.3: Performance of interpolation methods and of the proposed ANNs with and without removal of outlier from the training set. Evaluation speeds are measured on the full set of L lines for $1\,000$ random points. The measurements are performed on a personal laptop equipped with 8 logical cores running at 3.00 GHz. Error factors are evaluated on the test set. For ANN architectures:
– R: design of the last hidden layer using PCA,
– P: polynomial transform of the input up to degree 3,
– C: line clustering and parallel training of four specialist ANNs,
– D: dense architecture.

	Method		Error factor mean	Error factor 99th per.	Error factor max	Memory (MB)	Speed (ms)
No outlier removal	near. neighbor		×13.1	×11.3	×3e5	1 650	62
	linear		15.7	×2.3	×143	1 650	1.5e3
	spline	linear	15.7	×2.3	×144	1 650	...
		cubic	11.2	×2.2	×122	1 650	...
		quintic	19.1	×2.9	×304	1 650	...
	RBF	linear	10.2	96.8	×99	1 650	1.1e4
		cubic	10.4	×2.1	×112	1 650	1.1e4
		quintic	10.9	×2.1	×118	1 650	1.1e4
	ANN	R	7.3	64.8	×**81**	**118**	**12**
		R+P	**6.2**	**49.7**	×84	**118**	13
Outlier removal on training set	near. neighbor		×13.1	×11.6	×3e5	1 650	62
	linear		15.9	×2.4	×143	1 650	1.5e3
	spline	linear	15.9	×2.4	×144	1 650	...
		cubic	11.1	×2.2	×120	1 650	...
		quintic	20.0	×2.7	×285	1 650	...
	RBF	linear	10.3	97.3	×97.5	1 650	1.1e4
		cubic	10.5	×2.0	×106	1 650	1.1e4
		quintic	10.9	×2.0	×114	1 650	1.1e4
	ANN	R	5.1	42.0	×**32.8**	118	12
		R+P	5.5	42.3	×41	118	13
		R+P+C	4.9	44.5	×44	51	14
		R+P+D	**4.5**	**33.1**	×33.8	125	**11**
		R+P+C+D	4.8	37.9	×37.6	**43**	14

between 43 MB and 118 MB. Finally, the proposed ANNs are roughly twice as accurate as the best interpolation methods on average and between two and three times as accurate with respect to the percentile 99th. Overall, the proposed ANNs are the only surrogate models that yield a mean error factor lower than 10% and thus that are suited to a comparison with actual observations.

4.7.2 Removing outliers is crucial

When the outlier removal step is not applied, the distribution of errors is highly skewed for all surrogate models. For interpolation methods, the 99th percentile reveals that around 99% of the predictions correspond to errors lower than a factor of 2. For the best neural network (R+P), it reveals that 99% of the errors are lower than 49.7%. However, for all methods, the maximum error is at least 80 times higher than the 99th percentile and reaches unacceptable values. An inspection of the highest errors reveals that they are close to training points with outliers, which indicates that these outliers significantly deteriorate the fit.

After removing outliers from the training set, interpolation methods do not show average accuracy improvement. Only little improvements can be observed on the 99th percentile and maximum EF, especially for the RBF interpolation methods. Replacing outliers with interpolated values is therefore not relevant to derive surrogate models based on interpolation methods in this case. In contrast, the two ANNs trained both with and without the outlier removal step (R and R+P) show consequent improvements. With outlier removal, the mean EF decreased from 7.3% and 6.2% to 5.1% and 5.5%, respectively. Similarly, the 99th percentile dropped from 64.8% and 49.7% to 42% and 42.3%. Finally, the maximum error is reduced by more than a factor of 2. These important improvements demonstrate the interest of filtering outliers from the training set before training ANNs.

4.7.3 The importance of the polynomial feature augmentation

The polynomial transform improves the accuracy in presence of outliers in the training set, but deteriorates it after the outlier removal step. It provides flexibility to learn abrupt nonlinearities caused by outliers. However, with outliers removed, the function to learn is smoother. The EF on the masked training set is 1.44% without the polynomial transform (R) and 0.77% with it (R+P), while the EF on the test set is lower without the polynomial transform. This improvement on the training set does not lead to an improvement on the test set, suggesting an overfit. The polynomial transform therefore requires additional strategies to better reproduce data unused during the training phase.

Both the clustering step and dense architecture, used with the polynomial transform, led to better accuracy. The surrogate model that exploits the line clustering but not the dense architecture (R+P+C) improves the mean accuracy by 0.2 percentage points, while requiring 57% fewer parameters than the first two networks (R and R+P). A potential cause of the average error factor improvement is the separation of the trainings of each specialized ANN. Since H_2 lines represent 61% of all L lines, they dominate the loss function and thus are learned in priority. To separate them from other clusters might have improved performance on those other clusters.

The surrogate model based on a single network with dense architecture (R+P+D) is the most accurate on average and provides the lowest error upper bound for the robust 99th percentile estimator. Even with more trainable parameters than the first two networks, it does not overfit. It is also the fastest model as reusing intermediate values reduces the number of computations.

Finally, combining both line clustering and dense architectures (R+P+C+D) yields the lowest memory usage with only 2.7 million parameters, i.e., 43.2 MB, which is 38 times lighter than for interpolation methods. It also provides very good accuracy, both on average and for upper bounds.

Overall, a dense architecture and the line clustering effectively limit overfitting and thus perform better on data not used during the training phase. The line clustering leads to the

lightest models regarding memory requirements, and the dense architecture to the most accurate models.

4.8 Conclusion

The interpretation of observations of atomic and molecular tracers in galactic and extragalactic interstellar medium requires comparison with state-of-the-art astrophysical models to infer physical conditions. Such inference procedure requires numerous evaluations of the numerical model, particularly so for Bayesian approaches. Inference on large observation maps – that are becoming more and more common – farther relies on many evaluations. ISM models are often too slow to perform such inference and are generally emulated using grids of precomputed models. This emulation approach induces errors that are seldom quantified in the literature. Besides, these methods can have high evaluation time and memory costs.

In this chapter, the general problem of deriving a fast, accurate and memory-light surrogate model for a time-consuming ISM numerical model has been addressed. The proposed approach has been assessed in the case of the Meudon PDR code, a state-of-the-art ISM code. Four common families of interpolation methods – nearest-neighbor, linear, spline and RBF – have been compared to specifically designed ANNs. We found that ANNs outperform all interpolation methods by a large margin in terms of accuracy, speed and memory usage.

Attaining this performance level for an ISM model such as the Meudon PDR code requires addressing their specificities. First, ISM models usually predict many statistically independent observables – e.g., line intensities of many species – from few parameters – e.g., gas density or temperature. This setting is unusual in ANN applications – except for ANNs that generate structured data such as images, text or time series. Second, due to numerical instabilities or physical bistabilities or multistabilities, such models sometimes produce outliers that harm the training process. In this chapter, we proposed and combined five strategies to design and train adapted ANNs:

- To identify outliers, we train a first ANN with a loss function robust to large errors. Training points corresponding to large errors are manually reviewed. Identified outliers are removed from the training set.
- Lines are clustered into homogeneous subsets that are simpler to emulate: for each cluster one ANN is defined and trained.
- A dimension reduction technique, PCA, is used to determine an adequate size of hidden layers.
- A polynomial transform of the input physical parameters provides precomputed nonlinearities to the network, which permits the learning of nonlinearities with a limited number of hidden layers.
- A dense architecture exploits intermediate computations and thus limits redundant computations. Using such an architecture instead of the standard feedforward neural network architecture improves speed and avoids overfitting.

With the proposed strategies, ANNs achieve 4.5% average accuracy while the best interpolation method, RBF, attains 10.2%. Upper bound on the errors, quantified using their 99th percentile, reaches 33.1% for our ANNs compared to 97% for RBF. Besides, our ANNs are $1\,000$ times faster than RBF and are more than 10 times lighter in terms of memory. This chapter focuses on an application to the Meudon PDR code, motivated by the inverse problem at the core of this thesis. However, the proposed strategies are sufficiently general to be applicable to many other ISM models.

As we will show in Chapter 5, the fast and accurate ANN emulators obtained in this chapter enables performing sampling-based inference on observation maps using the Meudon PDR code, a physically comprehensive model. As we will show in the Chapter 6, it will also permit efficient analyses of large observations maps produced by today's instruments such as the JWST, ALMA, or the ORION-B dataset observed by the IRAM 30m (Pety et al., 2017).

Appendix 4.A Automatic outlier detection procedure

In this Section, we describe one possible automatic method for outlier identification along with the reasons why we chose not to use it. This method consists in simultaneously identifying outliers and training the ANN with the corresponding masked loss function. A similar approach was used in Gratier et al. (2016), where the likelihood combined two generative models: one for outliers values and one for non-outliers values. An additional parameter – to be tuned during training – determines how likely a value is to be an outlier. Formally, the mask $\mathbf{M}$ would be considered as a continuous variable in $[0,1]^{N\times L}$ optimized along with the parameters $\boldsymbol{\psi}$. Using a masked squared error function, one loss function option could be

$$\mathcal{L}\left(\widetilde{\mathbf{f}},\mathbf{M};\mathcal{D}\right)=\left[\frac{1}{NL}\sum_{n=1}^{N}\sum_{\ell=1}^{L}(1-m_{n\ell})(\ln\widetilde{f}_\ell(\boldsymbol{\theta}_n)-\ln y_{n\ell})^2\right]+\lambda_1 r_1(\mathbf{M})+\lambda_2 r_2(\mathbf{M}), \quad (4.8)$$

where r_1 is a regularization function that favors $m_{n\ell}$ values close to 0 or 1, where r_2 is a regularization function that favors masks $\mathbf{M}$ that agree with a priori knowledge on outliers, and where $\lambda_1,\lambda_2>0$ are regularization weights to be tuned. These properties are satisfied for instance with

$$r_1(\mathbf{M})=\iota_{[0,1]}(m_{n\ell})+\frac{1}{NL}m_{n\ell}(1-m_{n\ell}) \quad (4.9)$$

where $\iota_{[0,1]}$ is the indicator function on the $[0,1]$ interval, i.e., where $\iota_{[0,1]}(x)=0$ if $x\in[0,1]$ and $+\infty$ otherwise, and

$$r_2(\mathbf{M})=\|\mathbf{M}\|_1=\sum_{n=1}^{N}\sum_{\ell=1}^{L}|m_{n\ell}|. \quad (4.10)$$

Figure 4.12 shows the corresponding loss function for one value $m_{n\ell}$. The total regularization has a global minimum of value 0 at $m_{n\ell}=0$, i.e., when $y_{n\ell}$ is not considered as an outlier. It also has a local minimum at $m_{n\ell}=1$, i.e., when $y_{n\ell}$ is considered as an outlier. This local minimum has a higher value to limit the number of identified outliers. Higher values of λ_2 result in higher values attained at this local minimum, which in turn result in less identified outliers. Intermediate values are penalized to encourage the mask values towards 0 and 1.

The main drawback of this approach is that the model might classify important and physically consistent points that are challenging to emulate as outliers. The presented function r_2 considers all couples (n,ℓ) independently, which would cause any minimization algorithm to mask only the hardest values to reproduce with an ANN. These hardest values to reproduce might not be outliers and carry significant physical content that a relevant surrogate model must learn. For instance, values close to an abrupt change of regime may be hard to reproduce, but an emulator that disregards them would be irrelevant in observation analyses. More subtle regularization functions r_2 incorporating a priori knowledge on the outliers should thus be designed. For instance, r_2 could encode the hypothesis stated in Section 4.5.1, that is, if a line of a precomputed model is identified as an outlier, then it is likely that this precomputed model contains other outliers, especially among the lines emitted by the same species and its isotopologues. The design of such regularization functions is not trivial, and will then require tuning the associated parameters λ_1 and λ_2, which is difficult as well. For this reason, we chose to use the method described in Section 4.6.1 for our case, with its required manual review of a part of the training set. In cases where informative prior knowledge on outlier location or distribution is accessible, and where the mathematical formulation of such prior knowledge is manageable, this type of approach can bypass the need for a manual review.

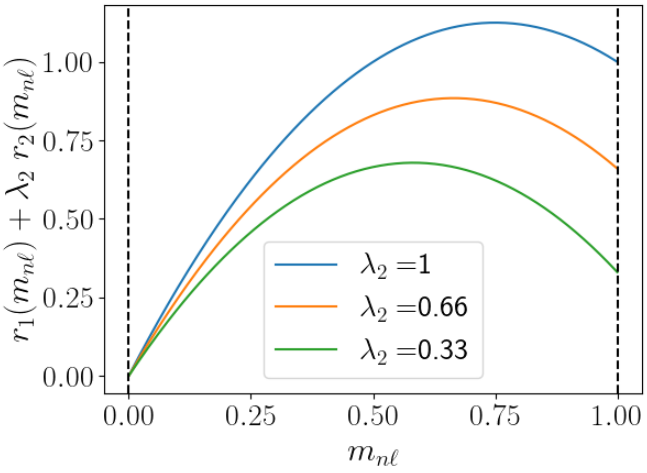


Figure 4.12: Regularization on the mask $\mathbf{M}$ with the r_1 and r_2 functions proposed in Eq. 4.9 and Eq. 4.10, respectively, and $\lambda_1 = 1$. Values out of the $[0, 1]$ interval are $+\infty$ due to the indicator function.

Appendix 4.B Content of clusters of lines

Figure 4.10 presents the four clusters obtained in Section 4.6.2. We now describe the line content of the four clusters. All species have lines distributed in at most 3 clusters, except for ^{12}CO that has lines in each of the 4 clusters. CO lines are indexed with 2 quantum numbers: the rotational number J and the vibrational number v.

Cluster 1 – The first cluster gathers lines that are emitted from the most external UV illuminated layers of the cloud and trace hot chemistry. It includes all H_2 lines but three, and is thus the largest. It also contains all lines from OH^+ (209), CH^+ (10), all 8 lines of S^+, all 5 lines of N, all 3 lines of O and the $158\,\mu$m line of C^+, and CO transitions with low J values for $v = 1 - 0$ and $v = 1 - 1$. The line intensities of this cluster are highly and positively correlated to G_0, and not correlated at all with A_V^{tot}.

Cluster 2 – The second cluster contains 99% of the 655 lines from water $H_2^{16}O$ and its isotopologue $H_2^{18}O$. It also contains lines from high energy levels for several molecules (HD, CO, ^{13}CO, $C^{18}O$, $^{13}C^{18}O$, HNC, HCN, HCO^+, SO, CN, SH^+, C_2H, OH, and CS), and transitions from the neutral atoms C, Si and S. Line intensities in this cluster are positively correlated with P_{th} and G_0 and not at all with A_V^{tot}.

Cluster 3 – The third cluster contains mostly c-C_3H_2 lines, and some C_2H lines. It also includes two transitions of CO with moderate J values in the lowest vibrational level $v = 0$ ($J = 3 - 2$ and $J = 4 - 3$). Its line intensities are overall positively correlated with P_{th} and A_V^{tot} and negatively correlated with G_0.

Cluster 4 – The fourth cluster contains the low energy lines of ^{13}CO, $C^{18}O$, $^{13}C^{18}O$, HNC, HCN, HCO^+, SO and c-C_3H_2, and the lowest temperature transitions of CO ($J = 1 - 0$ and $J = 2 - 1$). Its line intensities are overall positively correlated with P_{th}, strongly positively correlated with A_V^{tot} and negatively correlated with G_0.

Chapter 5

An MCMC algorithm for efficient inversion with quantified uncertainty

> " *At first glance, it might appear surprising that a trivial mathematical result obtained by an obscure minister over 200 years ago ought still to excite so much interest across so many disciplines, from econometrics to biostatistics, from financial risk analysis to cosmology.* "
>
> Trotta (2008)

Contents

In Part I, we depicted the interstellar medium (ISM), presented the Bayesian approach to model and solve an inverse problem, and reviewed many statistical inference applications in ISM

studies. In Chapter 4, we derived a fast, light and accurate artificial neural network (ANN) emulator of the Meudon PDR code, i.e., of the forward model considered in this thesis inverse problem. By construction, this ANN is twice continuously differentiable.

As outlined in Chapter 3, the proposed observation model combines additive Gaussian noise, multiplicative lognormal calibration errors, and censorship. In this chapter, we formally set the inverse problem studied in this thesis. Our goal is to derive large maps of physical parameters $\boldsymbol{\Theta} \in \mathbb{R}^{N \times D}$ from large maps of observations $\mathbf{Y} \in \mathbb{R}^{N \times L}$, with N the number of beams or pixels in the maps, D the number of considered physical parameters (e.g., the thermal pressure or the visual extinction) and L the number of observed atomic or molecular integrated emission lines. This inverse problem is solved with a new MCMC sampler. We resort to an MCMC algorithm to provide credibility intervals associated with the estimations. This quantifies the uncertainty for applications such as astrophysics where no ground truth is available. We exploit the map structure and resort to a spatial regularization prior. As the resulting posterior distribution is potentially multimodal, we design a dedicated MCMC sampler to explore it. This sampler combines two sampling kernels: one responsible for global exploration, and one for efficient local exploration.

This chapter is based on Palud et al. (2023b), a journal article published in IEEE *transactions on signal processing (TSP)* that describes the model and the sampler. Two conference articles also introduce the model and the sampler: Palud et al. (2022a), published in the Gretsi conference, and Palud et al. (2022b). Finally, the Bayesian hypothesis testing was published in a Gretsi conference article, Palud et al. (2023a).

In Section 5.1, we derive an approximation of the likelihood function with controlled error. The prior and posterior distributions are also introduced. Section 5.2 introduces the proposed sampler. In the illustrative Section 5.4, the proposed sampler is applied to two classical synthetic multimodal use cases, namely a Gaussian mixture and a sensor localization problem. Results are compared with other state-of-the-art samplers outlined in Chapter 2 (Section 2.A). In the illustrative Section 5.4.2, the sampler is validated on the astrophysical inverse problem of interest on synthetic data.

5.1 Proposed statistical model

This section first proposes an approximation of the likelihood with controlled error based on a combination of purely Gaussian additive and purely lognormal multiplicative approximations. Then, it introduces the considered prior distribution. Finally, we combine the prior distribution with the likelihood function to define the posterior distribution.

5.1.1 Approximation of the likelihood function

Summary of the observation model – The observation model presented in Chapter 3 (Section 3.4) is briefly recalled. Individual observations $\mathbf{y} = (y_\ell)_{\ell=1}^L$ gather L lines. They are considered to be generated from some parameter $\boldsymbol{\theta} \in \mathbb{R}^D$ and the Meudon PDR code $\mathbf{f} : \mathbb{R}^D \to \mathbb{R}^L$, where D is assumed to remain moderate, e.g. $D \lesssim 10$. The Meudon PDR code prediction for a channel ℓ is denoted by f_ℓ, so that for any $\boldsymbol{\theta} \in \mathbb{R}^D$, $\mathbf{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \cdots, f_L(\boldsymbol{\theta}))$. As a single evaluation of the Meudon PDR code $\mathbf{f}$ is slow and computationally expensive, it is replaced by the neural network-based surrogate model $\widetilde{\mathbf{f}}$ derived in Chapter 4. Individual observations and parameters are grouped in maps $\mathbf{Y} = (\mathbf{y}_n)_{n=1}^N$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)_{n=1}^N$ of N pixels, with N potentially $\mathcal{O}(10^4)$. The sensors are assumed to have a lower limit of sensitivity $\omega_{n\ell} \in \mathbb{R}$ below which an observation is considered censored. Both an additive and multiplicative noise degrade the observations. The resulting observation model is, for $n \in [\![1, N]\!]$ and $\ell \in [\![1, L]\!]$,

$$y_{n\ell} = \max\left\{\omega_{n\ell},\ \varepsilon_{n\ell}^{(m)} \widetilde{f}_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}^{(a)}\right\}, \tag{5.1}$$

where $\varepsilon_{n\ell}^{(a)} \sim \mathcal{N}(0, \sigma_{a,n\ell}^2)$ is an additive Gaussian white noise, and $\varepsilon_{n\ell}^{(m)} \sim \text{Lognormal}(-\sigma_m^2/2, \sigma_m^2)$ is a lognormal multiplicative noise such that $\mathbb{E}[\varepsilon_{n\ell}^{(m)}] = 1$. The noise terms $\varepsilon_{n\ell}^{(a)}$ and $\varepsilon_{n\ell}^{(m)}$ are assumed independent with known standard deviations $\sigma_{a,n\ell}$ and σ_m, respectively. For low ratios $\widetilde{f}_\ell(\boldsymbol{\theta}_n)/\sigma_{a,n\ell}$ low intensities, the additive noise dominates; for high ratios observations are mainly damaged by the multiplicative noise.

Taking a mixture of noises into account – The likelihood function associated with the observation model (Eq. 5.1) has no simple closed-form expression due to the mixture of noises. We first consider the uncensored part of the model from Eq. 5.1. In statistics, there are three main approaches to handle such a mixture of noises.

Approach 1: neglecting one source of noise – The first approach simply consists in neglecting one source. For instance, similar mixtures of additive and multiplicative noises also occur in medical ultrasound imaging (Krissian et al., 2007) or laser imaging and synthetic aperture radars (Durand et al., 2010). In these fields, when turning to inference, one of the two noises is generally neglected for sake of tractability (Krissian et al., 2007). As already mentioned, the Meudon PDR code spans several decades which causes the nature of the dominant noise to depend on the amplitude of $\widetilde{f}_\ell(\boldsymbol{\theta})$. The additive noise can be neglected when $\widetilde{f}_\ell(\boldsymbol{\theta}_n) \gg \sigma_{a,n\ell}$, while the multiplicative noise becomes negligible when $\widetilde{f}_\ell(\boldsymbol{\theta}_n) \ll \sigma_{a,n\ell}$. As each of the two noise sources dominates in a physical regime, both need to be taken into account in the inversion.

Approach 2: using a hierarchical model – A second approach, applied e.g., in Kelly et al. (2012), relies on hierarchical models. An auxiliary variable $\mathbf{U} \in R^{N\times L}$ such that $u_{n\ell} = \varepsilon_{n\ell}^{(m)} \widetilde{f}_\ell(\boldsymbol{\theta}_n)$ is included. The observation is then rewritten as

$$\begin{cases} \ln u_{n\ell} = \ln \widetilde{f}_\ell(\boldsymbol{\theta}_n) + \ln \varepsilon_{n\ell}^{(m)}, \quad \ln \varepsilon_{n\ell}^{(m)} \sim \mathcal{N}(-\sigma_m^2/2, \sigma_m^2), \\ y_{n\ell} = u_{n\ell} + \varepsilon_{n\ell}^{(a)}, \quad \varepsilon_{n\ell}^{(a)} \sim \mathcal{N}(0, \sigma_{a,n\ell}^2), \end{cases} \tag{5.2}$$

and $\mathbf{U}$ needs to be sampled and inferred along with $\boldsymbol{\Theta}$, yielding the augmented posterior

$$\pi(\boldsymbol{\Theta}, \mathbf{U}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\mathbf{U})\pi(\mathbf{U}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta}). \tag{5.3}$$

The two likelihood terms $\pi(\mathbf{Y}|\mathbf{U})$ and $\pi(\mathbf{U}|\boldsymbol{\Theta})$ have simple closed-form expressions, as they are Gaussian and lognormal, respectively. This approach permits the use of the exact likelihood model. Such a model is generally sampled with a Gibbs algorithm, i.e., that alternatively samples from $\pi(\boldsymbol{\Theta}|\mathbf{U}, \mathbf{Y})$ and $\pi(\mathbf{U}|\boldsymbol{\Theta}, \mathbf{Y})$. It doubles the number of parameters to sample, which can cause memory issues with large observation maps. This divide-to-conquer approach compensates for this additional cost if sampling from the conditional distributions is simple, i.e., if the coupling between $\mathbf{U}$ and $\boldsymbol{\Theta}$ is limited.

Preliminary results on this approach with the considered case, not reported in the manuscript, suggested that sampling is hard in practice although conceptually simple. Unlike in Kelly et al. (2012), $\widetilde{f}_\ell$ is not gradient Lipschitz continuous and thus neither is $-\ln \pi(\mathbf{U}|\boldsymbol{\Theta})$. This absence of gradient Lipschitz regularity and the changes in dominant noise creates a high coupling between $\boldsymbol{\Theta}$ and $\mathbf{U}$, which makes the extended posterior difficult to sample. In the region of the parameter space where the additive noise dominates, exploring the distribution on $\boldsymbol{\theta}_n$ is likely to be slow because the multiplicative noise is negligible. Similarly, exploring the distribution on $u_{n\ell}$ is likely to be slow when the additive noise is negligible. Avoiding a hierarchical model permits to always consider the dominant noise.

A possibility we did not explore yet is to resort to the ancillarity-sufficiency interweaving strategy (ASIS) sampling approach, dedicated to reducing coupling in hierarchical models. This has already been applied in dust studies, including in Kelly et al. (2012). Another option to reduce the coupling would be to resort to a splitting algorithm – see e.g., Vono et al. (2019).

Approach 3 (proposed approach): approximating the likelihood function – The third approach approximates the full likelihood function. For instance, Nicholson and Kaipio (2020) approximates the mixture using a purely additive Gaussian model. The approach we propose builds on this approximation. The additive noise $\varepsilon_{n\ell}^{(a)}$ in Eq. 5.1 can be neglected when $\widetilde{f}_\ell(\boldsymbol{\theta}_n) \to \infty$, while the multiplicative noise $\varepsilon_{n\ell}^{(m)}$ becomes negligible as $\widetilde{f}_\ell(\boldsymbol{\theta}_n) \to 0$. Therefore, for each observation $y_{n\ell}$, the true likelihood is approximated using three different regimes: low, intermediate and high values of $\widetilde{f}_\ell(\boldsymbol{\theta}_n)$. In the low value regime, the true likelihood function $\pi(y_{n\ell}|\boldsymbol{\theta}_n)$ is approximated by an additive Gaussian approximation $\pi^{(a)}(y_{n\ell}|\boldsymbol{\theta}_n)$ corresponding to

$$y_{n\ell} \simeq \widetilde{f}_\ell(\boldsymbol{\theta}_n) + e_{n\ell}^{(a)}, \quad e_{n\ell}^{(a)} \sim \mathcal{N}(m_{a,n\ell}, s_{a,n\ell}^2), \tag{5.4}$$

where $m_{a,n\ell}$ and $s_{a,n\ell}^2$ are obtained by matching the two first moments with the model from Eq. 5.1, which yields

$$\begin{cases} m_{a,n\ell} = 0, \\ s_{a,n\ell}^2 = \widetilde{f}_\ell(\boldsymbol{\theta}_n)^2(e^{\sigma_m^2} - 1) + \sigma_{a,n\ell}^2. \end{cases} \tag{5.5}$$

Conversely, in the high value regime, a multiplicative lognormal approximation $\pi^{(m)}(y_{n\ell}|\boldsymbol{\theta}_n)$ is used. It reads

$$y_{n\ell} \simeq e_{n\ell}^{(m)} \widetilde{f}_\ell(\boldsymbol{\theta}_n), \quad e_{n\ell}^{(m)} \sim \text{Lognormal}\left(m_{m,n\ell}, s_{m,n\ell}^2\right), \tag{5.6}$$

where moment matching with Eq. 5.1 yields:

$$\begin{cases} m_{m,n\ell} = -\frac{1}{2}\left\{\sigma_m^2 + \ln\left[1 + \frac{\sigma_{a,n\ell}^2}{\widetilde{f}_\ell(\boldsymbol{\theta}_n)^2 e^{\sigma_m^2}}\right]\right\}, \\ s_{m,n\ell}^2 = -2\, m_{m,n\ell} \quad \text{so that } \mathbb{E}[e_{n\ell}^{(m)}] = 1. \end{cases} \tag{5.7}$$

For the intermediate regime, for each channel ℓ, we introduce parameters $\boldsymbol{a}_\ell = (a_{\ell,0}, a_{\ell,1}) \in \mathbb{R}^2$. $a_{\ell,0}$ pinpoints the low to intermediate value transition and $a_{\ell,1}$ the intermediate to high value transition. In this intermediate regime, i.e., $a_{\ell,0} \leq \widetilde{f}_\ell(\boldsymbol{\theta}_n) \leq a_{\ell,1}$, we propose to use a geometric average of the two likelihood approximations $\pi^{(a)}(y_{n\ell}|\boldsymbol{\theta}_n)$ and $\pi^{(m)}(y_{n\ell}|\boldsymbol{\theta}_n)$ with weights $1-\lambda$ and λ, respectively, see the first term of Eq. 5.9 below. The weight function λ is defined as a twice differentiable sigmoid with values in $[0,1]$:

$$\lambda(\boldsymbol{\theta}_n, \boldsymbol{a}_\ell) = \begin{cases} 0 & \text{if } \widetilde{f}_\ell(\boldsymbol{\theta}_n) \leq a_{\ell,0} \\ 1 & \text{if } \widetilde{f}_\ell(\boldsymbol{\theta}_n) \geq a_{\ell,1} \\ Q\left(\frac{\ln \widetilde{f}_\ell(\boldsymbol{\theta}_n) - \ln a_{\ell,0}}{\ln a_{\ell,1} - \ln a_{\ell,0}}\right) & \text{otherwise} \end{cases}, \tag{5.8}$$

where Q is a polynomial such that $Q(0) = 0$, $Q(0) = 1$ and $Q^{'}(0) = Q^{'}(1) = Q^{''}(0) = Q^{''}(1) = 0$ for λ to be $\mathscr{C}^2$. One of the simplest such polynomials is $Q(u) = u^3(6u^2 - 15u + 10)$. Figure 5.1 illustrates the λ function. The accuracy of this likelihood approximation clearly depends on the choice of the parameter $\boldsymbol{a}_\ell$. Appendix 5.B explains how to optimize $\boldsymbol{a}_\ell$ to maximize the likelihood approximation quality.

Censorship – To take censorship into account, let $\mathbf{C} = (c_{n\ell})_{n,\ell} \in \{0,1\}^{NL}$ be a matrix such that $c_{n\ell} = 1$ for a censored observation, and $c_{n\ell} = 0$ otherwise. Let $F^{(a)}(\cdot|\boldsymbol{\theta}_n)$ and $F^{(m)}(\cdot|\boldsymbol{\theta}_n)$ be the cumulative density function (cdf) of $\pi^{(a)}(\cdot|\boldsymbol{\theta}_n)$ and $\pi^{(m)}(\cdot|\boldsymbol{\theta}_n)$, respectively. The likelihood of censored data involves $F^{(a)}(\omega_{n\ell}|\boldsymbol{\theta}_n)$ and $F^{(m)}(\omega_{n\ell}|\boldsymbol{\theta}_n)$. The proposed likelihood approximation of the model from Eq. 5.1 finally reads

$$\begin{aligned} \widetilde{\pi}(y_{n\ell}|\boldsymbol{\theta}_n, \boldsymbol{a}_\ell) \propto & \left[\pi^{(a)}(y_{n\ell}|\boldsymbol{\theta}_n)^{1-\lambda(\boldsymbol{\theta}_n,\boldsymbol{a}_\ell)}\ \pi^{(m)}(y_{n\ell}|\boldsymbol{\theta}_n)^{\lambda(\boldsymbol{\theta}_n,\boldsymbol{a}_\ell)}\right]^{1-c_{n\ell}} \\ & \times \left[F^{(a)}(\omega_{n\ell}|\boldsymbol{\theta}_n)^{1-\lambda(\boldsymbol{\theta}_n,\boldsymbol{a}_\ell)}\ F^{(m)}(\omega_{n\ell}|\boldsymbol{\theta}_n)^{\lambda(\boldsymbol{\theta}_n,\boldsymbol{a}_\ell)}\right]^{c_{n\ell}}. \end{aligned} \tag{5.9}$$

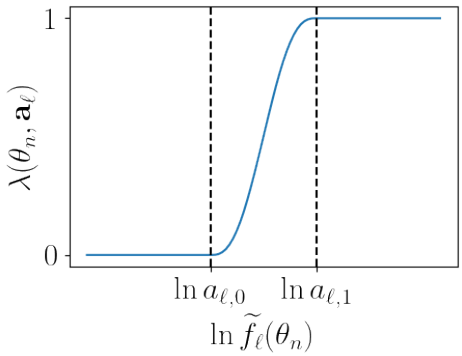


Figure 5.1: Illustration of the λ function from Eq. 5.8.

5.1.2 Prior distribution

In this thesis we consider maps of physical parameters $\boldsymbol{\Theta} \in \mathbb{R}^{N \times D}$, with N the number of pixels in the map, and D the number of physical parameters per pixel. We combine two penalties to build the prior distribution on $\boldsymbol{\Theta}$. The first provides validity intervals on the physical parameters, as is common in ISM studies (see Chapter 3, Section 3.1.3). The second exploits the map structure in $\boldsymbol{\Theta}$ with a spatial regularization.

Smooth approximation of uniform distribution – The first penalty term encodes the validity of $\widetilde{\mathbf{f}}$ on a compact set $\mathcal{C} = [l_1, u_1] \times \cdots \times [l_D, u_D]$. Note that the reduced model may be defined out of $\mathcal{C}$ but will not be considered as valid since it was not trained on such points. The most natural approach would be to use the indicator function $\iota_{\mathcal{C}^N}$ of the set $\mathcal{C}^N$, where $\iota_{\mathcal{C}^N}(\boldsymbol{\Theta}) = 0$ if $\boldsymbol{\Theta} \in \mathcal{C}^N$, and $+\infty$ otherwise. The PMALA kernel to be introduced in Section 5.2.1 requires twice differentiability. The indicator function being non-continuous, it can not be used as is. We use a regularized version. In optimization, one common approach is to replace the indicator function with an exterior penalty function (Nocedal and Wright, 2006). Arguably one of the most common exterior penalty is the quadratic penalty $\theta_{nd} \mapsto [\max(0, \theta_{nd} - u_d, l_d - \theta_{nd})]^2$ (Nocedal and Wright, 2006) which is differentiable and gradient Lipschitz, but not twice differentiable. To obtain a twice differentiable approximation of $\iota_{\mathcal{C}^N}$, we use the quartic penalty:

$$\widetilde{\iota}_{\mathcal{C}^N} : \boldsymbol{\Theta} \mapsto \sum_{n=1}^{N} \sum_{d=1}^{D} \left[\max(0, \theta_{nd} - u_d, l_d - \theta_{nd})\right]^4, \tag{5.10}$$

which is not gradient Lipschitz.

Spatial regularization – The second penalty term favors the spatial regularity of estimations. It is based on a local regularizer $h : \mathbb{R}^N \to \mathbb{R}_+$ applied to each map $\boldsymbol{\Theta}_{\cdot d} = (\theta_{nd})_{1 \leq n \leq N}$, with $d \in [\![1, D]\!]$. The regularizer can be the Euclidean norm of the usual gradient or Laplacian of the component map, with regularization parameter $\tau_d > 0$. In this thesis, for each of the D maps, we use a L_2 norm on the map Laplacian $h(\boldsymbol{\Theta}_{\cdot d}) = \|\Delta \boldsymbol{\Theta}_{\cdot d}\|_2^2$. This regularization function h is twice differentiable. This spatial regularization was already used in the ISM community, e.g., in the ROHSA code (Marchal et al., 2019).

Overall, the resulting prior distribution is given by

$$\pi(\boldsymbol{\Theta}) \propto \exp\left(-\xi\, \widetilde{\iota}_{\mathcal{C}^N}(\boldsymbol{\Theta}) - \sum_{d=1}^{D} \tau_d\, h(\boldsymbol{\Theta}_{\cdot d})\right), \tag{5.11}$$

where $\xi > 0$ is a penalty parameter. The higher ξ, the better the approximation of the indicator function and the higher the penalty out of $\mathcal{C}$.

Hyperparameters – This prior relies on two hyperparameters: $\xi > 0$ weighs the smooth indicator function, and $\boldsymbol{\tau}$ contains the D spatial regularization weights of the D maps to be inferred. To learn the spatial regularization weights from the data as in Galliano (2018) would avoid the tedious usual manual setting. In Appendix 5.A, we present a hierarchical model to learn the prior parameters from the data, along with the reasons why we did not use it. Therefore, in the remainder of this thesis, the prior parameters are set manually and iteratively to obtain a trade-off between physically and visually consistent maps. To avoid underestimating uncertainties and favoring physical consistency, the spatial regularization is set to low values, i.e., $\tau_d \in [1, 10]$ for each physical parameter d. By default, τ_d is set to 1.

5.1.3 Posterior distribution

The posterior distribution combines NL independent likelihoods (Eq. 5.9) and the prior (Eq. 5.11):

$$\pi(\boldsymbol{\Theta}|\mathbf{Y}) \propto \left[\prod_{n=1}^{N}\prod_{\ell=1}^{L} \widetilde{\pi}(y_{n\ell}|\boldsymbol{\theta}_n, \boldsymbol{a}_\ell)\right] \pi(\boldsymbol{\Theta}). \tag{5.12}$$

This posterior distribution is hard to manipulate as is, and requires sampling to derive estimators and credibility intervals. However, drawing samples from this posterior is challenging since it is non-log-concave, potentially multimodal. Besides, no gradient Lipschitz continuity is assumed for the log-posterior. In the following and as in Chapter 2, the negative log-posterior pdf $-\ln \pi(\boldsymbol{\Theta}|\mathbf{Y})$ is denoted $\mathcal{L}(\boldsymbol{\Theta})$.

5.2 Proposed MCMC algorithm

As the forward model spans several decades, the gradient of the negative log-likelihood $\nabla\mathcal{L}$ has a potentially very large Lipschitz constant, if any. Besides, the smooth uniform component of the prior is not gradient Lipschitz. The negative log posterior is therefore not gradient Lipschitz either. In addition, the posterior distribution (Eq. 5.12) is in general non-log-concave, potentially multimodal, which makes the sampling task challenging.

To address these two challenges, a new transition kernel is proposed as a combination of two kernels: PMALA (Xifara et al., 2014) and MTM (Liu et al., 2021). PMALA tackles the regularity issue to efficiently explore the neighborhood of a local mode, whereas MTM permits jumps between modes.

5.2.1 PMALA transition kernel

MALA and HMC rely on a step size inversely proportional to the Lipschitz constant of $\nabla\mathcal{L}$, if it exists. Here the forward model $\widetilde{\mathbf{f}}$ covers several decades so that this Lipschitz constant is potentially very large or even infinite. Therefore, MALA and HMC will typically fail to efficiently explore the posterior (Eq. 5.12). To accelerate the exploration, a preconditioned MALA equipped with RMSProp (Tieleman and Hinton, 2012) is introduced to perform larger steps. To simplify notation, we temporarily use the vector version of $\boldsymbol{\Theta}$ in lexicographic order so that $\boldsymbol{\Theta} \in \mathbb{R}^{ND}$.

A transition kernel that handles such situations relies on extensions of HMC and MALA to Riemannian manifolds, introduced in Chapter 2 (Section 2.2.2.5). The Riemannian manifold MALA version was improved in Xifara et al. (2014), resulting in the so-called preconditioned Metropolis adjusted Langevin algorithm (PMALA) kernel. It permits to exploit local information geometry thanks to a position-dependent preconditioner. We propose to use the RMSProp preconditioner (Tieleman and Hinton, 2012) that was initially defined in the deep learning literature for fast neural networks training. It adaptively estimates a local variance of the gradient $\nabla\mathcal{L}$ by keeping memory of former proposals $\boldsymbol{\Theta}_c^{(t)}$. At each step t, it updates a surrogate gradient variance

vector $\mathbf{v}^{(t)} \in \mathbb{R}^{ND}$ such that for all $i \in [\![1, ND]\!]$,

$$v_i^{(t)} = a v_i^{(t-1)} + (1-a)\left[\frac{\partial \mathcal{L}}{\partial \theta_i}\left(\boldsymbol{\Theta}_c^{(t)}\right)\right]^2 = (1-a)\sum_{j=1}^{t} a^{t-j}\left[\frac{\partial \mathcal{L}}{\partial \theta_i}\left(\boldsymbol{\Theta}_c^{(t-j)}\right)\right]^2, \tag{5.13}$$

where $a \in]0,1[$ is an exponential decay rate. Note that the variance vector $\mathbf{v}^{(t)}$ relies on candidates $\boldsymbol{\Theta}_c^{(t)}$ instead of iterates $\boldsymbol{\Theta}^{(t)}$: candidates might not be kept in the Markov chain, but they still contain important information about the shape of the distribution. The RMSProp preconditioner is defined as (Tieleman and Hinton, 2012)

$$\mathbf{G}^{(t)} = \operatorname{diag}\left(\frac{1}{\epsilon + \sqrt{\mathbf{v}^{(t)}}}\right) \in \mathbb{R}^{ND \times ND}, \tag{5.14}$$

with ϵ a small damping parameter. This preconditioner has already been used in a MCMC context (Li et al., 2016) within an approximate sampler. The goal in Li et al. (2016) was to sample from a distribution defined over the parameters of a neural network trained on a large dataset. Accept or reject steps were omitted as they would have required expensive computations on the full dataset. Additionally, the discretization of the Langevin diffusion process equipped with a position-dependent preconditioner comes with an additional drift term (Xifara et al., 2014) that was neglected in Li et al. (2016). We correct these two approximations to sample exactly from Eq. 5.12.

Following Xifara et al. (2014), the proposal distribution corresponding to PMALA with the RMSProp preconditioner is the Gaussian distribution:

$$q\left(\boldsymbol{\Theta}_c^{(t)} | \boldsymbol{\Theta}^{(t-1)}\right) = \mathcal{N}\left(\boldsymbol{\Theta}_c^{(t)} | \boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}^{(t)}\right) \tag{5.15}$$

with

$$\begin{cases} \boldsymbol{\mu}^{(t)} = \boldsymbol{\Theta}^{(t-1)} - \eta\, \mathbf{G}^{(t-1)} \nabla \mathcal{L}\left(\boldsymbol{\Theta}^{(t-1)}\right) + 2\eta\, \boldsymbol{\gamma}^{(t-1)}, \\ \boldsymbol{\Lambda}^{(t)} = 2\eta\, \mathbf{G}^{(t-1)}, \end{cases} \tag{5.16}$$

where η is a step size and $\boldsymbol{\gamma}^{(t-1)}$ is the additional drift term due to the position-dependent preconditioner (Xifara et al., 2014). In full generality, for all $i \in [\![1, ND]\!]$,

$$\gamma_i^{(t-1)} = \frac{1}{2}\sum_{j=1}^{ND} \frac{\partial G_{ij}^{(t-1)}}{\partial \theta_j^{(t-1)}}. \tag{5.17}$$

However, the RMSProp preconditioner is diagonal so that the sum in Eq. 5.17 reduces to the $j = i$ term only. Note that $\boldsymbol{\gamma}^{(t-1)}$ is defined from a differentiation with respect to iterate $\boldsymbol{\Theta}^{(t-1)}$ while the variance vector $\mathbf{v}$ in Eq. 5.13 is defined from candidates. Since all iterates start as candidates, let $j^{(t)}$ be the number of iterations since last accept: $j^{(t)} = \min\{j \geq 0 | \boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}_c^{(t-j)}\}$. The correction terms $\gamma_i^{(t-1)}$ are then given by

$$\gamma_i^{(t-1)} = -\frac{(1-a)a^{j^{(t-1)}}}{2\sqrt{v_i^{(t-1)}}\left(\epsilon + \sqrt{v_i^{(t-1)}}\right)^2}\left(\frac{\partial \mathcal{L}}{\partial \theta_i} \cdot \frac{\partial^2 \mathcal{L}}{\partial \theta_i^2}\right)\left(\boldsymbol{\Theta}^{(t-1)}\right). \tag{5.18}$$

We now can compute the four components involved in the accept-reject probability $\rho^{(t)}$ (Eq. 2.34) except $q(\boldsymbol{\Theta}^{(t-1)} | \boldsymbol{\Theta}_c^{(t)}) = \mathcal{N}(\boldsymbol{\Theta}^{(t-1)} | \boldsymbol{\mu}_c^{(t)}, \boldsymbol{\Lambda}_c^{(t)})$. To compute this last term, one needs to update the variance $\mathbf{v}^{(t)}$ and preconditioner $\mathbf{G}^{(t)}$ and evaluate the candidate additional drift term $\boldsymbol{\gamma}_c^{(t)}$. By definition, $j^{(t)} = 0$ for candidates, so for all $i \in [\![1, ND]\!]$,

$$\gamma_{c,i}^{(t)} = -\frac{(1-a)}{2\sqrt{v_i^{(t)}}\left(\epsilon + \sqrt{v_i^{(t)}}\right)^2}\left(\frac{\partial \mathcal{L}}{\partial \theta_i} \cdot \frac{\partial^2 \mathcal{L}}{\partial \theta_i^2}\right)\left(\boldsymbol{\Theta}^{(t-1)}\right). \tag{5.19}$$

The parameters $\boldsymbol{\mu}_c^{(t)}, \boldsymbol{\Lambda}_c^{(t)}$ are thus given by

$$\begin{cases} \boldsymbol{\mu}_c^{(t)} = \boldsymbol{\Theta}_c^{(t)} - \eta\, \mathbf{G}^{(t)} \nabla \mathcal{L}\left(\boldsymbol{\Theta}_c^{(t)}\right) + 2\eta\, \boldsymbol{\gamma}_c^{(t)}, \\ \boldsymbol{\Lambda}_c^{(t)} = 2\eta\, \mathbf{G}^{(t)}. \end{cases} \tag{5.20}$$

Algorithm 5.1 describes the proposed PMALA kernel with RMSProp preconditioner. It relies on three scalar parameters: a damping parameter ϵ, an exponential decay rate a and a step size η. The first two are generally set to $\epsilon = 10^{-5}$ and $a = 0.99$ (Li et al., 2016). The step size is chosen empirically. MALA achieves optimal convergence rates with an acceptance rate equal to 0.574 when the components of $\boldsymbol{\Theta}$ are independent (Robert and Casella, 2004). Despite the interdependencies in the posterior distribution, we also set η to obtain an average acceptance rate close to 0.574, which yields good results in practice.

Algorithm 5.1: PMALA kernel $\mathcal{K}_1$ at step t

Input: $\boldsymbol{\Theta}^{(t-1)}$, $\mathbf{v}^{(t-1)}$, $j^{(t-1)}$
Output: $\boldsymbol{\Theta}^{(t)}$, $\mathbf{v}^{(t)}$, $j^{(t)}$

// Propose candidate
$\mathbf{G}^{(t-1)}$ and $\boldsymbol{\gamma}^{(t-1)}$ // using Eq. 5.14 and Eq. 5.18
$\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Lambda}^{(t)}$ // using Eq. 5.16
$\boldsymbol{\Theta}_c^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}^{(t)})$
// Accept or reject
$\mathbf{v}^{(t)}$, $\mathbf{G}^{(t)}$ and $\boldsymbol{\gamma}_c^{(t)}$ // using Eq. 5.13, Eq. 5.14, and Eq. 5.19
$\boldsymbol{\mu}_c^{(t)}$, $\boldsymbol{\Lambda}_c^{(t)}$ and $\rho^{(t)}$ // using Eq. 5.20, and Eq. 2.34
Draw $\zeta \sim \text{Unif}\,(0, 1)$
if $\zeta \leq \rho^{(t)}$ **then** $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}_c^{(t)}$, $j^{(t)} = 0$
else $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}^{(t-1)}$, $j^{(t)} = j^{(t-1)} + 1$

Accounting for the correction term $\boldsymbol{\gamma}$ is necessary in stochastic gradient MCMC (SG-MCMC) as it does not apply accept-reject steps. Alternatively, disregarding it, as in Li et al. (2016), leads to a controlled error on estimations. In our case, as we do apply a MH accept-reject step, computing the correction term is not mandatory to obtain an MCMC algorithm that converges to the correct stationary distribution $\pi(\boldsymbol{\Theta}|\mathbf{Y})$. However, we keep this term to obtain a correct discretization of the associated Langevin diffusion process. As stated in Chapter 2 (Section 2.2.2.4), better discretization of the Langevin process often leads to better candidates.

5.2.2 MTM transition kernel

The non-log-concavity and potential multimodality of the posterior (Eq. 5.12) is the second major difficulty to be addressed. As explained in Chapter 2, samplers such as MH, MALA, HMC or even PMALA fail to explore the full distribution when modes are far away: they get stuck in one. In Section 2.A, we reviewed existing algorithms dedicated to multimodal distributions. None of these methods can efficiently address the very high dimensionality $\mathcal{O}(10^4)$ to be encountered in some of the applications considered in this thesis.

We define a kernel that can escape a local mode and explore other ones without any knowledge about the number, positions or variances of the modes. To do so, we propose to exploit two particularities of the considered posterior, namely its map structure and the speed of evaluation of its forward model - and thus of its full pdf. Instead of sampling the whole vector $\boldsymbol{\Theta} \in \mathbb{R}^{ND}$ at once, it uses a Metropolis-within-Gibbs sampler to decompose it into N individual $\boldsymbol{\theta}_n$ (see Chapter 2, Section 2.2.2.6). For each conditional distribution, it harnesses an independent multiple-try Metropolis (I-MTM) approach (Liu et al., 2021; Martino, 2018) (see Chapter 2, Section 2.2.2.7).

This method generates $K \geq 1$ candidates $(\boldsymbol{\theta}_n^{(k)})_{k=1}^K$ independently of $\boldsymbol{\theta}_n^{(t-1)}$. This divide-to-conquer approach permits to consider N conditional distributions $\pi(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\Theta}_{\setminus n}^{(t-1+(n-1)/N)})$ of small dimension, where $\boldsymbol{\Theta}_{\setminus n}^{(t-1+(n-1)/N)} = (\boldsymbol{\theta}_1^{(t)}, \dots, \boldsymbol{\theta}_{n-1}^{(t)}, \boldsymbol{\theta}_{n+1}^{(t-1)}, \dots, \boldsymbol{\theta}_N^{(t-1)})$. Candidates are sampled from a proposal distribution $q(\boldsymbol{\theta}_n|\boldsymbol{\Theta}_{\setminus n}^{(t-1)})$ that should be permissive enough to generate candidates in all modes of $\pi(\boldsymbol{\theta}_n|\mathbf{y}_n, \boldsymbol{\Theta}_{\setminus n}^{(t-1)})$. Then, using the importance weight function w from Eq. 2.42 adapted to this case,

$$w(\boldsymbol{\theta}_n^{(k)}) = \frac{\pi(\boldsymbol{\theta}_n^{(k)}|\mathbf{y}_n, \boldsymbol{\Theta}_{\setminus n}^{(t-1)})}{q(\boldsymbol{\theta}_n^{(k)}|\boldsymbol{\Theta}_{\setminus n}^{(t-1)})}, \tag{5.21}$$

one candidate is selected using a categorical distribution with selection probability w_k for candidate k

$$w_k = \frac{w(\boldsymbol{\theta}_n^{(k)})}{\sum_{j=1}^K w(\boldsymbol{\theta}_n^{(j)})}. \tag{5.22}$$

The MH step is then performed with the selected candidate i and the following generalized acceptance probability, adapted from Eq. 2.44,

$$\widetilde{\rho}_n^{(t)} = \min\left(1, \frac{w(\boldsymbol{\theta}_n^{(i)}) + \sum_{j=1, j\neq i}^K w(\boldsymbol{\theta}_n^{(j)})}{w(\boldsymbol{\theta}_n^{(t-1)}) + \sum_{j=1, j\neq i}^K w(\boldsymbol{\theta}_n^{(j)})}\right). \tag{5.23}$$

Algorithm 5.2 summarizes the proposed Gibbs and MTM kernel. Note that due to the Gibbs approach, it updates one component at a time and returns the result of all updates. A succession of intermediate updates $(\boldsymbol{\Theta}^{(t-1+n/N)})_{n=1}^N$ is therefore introduced (see Algorithm 2.3). This transition kernel relies on the choice of the proposal distribution q and on the number of candidates K generated at each step. This parameter is chosen as a trade-off between computational intensity and average acceptance probability: the higher K, the higher the acceptance probability, the mixing capability but also the computational cost.

Algorithm 5.2: MTM kernel $\mathcal{K}_2$ at step t

Input: $\boldsymbol{\Theta}^{(t-1)}$
Output: $\boldsymbol{\Theta}^{(t)}$

for $n = 1$ **to** N **do**
  `// Propose candidates, select one`
  $\boldsymbol{\theta}_n^{(k)} \sim q\left(\boldsymbol{\theta}_n \middle| \boldsymbol{\Theta}_{\setminus n}^{\left(t-1+\frac{n-1}{N}\right)}\right)$ for $k = 1$ to K
  $w(\boldsymbol{\theta}_n^{(k)})$ for $k = 1$ to K `// using Eq. 5.21`
  w_k for $k = 1$ to K `// using Eq. 5.22`
  $i \sim \text{Cat}(w_1, \cdots, w_K)$
  `// Accept or reject`
  $\widetilde{\rho}_n^{(t)}$ `// using Eq. 5.23`
  Draw $\zeta_n \sim \text{Unif}(0, 1)$
  if $\zeta_n \leq \widetilde{\rho}_n^{(t)}$ **then**
    $\boldsymbol{\theta}_n^{\left(t-1+\frac{n}{N}\right)} = \boldsymbol{\theta}_n^{(i)}$, $\boldsymbol{\Theta}_{\setminus n}^{\left(t-1+\frac{n}{N}\right)} = \boldsymbol{\Theta}_{\setminus n}^{\left(t-1+\frac{n-1}{N}\right)}$
  else $\boldsymbol{\Theta}^{\left(t-1+\frac{n}{N}\right)} = \boldsymbol{\Theta}^{\left(t-1+\frac{n-1}{N}\right)}$

In image inverse problems, many common spatial priors are based on a local operator such as the image gradient or Laplacian. In such cases, many components $\boldsymbol{\theta}_n$ are conditionally independent. They can be sampled in parallel using a Chromatic Gibbs sampler (Gonzalez et al., 2011), which can significantly speed up computations. Note that this Gibbs sampling and chromatic Gibbs sampling can also be adopted in the PMALA kernel to farther accelerate local explorations.

5.2.3 Proposed sampler and implementation details

To combine a good local exploration of modes as well as jumps between modes, the proposed kernel mixes the PMALA and MTM transition kernels above. At every step t, the MTM kernel is selected with probability p_{MTM}, and the PMALA kernel with probability $1 - p_{\text{MTM}}$. Since the MTM kernel divides the parameter space in N D-dimensional subspaces, the PMALA global integer $j^{(t)} \geq 0$ is replaced by a vector $\boldsymbol{j}^{(t)} \in \mathbb{N}^N$, where $j_n^{(t)}$ counts the number of steps since last acceptance for component $\boldsymbol{\theta}_n$. When a component $\boldsymbol{\theta}_n$ is accepted by the MTM kernel, the counter j_n is reset to 0 and the variance component $\boldsymbol{v}_n \in \mathbb{R}^D$ is updated as in Eq. 5.13 with $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_n}(\boldsymbol{\Theta}^{(t)})$.

Algorithm 5.3 reports the complete proposed sampler. Similarly to RDMC and WHMC, the proposed sampler mixes a kernel dedicated to local exploration – PMALA – and another to jump between modes – MTM. The decomposition of the parameter space into N D-dimensional subspaces makes the sampler much simpler than previous approaches. It will perform well in structured problems that allow such decomposition, e.g., images and graphs, and poorly in high-dimensional problems that do not, e.g., Gaussian Mixtures over the full space. In particular, resorting to an Gibbs and MTM strategy leads to fast convergence to the high probability regions. Therefore, the burn-in phase is shorter than with usual MCMC algorithms. Using the Gibbs approach in the PMALA kernel also accelerates convergence and local exploration.

An optimization algorithm can be obtained from this sampler. It combines preconditioned gradient descent and simulated annealing-like component-wise updates that can escape from local minima.

Algorithm 5.3: Proposed sampler: PMALA and MTM

Input: number of iterations T_{MC}, starting point $\boldsymbol{\Theta}^{(0)}$
Output: Markov chain $\{\boldsymbol{\Theta}^{(t)}\}_{t=1}^{T_{\text{MC}}}$

Initialize $v_{nd}^{(0)} = \left[\frac{\partial \mathcal{L}}{\partial \theta_{nd}}(\boldsymbol{\Theta}^{(0)})\right]^2$ for all n and d
Initialize $\boldsymbol{j}^{(0)} = \mathbf{0}_N$
for $t = 1$ **to** T_{MC} **do**
  Draw $\zeta \sim \text{Unif}(0, 1)$
  if $\zeta > p_{MTM}$ **then** `// PMALA kernel (Algo. 5.1)`
    $\boldsymbol{\Theta}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{j}^{(t)} = \mathcal{K}_1(\boldsymbol{\Theta}^{(t-1)}, \mathbf{v}^{(t-1)}, \boldsymbol{j}^{(t-1)})$
  else `// MTM kernel (Algo. 5.2)`
    $\boldsymbol{\Theta}^{(t)} = \mathcal{K}_2(\boldsymbol{\Theta}^{(t-1)})$
    `// Update PMALA parameters`
    for $n = 1$ **to** N **do**
      if *candidate for* $\boldsymbol{\theta}_n$ *was accepted* **then** $\forall d$,
        $v_{nd}^{(t)} = a v_{nd}^{(t-1)} + (1 - a)\left[\frac{\partial \mathcal{L}}{\partial \theta_{nd}}(\boldsymbol{\Theta}^{(t)})\right]^2$,
        $j_n^{(t)} = 0$, $\mathbf{v}_{\setminus n}^{(t)} = \mathbf{v}_{\setminus n}^{(t-1)}$, $\boldsymbol{j}_{\setminus n}^{(t)} = \boldsymbol{j}_{\setminus n}^{(t-1)}$

Regarding theoretical properties, the PMALA kernel satisfies the detailed balance property – from Robert and Casella (2004, theorem 7.2) – and produces ergodic Markov chains – from Robert and Casella (2004, corollary 7.5). The proposed MTM kernel is a Metropolis-within-Gibbs algorithm with propositions independent to the current location and with multiple candidates K. In the particular case where $K = 1$, it satisfies the detailed balance property and produces uniformly ergodic Markov Chains – from Jones et al. (2014, theorem 7). Using $K > 1$ candidates in a MTM framework maintains detailed balance and ergodicity (Martino, 2018). As a mixture of kernels having the same stationary distribution, the proposed kernel also admits the posterior as

a stationary distribution – from Robert and Casella (2004, chapter 10). As the MTM kernel produces uniformly ergodic Markov chains, so does the proposed mixture kernel – from Robert and Casella (2004, proposition 10.20). These results of convergence towards the posterior are mostly asymptotic and also hold for simpler algorithms such as RWMH (Roberts and Tweedie, 1996). A comparative theoretical study of non-asymptotic properties that could demonstrate a faster convergence of the proposed sampler is beyond the scope of this manuscript. However, Section 5.2.4 presents empirical results showing that the proposed sampler yields state-of-the-art performance on multimodal distributions in low-dimensional settings. As reported in Section 5.4, the proposed sampler also yields state-of-the-art performance on higher dimensional applications with relevant low-dimensional conditional distributions.

5.2.4 Illustration: 2D Gaussian mixture model

The proposed sampler is applied to a two-dimensional Gaussian mixture model (GMM) restricted to the square $\mathcal{C} = [-15, 15]^2$. This simple multimodal distribution, shown on Figure 5.2 (top left), is set to contain 15 modes $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. It will demonstrate the ability of the proposed sampler to jump between modes. For simplicity, all the modes have an equal weight in the mixture

$$\pi(\boldsymbol{\theta}) \propto \left[\sum_{i=1}^{15} \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right] \exp\left(-\xi\, \widetilde{\imath}_{\mathcal{C}}(\boldsymbol{\theta})\right), \tag{5.24}$$

with $\xi = 10^4$. No natural structure decomposition exists for a GMM since each observation consists of $N = 1$ point only in dimension $D = 2$. A Markov chain composed of $T_{\text{MC}} = 10\,000$ samples is considered, including $T_{\text{BI}} = 100$ burn-in samples. To illustrate the role of each of the two kernels in the proposed sampler, two different values are considered for the probability of selecting the MTM kernel: $p_{\text{MTM}} = 0.1$ or $p_{\text{MTM}} = 0.9$. The number of candidates of the MTM kernel is set to $K = 50$, and the proposal distribution q is the smooth uniform prior on $\mathcal{C}$ – see Appendix 5.C. The MTM candidates weights $w(\boldsymbol{\theta}_n^{(k)})$ in Eq. 5.21 are then equal to the likelihood term, i.e., the sum of Gaussian pdfs. The default values $a = 0.99$ and $\epsilon = 10^{-5}$ are considered for the exponential decay and damping factor of the PMALA kernel (Li et al., 2016), and its step size is set to $\eta = 0.25$. The proposed approach is compared to the state-of-the-art wormhole Hamiltonian Monte Carlo (WHMC) (Lan et al., 2014), using the same number of samples. Note that WHMC needs the prior knowledge of mode positions $(\boldsymbol{\mu}_i)_{1\leq i\leq 15}$, while the proposed kernel does not.

The three sampling algorithms widespread in astrophysics introduced in Chapter 2 (Appendix 2.A) are also applied. The affine-invariant sampler (Goodman and Weare, 2010) is run with 4 parallel chains, using the EMCEE implementation (Foreman-Mackey et al., 2013). The sequential MC (SMC) algorithm (Del Moral et al., 2006) is run with the state-of-the-art PyMC code[1] (Abril-Pla et al., 2023). The MultiNest algorithm (Feroz and Hobson, 2008) is run with its Python implementation PyMultiNest (Buchner, 2016a). All three algorithms are run with their respective default parameters. Note that these three samplers are not MCMC algorithms.

Figure 5.2 shows the 2D histograms obtained with the six samplers. The affine-invariant sampler is the only sampler that fails to explore all the modes. The five others visit all the modes, approximately give them equal weights in the histogramm, and their local dispersion obeys the covariance structures equally well. Table 5.1 compares their bias[2] and effective sample size (ESS) for the MCMC algorithms – see Chapter 2 (Section 2.2.2) for the definition of the ESS. When $p_{\text{MTM}} = 0.9$, the proposed sampler achieves the second lowest bias and the highest ESS. In particular, it achieves better performances than WHMC, despite the absence of information about

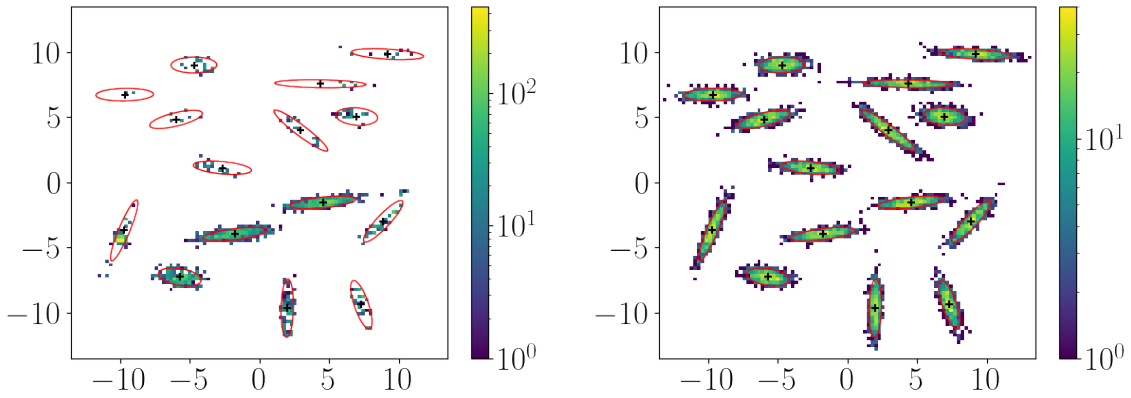
[1] https://www.pymc.io/welcome.html

[2] To compute the bias, we consider that $\mathbb{E}[\boldsymbol{\theta}] \simeq \frac{1}{15}\sum_{i=1}^{15} \boldsymbol{\mu}_i$, which is inexact due to the the smooth indicator term. However, the probability mass of the Gaussian mixture out of the validity intervals is numerically negligible. Therefore, the error due to this approximation is negligible as well.

the position of the modes $\boldsymbol{\mu}_i$. The high ESS values result from the 85% acceptance rate of the MTM kernel for $K = 50$. However, the MTM kernel with a fixed number of candidates K would not scale up to much higher dimensions. The probability to jump between modes is proportional to the volume of the high probability regions compared to the volume of $\mathcal{C}$, and thus decreases exponentially with the dimension of the problem. The proposed sampler would therefore fail to reach isolated modes in a high-dimensional GMM, whereas WHMC would succeed to do so by exploiting its additional information about the modes. However, the proposed approach focuses on scenarios where the parameter space can be partitioned into a collection of N subspaces of limited dimension D, typically $D \lesssim 10$. The MTM kernel thus remains out of reach from the curse of dimension thanks to the structure of the problem. As in this simple GMM example, the proposed sampler can then outperform WHMC, even without any prior information on the modes of a multimodal distribution.

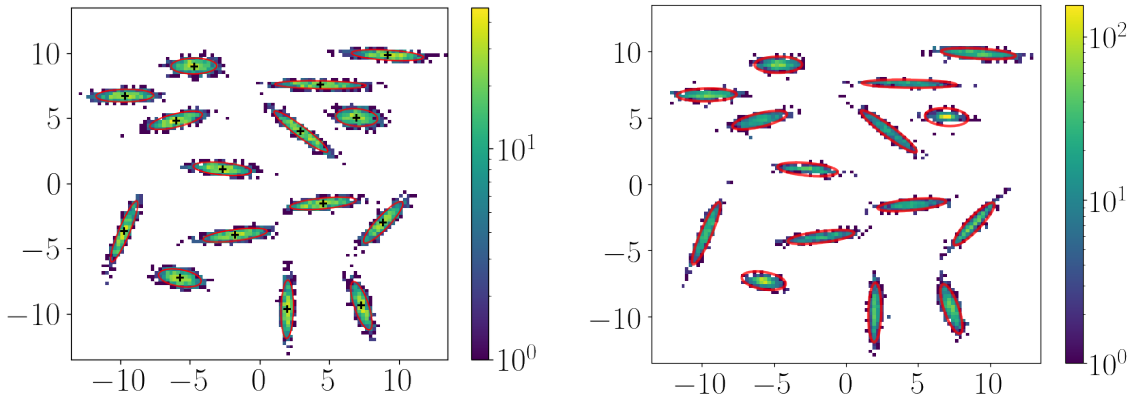
Table 5.1: Samplers comparison on 2D Gaussian mixture model.

Algorithm	Bias $\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{MMSE}} - \mathbb{E}[\boldsymbol{\theta}]\right\|$	ESS θ_1	ESS θ_2
Affine-invariant sampler (Goodman and Weare, 2010)	$4.32 \cdot 10^{0}$	–	–
sequential MC (Del Moral et al., 2006)	$1.09 \cdot 10^{-1}$	–	–
MultiNest (Feroz and Hobson, 2008)	$\mathbf{3.22 \cdot 10^{-2}}$	–	–
WHMC (Lan et al., 2014)	$1.28 \cdot 10^{-1}$	2 753	2 993
Proposed, $p_{\mathsf{MTM}} = 0.1$	$7.02 \cdot 10^{-1}$	395	444
Proposed, $p_{\mathsf{MTM}} = 0.9$	$4.61 \cdot 10^{-2}$	**6 157**	**5 780**



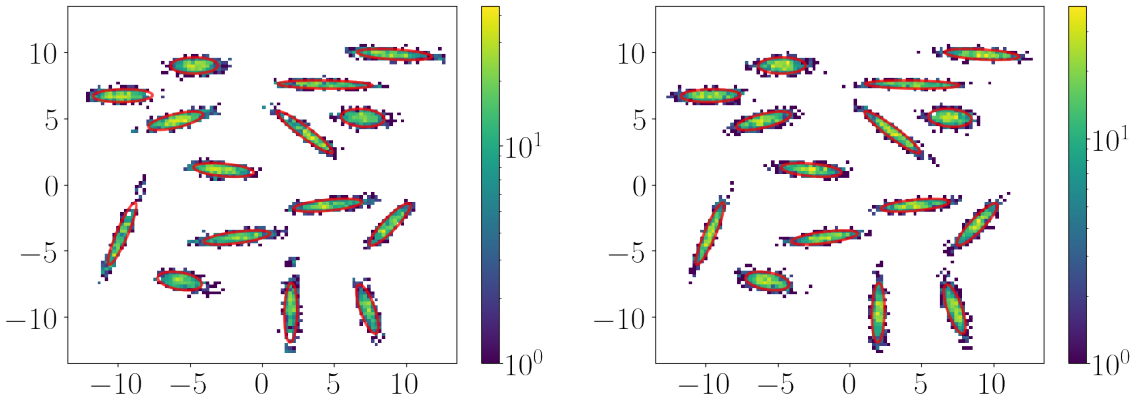
(a) Affine-invariant sampler

(b) SMC



(c) MultiNest

(d) WHMC



(e) Proposed kernel with $p_{\mathrm{MTM}} = 0.1$

(f) Proposed kernel with $p_{\mathrm{MTM}} = 0.9$

Figure 5.2: Comparison of samplers on a 2D Gaussian mixture model. The red ellipses show the probability level at 2σ. All histograms are in logarithmic norm.

5.3 Model checking using a predictive posterior p-value

This section presents a method that checks whether the observation model $\mathcal{M}$ can reproduce the observation $\mathbf{Y}$. As in Chapter 2 (Section 2.1), we consider that the model $\mathcal{M}$ gathers the forward model and the noise model. In astrophysics applications, the forward model is the emulator of the Meudon PDR code defined in Chapter 4. The noise model is described in Chapter 3 (Section 3.4).

To perform this model checking, we resort to the predictive posterior p-value introduced in Chapter 2 (Section 2.3.2). As we showed in Chapter 3 (Section 3.3.2), this approach was already used in astrophysics. We improve it in two ways. First, we set the discrepancy measure to the negative log-likelihood for generality. As the resulting predictive posterior p-value is intractable, we approximate it with a Monte Carlo (MC) estimator. Second, we control the associated approximation error to avoid making wrong decisions because of insufficient number of iterates in our MC estimator.

5.3.1 Definition of selected p-value

For inverse problems relying on a non-linear forward model or on a posterior distribution instead of a point estimate $\widehat{\boldsymbol{\Theta}}$, there exists no general test statistic T that leads to an exact evaluation of the p-value. The choice of the test statistic should then enforce consistent behavior between the observation model and the p-value. We propose to use the negative log likelihood function as a test statistic,

$$T(\widetilde{\mathbf{Y}}, \boldsymbol{\Theta}) = -\ln \pi\left(\widetilde{\mathbf{Y}} | \boldsymbol{\Theta}, \mathcal{M}\right), \tag{5.25}$$

where $\widetilde{\mathbf{Y}}$ are reproductions of observations following the observation model. In addition, as the observation model handles pixels independently, we propose to evaluate one p-value per pixel. Overall, a map $\mathbf{p} = (p_n)_{n=1}^N$ of N p-values is evaluated on the marginal posterior predictive distributions $\widetilde{\mathbf{y}}_n|\mathbf{Y}$. The p-value p_n corresponds to the measure of the set

$$I_n = \{(\widetilde{\mathbf{y}}_n, \boldsymbol{\Theta}) \mid \pi\left(\widetilde{\mathbf{y}}_n | \boldsymbol{\Theta}, \mathcal{M}\right) \leq \pi\left(\mathbf{y}_n | \boldsymbol{\Theta}, \mathcal{M}\right)\}, \tag{5.26}$$

i.e.,

$$p_n = \int \mathbb{1}_{I_n}(\widetilde{\mathbf{y}}_n, \boldsymbol{\Theta})\, \pi(\widetilde{\mathbf{y}}_n | \boldsymbol{\Theta}, \mathcal{M})\, \pi(\boldsymbol{\Theta} | \mathbf{Y}, \mathcal{M})\, d\boldsymbol{\Theta} d\widetilde{\mathbf{y}}_n. \tag{5.27}$$

Small p-values p_n enable detecting anomalies in observed maps, e.g., regions that are not well modeled by a PDR. This definition is general, as it does not rely on restraining assumptions on the observation model. The only two necessary assumptions are 1) that the observation model can be used to generate observations from a given $\boldsymbol{\Theta}$ and 2) that the likelihood function can be evaluated.

5.3.2 p-value approximation and associated uncertainties

As detailed in Chapter 2 (Section 2.3.2), for general models the p-value is evaluated with a Monte Carlo (MC) estimator $\widehat{p}$ (Eq. 2.57). The map of p-values $\widehat{\boldsymbol{p}}^{(t)} = (\widehat{p}_n^{(t)})_{n=1}^N$ can be evaluated after the burn-in phase of size $T_{\mathsf{BI}} \geq 0$:

$$\widehat{p}_n^{(t)} = \frac{1}{t - T_{\mathsf{BI}}} \sum_{\tau=T_{\mathsf{BI}}+1}^{t} \mathbb{1}_{I_n}\left(\widetilde{\mathbf{y}}_n^{(\tau)}, \boldsymbol{\Theta}^{(\tau)}\right). \tag{5.28}$$

Algorithm 5.4 details how this estimator can be effortlessly included in an MCMC algorithm. In case of a point estimator $\widehat{\boldsymbol{\Theta}}$, the Markov chain is replaced by $\boldsymbol{\Theta}^{(0)} = \ldots = \boldsymbol{\Theta}^{(T_{\mathsf{MC}})} = \widehat{\boldsymbol{\Theta}}$, and $T_{\mathsf{BI}} = 0$.

Accounting for the uncertainty inherent to a Monte Carlo estimation of the p-values makes hypothesis testing more robust. The probability of null hypothesis rejection with confidence level

Algorithm 5.4: MCMC sampling with p-value computation

Input: Starting point $\boldsymbol{\Theta}^{(T_{\text{BI}})}$, sampling kernel $\mathcal{K}$, numbers of iterations T_{BI} et T_{MC}

Initialization: map of p-values $\widehat{\boldsymbol{p}}^{(T_{\text{BI}})} = \mathbf{0}_N$, $\boldsymbol{u}^{(T_{\text{BI}})} = \mathbf{0}_N$

for $t = T_{BI} + 1, \dots, T_{MC}$ `// After burn-in phase`

do

$\boldsymbol{\Theta}^{(t)}, \left(\pi(\boldsymbol{y}_n|\boldsymbol{\Theta}^{(t)}, \mathcal{M})\right)_{n=1}^{N} = \mathcal{K}(\boldsymbol{\Theta}^{(t-1)})$ `// Sample from` $\mathcal{K}$

$\widetilde{\mathbf{Y}}^{(t)} \sim \pi\left(\cdot|\boldsymbol{\Theta}^{(t)}, \mathcal{M}\right)$ `// generate observation reproduction`

$u_n^{(t)} = u_n^{(t-1)} + \mathbb{1}_{I_n}\left(\widetilde{\mathbf{y}}_n^{(t)}, \boldsymbol{\Theta}^{(t)}\right)$ for $n \in [\![1, N]\!]$ `// update` p`-values`

$\widehat{\boldsymbol{p}}^{(T_{\text{MC}})} = \frac{1}{T_{\text{MC}} - T_{\text{BI}}} \boldsymbol{u}^{(T_{\text{MC}})}$

Output: Markov chain $\{\boldsymbol{\Theta}^{(t)}\}_{t=T_{\text{BI}}+1}^{T_{\text{MC}}}$, map of p-values $\widehat{\boldsymbol{p}}^{(T_{\text{MC}})}$

α then corresponds to the cdf of p_n evaluated on α, $\mathbb{P}[p_n \leq \alpha]$. Using a threshold $\delta \in [0, 0.5]$, $\mathbb{P}[p_n \leq \alpha] \in [\delta, 1-\delta]$ implies that more samples $(\boldsymbol{\Theta}^{(t)}, \widetilde{\mathbf{y}}_n^{(t)})$ are necessary to make a decision with the desired confidence level. The model $\mathcal{M}$ is rejected for pixel n when $\mathbb{P}[p_n^{(t)} \leq \alpha] > 1-\delta$, and is not rejected when $\mathbb{P}[p_n^{(t)} \leq \alpha] < \delta$. In other words,

$$\begin{cases} \mathbb{P}[p_n^{(t)} \leq \alpha] > 1-\delta \implies \text{model } \mathcal{M} \text{ is rejected for pixel } n \text{ with confidence level } \alpha. \\ \mathbb{P}[p_n^{(t)} \leq \alpha] < \delta \implies \text{model } \mathcal{M} \text{ is not rejected for pixel } n \text{ with confidence level } \alpha. \\ \mathbb{P}[p_n^{(t)} \leq \alpha] \in [\delta,\ 1-\delta] \implies \text{the sample size is insufficient to make a decision.} \end{cases} \tag{5.29}$$

In general, evaluating the cdf of the p-value p_n would require an additional MCMC layer. We avoid this additional step and maintain a reasonable cost for this test. First, note from Eq. 5.27 that $\mathbb{1}_{I_n}(\widetilde{\mathbf{y}}_n, \boldsymbol{\Theta})$ is a binary random variable that equals 1 with probability p_n and 0 with probability $1-p_n$. In other words, it follows a Bernoulli distribution of parameter p_n. A natural prior distribution for p_n is a uniform distribution on $[0,1]$, as p_n is a probability. This is a special case of a Beta distribution, Beta$(1,1)$, which is the conjugate prior of the Bernoulli distribution (Gelman et al., 2015, chapter 2, section 2.4). The uncertainty on p_n thus can be reasonably modeled by an a posteriori Beta distribution with closed-form parameters

$$p_n^{(t)} \sim \text{Beta}\left(1 + N_n^{(t)}\, \widehat{p}_n^{(t)}, 1 + N_n^{(t)}\,(1 - \widehat{p}_n^{(t)})\right), \tag{5.30}$$

for any $t > T_{\text{BI}}$, with $N_n^{(t)}$ the number of independent samples. When the posterior on $\boldsymbol{\Theta}$ is reduced to a Dirac on an estimator $\widehat{\boldsymbol{\Theta}}$, then all samples are independent and $N_n^{(t)} = t - T_{\text{BI}}$. When a Markov chain is generated to sample from the posterior distribution $\pi(\boldsymbol{\Theta}|\mathbf{Y}, \mathcal{M})$, the samples $\boldsymbol{\Theta}^{(t)}$ are correlated. To compensate this correlation and avoid underestimating uncertainties, we set $N_n^{(t)}$ to the effective sample size (ESS) on $\boldsymbol{\Theta}$, i.e., $N_n^{(t)} = \text{ESS}_n^{(t)}$ This approach permits accounting for errors associated with the numerical evaluation of the p-values, which makes the test more robust.

Figure 5.3 shows three cases using this definition of the p-value. In this case, In the first case, the Beta distribution has most of its weight on p-values higher than α. The cdf at α is therefore very small. An associated model $\mathcal{M}$ would not be rejected. In the second case, the Beta distribution has most of its weight on p-values lower than α. The cdf at α is therefore very high, and higher than the threshold δ. An associated model $\mathcal{M}$ would therefore be rejected. Finally, in the third case, the Beta distribution has significant weight both below and above the confidence level α. As a consequence, the cdf evaluated at α falls in the $[\delta, 1-\delta]$ interval. In this case, more samples would be necessary to make a decision.

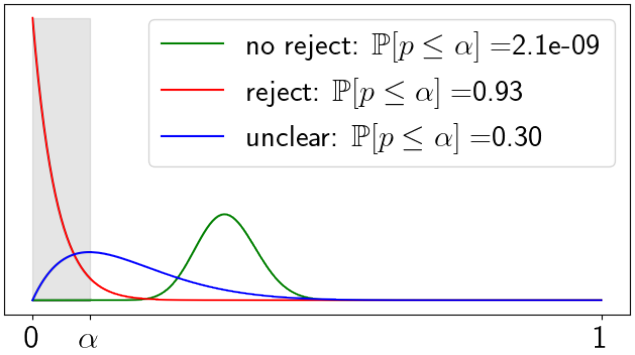


Figure 5.3: Illustration of the three-case test for model assessment. In this case, $\alpha = 0.1$ and $\delta = 0.1$.

5.3.3 Illustration: 2D Gaussian distribution

As already shown in Chapter 2 (Section 2.3.2), the Bayesian p-value can be efficiently estimated in the case of a point estimate $\widehat{\boldsymbol{\Theta}}$ and of an additive Gaussian noise model. In this particular case, the test statistic T follows a χ^2_{NL} distribution, with N the number of pixels and L the number of observables per pixel. This subsection illustrates this equivalence and shows that the proposed test behaves as expected. In particular, we illustrate the convergence of the distribution on $p^{(t)}$ to the theoretical value p at speed $1/\sqrt{t}$. We consider an inverse problem such that $N = 1$ et $L = 2$, i.e., with an observation $\mathbf{y} \in \mathbb{R}^2$. The forward model is set to the identity function $\boldsymbol{f} = \mathrm{id}_2$. The noise is assumed Gaussian, additive and with unit variance. We consider three estimators $\widehat{\boldsymbol{\theta}}^{(i)} \in \mathbb{R}^2$. The confidence levels are set to $\alpha = 0.05$ and $\delta = 0.1$.

Figure 5.4 shows the convergence of the three p-values estimated with Eq. 5.28 to the theoretical values. With $T_{\mathrm{MC}} = 10^3$ independent samples drawn from the posterior predictive distribution $\pi(\widetilde{\mathbf{y}}|\widehat{\boldsymbol{\theta}}^{(i)}, \mathcal{M})$, the point estimate $\widehat{p}$ from Eq. 5.28 leads to a rejection of the null hypothesis from $\widehat{\boldsymbol{\theta}}^{(1)}$ and from $\widehat{\boldsymbol{\theta}}^{(2)}$, and to a non-rejection from $\widehat{\boldsymbol{\theta}}^{(3)}$. These decisions are in agreement with those corresponding to the theoretical p-values for $\widehat{\boldsymbol{\theta}}^{(1)}$ and $\widehat{\boldsymbol{\theta}}^{(3)}$. However, for $\widehat{\boldsymbol{\theta}}^{(2)}$, the rejection is in disagreement with the theoretical p-values, which is higher than the confidence level α. This rejection is caused by the uncertainties inherent to using a Monte Carlo estimator of the p-value with insufficient number of samples. Using the Beta distribution (Eq. 5.30) to quantify uncertainty, one obtains $\mathbb{P}[p_n^{(t)} \leq \alpha] = 0.83$. Since $\mathbb{P}[p_n^{(t)} \leq \alpha] \in [\delta, 1-\delta]$, the proposed three-case test (Eq. 5.29) identifies this issue and recommends drawing more samples. With $T_{\mathrm{MC}} = 10^4$, $\mathbb{P}[p_n^{(t)} \leq \alpha] = 0.096 < \delta$. The proposed three-case test then considers that enough samples have been drawn to make a confident decision, and the model associated with $\widehat{\boldsymbol{\theta}}^{(2)}$ is not rejected. This decision is consistent with the corresponding theoretical p-value. The proposed three-case test (Eq. 5.29) is therefore robust to using a Monte Carlo estimator of the p-values.

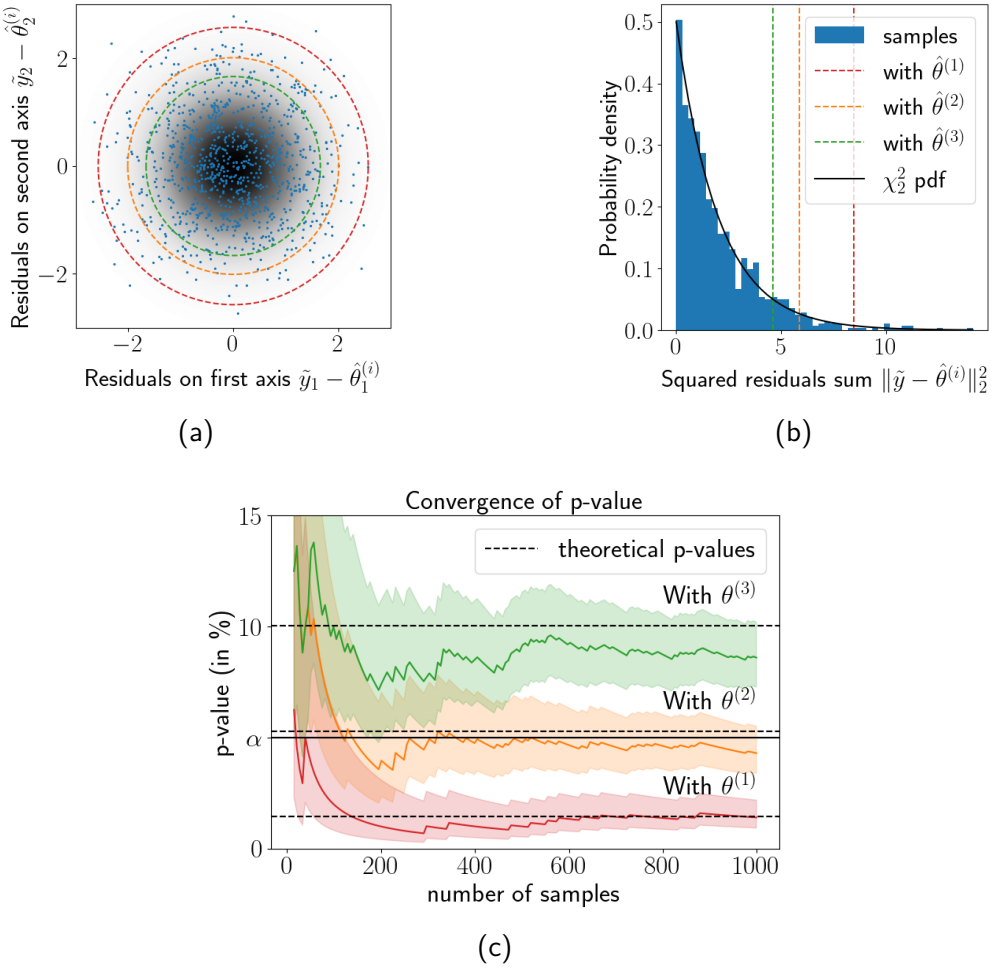


Figure 5.4: Application of Bayesian p-value with three-case test on a simple case, with Gaussian noise and point estimate $\widehat{\boldsymbol{\theta}}$, and comparison with corresponding theoretical p-values. (Top left) pdf of residuals $\widetilde{\mathbf{y}} - \widehat{\boldsymbol{\theta}}$ (in gray levels) and $T_{\text{MC}} = 1\,000$ independent samples $\widetilde{\mathbf{y}} \sim \pi(\widetilde{\mathbf{y}}|\widehat{\boldsymbol{\theta}}^{(i)})$. The contours correspond to constant values of $\|\widetilde{\mathbf{y}} - \widehat{\boldsymbol{\theta}}\|_2^2$. (Top right) Distribution χ^2_2 of $\|\widetilde{\mathbf{y}} - \widehat{\boldsymbol{\theta}}\|_2^2$. The vertical lines correspond to the contour lines in Figure 5.4a. (Bottom) Convergence of p-value estimator (Eq. 5.28) and of the 90% credibility intervals, obtained from Eq. 5.30.

5.4 Applications

In this section, the proposed sampler is applied to two inverse problems: a sensor localization problem and a synthetic example of the general astrophysics problem addressed in this thesis.

5.4.1 Sensor localization

The sensor localization problem introduced in Ihler et al. (2005) is a common test case in multimodal sampling, e.g., in Ahn et al. (2013); Lan et al. (2014); Pompe et al. (2020). Three sensors have known locations and will serve as a reference to avoid ambiguities with respect to translation, rotation and negation. The goal is to estimate the unknown positions $\boldsymbol{\Theta} \in \mathbb{R}^{ND}$ of $N = 8$ sensors in dimension $D = 2$. The observation matrix $\mathbf{Y} \in \mathbb{R}^{NL}$ collects noisy and partially censored pairwise distances, where $L = N + 3$ is the total number of sensors. The distance to sensor ℓ feeds channel ℓ, so that the forward model is $f_\ell(\boldsymbol{\theta}_n) = \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\ell\|$. Note that only $N+2$ distances will really be used since $f_\ell(\boldsymbol{\theta}_\ell) = 0$, and that we set $y_{n\ell} = 0$ by convention. The probability of communication from sensor $\ell \in [\![1, L]\!]$ to sensor $n \in [\![1, N]\!]$ is set to $\exp\left\{-\frac{f_\ell(\boldsymbol{\theta}_n)^2}{2R^2}\right\}$ with $R = 0.3$. In absence of communication, the observation is censored, which is encoded by the binary latent variable $c_{n\ell} = 1$. Otherwise, $c_{n\ell} = 0$ when the observation occurs and is corrupted by a white Gaussian noise

$$y_{n\ell} = f_\ell(\boldsymbol{\theta}_n) + \varepsilon_{n\ell}, \quad \text{with } \varepsilon_{n\ell} \sim \mathcal{N}(0, \sigma^2), \tag{5.31}$$

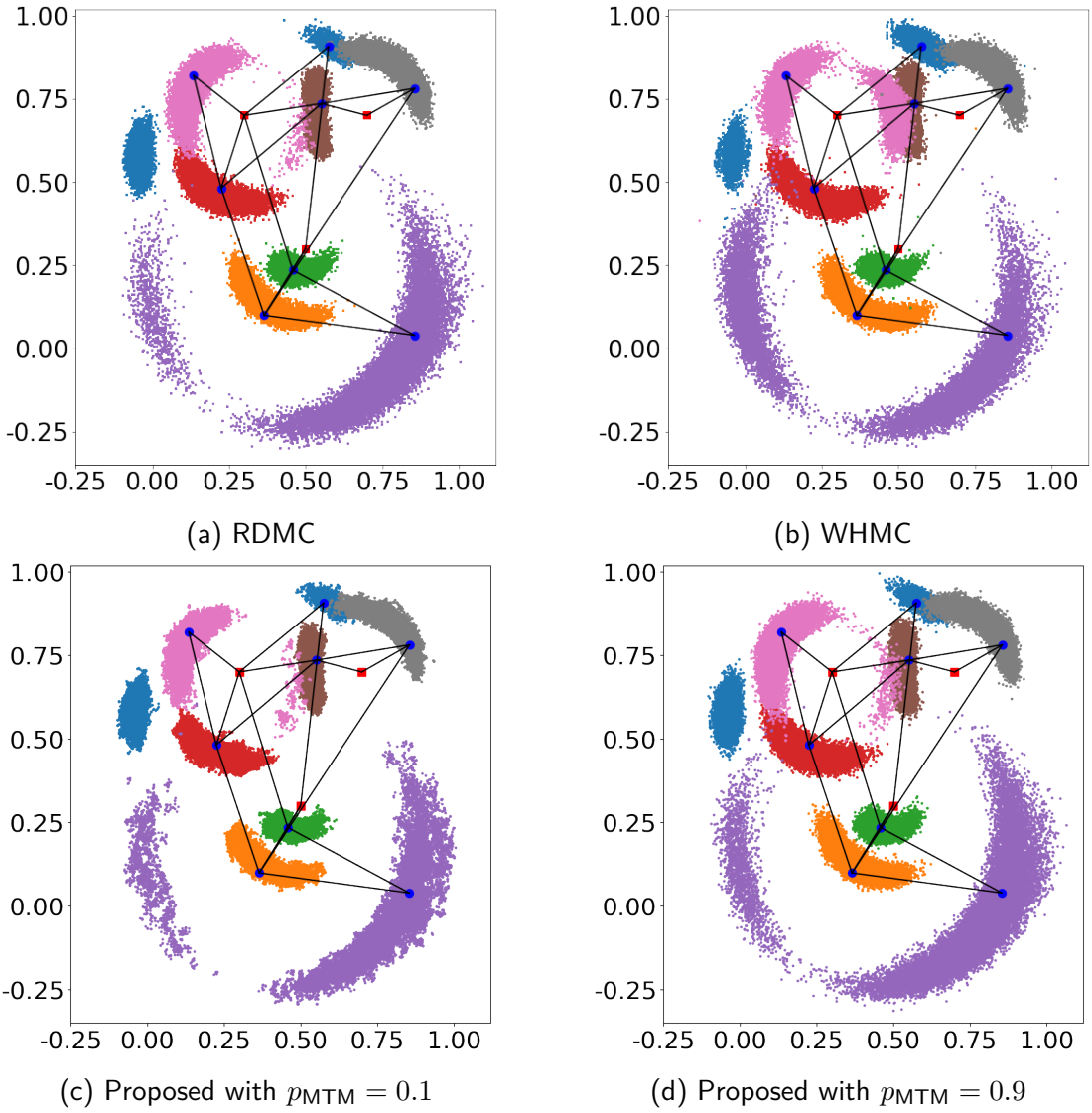
with $\sigma = 0.02$, leading to

$$\begin{aligned} -\ln \pi\left(\mathbf{Y}|\boldsymbol{\Theta}\right) = \sum_{n=1}^{N}\sum_{\ell=1}^{L}(1-c_{n\ell})\left[\frac{(f_\ell(\boldsymbol{\theta}_n) - y_{n\ell})^2}{2\sigma^2} + \frac{f_\ell(\boldsymbol{\theta}_n)^2}{2R^2}\right] \\ + c_{n\ell}\ln\left[1 - \exp\left(-\frac{f_\ell(\boldsymbol{\theta}_n)^2}{2R^2}\right)\right], \end{aligned} \tag{5.32}$$

The smoothed uniform prior on the square $\mathcal{C} = [-0.35, 1.2]^2$ is used as a prior on the location of each sensor. The corresponding penalty parameter ξ introduced in Eq. 5.11 is set to 10^4. This prior is non-informative enough to match the results shown in Ahn et al. (2013); Lan et al. (2014); Pompe et al. (2020). The proposed sampler is compared to both regeneration darting MC (RDMC) (Ahn et al., 2013) and WHMC. A Markov chain of size $30\,000$ is generated by each algorithm, including $5\,000$ burn-in samples. The parameters of the PMALA kernel are set to $a = 0.99$, $\epsilon = 10^{-5}$ and $\eta = 1.5 \times 10^{-3}$. The MTM kernel is selected with $p_{\text{MTM}} = 0.1$ or $p_{\text{MTM}} = 0.9$. Its proposal distribution q is set to the smooth uniform prior on $\mathcal{C}$. For each sensor, the high probability regions are small compared to $\mathcal{C}$. To obtain high acceptance rates for the MTM kernel, the number of candidates is set to $K = 1\,000$. Better proposal distributions can be obtained for this specific problem, which is beyond the scope of this experiment.

Figure 5.5 shows the marginal distributions of each sensor position. The four samplers identified the same modes. Table 5.2 compares the samplers in terms of ESS. With $p_{\text{MTM}} = 0.9$, the proposed sampler yields better mixing capability than WHMC and RDMC. This is due to the partition of the $ND = 16$-dimensional problem into $N = 8$ simpler $D = 2$-dimensional problems. This divide-to-conquer strategy exploits the problem structure to fight the curse of dimension.

Bayesian model assessment – For sake of illustration, one observed distance is divided by a factor 10 to make the observation set inconsistent. The resulting posterior is sampled using the same parameters as above and $p_{\text{MTM}} = 0.9$. Figure 5.6 indicates which observed distance was modified, and compares the marginal posterior distributions with and without this alteration. It appears that few sensors are affected, as most of the marginals are very similar in the two cases. If the ground truth was not known, one could not detect the issue from the marginal posterior distributions alone.



(a) RDMC (b) WHMC

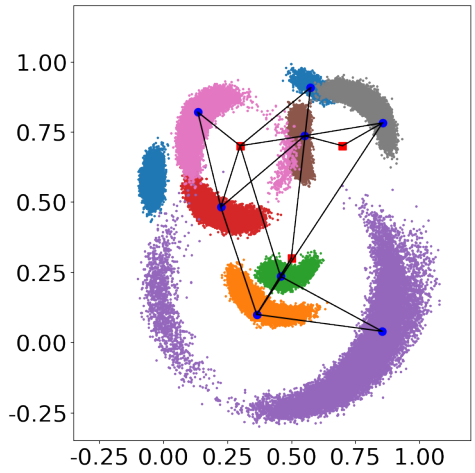
(c) Proposed with $p_{\text{MTM}} = 0.1$ (d) Proposed with $p_{\text{MTM}} = 0.9$

Figure 5.5: Marginal distributions of the sensors positions for the considered samplers. The graph shows the true position of all sensors. The sensors with a known position are in red and those whose position is inferred are in blue. The edges of the graph indicate which pairs of sensors are observed.

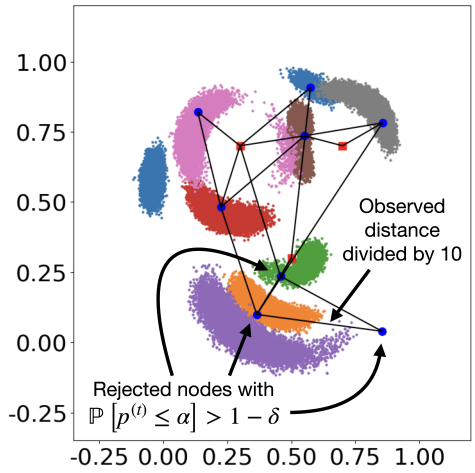
Table 5.2: Effective Sample Size (ESS) on the sensor localization problem.

MCMC sampler	ESS		
	min	mean	max
WHMC (Lan et al., 2014)	29	1 026	5 753
RDMC (Ahn et al., 2013)	168	3 354	11 192
Proposed, $p_{\text{MTM}} = 0.1$	29	329	1 235
Proposed, $p_{\text{MTM}} = 0.9$	**299**	**3 561**	**16 789**

The proposed extended Bayesian model assessment computes one p-value per sensor. In the non altered case, the observations are derived from the observation model and are thus compatible with it. The smallest estimated p-value is $\widehat{p}_n^{(T_{\text{MC}})} = 0.58$ and the largest probability of rejection $\mathbb{P}\left[p^{(T_{\text{MC}})} \leq \alpha\right] = 8 \times 10^{-99} \leq \delta$. The model assessment test thus does not reject any position estimation. In the altered case, most sensors are not rejected, as they are little to not affected by this observation modification. Indeed, for these unaffected sensors, the smallest estimated p-value is $\widehat{p}_n^{(T_{\text{MC}})} = 0.60$ and the largest probability of rejection $\mathbb{P}\left[p^{(T_{\text{MC}})} \leq \alpha\right] = 5 \times 10^{-159} \leq \delta$. However, the model assessment test rejects the three sensors affected by the alteration. For these three sensors, the largest estimated p-value is $\widehat{p}_n^{(T_{\text{MC}})} = 3 \times 10^{-5}$ and the smallest probability of rejection $\mathbb{P}\left[p^{(T_{\text{MC}})} \leq \alpha\right] = 1 - 3 \times 10^{-6} > 1 - \delta$. Therefore, the model assessment approach detects the generated incompatibility between the observations and the observation model. In addition, it pinpoints which sensors are problematic, which helps in identifying the origin of the incompatibility.



(a) Marginals with the true observations. The model assessment test does not reject any position estimation.



(b) Marginals with one observation divided by 10. The model assessment test rejects the three sensors affected by this observation. The other sensors are not rejected, as their positions are little to not affected by this observation modification.

Figure 5.6: Application of model assessment on sensor localization problem.

5.4.2 Realistic astrophysical data

The overall approach is now applied to a synthetic yet realistic case of the general inverse problem addressed in this thesis. The goal is to reconstruct maps of physical parameters of a molecular cloud from radio wave multispectral intensity maps. Each observation map contains $N = 4\,096$ pixels. Each pixel is associated with $D = 4$ physical parameters $\boldsymbol{\theta} = (\kappa, P_{\text{th}}, G_0, A_V^{\text{tot}})$, so that the aim is to infer a set of parameters $\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)_{n=1}^N$ in dimension $N \times D = 16\,536$. The angle φ is set to 0 deg. As κ is a nuisance parameter related to the conditions of observations, its ground truth value is set to 1 over the whole map. The main parameters of interest are the thermal pressure P_{th}, the intensity of a UV radiative field G_0 – measured in reference to the Habing ISRF – and the visual extinction A_V^{tot}, related to the cloud depth along the line of sight. The physical parameters $\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)$ undergo the preprocessing described in Chapter 4 to have similar dynamics. As an abuse of notation, the preprocessed physical parameters are also denoted $\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)$. In the following, all distributions are defined on the preprocessed physical parameters.

Figure 5.7 shows the ground truth parameters $\boldsymbol{\Theta}^*$ in original scale. These maps are chosen

according to a plausible astrophysical scenario (Pety et al., 2017). They correspond to a PDR seen edge-on, with a UV source illuminating the right of the image. This scenario covers a wide variety of physical environments such as an atomic diffuse medium on the right of the map, an actual PDR on the right, a deep molecular cloud with a buried source on the left.

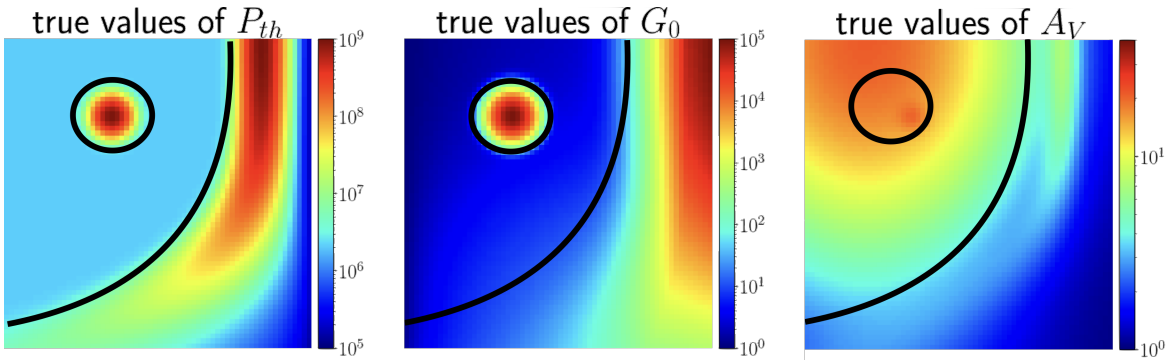


Figure 5.7: Structure of the synthetic molecular cloud and associated true maps of physical parameters $\boldsymbol{\Theta}$. The circle on the top left of the maps indicates a buried source. The surrounding region is a deep molecular cloud. The region on the right is a PDR illuminated by a UV source out of the image.

From the true (preprocessed) maps $\boldsymbol{\Theta}^*$, the emulator $\widetilde{\mathbf{f}}$ of the Meudon PDR code derived in Chapter 4 generates observation maps of $L = 10$ emission lines. These lines are ^{12}CO lines of mid-J rotational transitions, from $J = 4 - 3$ to $J = 13 - 12$. For each line ℓ, $\widetilde{f}_\ell$ ranges from 10^{-18} to 10^{-2} erg cm^{-2} s^{-1} sr^{-1}. These maps are deteriorated according to the observation model from Eq. 5.1. In other words, they are affected by an Gaussian additive uncorrelated noise, a lognormal multiplicative uncorrelated noise and censorship. Model misspecification noise is not considered in this example, as the forward model $\widetilde{\mathbf{f}}$ is used both to define the observation $\mathbf{Y}$ and to infer the physical parameters $\boldsymbol{\Theta}$. The standard deviation of the multiplicative noise is set to $\sigma_m = \log(1.1)$, which roughly represents a 10% alteration in average that corresponds to a calibration error. For the additive noise, $\sigma_{a,n\ell} = \sigma_a = 1.38715 \cdot 10^{-10}$ erg cm^{-2} s^{-1} sr^{-1} for all pixel n and line ℓ so that the signal-to-noise ratio (SNR) varies between -81 and 79 dB. The censorship level is set to $\omega_{n\ell} = \omega = 3\sigma_a$ for all pixel n and line ℓ. Figure 5.8 shows the observation maps of two lines and the spatial distribution of censorship importance.

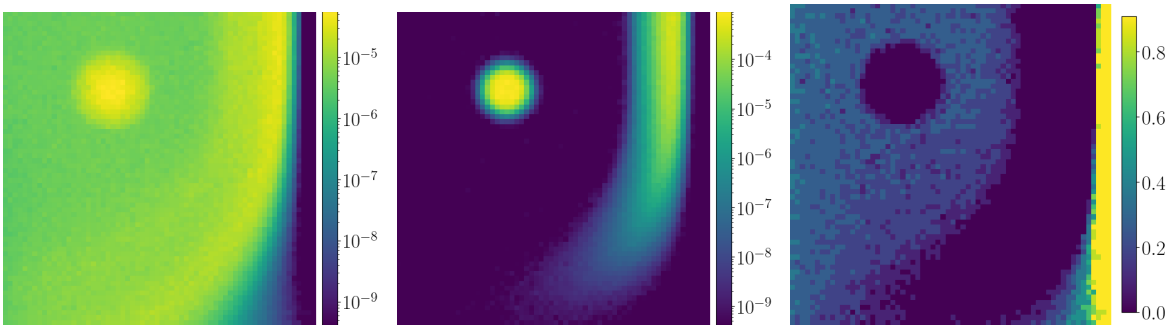


Figure 5.8: Some observation maps of the astrophysical experiment. From left to right: line $\ell = 1$, line $\ell = 10$, proportion of censored lines per pixel.

The likelihood approximation is obtained as indicated in Section 5.1.1, and its parameters $\boldsymbol{a}_\ell$ are adjusted as described in Appendix 5.B. The validity set $\mathcal{C}$ of physical parameters is set as in Wu et al. (2018), and the penalty parameter ξ of the smooth uniform prior is set to 10^4. Given

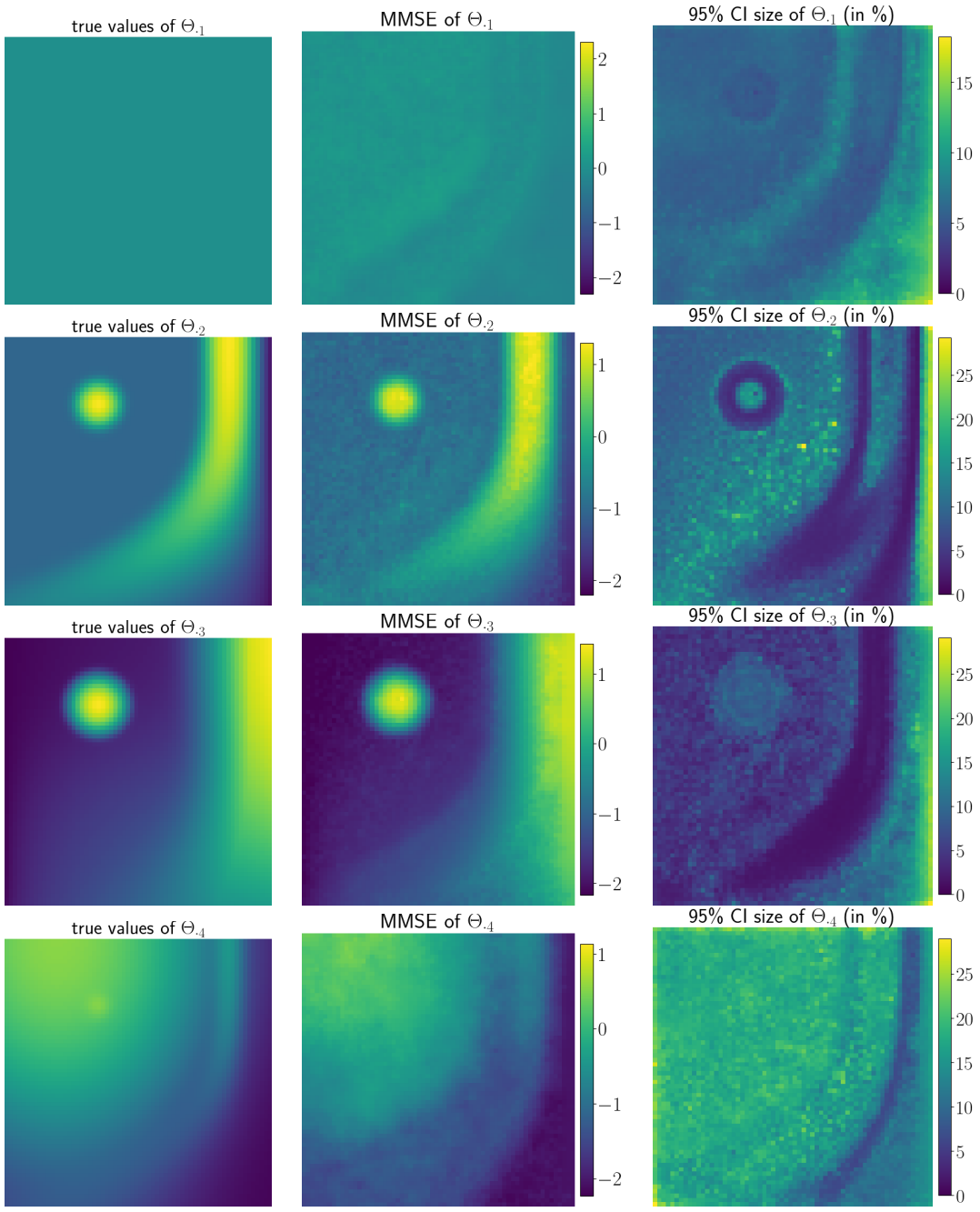


Figure 5.9: Inference results: (left) ground truth $\boldsymbol{\Theta}^*$; (middle) MMSE estimate from the proposed transition kernel; (right) size of the 95% credibility interval (CI), in % of the size of the validity intervals. All values of $\boldsymbol{\Theta}$ are displayed in the preprocessed space.

the smoothness of the true maps, for each d the chosen spatial regularizer h is taken as

$$h(\boldsymbol{\Theta}_{\cdot d}) = \|\Delta \boldsymbol{\Theta}_{\cdot d}\|_2^2 = \sum_{n=1}^{N} \sum_{i \in V_n} (\theta_{nd} - \theta_{id})^2, \tag{5.33}$$

where Δ is the discrete 5-point based 2D Laplacian operator, and V_n is the set of neighbors of pixel n induced by Δ. The hyperparameter $\boldsymbol{\tau}$ from Eq. 5.11 is fixed to $\boldsymbol{\tau} = (10, 2, 3, 4)$.

Inference is carried out using $10\,000$ iterations of a Markov chain including $1\,500$ burn-in samples. The parameters of the proposed sampler are set to $a = 0.99$, $\epsilon = 10^{-5}$ and $\eta = 5 \times 10^{-7}$ for PMALA, and to $p_{\text{MTM}} = 0.5$ and $K = 50$ for MTM. Since the operator Δ only compares a pixel to its four neighbors and since the indicator prior and likelihood are pixel-wise, the set of pixels can be partitioned into two conditionally independent subsets of pixels. A two sites Chromatic Gibbs sampling (Gonzalez et al., 2011) is therefore performed in the MTM kernel to speed up computations. Note that using the smooth uniform prior as a proposal distribution in MTM is inefficient due to the small size of high probability regions compared to the volume of $\mathcal{C}$. The proposal distribution q is based on the spatial prior (Eq. 5.33) instead. For any pixel, one can show that the conditional spatial prior is a Gaussian distribution centered on the mean of the set of neighboring pixels V_n. Since maps are assumed to be smooth, the likelihood functions for a pixel n and its neighbors should correspond to similar modes in the parameters' domain. If the neighbors are not all in the same mode, the mean of the neighbors will in general not fall in a high probability region. Therefore, for a pixel n, the proposal distribution is defined as a Gaussian mixture whose modes are all the means of non-empty subsets $V \in \mathcal{P}(V_n)$ of V_n:

$$q(\boldsymbol{\theta}_n | \boldsymbol{\Theta}_{\setminus n}) \propto \prod_{d=1}^{D} \sum_{V \in \mathcal{P}(V_n)} \exp\left[-2\tau_d \sum_{i \in V} (\theta_{nd} - \theta_{id})^2\right] \tag{5.34}$$

$$\propto \prod_{d=1}^{D} \sum_{V \in \mathcal{P}(V_n)} \exp\left[-2\tau_d |V| \left(\theta_{nd} - \frac{1}{|V|} \sum_{i \in V} \theta_{id}\right)^2\right]. \tag{5.35}$$

Performance is assessed for the minimum mean square error (MMSE) estimate $\widehat{\boldsymbol{\Theta}}$. Recall that the inferred parameters $\boldsymbol{\Theta}$ correspond to normalized logarithms of physical parameters $\boldsymbol{\Theta}$. Therefore, prediction errors on the D parameter maps $\boldsymbol{\Theta}_{\cdot d}$ are comparable. The quality of the reconstruction is quantified with the mean squared error (MSE) $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2^2$ and the reconstruction signal-to-noise ratio (R-SNR) $20 \log_{10}\left(\frac{\|\boldsymbol{\Theta}^*\|}{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|}\right)$.

Figure 5.9 shows the estimation results. The MMSE estimate $\widehat{\boldsymbol{\Theta}}$ (middle) is very close to the ground truth $\boldsymbol{\Theta}^*$ (left). The reconstructions are qualitatively very consistent with the underlying physics. The parameter $\boldsymbol{\Theta}_{\cdot 4}$, corresponding to A_V^{tot}, is known by astrophysicists to be the most difficult to retrieve from ^{12}CO low and mid-J lines. Indeed, these lines get optically thick, which means that observed photons come from the surface of the cloud, and that the integrated intensities saturate with $A_V^{\text{tot}} \gtrsim 7$ mag. Such pixels appear in the top left corner of the ground truth map.

Table 5.3 shows the MSE and the R-SNR for each parameter $\boldsymbol{\Theta}_{\cdot d}$, and the relative size of the credibility intervals with respect to the associated (normalized) validity interval $\mathcal{C}$. As expected, the MSE is larger for $\boldsymbol{\Theta}_{\cdot 4}$ ($\leftrightarrow A_V^{\text{tot}}$), and the relative size of its credibility intervals are overall the largest, about 16.2%. The problem is also very ill-posed for all parameters in pixels with very low SNR, where most of the lines are censored, see Figure 5.8 (right). To interpret the results from an astrophysical viewpoint, performances are computed over two subsets of pixels with either less or more than 50% of censored lines. As expected, the credibility intervals of the latter are about twice as large as the former. Finally, all the parameters but A_V^{tot} are well constrained for pixels with less than 50% of censored lines. The inference remains challenging since the posterior contains many local modes with high $\mathcal{L}$ values, but the proposal distribution q permits the Markov

chain to successfully reach the mode of interest. The relative quadratic error results in an R-SNR between 15.5 dB and 23.4 dB. Credibility intervals at 95% level remain small, ranging from 5.7% to 9.3% of the admissible interval $\mathcal{C}$.

Table 5.3: Reconstruction metrics and relative size of credible intervals for the astrophysics experiment. The R-SNR is not defined for $\boldsymbol{\Theta}_{\cdot 1}$, as its ground truth is 0 everywhere.

	MMSE		Mean 95% credibility intervals size		
	MSE	R-SNR (dB)	censorship $\leq$ 50%	censorship $>$ 50%	overall
$\boldsymbol{\Theta}_{\cdot 1}$	0.017	–	6.1 %	11.9 %	6.8 %
$\boldsymbol{\Theta}_{\cdot 2}$	0.019	16.8	9.3 %	20.6 %	9.9 %
$\boldsymbol{\Theta}_{\cdot 3}$	0.009	23.4	5.7 %	19.8 %	6.5 %
$\boldsymbol{\Theta}_{\cdot 4}$	0.034	15.5	16.3 %	14.5 %	16.2 %

Combining all the difficulties addressed in the general inverse problem addressed in this thesis, this synthetic inverse problem illustrates the good performances of the proposed approach in a challenging scenario. The proposed likelihood approximation enabled handling the censorship and mixture of noises present in the observation model. Dealing with a multimodal posterior distribution, the MTM kernel allows the different modes to be visited, while the PMALA kernel permits to explore them efficiently. The proposed sampler provides high quality estimates and informative credibility intervals.

5.5 Conclusion

In this chapter, we addressed a family of inverse problems that combine several difficulties: a non-linear black-box forward model, potentially non-injective, that covers multiple decades; observations damaged by both censorship and a mixture of additive and multiplicative noises. The proposed approach takes into account as many sources of uncertainty as possible. The likelihood is intractable and leads to a potentially multimodal posterior distribution. An approximation of the likelihood was proposed, based on a model reduction and an approximate parametric noise mixture model with controlled error. The prior distribution combined a spatial regularization and a smoothed uniform distribution encoding validity constraints on the physical parameters.

To efficiently sample from the resulting multimodal posterior, an original MCMC algorithm combining two kernels was proposed. The Gibbs-like MTM kernel permits jumps between modes, while the PMALA kernel efficiently explores the local geometry of each mode. The proposed sampler was shown to be competitive with state-of-the-art multimodal sampling methods on a Gaussian mixture model and a sensor localization problem. A realistic application to a challenging inverse problem on a large observation map has shown the interest and the good performances of the proposed approach. Estimation errors remain small and uncertainties are quantified.

In addition to the MCMC algorithm, the Bayesian hypothesis testing procedure was extended. Uncertainties due to the Monte Carlo evaluation of the p-value are accounted for in the rejection or non-rejection decision. This approach permits in real applications to assess the compatibility between the observations and the observation model.

In the next chapter, the proposed approach is applied to real observations of photodissociation regions.

Appendix 5.A Tuning automatically the prior hyperparameters

The considered prior from Eq. 5.11 is reminded here:

$$\pi(\boldsymbol{\Theta}) \propto \exp\left(-\xi\, \widetilde{\iota}_{\mathcal{C}^N}(\boldsymbol{\Theta}) - \sum_{d=1}^{D} \tau_d\, h(\boldsymbol{\Theta}_{\cdot d})\right).$$

This prior relies on two hyperparameters: $\xi > 0$ weighs the smooth indicator function, and $\boldsymbol{\tau}$ contains the D spatial regularization weights of the D maps to be inferred. The indicator prior parameter ξ has little impact on estimations. However, the values of the spatial regularization weights $\boldsymbol{\tau}$ have a dramatic impact on the posterior distribution. In some cases, this parameter can be learned from the data using a hierarchical model, as in Galliano (2018). For instance, Pereyra et al. (2015) proposes a general approach to estimate the best regularization parameter along with the physical parameters $\boldsymbol{\Theta}$. However, this approach relies on a scale invariance. In our case, the validity intervals violate this invariance. Using hierarchical model and sampling from the distribution of is thus much harder.

We tried applying the maximum marginal likelihood estimator presented in Vidal et al. (2020) to automatically tune $\boldsymbol{\tau}$. However, the additive noise variance is sometimes overestimated in astrophysics applications, as we will show in Chapter 6. In such cases, the spatial regularization weights τ_d diverge, leading to constant estimated maps. Such maps are unrealistic for the environments considered in Chapter 6. To avoid such divergences, we chose to set manually the model hyperparameters, depending on the expected smoothness in the final estimated maps and with inversions with different values.

Appendix 5.B Optimization of the approximation parameter

For each channel ℓ, the parameter $\boldsymbol{a}_\ell = (a_{\ell,0}, a_{\ell,1})$ locates the frontiers between low, intermediate and high values regimes of $\ln \widetilde{f}_\ell$ in the definition of λ (Eq. 5.8). It has a critical influence on the approximation quality. It should be adjusted to $\ln \widetilde{f}_\ell$, σ_a and σ_m. For simplicity, in this subsection, likelihood functions are conditioned with respect to $z = \ln \widetilde{f}_\ell(\boldsymbol{\theta}) \in \mathbb{R}$ instead of $\boldsymbol{\theta} \in \mathbb{R}^D$. The true likelihood is not explicit, but the model in Eq. 5.1 can be easily sampled from, and the approximation (Eq. 5.9) is known.

The parameter $\boldsymbol{a}_\ell$ is set to obtain an approximation as close as possible to the true likelihood, with respect to some divergence criterion. The Kullback-Leibler (KL) divergence would be a natural choice. However, due to the number of decades spanned, the standard deviation of KL estimators is in practice larger than the quantity of interest (Kraskov et al., 2004), which prevents from performing optimization. The Kolmogorov-Smirnov (KS) distance is not affected by this property: for a given z, it only requires ordered samples $(y^{(i)})_{i=1}^{M}$. It reads

$$\widehat{D}_{\mathsf{KS}}(z, \boldsymbol{a}_\ell) = \sup_{y \in \mathbb{R}} \left| \widehat{F}_M(y|z) - \widetilde{F}(y|z, \boldsymbol{a}_\ell) \right|, \tag{5.36}$$

where $\widehat{F}_M(\cdot|z)$ is the empirical cdf of the true likelihood $\pi(\cdot|z)$ estimated from M samples $y^{(i)}$, and $\widetilde{F}(\cdot|z, \boldsymbol{a}_\ell)$ is the cdf of the proposed approximation (Eq. 5.9). Assuming that $\boldsymbol{\theta}$ follows a uniform distribution on $\mathcal{C}$ yields a distribution on z with pdf $\pi(z)$ which can be estimated by kernel density estimation (KDE). The function to minimize is

$$\varphi(\boldsymbol{a}_\ell) = \mathbb{E}_z\left[\widehat{D}_{\mathsf{KS}}\left(z, \boldsymbol{a}_\ell\right)\right] = \int \widehat{D}_{\mathsf{KS}}(z, \boldsymbol{a}_\ell)\pi(z)dz. \tag{5.37}$$

An estimator $\widehat{\varphi}$ can be obtained using numerical integration on z over S bins. The higher M and S, the better the estimation accuracy. Minimizing $\widehat{\varphi}$ can be performed using a grid search, which is quite computationally intensive. A cheaper alternative is to use a Bayesian optimization (BO) procedure (Shahriari et al., 2016). This optimization was applied for each channel in the

astrophysical application described in Section 5.4.2. Both grid search and BO approaches were used. The KDE of $\pi(z)$ was performed from $810\,000$ samples. The BO procedure was run with $S = 100$ and $M = 250\,000$ using Nogueira (2014–) with default parameters. Figure 5.10 shows the results for one channel. The proposed approximation with adjusted $\boldsymbol{a}_\ell$ is closer to the true likelihood than a purely additive Gaussian approximation, i.e., $a_{\ell,0} > \max_j z^{(j)}$, or a purely multiplicative lognormal approximations, i.e., $a_{\ell,1} < \min_j z^{(j)}$.

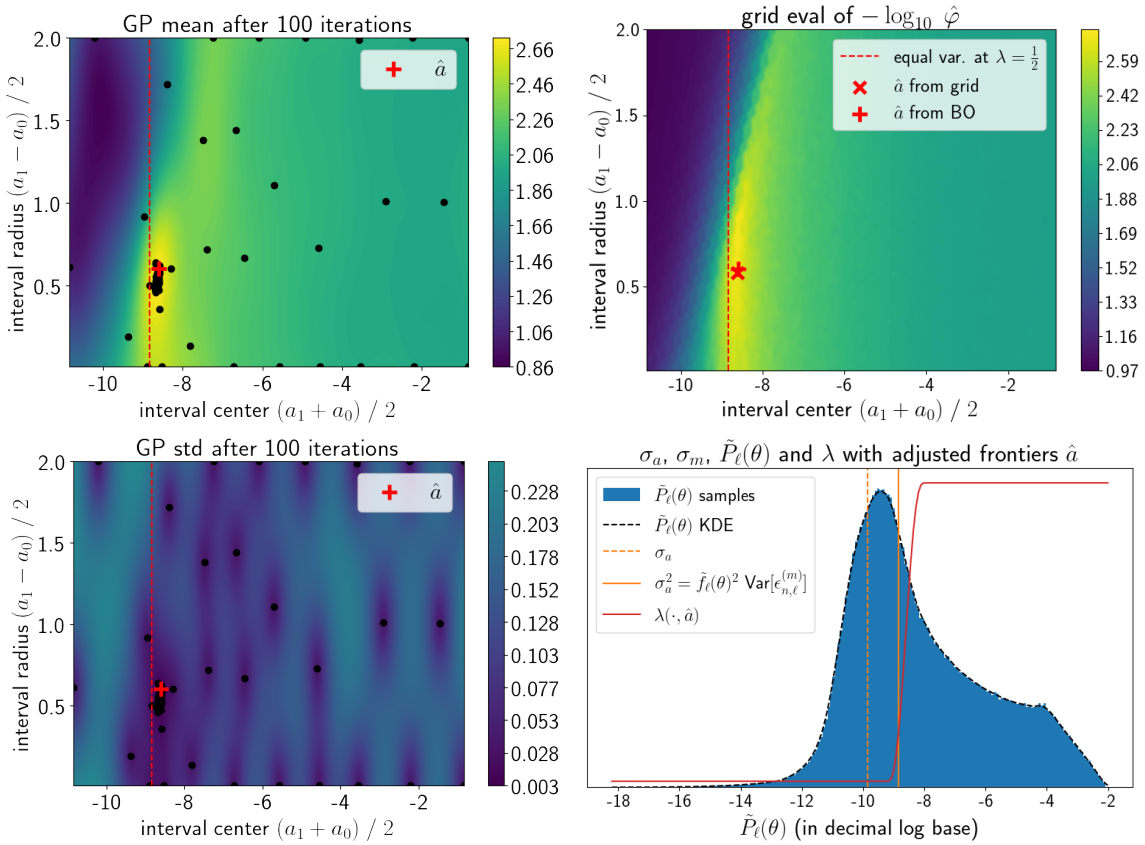


Figure 5.10: Maximization of $-\log_{10}\widehat{\varphi}$ using both Bayesian optimization (BO) and grid search for one channel of the astrophysical case detailed in 5.4.2. In BO, a Gaussian process (GP) replaces the function to optimize (left column). The red dashed vertical bar represents the value of $\frac{a_0+a_1}{2}$ for which the additive and multiplicative noises have equal variances, i.e. $\sigma_a^2 = \widetilde{f}_\ell(\boldsymbol{\theta})^2 \mathrm{Var}[\varepsilon_{n\ell}^{(m)}]$, at $\lambda = \frac{1}{2}$. For clarity, all scales are displayed in $\log_{10}$ scale, while computations are done in $\ln$ scale.

Appendix 5.C Sampling from the smoothed indicator distribution

This section describes the algorithm to draw samples from the real-valued probability distribution with density $\pi(\theta) \propto \exp\left(-\xi\, \widetilde{\iota}_{[l,u]}(\theta)\right)$, with $l < u$ and $\widetilde{\iota}_{[l,u]}$ introduced in Eq. 5.10. To this aim, consider the generalized normal distribution $G\mathcal{N}(0, 1/\xi^4, 4)$ of pdf (Nadarajah, 2005)

$$p_{G\mathcal{N}}(\theta) = \frac{2\xi^{\frac{1}{4}}}{\Gamma(1/4)} \exp\left(-\xi\,\theta^4\right). \tag{5.38}$$

Note that $\pi(\theta)$ is a continuous extension of a uniform distribution and of this generalized normal distribution at 0.

$$\pi(\theta) \propto \begin{cases} p_{G\mathcal{N}}(\theta - l) & \text{if } \theta < l, \\ p_{G\mathcal{N}}(0) & \text{if } \theta \in [l, u], \\ p_{G\mathcal{N}}(u - \theta) & \text{if } \theta > u. \end{cases} \tag{5.39}$$

The normalizing constant of $\pi(\theta)$ is $1 + p_{G\mathcal{N}}(0)(u - l)$. The weight of the uniform section in the combination is therefore

$$w_{\text{Unif}} = \frac{1}{1 + \frac{\Gamma(1/4)}{2} \frac{1}{\xi^{\frac{1}{4}}(u-l)}} \,. \tag{5.40}$$

Algorithm 5.5 summarizes the procedure to sample from $\pi(\theta)$.

Algorithm 5.5: Sampling from the smooth distribution in Eq. 5.39

Input: scale factor ξ, bounds $l, u \in \mathbb{R}$ such that $l < u$
Output: sample θ

w_{Unif} // using Eq. 5.40
$z \sim \mathcal{B}(w_{\text{Unif}})$
if $z = 1$ **then** $\theta \sim \text{Unif}(l, u)$
else
  $\theta \sim G\mathcal{N}(0, 1/\xi^4, 4)$ // using Nardon and Pianca (2009)
  if $\theta < 0$ **then** $\theta = \theta + l$ **else** $\theta = \theta + u$

Chapter 6

Application to real data

" In theory, there is no difference between theory and practice. In practice there is. "

Benjamin Brewster, "The Yale Literary Magazine", February 1882

Contents

In Chapter 4, we derived an approximation of the Meudon PDR code, that we now use to model photodissociation regions (PDRs). Chapter 5 presented the full observation model, prior and posterior distributions, and the proposed method to sample from the posterior. It also introduced the Bayesian test of hypothesis that permits to assess the compatibility of the model with the observations. In this Chapter, we apply the full inference workflow to real observations of photodissociation regions in star forming regions in order to determine the physical conditions in these environments. A journal article is currently in preparation.

Section 6.1 summarizes the full inversion procedure. In Section 6.2, we apply our method to observations of NGC 7023 ($N = 1$ and $L = 17$) studied in Joblin et al. (2018). Appendix 6.A describes a similar analysis on the Orion bar, that was also studied in Joblin et al. (2018). Section 6.3 presents a first application to real-life multi-pixel observation maps. It studies the Carina nebula maps analyzed in Wu et al. (2018) ($N = 176$ pixels and $L = 12$ lines for inversion). These first two analyses demonstrate that our results are consistent with those of Joblin et al. (2018) and Wu et al. (2018). They also show that our approach is richer as it provides complete uncertainty quantification in addition to the point estimates. Section 6.4 studies the OMC-1 observations introduced in Goicoechea et al. (2019) ($N = 2\,475$ pixels and $L = 4$ lines for inversion). The proposed analysis is the first for these OMC-1 observations.

6.1 Summary of the inversion procedure

This section summarizes the inversion procedure presented in Chapters 4 and 5. Figure 6.1 illustrates the full procedure.

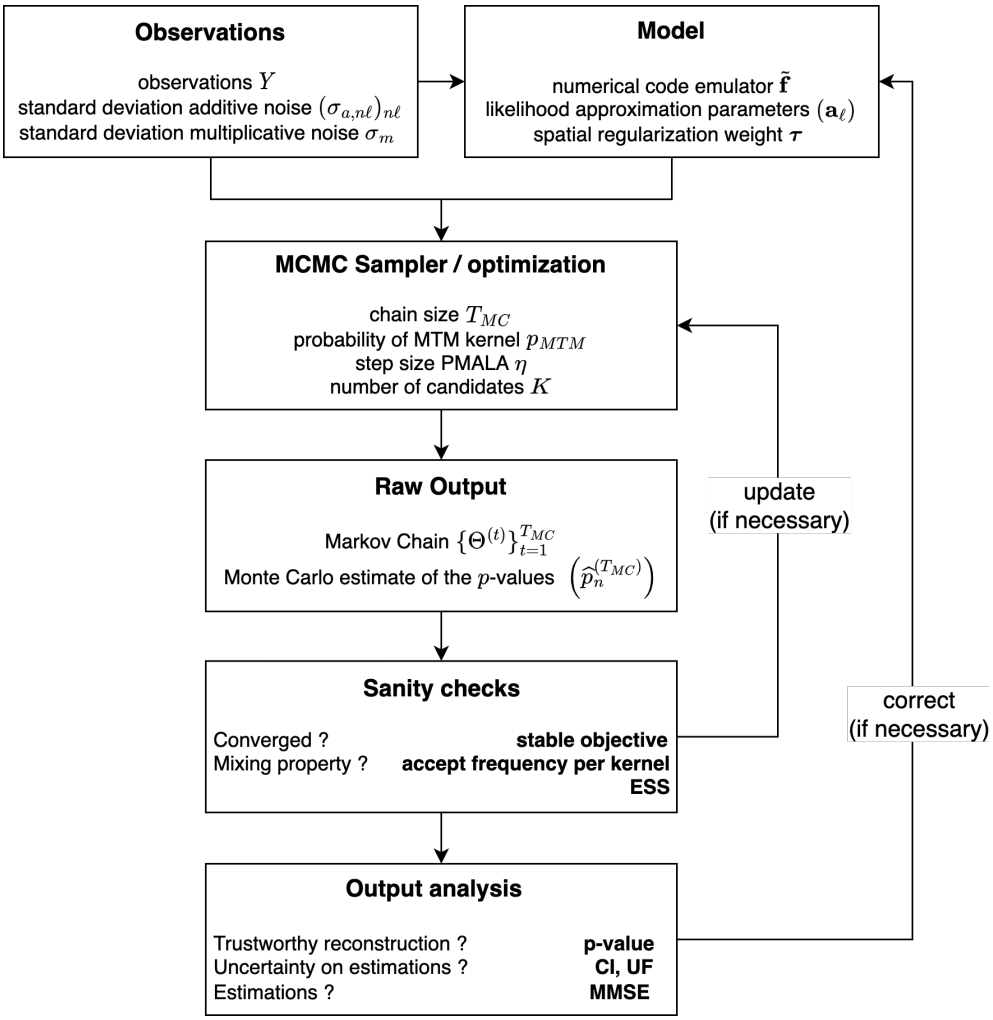


Figure 6.1: Full inference workflow. Observations and noise standard deviations are assumed known. The indicated model and sampler parameters need to be adjusted for each application.

Setting the posterior distribution – In this chapter, the posterior distribution is defined on maps of physical parameters $\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)_{n=1}^N \in \mathbb{R}^{N\times D}$. The physical parameter vectors contain $D = 4$ elements: $\boldsymbol{\theta} = (\kappa, P_{\text{th}}, G_0, A_V^{\text{tot}})$, with κ a scaling parameter, P_{th} the cloud thermal pressure, G_0 the intensity of the incident radiative field and A_V^{tot} the cloud total visual extinction – see Chapter 1 (Section 1.1.4) for a description of these parameters. The posterior distribution relies on a likelihood function and on a prior distribution.

The likelihood function relies on a forward model and a noise model. The noise model described in Chapter 3 (Section 3.4) combines a Gaussian additive noise and a lognormal multiplicative noise. The corresponding standard deviations, $(\sigma_{a,n\ell})$ and σ_m, are assumed known. The forward model $\mathbf{f}$ is the Meudon PDR code, which requires a few hours for each evaluation. As inference requires many evaluations, we replace it with a fast and accurate neural network-based emulator $\widetilde{\mathbf{f}}$ (Chapter 4). The resulting likelihood function has no simple closed-form expression. Chapter 5 (Section 5.1.1) proposed a simple approximation with controlled error. This approximation relies on parameters $(\boldsymbol{a}_\ell)$ to be adjusted to minimize the error – see Chapter 5 (Appendix 5.B).

The prior distribution combines an indicator term that restricts values to an acceptable range for each parameter, and a spatial regularization term that favors smooth reconstructed maps for

multi-pixel observations. In this chapter, the indicator prior is set on the validity intervals used to train the emulator $\widetilde{\mathbf{f}}$ in Chapter 4, i.e., $P_{\text{th}} \in [10^5, 10^9]\,\text{K cm}^{-3}$, $G_0 \in [1, 10^5]$[1], and $A_V^{\text{tot}} \in [1, 40]$ mag. The scaling parameter κ is limited to $[0.1, 10]$. The smoothing of the indicator prior at the edges of the domain is performed using $\xi = 10^4$. The spatial regularization term relies on a weight parameter $\boldsymbol{\tau} \in \mathbb{R}^4$, that will be adjusted for multi-pixel observations.

Performing inference with the proposed sampler – Markov chain Monte Carlo (MCMC) algorithms produce sets of $T_{\text{MC}} \geq 1$ correlated samples $(\boldsymbol{\Theta}^{(t)})_{t=1}^{T_{\text{MC}}}$ from the posterior distribution. In each application, the number of iterates T_{MC} is chosen so that the posterior distribution is well sampled. In particular, the duration of the burn-in phase T_{BI} is set so that estimators are only evaluated from iterates in the Markov chain stationary phase.

The MCMC proposed in Chapter 5 (Section 5.2) combines two update steps. The first update step is called PMALA. It is an MCMC variant of a preconditioned gradient descent that performs efficient local exploration of the posterior distribution. We resort to the RMSProp diagonal preconditioner, which offers a good trade-off between computational costs and preconditioning efficiency – see Chapter 2 (Section 2.2.2.5). It relies on three parameters: a step size η, a damping parameter ϵ and an exponential decay rate a. In this chapter, the damping parameter and exponential decay rate are set to their default values, i.e., $\epsilon = 10^{-5}$ and $a = 0.99$. In each application, the step size is chosen so that the average acceptance probability is close to the MALA optimum acceptance probability, i.e., around 50%-60%.

The second update step is called MTM. It permits to globally explore the posterior and to escape local minima for multimodal distributions. It handles pixels individually. For each pixel, it generates $K \geq 1$ candidates from a proposal distribution q, and selects one to perform an accept-reject step. At each iteration t, one of the two update steps is selected with probability $p_{\text{MTM}} \in [0, 1]$ for MTM and $1 - p_{\text{MTM}}$ for PMALA. For each application, the number of candidates K and the selection probability p_{MTM} are set to explore efficiently the posterior distribution, i.e., to obtain high effective sample sizes (ESSs) – see Chapter 2 (Section 2.2.2.2).

The output correlated set $(\boldsymbol{\Theta}^{(t)})_{t=1}^{T_{\text{MC}}}$ can be used to perform checks including on the chain convergence – did the sampling reach high probability region of the posterior distribution? – and on the mixing performance – did the sampling explore well the posterior distribution? These tests were performed for all the applications presented in this chapter, but are not shown for conciseness.

Bayesian hypothesis testing – The ability of the Meudon PDR code and of the noise model to explain the observations is checked for each application. We resort to the model assessment method presented in Chapter 5 (Section 5.3). This method evaluates a p-value for each pixel n. A p-value below a confidence level α pinpoints an incompatibility between the model and the observation. The proposed sampling algorithm evaluates Monte Carlo (MC) estimators $\widehat{p}_n^{(T_{\text{MC}})}$ of the p-values – see Chapter 2 (Section 2.2.2) for a description on MC estimators. The uncertainty inherent to the MC estimator is accounted for, allowing us to evaluate a probability of rejection $\mathbb{P}[p_n \leq \alpha]$. We proposed a three-case rule by introducing a probability threshold δ to make the test robust to the MC estimator error. In this chapter, we choose $\alpha = 0.05$ and $\delta = 0.1$.

MMSE, credibility intervals and uncertainty factor – The set of correlated samples $(\boldsymbol{\Theta}^{(t)})_{t=1}^{T_{\text{MC}}}$ are used to evaluate MC estimators. In this chapter, two estimators are considered: the minimum mean square error (MMSE) and credibility intervals (CIs). The MMSE $\widehat{\boldsymbol{\Theta}}_{\text{MMSE}}$ is the sample mean and approximates the posterior expectation. Credibility intervals enable to quantify the uncertainty associated with each inferred physical parameter. See Chapter 2 (Section 2.1.2) for a description of these estimators.

The parameters inferred in this chapter cover multiple decades. Uncertainties on these parameters are thus better described with a multiplicative error. We define the uncertainty factor (UF)

[1] In this thesis, G_0 is defined in reference to the Habing ISRF. See Chapter 1 (Section 1.1.4) for more details.

to quantify this multiplicative error. For a confidence level α and a credibility interval $[l_\alpha, u_\alpha]$,

$$\mathsf{UF}_\alpha = \sqrt{\frac{u_\alpha}{l_\alpha}}, \tag{6.1}$$

so that the true value is expected to be within a factor UF_α below or above the estimated value. For a lognormal posterior distribution on a physical parameter, the $\mathsf{UF}_{68\%}$ corresponds to a 1σ error. Similarly, $\mathsf{UF}_{95\%}$ corresponds to a $\simeq 2\sigma$ error. By default, we denote $\mathsf{UF} = \mathsf{UF}_{68\%}$.

6.2 NGC 7023

Joblin et al. (2018) analyzed two one-pixel PDR observations: one of NGC 7023, and one of the Orion Bar. We applied the proposed inversion procedure to both. In this section, we only present results on NGC 7023 for conciseness. We demonstrate that the proposed inversion procedure yields consistent results with Joblin et al. (2018) and provides complete uncertainty quantification in addition to point estimates. Appendix 6.A presents the results for the Orion Bar.

Located in the Cepheus constellation, NGC 7023 is a PDR illuminated by the spectroscopic binary system HD 200775 [RA(2000) = 21h01m36.9s; Dec(2000) =+68 09047.800]. Its distance from the Sun was estimated at 320 ± 51 pc in Benisty et al. (2013). NGC 7023 is extensively studied to understand the effect of radiative feedback because of its edge-on geometry, brightness and proximity. The two stars in the binary system are classified as B3Ve and B5 (Alecian et al., 2008). Chokshi et al. (1988) estimated an intensity of $G_0 = 2\,600$ and a proton density of $n_H \simeq 4 \times 10^3$ cm^{-3} from the C^+ 158 µm and O 63 µm lines.

Joblin et al. (2018) studied the impact of radiative feedback in NGC 7023 from Herschel observations combined with some Spitzer-IRS observations for rotational lines of H_2. The resulting observations gathered $L = 17$ lines on $N = 1$ pixel, including ^{12}CO lines (from $J = 11 \to 10$ to $J = 19 \to 18$), rotational H_2 lines (from S(0) to S(5)) and low level CH^+ rotational lines (from $J = 1 \to 0$ to $J = 3 \to 2$). These observations were modeled with a Gaussian additive and uncorrelated noise model. The standard deviations $(\sigma_{a,\ell})_{\ell=1}^L$ were assumed known and included calibration error. The analysis was performed with the version 1.5.4 of the Meudon PDR code with a fixed value of $G_0 = 2\,600$. As the observed lines only trace the warm molecular layer of the PDR, the visual extinction A_V^{tot} was fixed to 10 mag. The observation angle φ was set to 60 deg to account for the edge-on geometry of the cloud – the Meudon PDR code cannot run for $\varphi = 90$ deg due to its infinite slab with finite thickness geometry, and $\varphi = 60$ deg is the closest to an edge-on inclination. The inference was performed on $\boldsymbol{\theta} = (\kappa, P_{\mathsf{th}})$ using a grid search on P_{th} and a simple continuous optimization for κ. They obtained $P_{\mathsf{th}} = 1 \times 10^8$ K cm^{-3} and $\kappa = 0.7$. This high P_{th} value indicates that the radiative feedback compresses the PDR.

Inversion setup – The posterior distribution is defined on the physical parameter vector $\boldsymbol{\theta} = (\kappa, P_{\mathsf{th}}, G_0, A_V^{\mathsf{tot}}) \in \mathbb{R}^4$. As in Joblin et al. (2018), we set the observation angle to $\varphi = 60$ deg. The calibration error is included in the additive error. The multiplicative noise source accounts only for the model misspecification. We set $\sigma_m = \ln 1.3$, where a 1σ error for this multiplicative term corresponds to a factor 1.3. The resulting observation model is

$$\forall \ell \in [\![1, L]\!], \quad y_\ell = \varepsilon_\ell^{(m)} \widetilde{f}_\ell(\boldsymbol{\theta}) + \varepsilon_\ell^{(a)}, \tag{6.2}$$

with $\varepsilon_\ell^{(m)} \sim \mathsf{Lognormal}(-\sigma_m^2/2, \sigma_m^2)$ and $\varepsilon_\ell^{(a)} \sim \mathcal{N}(0, \sigma_{a,\ell}^2)$ for all lines ℓ. The likelihood approximation parameters $\boldsymbol{a}_\ell$ are optimized using the Bayesian optimization procedure described in Chapter 5 (Appendix 5.B). The proposed sampler is run for $T_{\mathsf{MC}} = 20\,000$ iterations including $T_{\mathsf{BI}} = 500$ of burn-in. The MTM kernel has a selection probability of $p_{\mathsf{MTM}} = 0.5$. As the observation only contains one pixel, the proposal q is set to the smoothed uniform prior distribution and the number of candidates to $K = 2\,000$. The PMALA step size is set to $\eta = 0.05$.

Model assessment and Bayesian p-value – The proposed sampling algorithm led to an estimated p-value of $\widehat{p}^{(T_{\text{MC}})} = 0.56$ and to a rejection probability of $\mathbb{P}\left[p^{(t)} \leq \alpha\right] < 10^{-200} < \delta$, which does not lead to a rejection. This indicates that the Meudon PDR code and the noise model are compatible with the observations.

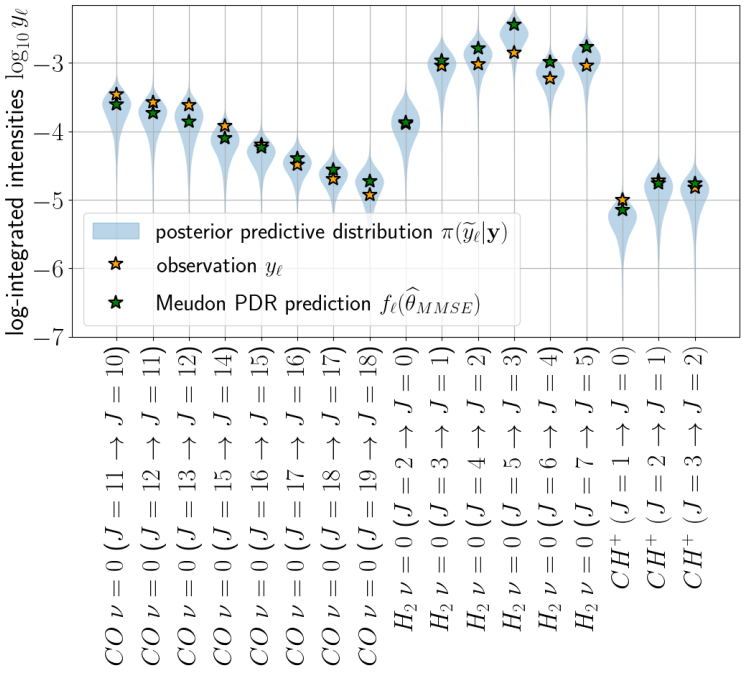
Figure 6.2 shows how the observations $\mathbf{y}$ compare to the marginal posterior predictive distributions $\pi(\widetilde{y}_{n\ell}|\mathbf{y})$ for lines used and not used for the inversion. It permits to visualize the compatibility between the observation $\mathbf{y}$ and the posterior distribution on the physical parameters $\boldsymbol{\theta}$. However, these plots are imperfect as the Gaussian additive noise can lead to negative predicted observations $\widetilde{y}_\ell$ in the low SNR regime. These negative values are not displayed. Underestimating the integrated intensity y_ℓ may thus not imply incompatibility between an observation y_ℓ and the observation model.

Figure 6.2a shows that for the L lines used for the inversion, the observations y_ℓ fall in high probability regions of the marginal predictive distributions $\pi(\widetilde{y}_{n\ell}|\mathbf{y})$, in agreement with the results of Joblin et al. (2018). In particular, the Meudon PDR prediction from the MMSE $\widetilde{\mathbf{f}}(\widehat{\boldsymbol{\theta}}_{\text{MMSE}})$ successfully reproduces all the lines at once. For lines that were not used in the inversion procedure, Figure 6.2b shows that this inversion reproduces well the low J lines of ^{12}CO, but underestimates the ^{13}CO and C^{18}O. This discrepancy was already noted in Joblin et al. (2018). It might be due to the fact that the considered grid of Meudon PDR code simulations does not include mutual radiative shielding between ^{12}CO and its isotopologues. Figure 6.2c shows satisfying compatibility for the marginals except for the O 3p $J = 1-2$ at 63 µm, which is overestimated by a factor $\simeq 10$. Joblin et al. (2018) showed the same overestimation for the O 63 µm line and underestimations for ^{13}CO and C^{18}O lines.

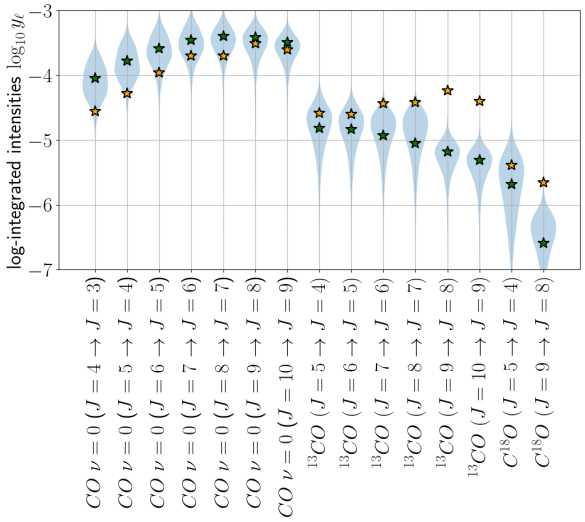
Inference results – Figure 6.3 shows the pairwise histograms of the posterior samples obtained with the proposed MCMC algorithm. We first note that posterior distribution indicates high uncertainties as it covers a large portion of the physical parameter space. The visual extinction A_V^{tot} is not constrained at all as the full interval $[1, 40]$ mag is covered. This is due to the fact that the observed lines are warm gas tracers and are thus emitted in the warm external layer of the PDR. Therefore, they do not trace the total column density of matter. However, the triplet $(\kappa,\, P_{\text{th}},\, G_0)$ is well constrained, although the individual credibility intervals are large. In particular, κ and G_0 are strongly anti-correlated. The G_0 thus has a large uncertainty. However, its MMSE and posterior mode are of a few 10^3, which is compatible with past estimations. The thermal pressure P_{th} is well constrained at $10^8\,\text{K}\,\text{cm}^{-3}$.

Joblin et al. (2018) identified a positive correlation among multiple sources between estimated P_{th} and G_0. This correlation could have been caused by a positive correlation in uncertainties on individual (P_{th}, G_0) estimations. Figure 6.3 shows a negative slope in the (P_{th}, G_0) joint histogram, which indicates that the correlation in uncertainties is negative. Figure 6.21 in Appendix 6.A shows that a similar inference on Orion bar observations leads to the same negative correlation in uncertainties. Therefore, the positive correlation among multiple sources should have a physical origin, as hypothesized in Joblin et al. (2018).

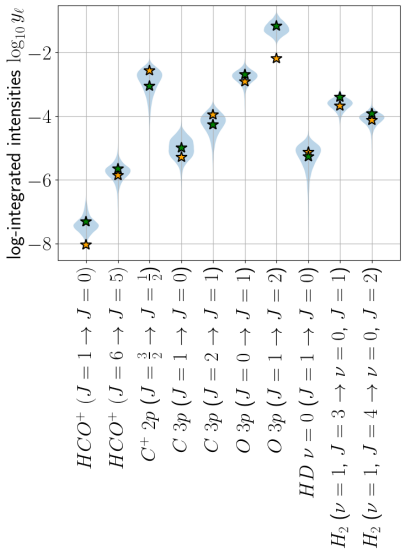
Table 6.1 shows the physical parameters estimated values and their credibility intervals. For P_{th}, G_0 and A_V^{tot}, the values from Joblin et al. (2018) fall in the credibility intervals, which means that the two estimations are consistent. In addition, the MMSE is close to the values from Joblin et al. (2018). However, this is mostly a coincidence considering how spread the posterior distribution is over the parameter space. The scaling parameter κ is the only parameter for which the two estimations are incompatible. In principle, κ corresponds to observational effects such as the beam filling factor and the inclination of the PDR with respect to the line of sight. However, the forward models differ in the two estimations. Joblin et al. (2018) relies on the version 1.5.4 of the Meudon PDR code, while we use an emulator of version 1.7 which contains better implementation of physical processes and updated atomic and molecular data. Besides, Joblin et al. (2018) exploited dust grain properties corresponding to dense gas, while we used average galactic grain properties. These differences appear to result in a multiplicative



(a) Lines used for the inversion.



(b) Additional lines of CO and isotopologues, unseen during the inversion.



(c) Additional lines of other molecules, atoms and ions, unseen during the inversion.

Figure 6.2: Posterior predictive assessment for NGC 7023. Comparison of observations $\mathbf{Y}$ and associated noise model with posterior predictive distributions on $\widetilde{\mathbf{f}}(\mathbf{\Theta})$, with $\mathbf{\Theta} \sim \pi(\mathbf{\Theta}|\mathbf{Y})$.

factor for the lines used in the inversion that was compensated for by κ.

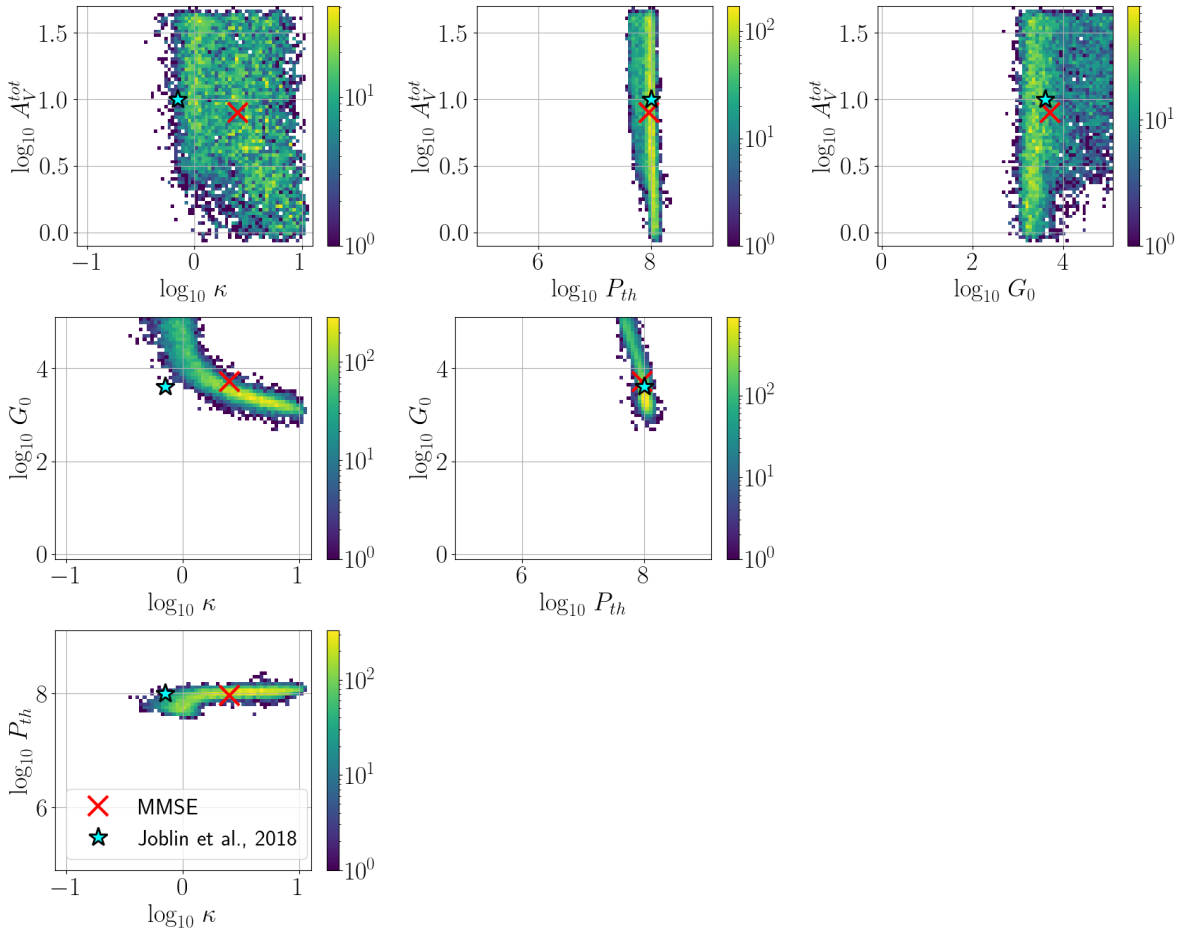


Figure 6.3: Inference results for NGC 7023. Two-dimensional marginal histograms in the physical parameters $\boldsymbol{\Theta}$ space. All histograms are in logarithmic norm.

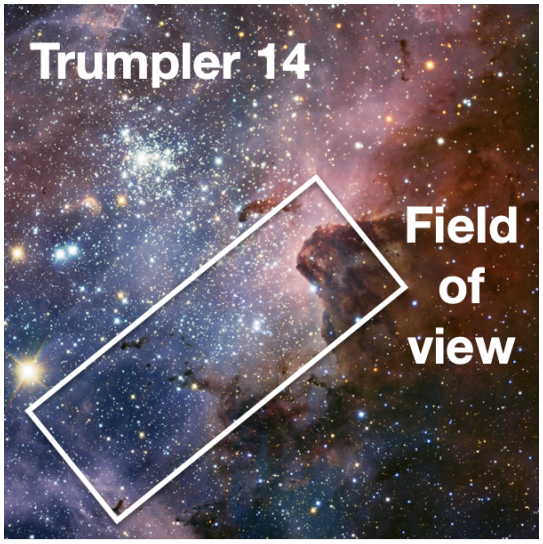
Table 6.1: Inference results on NGC 7023. Joblin et al. (2018) only infers κ and P_{th}.

		κ –	P_{th} (K cm^{-3})	G_0 –	A_V^{tot} (mag)
(Joblin et al., 2018)		0.7	1×10^8	2.6×10^3	10
MMSE		2.9	9.7×10^7	4.0×10^3	7.3
68% credibility interval	lower bound $l_{68\%}$	1.2	8.1×10^7	1.8×10^3	2.2
	upper bound $u_{68\%}$	6.0	1.2×10^8	1.0×10^4	24.9
	UF$_{68\%}$	2.2	1.2	2.4	3.4
95% credibility interval	lower bound $l_{95\%}$	0.8	4.7×10^7	1.3×10^3	1.1
	upper bound $u_{95\%}$	9.0	1.5×10^8	8.9×10^4	41.2
	UF$_{95\%}$	3.3	1.8	8.2	6.1

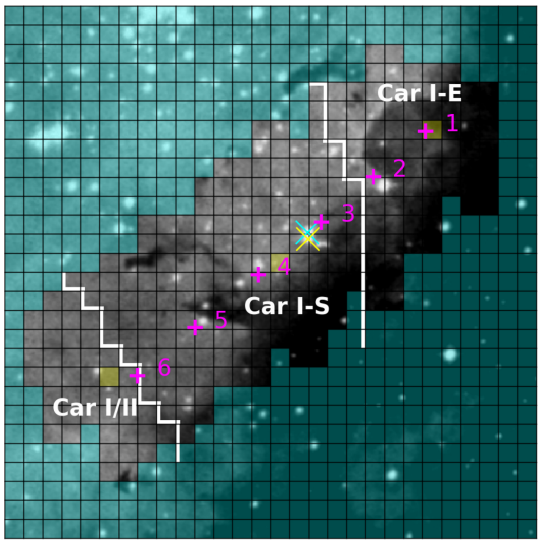
6.3 The Carina nebula

Having tested the method on known single-pixel observations, we now move to multi-pixel maps. In this section, we perform inference on an already studied observation of the Carina nebula. This analysis aims at completing the validation of the proposed method, at showing the relevance of the spatial regularization and of the sampling approach compared to optimization.

The Carina nebula is located in the Milky Way, at an estimated distance of $2.4\,$kpc from the Sun (Smith, 2006). Illuminated by the massive star clusters Trumpler 14 and Trumpler 16, it is the brightest nebula in the sky of the Southern hemisphere. It includes two major components, Car I and Car II, separated by $\simeq 7\,$pc. The structure of Car I is farther divided into three bright regions: Car I-W, Car I-E, and Car I-S, located in the west, east and south of Car I, respectively (Whiteoak, 1994). Wu et al. (2018) studied the CO and C emission observed by Herschel SPIRE/FTS. Figure 6.4 shows the field of view, which includes Car I-E, Car I-S, and a region at the intersection of Car I and Car II (Car I/II).



(a) Localization of the field of view with respect to the Trumpler 14 star cluster.



(b) Effective field of view (gray pixels). The background image is taken in the DSS2-red band. The magenta crosses indicate the pointings used during observation. The blue and yellow crosses pinpoint two massive stars. Taken from Wu et al. (2018).

Figure 6.4: Observation of the Carina nebula.

The G_0 in the Car I-E region was estimated of the order of 10^4 from stellar composition and FIR observations (Brooks et al., 2003; Mizutani et al., 2004). However, estimations with PDR models led to lower values: $G_0 = 1\,390$ according to Oberst et al. (2011) and $G_0 = 3\,200$ according to Kramer et al. (2008).

In Wu et al. (2018), the authors inferred multiple parameters – including G_0 – from $L = 12$ lines, including 10 ^{12}CO lines (from $J = 4 \to 3$ to $J = 13 \to 12$) and 2 atomic C lines (3p, $J = 1 \to 0$ and $J = 2 \to 1$). They used a RBF interpolation of a grid of the version 1.5.4 of the Meudon PDR model with PAHs in the dust population. The noise model involved an additive, uncorrelated Gaussian noise. The prior distribution relied on an indicator term with wide validity intervals on the physical parameters and on an ad hoc physics-informed regularization term. This ad hoc regularization term enforced the reconstructed integrated intensities to be decreasing for the CO lines $J = 11 \to 10$, $J = 12 \to 11$, and $J = 13 \to 12$, as this condition seems always satisfied in observations. This ad hoc constraint is not generalizable to observations of other lines that the high J rotational lines of ^{12}CO. The inversion estimated the maximum a posteriori (MAP) using a preconditioned gradient descent algorithm with the limited memory BFGS preconditioner. In the Car I-E region, they obtained $G_0 \simeq 2 \times 10^4$.

Inversion setup – The inference is performed with the same $L = 12$ lines as in Wu et al. (2018). Figure 6.5 shows the integrated intensity maps $\mathbf{Y}$. These maps contain $N = 176$ pixels.

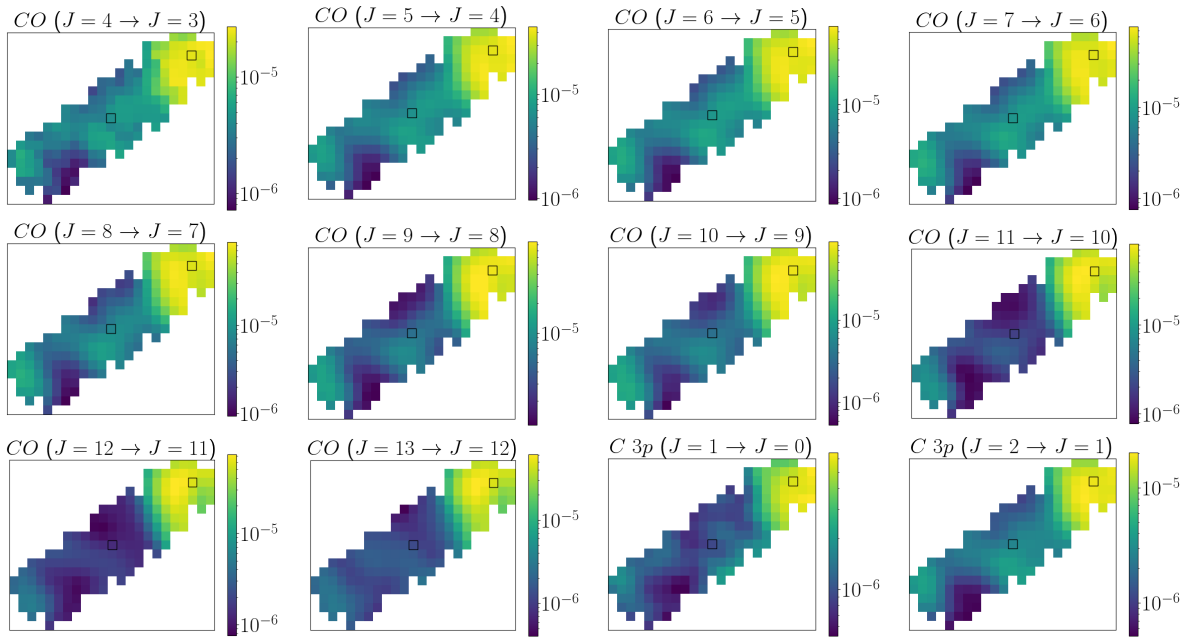


Figure 6.5: Observations of the Carina nebula. The Car I-E is brightest for all lines, as it is directly illuminated by the Trumpler 14 star cluster.

The inference is performed with the data used in Wu et al. (2018). The standard deviations $\sigma_{a,n\ell}$ of the additive noise already include multiplicative errors. As the signal-to-noise ratio is low in this map, we decided not to include additional multiplicative noise in the observation model. Besides, some standard deviations $\sigma_{a,n\ell}$ are missing. To avoid biasing the reconstructed maps, the associated observations $y_{n\ell}$ are disregarded for the inversion[2]. The resulting observation model is

$$y_{n\ell} = \widetilde{f}_\ell(\boldsymbol{\theta}_n) + \varepsilon^{(a)}_{n\ell}, \quad \varepsilon^{(a)}_{n\ell} \sim \mathcal{N}(0, \sigma^2_{a,n\ell}). \tag{6.3}$$

The spatial regularization parameter is set to $\tau_d = 1$ for each of the $D = 4$ physical parameters. We tried multiple values between 0.1 and 10 for each parameter, and selected the value that led

[2]The associated observations are disregarded by setting $\sigma_{a,n\ell} = 1\,\mathrm{erg\,cm^{-2}\,s^{-1}\,sr^{-1}}$, an arbitrarily high value.

to physically consistent maps that reproduce the structures in the observations. Note that we do not resort to the same prior as in Wu et al. (2018): the article used a line-specific pixel-wise ad hoc constraint on the predicted integrated intensities, while we rely on a more generally applicable spatial regularization term.

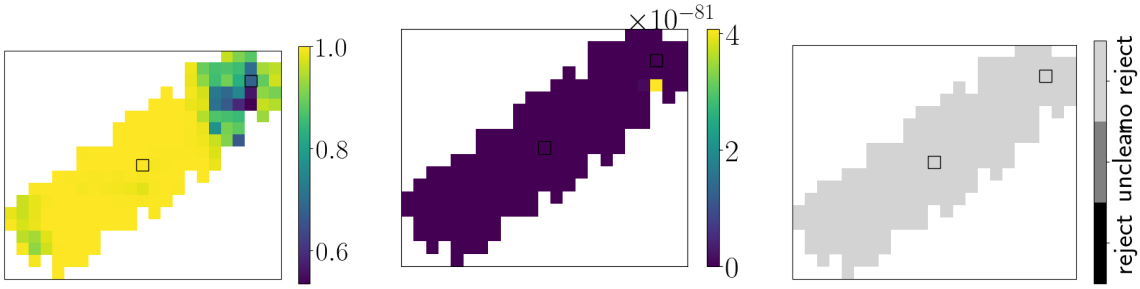
The observation angle φ varies in the field of view, as the Car I-E is considered to be edge-on while the Car I-S is face-on. For optically thin lines such as excited ^{12}CO lines, φ is degenerated with the scaling parameter κ. Therefore, we chose not to infer it, and set it to $\varphi = 0$ deg.

As this use case is the first map of real observations, we compare the maximum likelihood estimator (MLE) and the MMSE. The MLE is evaluated for comparison, as a naive estimator. It is the estimator most compatible with the observations for each pixel, and is thus very sensitive to noise. In particular, it does not exploit the spatial regularization and is not accompanied by an uncertainty quantification. The goal is to demonstrate the interest of the spatial regularization.

The MLE is evaluated using the optimization algorithm adapted from the proposed sampler mentioned in Chapter 5 (Section 5.2.3), run for 500 iterations. This variant combines two update steps. The first update step is a preconditioned gradient descent step with the RMSProp preconditioner. It was run with the step size $\eta = 0.03$. At each step t, it was selected with probability $1 - p_{\text{MTM}} = 0.99$. The second update step, selected with probability $p_{\text{MTM}} = 0.01$, is a simulated annealing-like variant of the MTM kernel. Unlike gradient descent algorithms, this second update step enables to escape from local minima. As the MLE does not include the spatial regularization, the proposal q is set to the smoothed uniform prior. It generates $K = 200$ candidates.

The MMSE and the credibility intervals are obtained with the approach we proposed in this thesis. Both are evaluated from a Markov chain with $T_{\text{MC}} = 10\,000$ iterates, including 500 of burn-in. The PMALA kernel relies on a step size $\eta = 0.003$. The MTM kernel is selected with probability $p_{\text{MTM}} = 0.5$. It generates $K = 20$ candidates from the proposal q described in Chapter 5 (Eq. 5.35).

Model assessment and Bayesian p-value – Figure 6.6 shows the maps of estimated p-values $\widehat{p}_n^{(T_{\text{MC}})}$, of rejection probabilities $\mathbb{P}[p_n \leq \alpha]$ and of the resulting rejection decision. All the estimated p-values are above 0.5. This indicates a good fit between the observations and the observation model. When accounting for the uncertainty on the p-values estimation, the highest model rejection probability is 4×10^{-81}. None of the pixels was rejected.



(a) Map of estimated p-values (b) Map of rejection probabilities (c) Decision for each pixel

Figure 6.6: Model assessment on the Carina observations from Wu et al. (2018).

Figure 6.7 compares the observations with the marginal posterior predictive distributions for the two highlighted pixels. For the Car I-E pixel, the bright ^{12}CO lines are well reproduced. However, the predicted integrated intensities for the atomic C lines differ from the true observations by up to a factor 5. Yet, these values for the C lines are considered compatible as the standard deviations of the additive noise are large. In the Car I-S pixel, the ^{12}CO lines are not as bright as in the Car I-E, and are roughly as bright as atomic C lines. The errors between predicted integrated intensities and observations are thus more uniformly distributed among ^{12}CO lines and

atomic C lines.

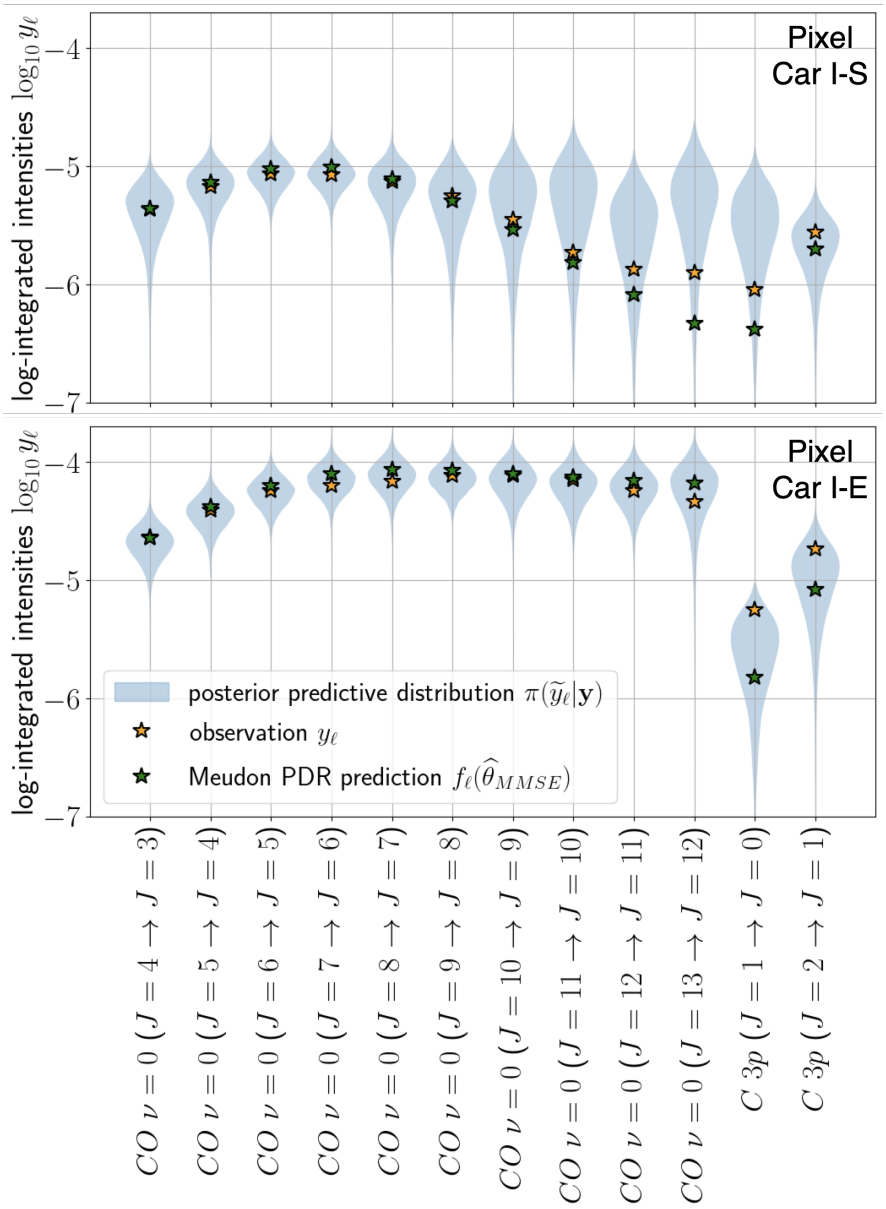


Figure 6.7: Posterior predictive assessment two pixels of the Carina nebula on the lines used for the inversion. Comparison of observations $\mathbf{Y}$ and associated noise model with posterior predictive distributions on $\widetilde{\mathbf{f}}(\mathbf{\Theta})$, with $\mathbf{\Theta} \sim \pi(\mathbf{\Theta}|\mathbf{Y})$. Note that these plots are imperfect as the Gaussian additive noise can lead to negative predicted observations $\widetilde{y}_\ell$ in the low SNR regime. These negative values are not displayed. For the lower intensities, underestimating the integrated intensity y_ℓ may thus be compatible with the additive Gaussian noise model.

Inference results: MLE, MMSE and UF – Figure 6.8 shows the obtained MLE, MMSE and UF. It also shows the estimated thermal pressure P_{th} and incident radiative field intensity G_0 maps from Wu et al. (2018). The maps of scaling factor κ and of visual extinction A_V^{tot} were not displayed in the article.

The MLE is visually not good. It presents unrealistic values and variations especially in P_{th} and A_V^{tot} because of its sensitivity to noise. Therefore, it cannot be exploited to extract meaningful information on the physical parameters in the observed region.

Conversely, the MMSE is smooth and spatially consistent. The maps of P_{th} and G_0 show a clear frontier between the Car I-E and Car I-S regions, with an increase by a factor $\sim 3-5$ from

Car I-S to Car I-E. They also show a blurry frontier between the Car I-S and Car I/II regions, with higher P_{th} and G_0 in the Car I/II region. These lower values in the Car I-S region are due to the fact that this region is farther in the back than the other two, and thus farther away from the Trumpler 14 star cluster. The scaling parameter κ ranges from 0.1 to 1.2, with its highest values attained in the Car I-E region. The lower values may correspond to smaller and spatially unresolved PDR edges. The higher values may be due to the more edge-on geometry and to a higher beam filling factor. Finally, the observed lines do not trace the total visual extinction for deep clouds as they are emitted mostly from the warm surface of the cloud. It is therefore expected not to recover well the spatial structure in the A_V^{tot} map. Figures 6.10 and 6.9 show pairwise two-dimensional histograms of the posterior distribution samples. Unlike in NGC 7023, there is no strong degeneracy between κ and G_0, although there is a negative correlation between the two parameters. In both pixels, there is also a negative correlation between P_{th} and G_0. In the bright Car I-E region, this correlation is higher, which might be due to the higher SNR and constraining power of the observed lines.

The UF maps show that the uncertainty is on average lower in the bright Car I-E region. This is due to the larger signal-to-noise ratio in this region. It also illustrates the constraining power of the spatial regularization. Indeed, on each map, the pixels with highest UF are on the border of the image and with only one neighbor.

The inferred maps from Wu et al. (2018) are spatially very similar to the MMSE. However, the G_0 values differ. In the bright Car I-E region, the article found $G_0 \simeq 2 \times 10^4$, while in the MMSE $G_0 \simeq 5 \times 10^3$. Figure 6.10 shows that the point from the article falls in a low posterior density region, even though it is close to the posterior mode. This difference is due to the Meudon PDR code version change – Wu et al. (2018) relied on version 1.5.4 with polycyclic aromatic hydrocarbons (PAH) while we resort to version 1.7 with its standard dust population, which does not include PAHs. To check whether the differences came from the difference in the model used and not from the inversion procedure, we trained a neural network emulator of the grid that was used in the article and performed the inversion again. The G_0 values in the obtained MMSE maps were similar to those of Wu et al. (2018).

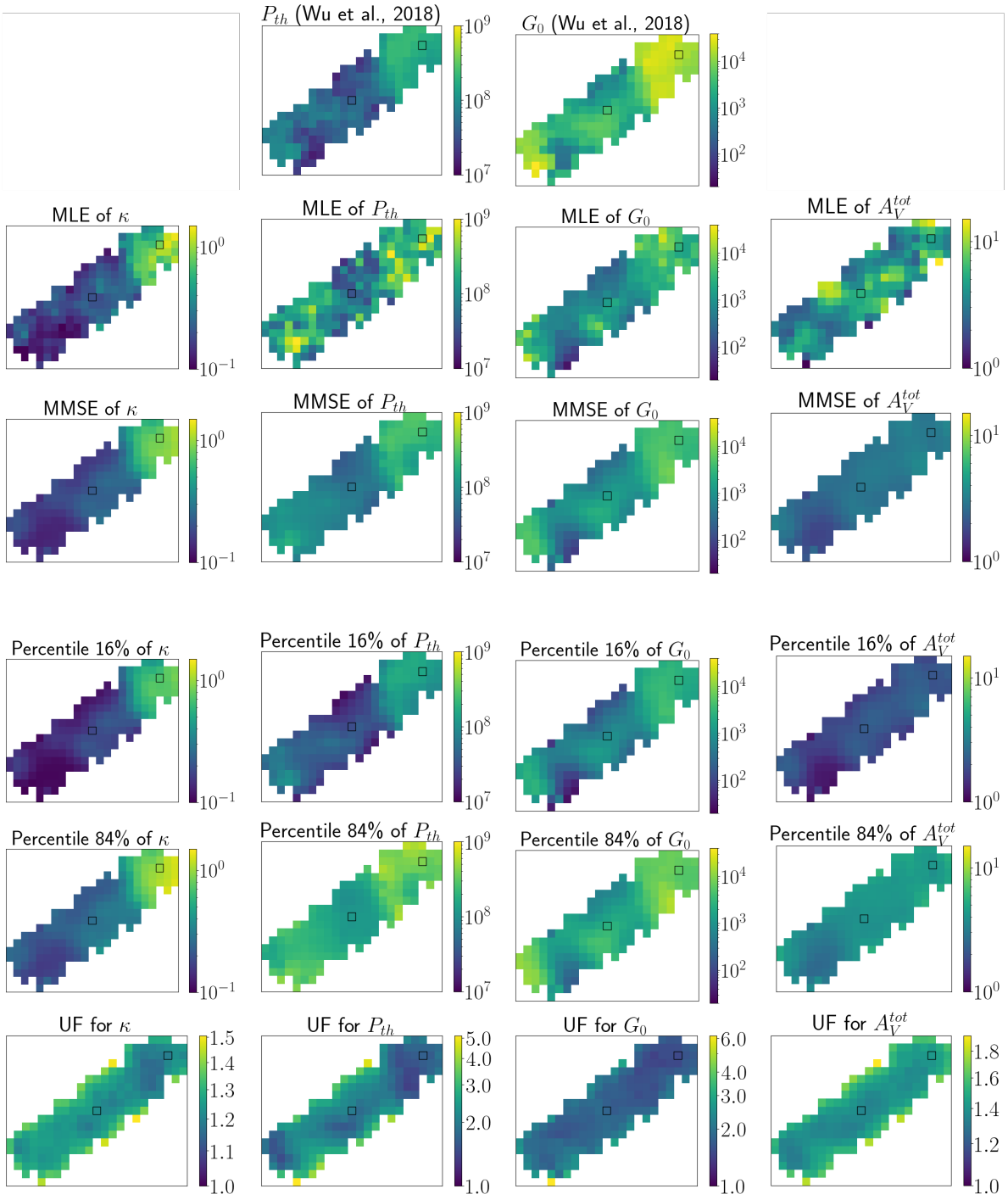


Figure 6.8: Inference results for the Carina nebula.

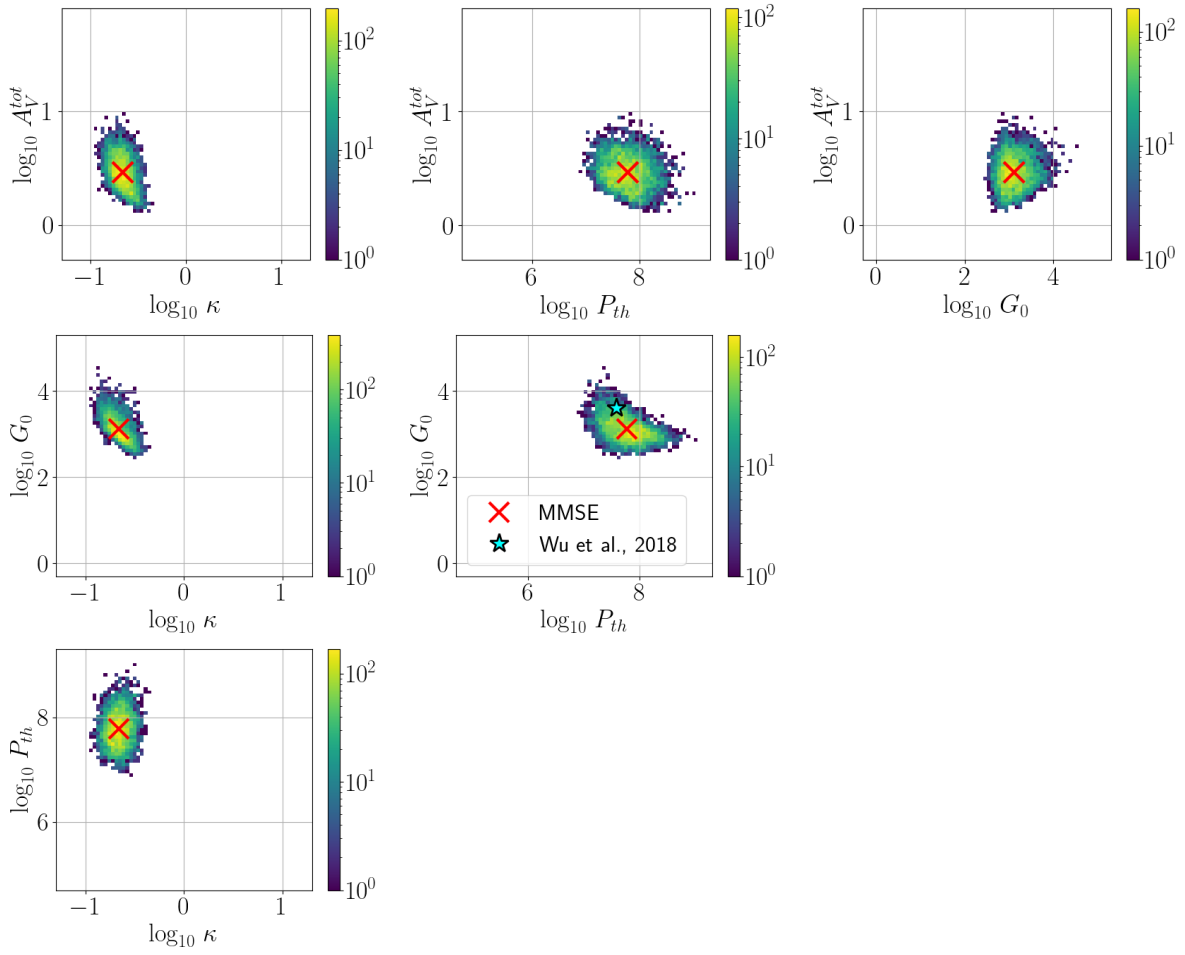


Figure 6.9: Inference results for the Car I-S pixel of the Carina nebula. Two-dimensional marginal histograms in the physical parameters $\boldsymbol{\Theta}$ space. All histograms are in logarithmic norm.

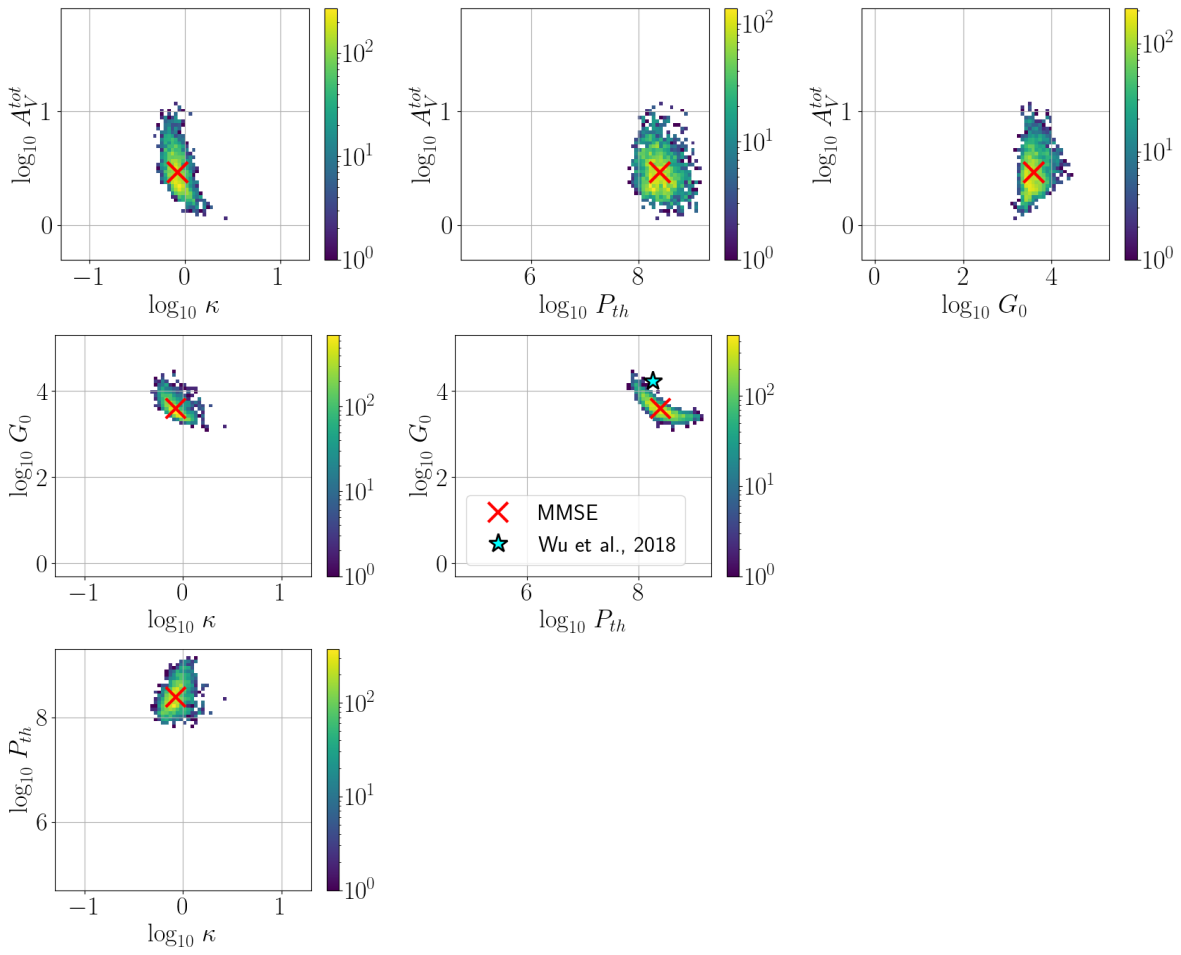
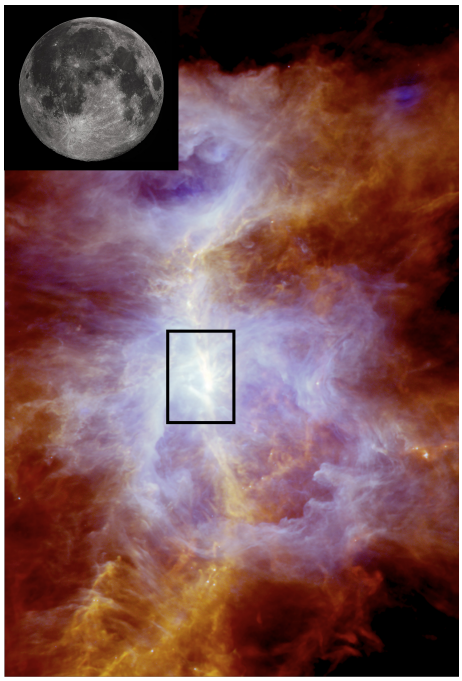


Figure 6.10: Inference results for the Car I-E pixel of the Carina nebula. Two-dimensional marginal histograms in the physical parameters $\boldsymbol{\Theta}$ space. All histograms are in logarithmic norm.

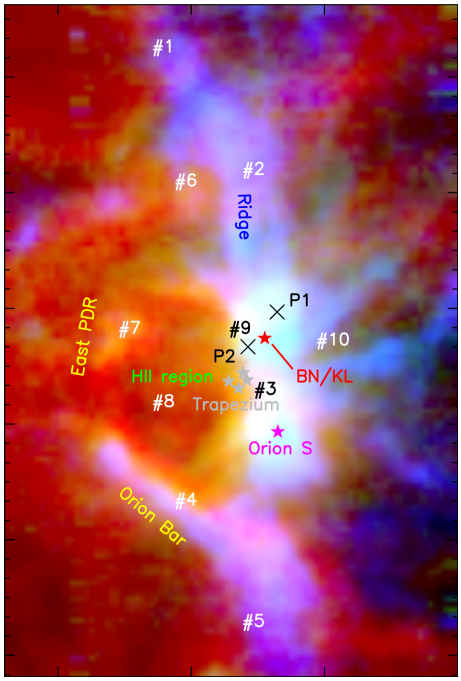
6.4 Orion molecular cloud 1 (OMC-1)

The first two applications considered observations that had already been previously studied with the Meudon PDR code. In this section, we apply our inference procedure on the Orion molecular cloud 1 (OMC-1) far infrared and millimetric maps described in Goicoechea et al. (2019). This inference is the first performed on the full OMC-1 map. Unlike in the two previous applications, it is not clear a priori whether the considered lines can constrain the physical parameters.

The OMC-1 is a bright region of the Orion A cloud. It is located at about $414\,\mathrm{pc}$ (Menten et al., 2007) from the Sun, making it the closest region of star formation for stars of intermediate and high masses. Figure 6.11 shows its position in Orion A and its general structure. OMC-1 is illuminated by the Trapezium star cluster, which contains many heavy stars. The cluster lies within the HII bubble. It heats and photodissociates the Orion bar and the East PDR. The two major star-forming sites are the Becklin-Neugebauer/Kleinmann-Low region (BN/KL) and Orion South (Orion S). The Orion bar is one of the most famous and studied PDRs.



(a) The Orion A star-formation cloud seen by ESA's Herschel space observatory. The black rectangle outlines the observed region of OMC-1, centered on the Trapezium star cluster. The image is a composite of the wavelengths of $70\,\mu\mathrm{m}$ (blue), $160\,\mu\mathrm{m}$ (green) and $250\,\mu\mathrm{m}$ (red), i.e., dust emission. It spans about 1.3×1.6 deg^2. The moon is shown for scale. Credits: ESA/Herschel/Ph. André, D. Polychroni, A. Roy, V. Könyves, N. Schneider for the Gould Belt survey Key Programme.



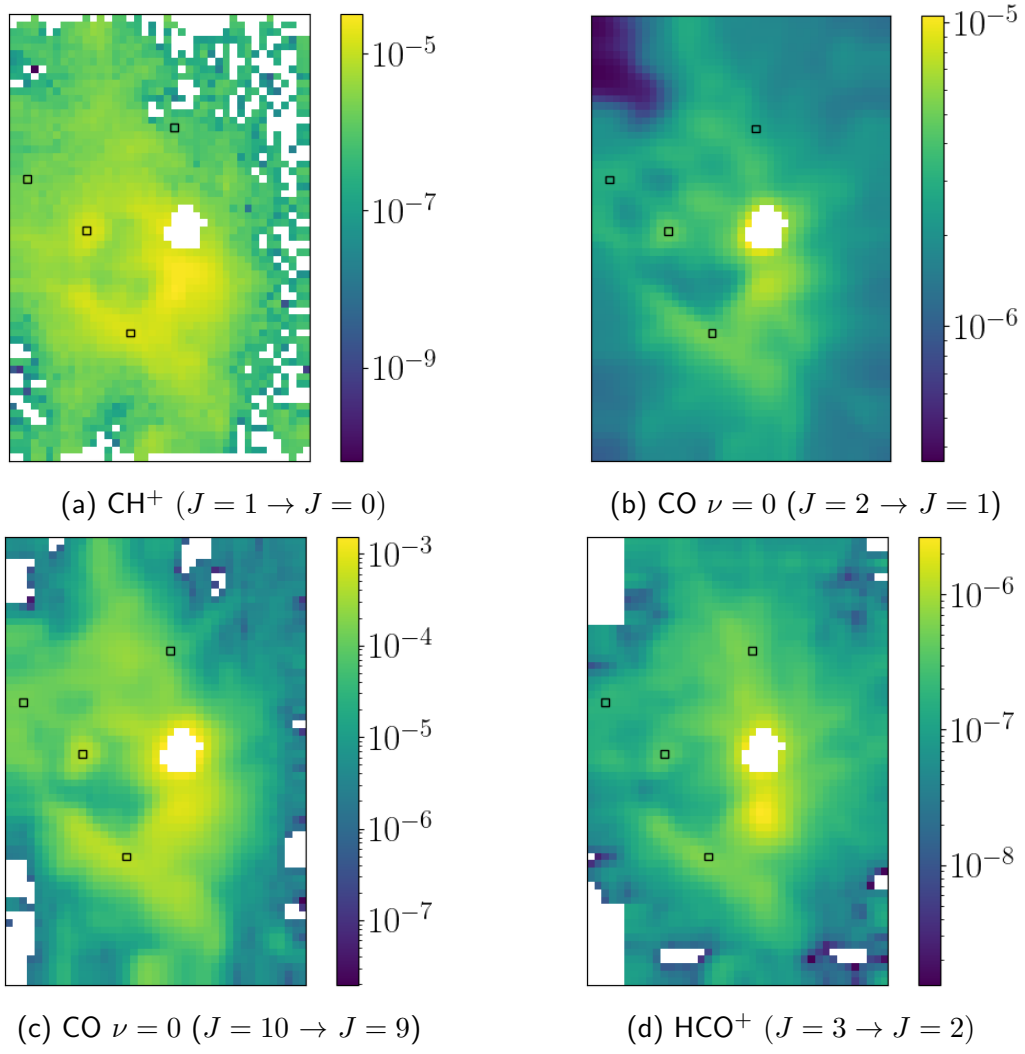
(b) OMC-1 composite image covering ~ 85 arcmin^2 at ~ 12" with three emission lines: the C$^+$ $158\,\mu\mathrm{m}$ line which traces the FUV-illuminated surface of the molecular cloud (red), the C^{18}O, $J = 2 \rightarrow 1$ line which traces cold dense gas (blue) and the HCO$^+$ $J = 3 \rightarrow 2$ line (green). Taken from Goicoechea et al. (2019).

Figure 6.11: Structure of the OMC-1 cloud.

Inversion setup – The considered observation $\mathbf{Y}$ contains $L = 4$ molecular emission lines: CH$^+$ $(J = 1 \rightarrow 0)$, CO $(J = 2 \rightarrow 1)$, CO $(J = 10 \rightarrow 9)$, and HCO$^+$ $(J = 3 \rightarrow 2)$. Other lines such as C$^+$ $158\,\mu\mathrm{m}$ and ^{13}CO $J = 2 \rightarrow 1$ were measured. We tried performing inversion with

these lines, which resulted in an incompatibility between the models and observations. Therefore, these lines are removed for this analysis.

Figure 6.12 shows the observation maps associated with the $L = 4$ lines used in the inversion. The BN/KL region is masked (in white, in the center of the maps) as it is known to be dominated by shocks. As such events are not included in PDRs models, this region is removed for inversion to avoid biasing the neighboring pixels with the spatial regularization. Dedicated models such as the Paris-Durham code (Godard et al., 2019) would better simulate this region. Note that the model assessment approach might have rejected the model for these pixels. Four pixels are highlighted in this figure: in the Orion bar (lowest square), in the East PDR (mid-height square), in the North-East border of the map (top left square) and in the North-West ridge edge (top right square). Results on these pixels will be detailed.



(a) CH^+ $(J = 1 \to J = 0)$ (b) CO $\nu = 0$ $(J = 2 \to J = 1)$

(c) CO $\nu = 0$ $(J = 10 \to J = 9)$ (d) HCO^+ $(J = 3 \to J = 2)$

Figure 6.12: Observations of OMC-1 used for inversion. The white region is the middle of the maps is the BN/KN region, dominated by shocks, and not considered in the inversion. The remaining white pixels correspond to negative integrated intensities due to the additive Gaussian noise. The inference results will be detailed for the four pixels highlighted with a black square.

The observation model is identical to Eq. 6.2. The integrated intensity maps $\mathbf{Y}$ and the associated additive noise standard deviations $(\sigma_{a,n\ell})_{n\ell}$ are computed directly from the hyperspectral cubes. The multiplicative noise is modeled as a lognormal distribution that combines two terms. The model misspecification is set so that a 3σ error corresponds to an error of a factor 3, i.e., $\sigma_{\text{mod}} = \frac{1}{3}\ln 3 \simeq \ln 1.44$. An estimated 8% calibration error is considered for the whole map, i.e., $\sigma_c = \ln 1.08$. The total multiplicative standard deviation is $\sigma_m = \sqrt{\sigma_{\text{mod}}^2 + \sigma_c^2} \simeq 0.373 \simeq$

$\ln 1.452$. The parameters $\boldsymbol{a}_\ell$ of the likelihood approximation should be optimized for each pixel, which is very computationally demanding for $N > 10^3$ pixels. For simplicity, we set $\boldsymbol{a}_\ell$ to high values so that only the additive Gaussian approximation is used. After testing different values, the spatial regularization parameter $\boldsymbol{\tau}$ is set to 1 for the four physical parameters to obtain smooth and physically consistent maps.

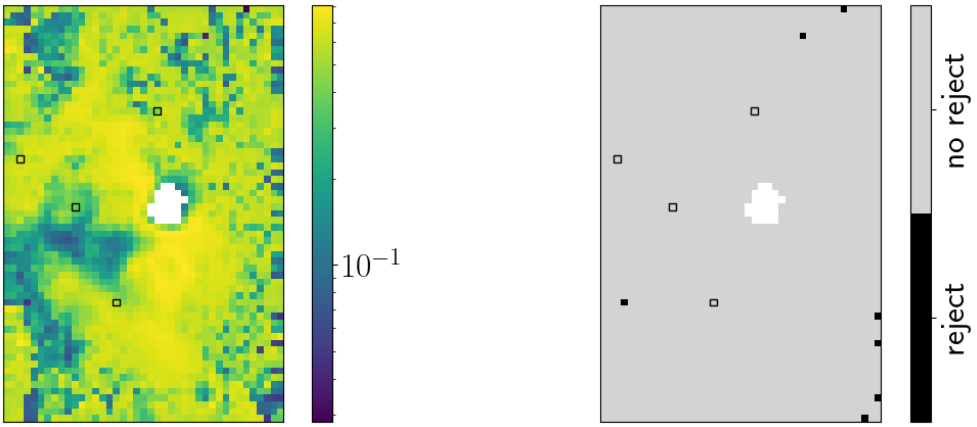
The observation angle φ between the surface of the cloud and the line of sight varies in the observed map. For instance, the Orion South (Orion S) region is considered to be mostly face-on ($\varphi \sim 0$ deg) while the Orion bar is close to edge-on ($\varphi \simeq 90$ deg). To avoid inferring more parameters than observed lines, we set $\varphi = 0$ deg. Like in the Carina nebula inversion, we assume that the inclination effect will be approximately captured by the scaling factor κ.

Like for the Carina nebula, we compare the MLE, as a naive estimator yet widespread in astrophysics – see Chapter 3 –, and the MMSE from the method proposed in this thesis. The MLE does not exploit spatial regularization and handles all pixels independently. It is obtained with the optimization algorithm used for the Carina nebula run for 500 iterations. At each step, the optimization variant of the MTM kernel is used with probability $p_{\text{MTM}} = 0.02$. The proposal distribution is set to the smooth uniform prior, and the number of candidates to $K = 50$. The preconditioned gradient descent update step size is set to $\eta = 0.03$.

The MMSE is evaluated from a Markov chain of $T_{\text{MC}} = 10\,000$ iterates, including $T_{\text{BI}} = 500$ of burn-in, generated with the proposed MCMC algorithm. At each iteration t, the MTM kernel is selected with probability $p_{\text{MTM}} = 0.5$. It generates $K = 50$ candidates for each pixel $\boldsymbol{\theta}_n$. As in the Carina inversion, the proposal q is set to the Gaussian mixture defined with the neighboring pixels described in Chapter 5 (Eq. 5.35). The PMALA step size is set to $\eta = 0.1$.

Model assessment and Bayesian p-value – Figure 6.13 shows the map of estimated p-values $(\widehat{p}_n^{(T_{\text{MC}})})_{n=1}^N$. As computing the effective sample size (ESS) is costly for large observation maps, the uncertainties due to the Monte Carlo estimation of the p-values is not accounted for. However, Figure 6.13a shows that the estimated p-values is larger than 0.5 on the regions of interest of the maps. As a result, as shown in Figure 6.13b, only 7 pixels out of $2\,475$ lead to a model rejection, i.e., less than 0.3%. In particular, none of the 4 highlighted pixels lead to a model rejection.

Figure 6.14 shows that the predicted integrated intensities are very close to the actual observations. The only exception is the CH^+ $J = 1 \to 0$ emission line for the pixel in the North-West ridge edge, where the MMSE is a factor $\sim 3 - 5$ too high. Note that the two pixels from PDRs contain brighter lines than the two other pixels, which is also visible in Figure 6.12.



(a) Estimated p-value map. (b) Associated decision.

Figure 6.13: Model assessment for OMC-1.

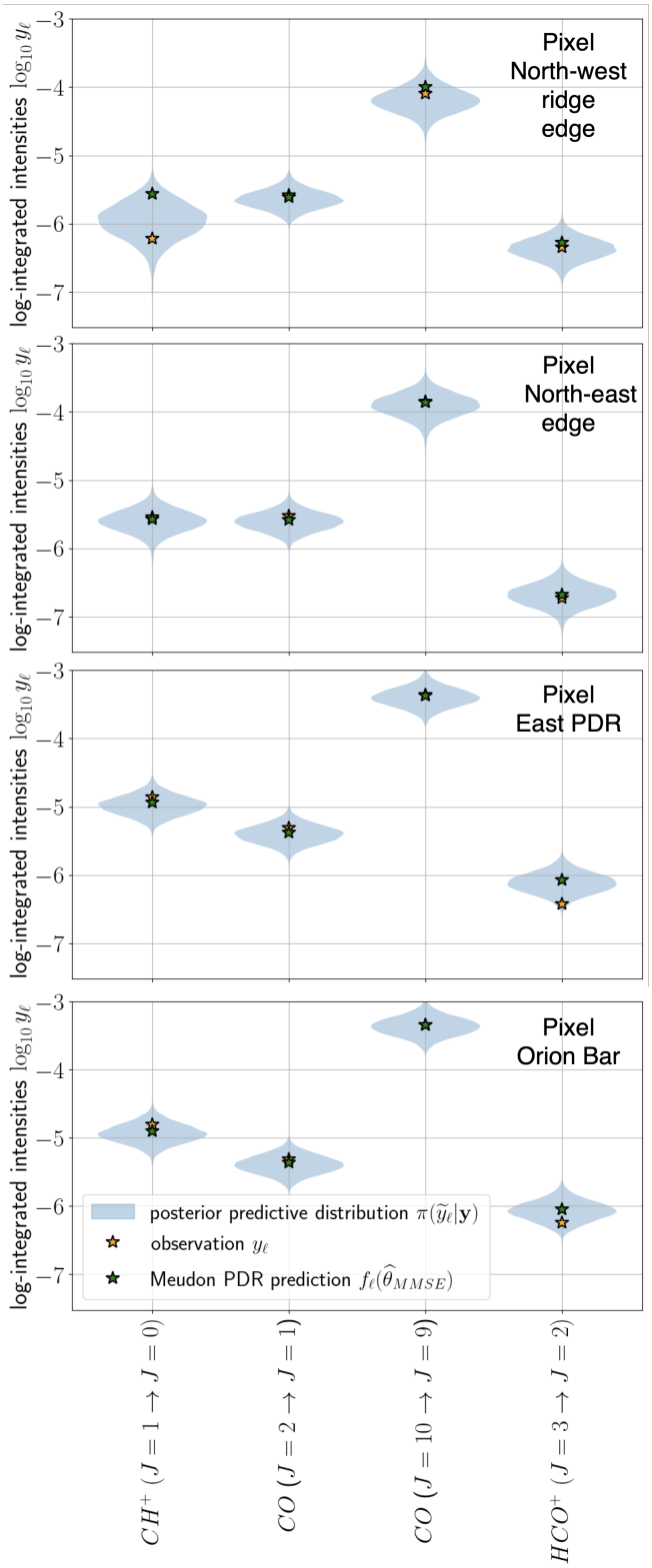


Figure 6.14: Posterior predictive assessment for four pixels of OMC-1 on the lines used for the inversion. Comparison of observations $\mathbf{Y}$ and associated noise model with posterior predictive distributions on $\widetilde{\mathbf{f}}(\mathbf{\Theta})$, with $\mathbf{\Theta} \sim \pi(\mathbf{\Theta}|\mathbf{Y})$.

Inference results: MLE, MMSE, and UF maps – Figure 6.15 shows inference results for both likelihood minimization and posterior sampling. The spatial structures in the observations are recovered with the MMSE when including the spatial smoothness prior, but hardly visible when using a simple MLE. Due to the low number of lines L, the spatial regularization plays a key role in recovering the correct spatial structure. As there are few lines and thus few constraints for the models, the likelihood is very sensitive to noise and the MLE presents unphysical variations. Conversely, the spatial regularization permits each pixel to access the information contained in its neighbors, which share similar physical conditions.

We now focus on the MMSE and UF maps. The MMSE permits to recover the overall structure of OMC-1. The Orion bar, the HII region, the Orion S region, the East PDR and the North-West ridge are visible on each of the $D = 4$ reconstructed maps. The scaling factor κ is close to 1 in the bright regions. In the Orion S region, in the East PDR and in the Orion bar, the thermal pressure P_{th} is estimated at $\simeq 4 \times 10^8\,\text{K}\,\text{cm}^{-3}$, and the radiative intensity factor G_0 at $\simeq 5 \times 10^4$. In these regions, the recovered visual extinction A_V^{tot} is roughly 10 mag. These values are consistent with the literature and with the analysis of the one-pixel observation of the Orion bar presented in Appendix 6.A. Our results show for the first time that the high thermal pressures previously found in the Orion Bar appear to be widespread in the PDRs of OMC-1.

Over the whole map, the UF is lower than 1.5 for the scaling factor κ, than 2 for the visual extinction A_V^{tot}. This means that the typical error is at most 50% for κ, and of a factor 2 for A_V^{tot}. Despite using only $L = 4$ lines to infer $D = 4$ parameters for each pixel, these two parameters are thus well constrained. Similarly, on most of the map, the UF is lower than 3 for the thermal pressure P_{th} and than 4 for the intensity of the incident radiative field G_0. These two parameters are thus relatively well constrained considering their large validity intervals and the low number of lines used for inversion ($L = 4$). The UF maps of P_{th} and G_0 show that the darker regions yield higher associated uncertainties. Besides, the UF attains a maximum in the North-West ridge edge, around one of the highlighted pixels.

Inference results: analysis of highlighted pixels – Figure 6.16 shows the pairwise histograms for the Orion bar pixel (lowest of the four squares). Figure 6.17 shows the pairwise histograms for the East PDR pixel (mid-height square). They show that the marginal posterior distributions for these two pixels are simple, as they could be well approximated by a Gaussian distribution with a diagonal covariance matrix. Uncertainties are large but do not show degeneracy between parameters. Figure 6.18 shows the North-East pixel (top left square). This pixel and its surrounding region yields a more moderate P_{th} ($\simeq 7 \times 10^7\,\text{K}\,\text{cm}^{-3}$) while the G_0 remains similar to what is found in the Orion Bar ($\simeq 5 \times 10^4$). This seems to contradict the correlation between P_{th} and G_0 described in Joblin et al. (2018) and Wu et al. (2018). Further investigation is thus required to better understand this specific region.

Finally, Figure 6.19 shows that the UF peak around the North-West ridge edge pixel (top right pixel) corresponds to a bimodality in the joint distribution (P_{th}, G_0). The sampler successfully identified the two modes in the (P_{th}, G_0) joint distributions, which is challenging for standard MCMC algorithms. The two modes correspond to two different environments. One has a very high $G_0 \simeq 3 \times 10^4$ and a moderate $P_{\text{th}} \simeq 3 \times 10^7\,\text{K}\,\text{cm}^{-3}$. The other, which is more likely, has a moderate $G_0 \simeq 5 \times 10^3$ and a very high $P_{\text{th}} \simeq 3 \times 10^8\,\text{K}\,\text{cm}^{-3}$. Different hypotheses may explain this bimodality. For instance, the existence of two solutions may be due to the low number of lines $L = 4$. In this case, adding other lines for the inversion may remove one mode. Another possibility is that this pixel mixes emissions from distinct environments at different positions along the line of sight. This is also possible as this pixel is on the edge of a ridge. Note that although the MMSE values represented on the reconstructed maps fall in a high probability region, it does not fall in a mode. In such a case, the MAP would be a more representative estimator of maps of P_{th} and G_0, but would require running an additional optimization algorithm.

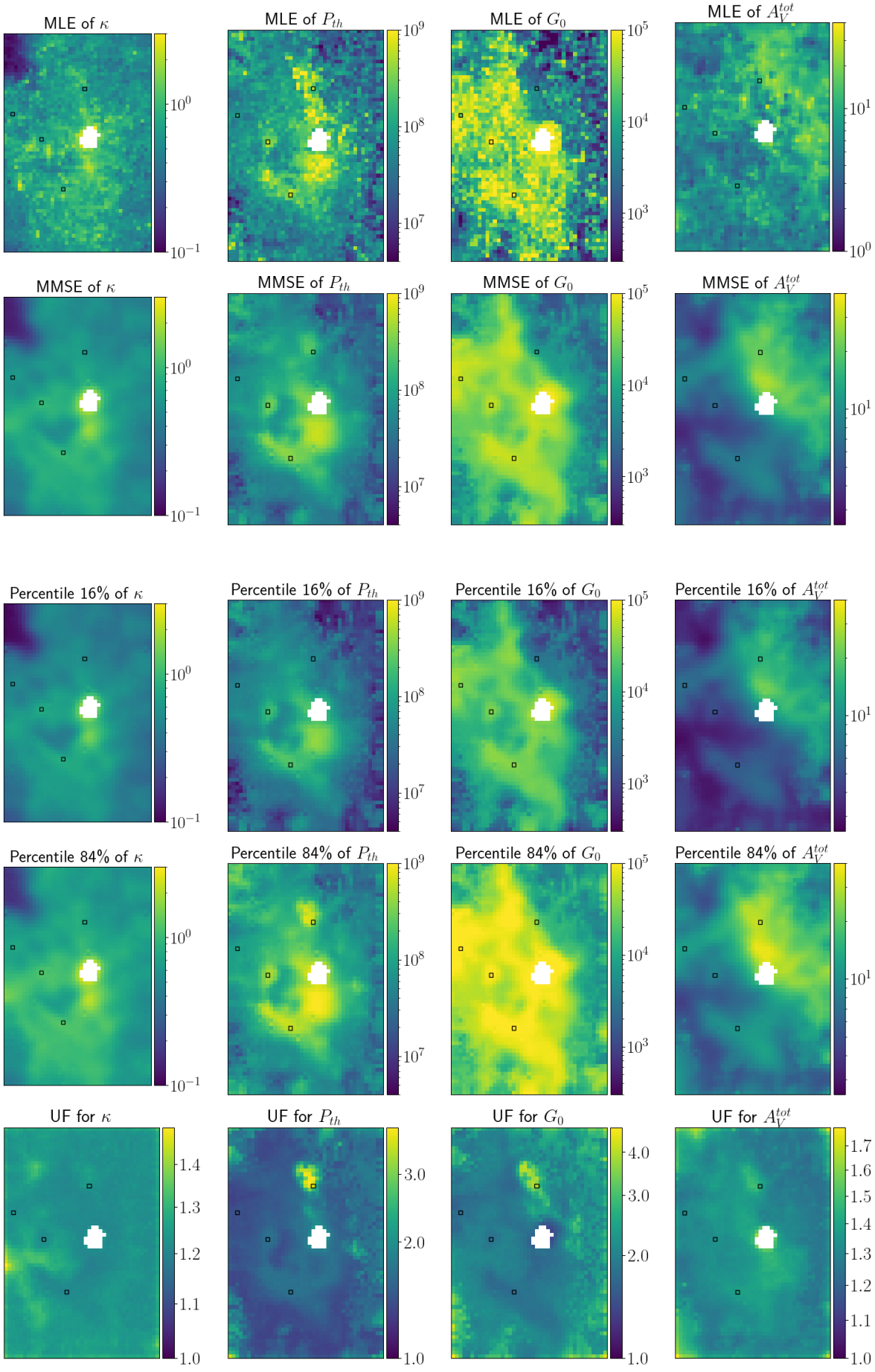


Figure 6.15: Inference results for OMC-1.

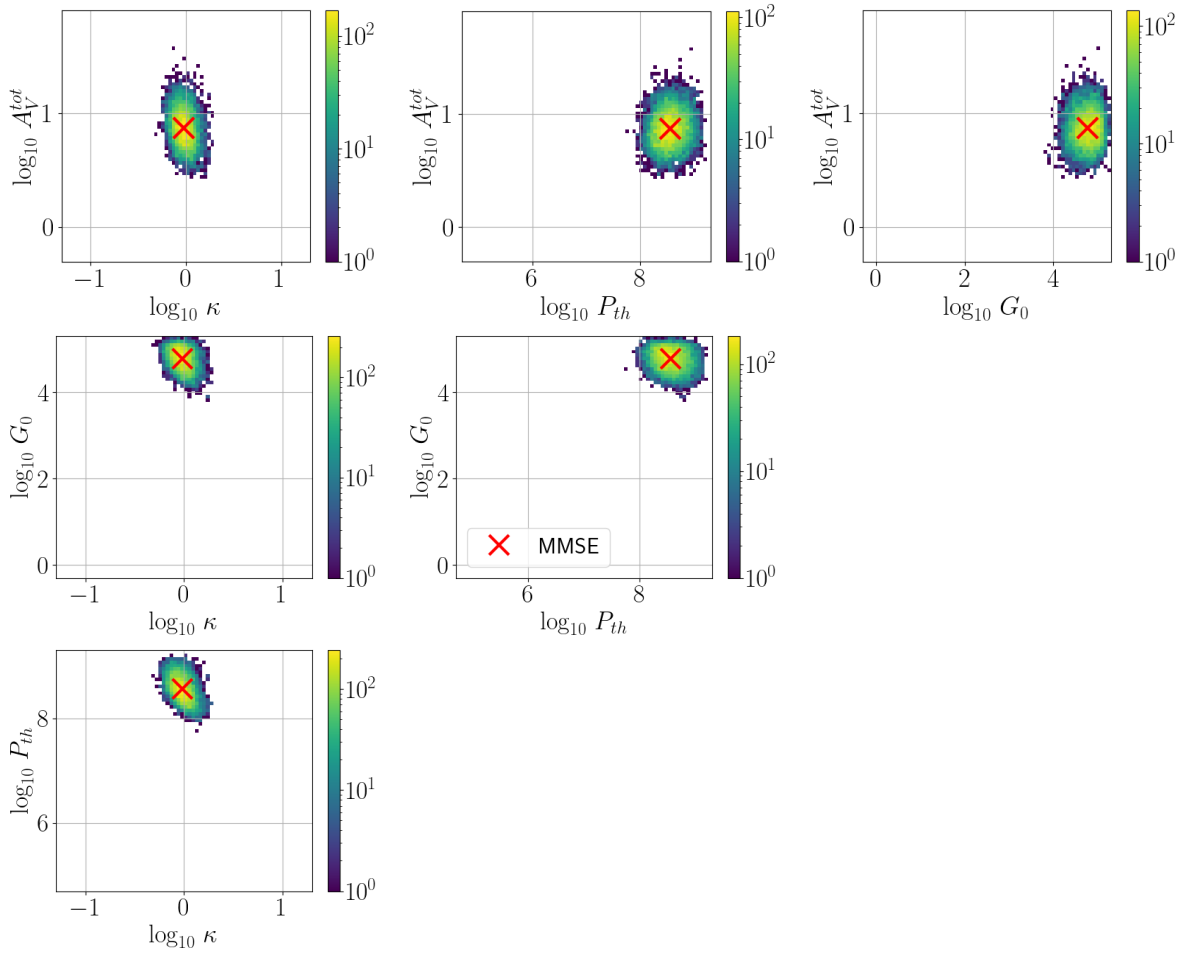


Figure 6.16: Inference results for OMC-1: Orion bar pixel. Two-dimensional marginal histograms in the physical parameters $\boldsymbol{\Theta}$ space. All histograms are in logarithmic norm.

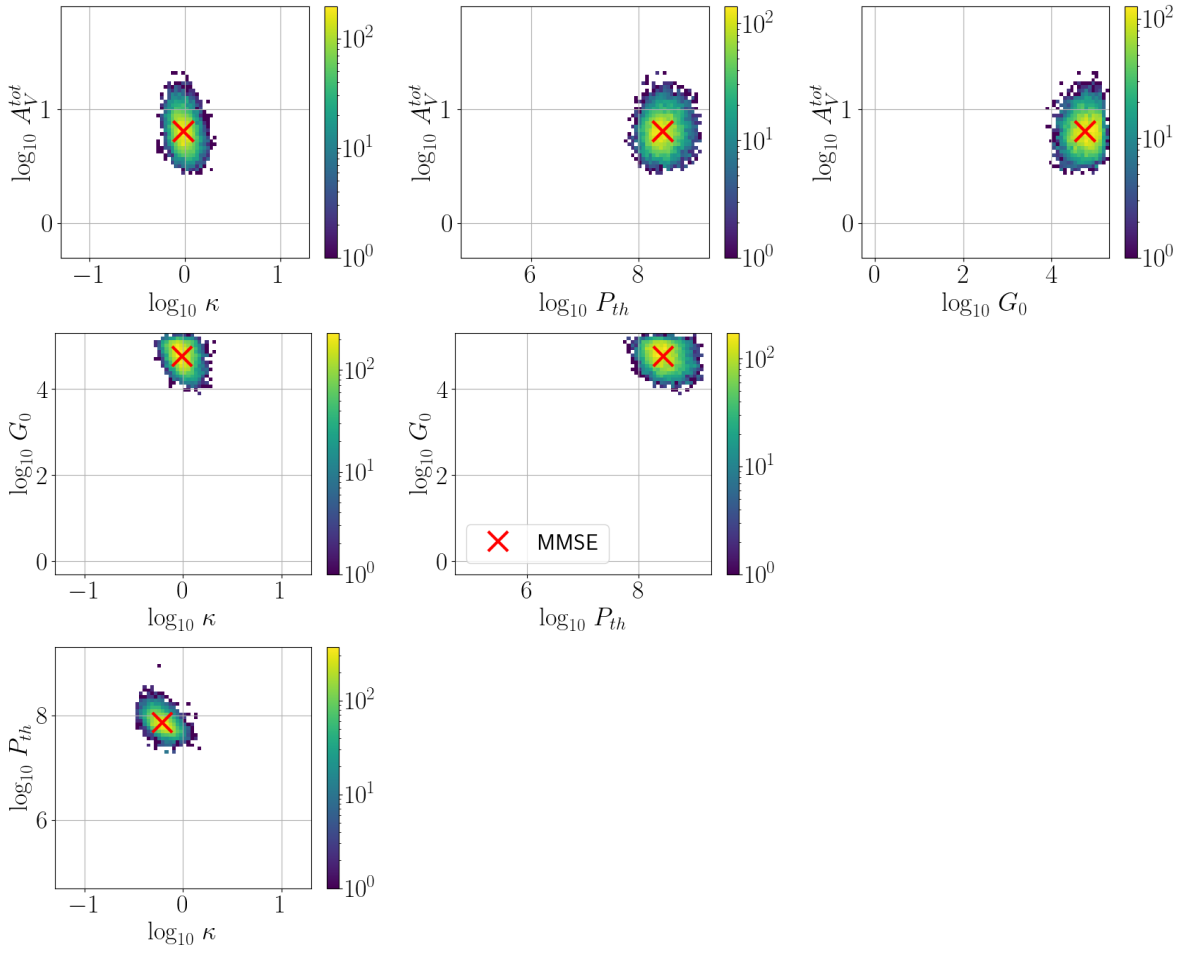


Figure 6.17: Inference results for OMC-1: East PDR pixel. Two-dimensional marginal histograms in the physical parameters $\boldsymbol{\Theta}$ space. All histograms are in logarithmic norm.

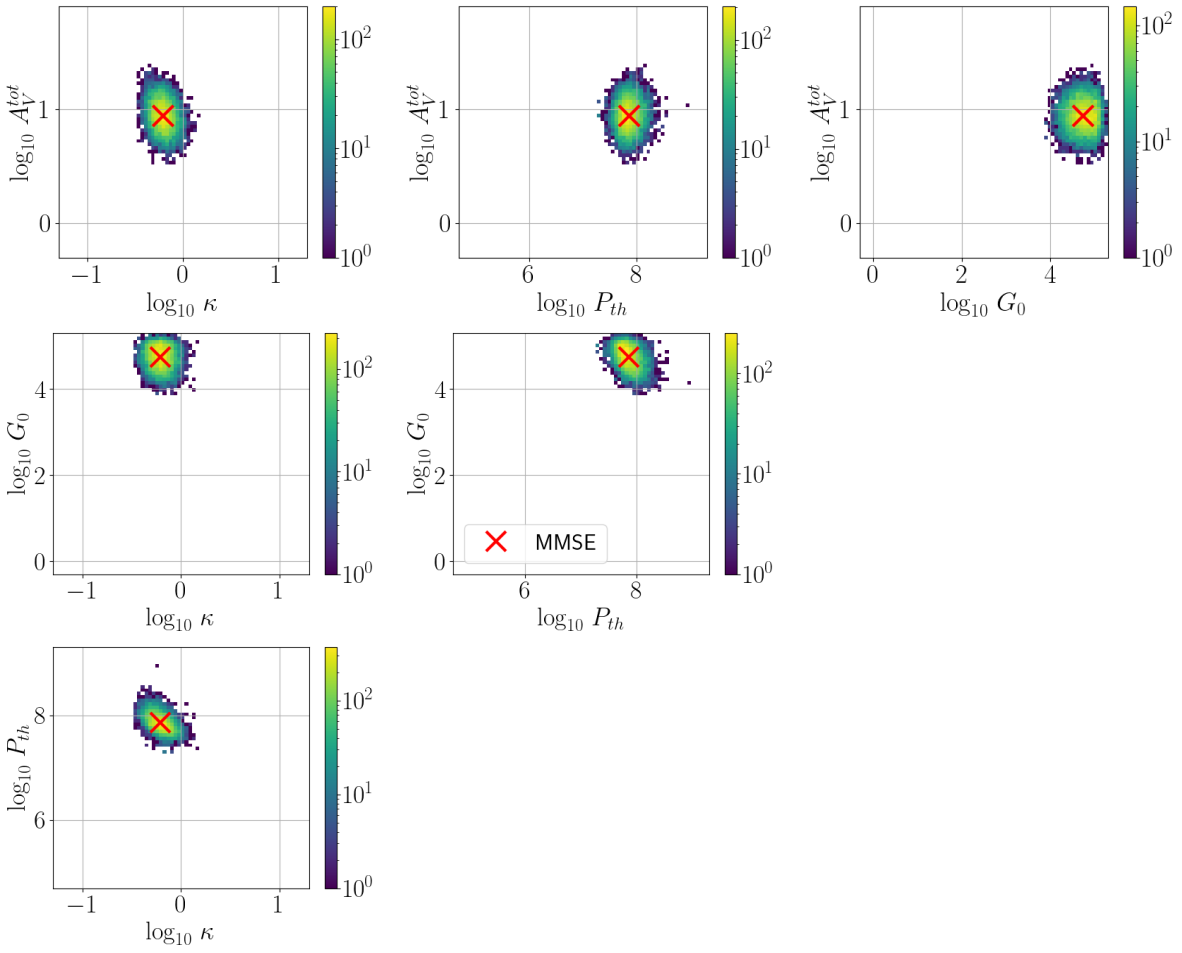


Figure 6.18: Inference results for OMC-1: North-East edge pixel. Two-dimensional marginal histograms in the physical parameters $\boldsymbol{\Theta}$ space. All histograms are in logarithmic norm.

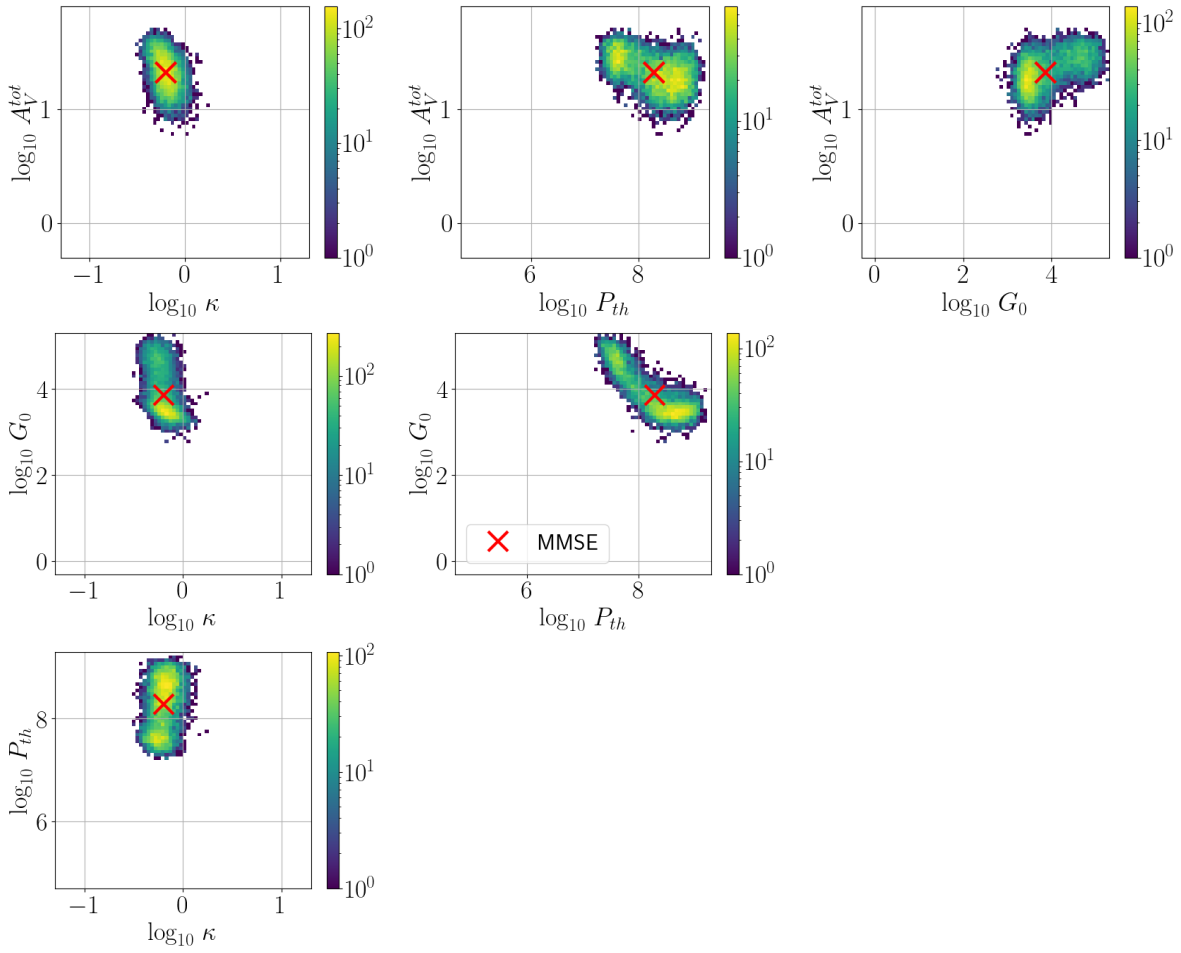


Figure 6.19: Inference results for OMC-1: North-West ridge edge pixel. Two-dimensional marginal histograms in the physical parameters $\boldsymbol{\Theta}$ space. All histograms are in logarithmic norm.

6.5 Conclusion

In this chapter, we applied the full inference procedure to three real observations – NGC 7023, the Carina nebula and OMC-1. The preliminary analysis of the presented inference results already provides interesting astrophysical insights. For instance, Joblin et al. (2018) and Wu et al. (2018) describe a positive correlation between point estimates of P_{th} and G_0 among multiple sources. In contrast, our inference results on NGC 7023, the Carina nebula and the Orion bar revealed a negative correlation between these two parameters in their joint distribution. Putting things together, this result indicates that the positive correlation among sources from the literature was not due to inference uncertainties, and is likely to have a physical origin. However, in view of the large estimation uncertainties, deriving a numerical relation between P_{th} and G_0 from point estimates seems inaccurate. In particular, the power law between proposed in Joblin et al. (2018) and Wu et al. (2018) may be improved by accounting for these inference uncertainties.

We emphasize that this is the first study of maps of ionic, atomic or molecular line observations that permits to infer maps of physical parameters and to quantify the associated uncertainties at once. In particular, the proposed method proved to be able to detect multiple modes and to compare their relative weights in the posterior distribution. It also proved to be able to detect correlations between inferred physical parameters. These properties lead to finer and more trustful results and interpretations.

Appendix 6.A Orion Bar

The Orion Bar PDR lies ~ 2' south-east of the Trapezium star cluster (Allers et al., 2005). This cluster of massive stars creates a HII region that penetrates into the parent molecular cloud. The distance between the Orion nebula and the Sun has been measured at 414 ± 7 pc (Menten et al., 2007). Because of its proximity and edge-on geometry, the Bar is one of the most studied PDRs.

Joblin et al. (2018) analyzed the Orion bar with an observation of $L = 24$ lines in $N = 1$ pixel. The observed lines used in the inversion include ^{12}CO lines (from $J = 11 \to 10$ to $J = 23 \to 22$), rotational H_2 lines (from S(0) to S(5)) and low level CH^+ rotational lines (from $J = 1 \to 0$ to $J = 6 \to 5$). The noise on the observation $\mathbf{y}$ was assumed additive, Gaussian and uncorrelated with known standard deviations $(\sigma_{a,\ell})_{\ell=1}^{L}$. Like NGC 7023, the authors inferred $\boldsymbol{\theta} = (\kappa, P_{\text{th}})$, while fixing the observation angle to $\varphi = 60$ deg and the total extinction to $A_V^{\text{tot}} = 10$ mag – as the observed lines did not provide sufficient constraint on A_V^{tot}. Based on Tielens and Hollenbach (1985) and Marconi et al. (1998), they set $G_0 = 2 \times 10^4$. Like NGC 7023, the fit was performed with a grid search on P_{th} and a simple continuous optimization on κ.

They obtained $P_{\text{th}} = 2.8 \times 10^8$ K cm^{-3} and $\kappa = 1.3$.

Inversion setup – In this analysis, we use the same observation model as for NGC 7023 (Eq. 6.2). The prior hyperparameters and validity intervals are identical. The parameters of the proposed sampler are the same as in the previous section, except for the step size of the PMALA kernel. We empirically set it to $\eta = 0.03$ to improve the acceptance probability.

Model assessment and Bayesian p-value – In Joblin et al. (2018), the authors stated that the fit did not reproduce the observations as well as for NGC 7023. Applying our inference procedure, we obtain an estimated p-value is $\widehat{p}^{(T_{\text{MC}})} = 0.52$, and a rejection probability of $\mathbb{P}\left[p^{(t)} \leq \alpha\right] < 10^{-200} < \delta$, which does not lead to a rejection. Therefore, the considered observation model is compatible with the observations. Figure 6.20 compares the marginal posterior predictive distributions with the true observations. Most H_2 lines and the CH^+ $J = 1 - 0$ line are relatively poorly reconstructed, with an error of a factor ~ 3 for the MMSE on these lines. However, this factor 3 remains compatible with the errors $\sigma_{a,\ell}$ and σ_m in the observation model.

Inference results – Figure 6.21 shows our estimation results. The anticorrelation between P_{th} and G_0 observed in NGC 7023 is also visible here. This is expected as we are using almost the same set of lines as observational constraints. As in NGC 7023, the results for P_{th}, G_0 and A_V^{tot} are overall consistent with the values from Joblin et al. (2018). Table 6.2 details the obtained values. The obtained $G_0 = 3.0 \times 10^4$ – with $G_0 \in [4.8 \times 10^3, 1.3 \times 10^5]$ with probability 95% – is consistent with the literature. The inferred thermal pressure $P_{\text{th}} = 1.1 \times 10^8$ K cm^{-3} – with $P_{\text{th}} \in [4.7 \times 10^7, 2.8 \times 10^8]$ K cm^{-3} with probability 95% – is slightly lower than the 2.8×10^8 K cm^{-3} from Joblin et al. (2018). Conversely, the scaling factor $\kappa = 5.2$ – with $\kappa \in [2.7, 10]$ with probability 95% – is much larger than 1.3. Like in NGC 7023, this might be due to the change of Meudon PDR version – 1.5.4 in Joblin et al. (2018), and 1.7 in our work. It may also be due to the difference in some secondary parameters.

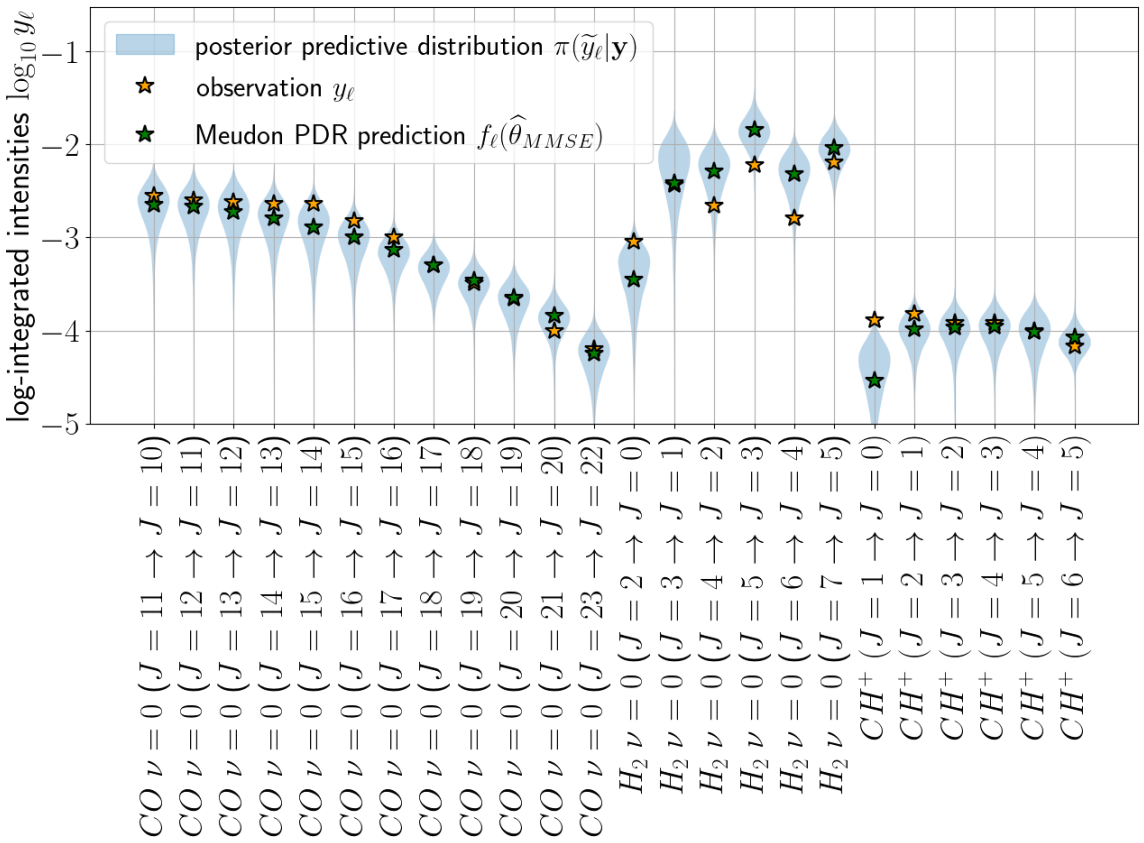


Figure 6.20: Posterior predictive assessment for the Orion Bar on the lines used for the inversion. Comparison of observations $\mathbf{Y}$ and associated noise model with posterior predictive distributions on $\widetilde{\mathbf{f}}(\boldsymbol{\Theta})$, with $\boldsymbol{\Theta} \sim \pi(\boldsymbol{\Theta}|\mathbf{Y})$.

Table 6.2: Inference results on Orion Bar. In Joblin et al. (2018), only the scaling parameter κ and the thermal pressure P_{th} are inferred.

		κ –	P_{th} (K cm^{-3})	G_0 –	A_V^{tot} (mag)
(Joblin et al., 2018)		1.3	2.8×10^8	2×10^4	10
MMSE		5.2	1.1×10^8	3.0×10^4	11.6
68% credibility interval	lower bound $l_{68\%}$	3.7	8.0×10^7	9.5×10^3	4.1
	upper bound $u_{68\%}$	7.5	1.6×10^8	8.5×10^4	31.5
	$\text{UF}_{68\%}$	1.4	1.4	3.0	2.8
95% credibility interval	lower bound $l_{95\%}$	2.7	4.7×10^7	4.8×10^3	1.9
	upper bound $u_{95\%}$	10.0	2.8×10^8	1.3×10^5	42.4
	$\text{UF}_{95\%}$	1.9	2.5	5.2	4.7

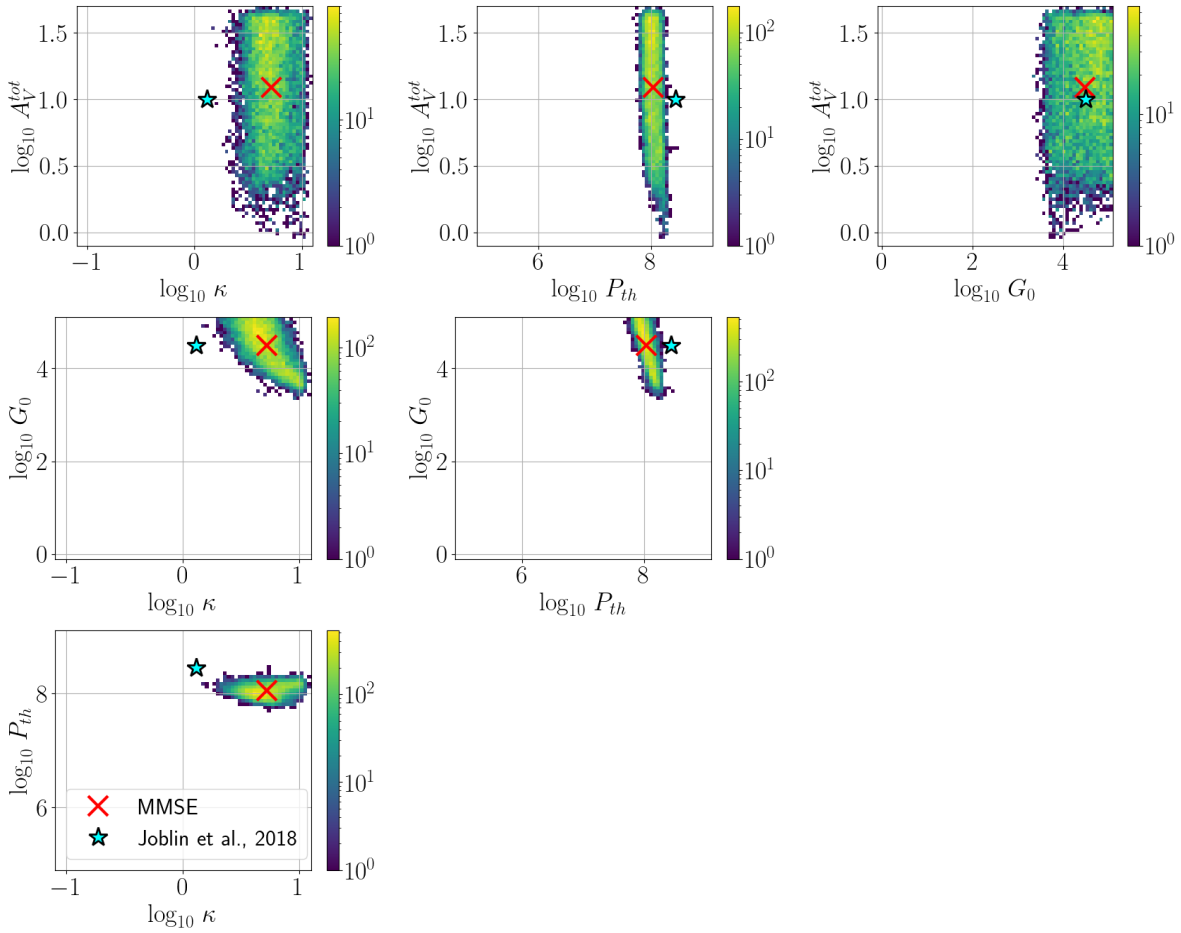


Figure 6.21: Inference results for the Orion Bar. Two-dimensional marginal histograms in the physical parameters $\boldsymbol{\Theta}$ space. All histograms are in logarithmic norm.

Conclusions and perspectives

"*People from different backgrounds approach a subject in different ways and ask different questions.*"

Jocelyn Bell Burnell

Contents

Conclusions

The study of the ISM carries fundamental questions such as the regulation of star formation or the development of molecular complexity, possibly leading to the formation of prebiotic molecules. In this thesis, we studied the impact of the radiative feedback of newborn massive stars on their parent molecular clouds. New facilities at IRAM, ALMA and the JWST might lead to breakthroughs in the coming years, thanks to the very rich hyperspectral data they provide. For instance, the IRAM-30m Large Program "Orion B" observed the Orion B cloud at dense core resolution, resulting in a million-pixel map, with $240\,000$ spectral bands containing emissions of dozens of ionic, atomic or molecular tracers (Pety et al., 2017). In this thesis, we addressed five problems to extract as much information as possible from such observations combined with an ISM numerical model such as the Meudon PDR code:

- Deriving fast, accurate and light emulators of the Meudon PDR code, an ISM numerical model, to be able to make inference scalable. This task was addressed in Chapter 4.
- Approximating the likelihood function to obtain a simple closed-form expression without neglecting a source of noise. The obtained approximation simplifies the sampling task while introducing a small and controlled error. This task was addressed in Chapter 5 (Section 5.1.1).
- Proposing a spatial regularization prior to improve the quality of estimations in low signal-to-noise ratio regions. This task was addressed in Chapter 5 (Section 5.1).
- Estimating physical parameter maps such as the thermal pressure and the total visual extinction from observations, to bring new insights on the interstellar medium (ISM) and star formation. The proposed Bayesian inference method accounts for as much physics as

possible – in the forward model, in the observation model, and in the prior distribution – and scales well enough to be applied to observation maps of up to $\mathcal{O}(10^4)$ pixels. It provides uncertainty quantification through credibility intervals. This task was addressed in Chapter 5 (Section 5.2).

- Assessing the compatibility between the observation model and the observations, to potentially provide feedback to the astrophysicists who build these models. This task was addressed in Chapter 5 (Section 5.3).

An additional task was also addressed. It consists in selecting the most informative observables to minimize uncertainty on inferred physical parameters. The resulting variable selection method is based on the conditional differential entropy – or, equivalently, on the mutual information. Although we already have preliminary results, this work is still in progress in collaboration with other members of the ORION-B consortium. This task was not presented in the body of this manuscript. The interested reader can find a description of the preliminary results on our side, i.e., on the numerical model side, in Appendix A.

As part of an interdisciplinary project, this thesis yielded contributions in both statistics and astrophysics for each of the considered problems.

Contributions in statistics

The main methodological contributions of this thesis are the proposition of a general method to build emulators, of a likelihood approximation, of a new Markov chain Monte Carlo (MCMC) algorithm, and of an extension of the Bayesian test of hypothesis.

Derivation of fast and accurate approximations of ISM numerical models – State-of-the-art ISM numerical models have prohibitively long evaluation times. They are thus often replaced with approximations for inference, which induces an error. Interpolation methods are often used in the ISM community (Wu et al., 2018; Ramambason et al., 2022), and the associated error is not always quantified. In Chapter 4, we proposed to apply general strategies to emulate a numerical model that predicts many observables at once and punctually produces outliers. We applied these strategies to emulate the Meudon PDR code, and compared the obtained artificial neural networks (ANNs) with interpolation methods with respect to memory requirements, speed, and accuracy. The proposed ANNs significantly outperformed all the considered interpolation methods.

Likelihood approximation – The observation model introduced in Chapter 3 (Section 3.4), which involves a multiplicative lognormal noise and an additive Gaussian noise, does not lead to a simple closed-form likelihood function. In Section 5.1.1, instead of neglecting one source of noise, we proposed a parametric closed-form approximation of the likelihood function. This approximation builds on that from Nicholson and Kaipio (2020), which is purely additive and Gaussian. We combined a similar additive and Gaussian approximation with a multiplicative and lognormal one using a weighted geometric mean. The weight of this geometric mean is a function of the physical parameters such that the multiplicative approximation dominates in the high SNR regime, and the additive approximation in the low SNR regime. In Section 5.B, we proposed a grid search and a Bayesian optimization approach to tune the weight parameters so that the approximation error is minimized.

A new MCMC algorithm – The posterior distribution presented in Chapter 5 (Section 5.1.3) is challenging to sample from. The log-posterior is considered to be non log-concave and non gradient Lipschitz continuous, or with an unusable Lipschitz constant. We proposed in Section 5.2 a new MCMC algorithm. This algorithm combines two sampling kernels, each addressing one of these two difficulties.

The first kernel is based on a preconditioned MALA algorithm (Xifara et al., 2014) that exploits information on the local geometry of the log-posterior to propose relevant candidates. We resorted to the RMSProp preconditioner (Tieleman and Hinton, 2012), which was already applied in an MCMC algorithm (Li et al., 2016). We improved the algorithm from Li et al. (2016) by performing accept-reject steps and by accounting for the correction term γ that emerges from a position-dependent preconditioner.

The second kernel combines three algorithms. A Gibbs sampling (Geman and Geman, 1984) divides the physical parameter of dimension $N \times D$ into N conditional distributions of dimension D that are easier to explore. A multiple-try Metropolis algorithm (Martino, 2018) increases the probability of acceptance, and uses a proposal distribution that allows escapes from local modes. A chromatic Gibbs sampling algorithm (Gonzalez et al., 2011) permits to perform the sampling of many pixels in parallel, which greatly accelerates the sampling.

Extension of Bayesian hypothesis testing for model assessment – Solving an inverse problem relies on the hypothesis of a possibly misspecified observation model. The Bayesian test described in Gelman et al. (1996) permits to assess the model compatibility with the observations. It relies on a test statistic that permits to compare true observations with observations reproduced from the observation model. For simple cases such as a point estimate, a Gaussian additive uncorrelated noise and the L_2 norm as the test statistic, the p-value associated with the test can be computed exactly. Otherwise, it is approximated by a Monte Carlo (MC) estimator, which induces an error that can lead to a wrong decision on the compatibility or incompatibility of the model with the observations. In Chapter 5 (Section 5.3), we extended the test to account for the uncertainty due to the MC approximation error. We considered the p-value as a random variable. Using a simple likelihood model and a conjugate prior, we obtained a simple uncertainty description of the p-value. This description represents a negligible additional computational cost with respect to the standard MC estimator. The proposed test can detect cases where the uncertainty on the MC estimator is too large to reject or not the numerical model. Too large uncertainties indicate that longer Markov chains should be used.

Contributions in astrophysics

Chapter 3 reviewed applications of statistical inference in ISM studies, and the position of our work with respect to the literature. We now review our contributions for the ISM community, namely the derivation of a fast, accurate and light surrogate for numerical models, the use of a spatial regularization prior, the use of Bayesian inference for high dimensional observation maps of integrated intensities, the application of the model assessment test, and the insights obtained from the application of our methods to real data.

Informative spatial regularization prior – To exploit the map structure of the physical parameters, we resorted to a spatial regularizing prior – introduced in Chapter 5 (Section 5.1.2). Like Marchal et al. (2019), we use a L_2 norm of the map Laplacian. However, this article aimed at fitting one mixture of Gaussian profiles per spectrum on a hyperspectral map, while we directly infer more complex physical parameters. A common approach in the ISM community for low SNR regions is to stack pixels to increase the SNR, thus reducing spatial resolution. This spatial regularization approach permits to better exploit low SNR observations, as it enables pixels to access the information contained in their neighbors. In addition, the spatial regularization is negligible when the likelihood function constrains well the physical parameters, i.e., in high SNR observations and good tracers of the physical parameters. It can thus be seen as a form of adaptive stacking.

Bayesian inference for high dimensional maps – When the observed lines constrain poorly the physical conditions or when the Signal-to-Noise Ratio is low, multiple solutions might reconstruct observations equivalently well. Moreover, the non-linearity of astrochemistry models leads

to non-convex and even multi-modal problems. Many methods used in interstellar astrophysics do not consider these degeneracies and only return one estimated map (Joblin et al., 2018; Wu et al., 2018). Most methods get trapped in local minima and estimations have no optimality guarantees. Heuristics can overcome this issue for small maps but would be unrealistically slow for larger ones. In Chapter 6, using a Bayesian approach, we obtained meaningful uncertainty quantification on inferred physical parameters. In the ISM community, dust studies such as Galliano (2018); Galliano et al. (2021) had already applied Bayesian sampling methods in high dimensional settings. However, sampling methods had never been used – to the best of our knowledge – to reconstruct maps of physical parameters from observations of integrated intensities of ionic, atomic or molecular emission lines.

Model assessment with Bayesian hypothesis test – In ISM studies, the compatibility of the model with the observations is rarely assessed. However, recently, Lebouteiller and Ramambason (2022) used multiple information criteria, the Bayesian evidence and a model assessment to estimate the number of distinct environments that are necessary to explain observations. Similarly, Galliano (2022) used the Bayesian hypothesis test to assess the quality of the model. In Chapter 6, we applied our extended test of hypothesis which includes the error that comes from the MC approximation. We also apply it to each pixel of our reconstructed maps, which permits to identify pixels corresponding to environments that are poorly modeled by the Meudon PDR code. Such pixels may not correspond to PDR, or may pinpoint an insufficiency from the Meudon PDR code. This feedback is therefore valuable for astrophysicists working on numerical models, as it can indicate potentially necessary future development.

New insights – Applications of the proposed method to real observations already brought new astrophysical insights. Here, we highlight two of them. First, in Joblin et al. (2018) and Wu et al. (2018), the authors outline a positive correlation between point estimates of the thermal pressure P_{th} and the intensity of the incident UV radiative field G_0 among multiple sources. In contrast, our inference results on NGC 7023, the Carina nebula and the Orion bar revealed a negative correlation between these two parameters in their joint distribution. This result indicates that the positive correlation among sources from the literature is not due to inference uncertainties, and is likely to have a physical origin. However, the large estimation uncertainties indicate that the numerical power law relations between P_{th} and G_0 proposed in Joblin et al. (2018) and Wu et al. (2018) may be improved by accounting for these inference uncertainties.

Second, the presented analysis on the OMC-1 cloud is the first inference performed on the full observation map. The results are consistent with the literature, with e.g., $P_{\text{th}} \simeq 3-5\times10^8$ Kcm^{-3} and $G_0 \simeq 3-5\times10^4$ in the Orion bar. New information about the region are provided, such as the fact that the pressure is high in all the PDRs around the Trapezium star cluster. Additional analyses are still required for a full physical interpretation of the inference results.

Perspectives

This manuscript represents a milestone in our research on the ISM and star formation with the ORION-B consortium. This work could now be extended in multiple directions. In the following, we list some promising directions.

Derivation of better emulators for ISM numerical models – The proposed neural network-based emulation strategy is quite general and could be applied to other ISM numerical models. In general, numerical models are based on a system of coupled equations, potentially partial differential equations. Deriving an emulator from a set of input-output pairs seems suboptimal as the emulator could in principle be trained to satisfy the system of equations. The *physics-informed neural networks* (Raissi et al., 2019) form a potentially promising approach to derive surrogate models for the ISM community. Such ANNs are trained directly with the system of physical

differential equations. Therefore, they have access to more information during training, and thus generalize better with limited datasets. This seems particularly promising for chemistry models, for which the differential equations are easy to write. This lead was considered for the Meudon PDR code. However, like many complex ISM codes, it includes many microphysical processes and thus many equations. The physics-informed ANNs were thus badly suited to this task.

A second potentially promising lead regarding the emulator is the management of outliers. We proposed a three-step approach that includes a semi-manual check. Preventing a numerical model from producing such outliers would be statistically simpler, and would improve the confidence in the model predictions. This is very challenging for a code as complex as the Meudon PDR code, in which bistability – the existence of two solutions for an equation system – can occur at different levels. Then, deriving a deterministic binary test that detects outliers would prevent from using any in the training of the emulator. Again, this is very challenging for the Meudon PDR code, as the number of profiles to check is high. Finally, if a statistical approach needs be used for outliers, a potentially better approach would be to learn the surrogate model and fit the binary mask on outliers at once, as detailed in Chapter 4 (Appendix 4.A). However, this approach would require careful hyperparameter tuning to avoid the classification of physically meaningful points that are difficult to reproduce as outliers.

Observation model – Several improvements of the observation model could be explored. First, using a hierarchical model instead of the proposed approximation of the likelihood function would permit to use the exact observation model in the inversion. We did try to adapt the proposed sampler to the hierarchical model, but using the approximation proved to considerably simplify the sampling. A possibility we did not explore yet is to resort to the ancillarity-sufficiency interweaving strategy (ASIS) sampling approach, already applied in dust studies, for instance in (Kelly, 2011; Kelly et al., 2012; Galliano, 2018).

Second, the observation model specified in Chapter 3 (Section 3.4) neglected all the correlations in the noise. This choice greatly simplified the definition of the likelihood function and the computation of its first and second order derivative, as it enabled to consider the observation elements $y_{n\ell}$ individually. However, it is known that the noise realizations in a telescope are not independent, neither spatially nor spectrally. For instance, in Einig et al. (2023), the authors describe the spatial and spectral noise structure in the Orion-B observations. Similarly, as we showed in Chapter 3 (Section 3.1.2), dust studies now frequently include non diagonal terms in the noise covariance. Accounting for these noise correlations would provide additional information in the observation model, which is likely to result in lower uncertainties on estimates.

Finally, the noise model presented in Section 3.4 and exploited for OMC-1 in Section 6.4 slightly overestimates the additive noise standard deviation. Indeed, we set a constant interval for the spectral integration of the intensities, and used large intervals to avoid cutting any signal. Using a line profile fitting algorithm such as ROHSA (Marchal et al., 2019) or CUBEFIT Paumard et al. (2022) may lead to more accurate integrated intensity estimates. Similarly, the denoiser proposed in (Einig et al., 2023) improves the SNR without assuming a line profile. However, these approaches may introduce a bias in the estimated integrated intensities. Besides, it is not simple how to describe the associated uncertainties, which is crucial for the inversion.

To improve the observation model with respect to the last two points, one could infer at once both 1) the multiline maps of integrated intensities from the hyperspectral maps with the complete noise model and 2) the physical parameters from the multiline maps of integrated intensities. The resulting complex posterior distribution would be written with a hierarchical model and sampled from using a Gibbs sampler. Such an approach might define a heavier inverse problem. However, the effective noise on the multiline maps would be reduced, which would lead to smaller uncertainties in inferred physical parameters.

Choice of a preconditioner – We chose the RMSProp preconditioner because it is a diagonal preconditioner that proved efficient in high dimensions both in theory and in practice (Dauphin

et al., 2015). As it is diagonal and only requires first order derivatives, the correction term γ in PMALA that comes from the position-dependent preconditioner can be written and evaluated with acceptable costs – see Chapter 5 (Section 5.2.1). This correction term ensures a good discretization of a Langevin diffusion process, which favors good candidates. However, accounting for this term is not necessary to ensure ergodicity or that the posterior is the invariant distribution of the generated Markov chain, even if the correction term is neglected. Other preconditioners that take into account the spatial regularization may lead to better mixing properties. As the spatial regularization prior is equivalent to a Gaussian distribution with a band precision matrix, implementing PMALA with such a preconditioner could exploit classic methods to sample from a Gaussian distribution (Vono et al., 2021). Such a preconditioner may perform better than RMSProp, e.g., for large spatial regularization parameter τ.

Analysis of larger observation maps – The proposed sampler was applied to cases with dimensions up to $ND = 4 \times 10^4$. Although this is already large, new and future maps are even larger. For instance, the Orion-B map (Pety et al., 2017) contains 10^6 pixels. There are two necessary and independent approaches to scale to these maps.

First, the Orion-B map contains multiple environments. Indeed, the Horsehead nebula is better modeled with a PDR model, but the dense cores in the cloud require a dark cloud time-dependent model. Similarly, the original observations of OMC-1 studied in Chapter 6 (Section 6.4) contain the BN/KL region, which is dominated by shocks. To be able to perform inference on these observations using the Meudon PDR code, some pixels were masked. In addition, observations cannot be spatially well resolved, which means that some pixels may contain multiple environments that require different models. Inference on such large maps would therefore require simultaneously unmixing the environments on each pixel and estimating the physical parameters of each environment. This is part of the PhD project of Antoine Zakardjian started in 2022, member of the ORION-B consortium, supervised by Jérôme Pety and Annie Hughes.

Second, even neglecting these mixtures of environments, the proposed algorithm is unlikely to scale to these dimensions without additional effort. Distributed algorithms are a common solution to divide one computationally very expensive task into smaller tasks addressed in parallel by different workers. Thanks to the simple regularization prior and diagonal noise correlation, the proposed MCMC algorithm could certainly be distributed, using for instance the framework presented in Thouvenin et al. (2023). This is also in part the PhD project of Maxime Bouton started in 2022, supervised by Pierre Chainais and Pierre-Antoine Thouvenin.

Appendix A

Entropy of probability distributions and line selection

"You're going to have to make a choice, Mr. Anderson."

The Matrix (1999), Lana and Lilly Wachowski

Contents

In Chapter 4, we derived a surrogate model of the full Meudon PDR code. This surrogate model emulates all the $5\,375$ emission lines predicted by the Meudon PDR code. Chapter 5 introduced the posterior distribution involved in the inverse problem considered in this thesis. In actual ISM observations, due to observational constraints, only $L \simeq 5$ to 30 lines are observed. Observations contain different lines depending on the telescope or observed wavelength range. In particular, observing lines in the infrared (IR) domain requires space telescopes, as these wavelengths are absorbed by the Earth atmosphere. In general, observing maps of many lines is extremely expensive. Besides, observing all lines at once is unfeasible due to the large differences in wavelengths, from near infrared to millimetric domains. It is therefore crucial to prepare observations to measure lines that permit to infer physical parameters of interest with low uncertainty.

In this chapter, we present a statistical method that determines, for a fixed number of emission lines, the set of lines that leads to the lowest uncertainty on inferred physical parameters. This uncertainty is quantified with the *conditional differential entropy* – or equivalently, the *mutual information*. We evaluate the best set of lines for a variety of true physical conditions $\boldsymbol{\theta} \in \mathbb{R}^D$ and for different sizes L of sets.

This project is a collaboration with Lucas Einig and Antoine Roueff, also members of the ORION-B consortium. The overall goal is to apply this variable selection strategy both on nu-

merical models and the Orion-B dataset (Pety et al., 2017), and to compare the results. In this appendix, we only present the theoretical developments of a line selection method. This is still ongoing work and an application to the Meudon PDR code with the noise model presented in Chapter 3 is currently under study.

Section A.1.2 introduces the mean squared error (MSE) considered in this context, the conditional differential entropy and the Fano bound that links the two. Section A.2 applies the Fano bound to a simple case with an affine forward model and an additive uncorrelated Gaussian noise to provide intuitive insights.

A.1 A line selection method that compares posterior distributions

In this chapter, we aim at determining the set of $K \in [\![1, L]\!]$ lines that leads to the lowest uncertainty on inferred physical parameters, i.e., that is the most informative. This uncertainty is quantified with a general mean squared error (MSE). As this MSE is in general not accessible, we resort to a lower bound on the MSE called the Fano bound. The Fano bound relies on the conditional differential entropy, that can be evaluated with a Monte Carlo (MC) estimator from samples. The set of K lines that minimizes the conditional differential entropy also minimizes the Fano bound.

A.1.1 Using the MSE as a quantitative criterion to rank sets of lines

Distribution on the observation $\pi(\mathbf{y}|\boldsymbol{\theta}^*)$ – The results of such a search depend on the type of observed environment. We account for this dependence and determine the most informative set of K lines for different physical conditions $\boldsymbol{\theta}^* \in \mathbb{R}^D$. Applying the observation model presented in Chapter 3 (Eq. 3.23) with the "true" physical parameters $\boldsymbol{\theta}^*$, $\omega = -\infty$ – no censorship –, and a set of lines $\mathcal{S} \in 2^{[\![1,L]\!]}$ yields

$$\forall \ell \in \mathcal{S}, \quad y_\ell = \varepsilon_{n\ell}^{(m)} \widetilde{f}_\ell(\boldsymbol{\theta}^*) + \varepsilon_{n\ell}^{(a)}, \quad \varepsilon_{n\ell}^{(a)} \sim \mathcal{N}(0, \sigma_a^2), \quad \varepsilon_{n\ell}^{(m)} \sim \text{Lognormal}\left(-\frac{\sigma_m^2}{2}, \sigma_m^2\right). \tag{A.1}$$

This observation model defines a distribution on observations denoted $\pi(\mathbf{y}|\boldsymbol{\theta}^*)$.

Joint distribution $\pi(\boldsymbol{\theta}, \mathbf{y})$ – From each observation $\mathbf{y}$, the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ is defined as in Chapter 5 (Eq. 5.12). In this case, as we consider a single pixel, there is no spatial regularization. The combination of the distribution on observations (Eq. A.1) and the posterior distribution defines a joint distribution on $(\boldsymbol{\theta}, \mathbf{y})$

$$\pi(\boldsymbol{\theta}, \mathbf{y}) = \pi(\boldsymbol{\theta}|\mathbf{y})\pi(\mathbf{y}|\boldsymbol{\theta}^*). \tag{A.2}$$

Sampling from this joint distribution is equivalent to solving one inverse problems per sample $\mathbf{y} \sim \pi(\mathbf{y}|\boldsymbol{\theta}^*)$, i.e., per realization of noise on an observation of the environment characterized by $\boldsymbol{\theta}^*$.

A general mean squared error (MSE) as quantitative criterion – To select a set of lines $s \subset [\![1, L]\!]$, we need to compare joint distributions on $(\boldsymbol{\theta}, \mathbf{y})$ with respect to a quantitative criterion. As already mentioned, the goal of this chapter is to determine the set of lines that best constrains the physical parameters. The MSE quantifies this notion of well constrained parameters. Here, the MSE is integrated with respect to both observation $\mathbf{y}$ and physical parameters $\boldsymbol{\theta}$,

$$\text{MSE} = \mathbb{E}_{(\boldsymbol{\theta},\mathbf{y})}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\|_2^2\right] = \int \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\|_2^2 \, \pi(\boldsymbol{\theta}, \mathbf{y}) \, d\boldsymbol{\theta} \, d\mathbf{y} \tag{A.3}$$

The MSE is minimized with one estimator $\widehat{\boldsymbol{\theta}} : \mathbf{y} \mapsto \widehat{\boldsymbol{\theta}}(\mathbf{y})$, the posterior expectation $\widehat{\theta}(\mathbf{y}) = \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}]$. Chapters 2 and 5 covered how to evaluate the posterior expectation for a single value of

$\mathbf{y}$ using an Monte Carlo (MC) estimator. When $\mathbf{y}$ follows a distribution, evaluating the posterior expectation would require running an MCMC algorithm for each $\mathbf{y}$ value, which is highly inefficient.

An alternative to using the MSE directly to rank sets of lines is to compute a lower bound on the MSE. In the following, we consider an entropy-based lower bound, called the Fano bound.

A.1.2 Differential entropy and the Fano Bound

Selecting the set of lines that best constrains the physical parameters is equivalent, in a sense, to selecting the set that brings most information on the physical parameters, i.e., that minimizes the uncertainties in inference. This notion of uncertainty can be quantified using the entropy of a distribution.

Differential entropy – The differential entropy is similar to a Shannon entropy counterpart for continuous random variables. For $\boldsymbol{\theta} \in \mathbb{R}^D$,

$$h(\boldsymbol{\theta}) = \mathbb{E}_\theta \left[-\ln \pi(\theta) \right] = -\int \ln \pi(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{A.4}$$

with $\pi(\boldsymbol{\theta})$ any distribution on $\boldsymbol{\theta}$. Table A.1 provides the entropy formula for a few common parametric distributions. The larger the entropy of a distribution, the larger the uncertainties it describes. For instance, the entropy of a univariate Gaussian distribution increases with the log-variance and is independent to its mean. Similarly, the entropy of a uniform distribution on a compact set is the log-volume of this set. The smaller the volume, the smaller the uncertainties on the associated random variable. These examples also show that the differential entropy is not always positive, unlike the Shannon entropy that applies to discrete random variables. For a univariate Gaussian distribution, when the variance tends to 0, the differential entropy tends to $-\infty$. For a uniform distribution on a compact set, the differential entropy tends to $-\infty$ when the volume of the set tends to 0.

Table A.1: Differential entropy for a few common distributions. The Vol set operator corresponds to the volume of a set. The $\det$ function corresponds to the determinant of a matrix.

	distribution	differential entropy h
univariate $\theta \in \mathbb{R}$	$\theta \sim \mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma > 0$	$\frac{1}{2}\ln(2\pi e) + \frac{1}{2}\ln\sigma^2$
	$\theta \sim \mathsf{Unif}\,(a, b),\ a < b$	$\ln(b-a)$
	$\theta \sim \mathsf{Lognormal}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma > 0$	$\mu + \frac{1}{2}\ln\left(2\pi e\,\sigma^2\right)$
multivariate $\boldsymbol{\theta} \in \mathbb{R}^D,\ D > 1$	$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$	$\frac{D}{2}\ln(2\pi e) + \frac{1}{2}\ln\det(\Sigma)$
	$\boldsymbol{\theta} \sim \mathsf{Unif}\,(\mathcal{C}),\ \mathcal{C} \subset \mathbb{R}^D$	$\ln \mathsf{Vol}\,\mathcal{C}$

Conditional differential entropy – The conditional differential entropy is defined for an observation variable $\mathbf{y} \in \mathbb{R}^L$ and a parameter $\boldsymbol{\theta} \in \mathbb{R}^D$,

$$h(\boldsymbol{\theta}|\mathbf{y}) = \mathbb{E}_{(\boldsymbol{\theta},\mathbf{y})} \left[-\ln \pi(\boldsymbol{\theta}|\mathbf{y}) \right] = -\int \ln \pi(\boldsymbol{\theta}|\mathbf{y})\, \pi(\boldsymbol{\theta}, \mathbf{y})\, d\boldsymbol{\theta}\, d\mathbf{y}. \tag{A.5}$$

One essential property of the conditional differential entropy is that

$$h(\boldsymbol{\theta}|\mathbf{y}) = -\int \ln \frac{\pi(\boldsymbol{\theta}, \mathbf{y})}{\pi(\mathbf{y}|\boldsymbol{\theta}^*)}\, \pi(\boldsymbol{\theta}, \mathbf{y})\, d\boldsymbol{\theta}\, d\mathbf{y} = h(\boldsymbol{\theta}, \mathbf{y}) - h(\mathbf{y}). \tag{A.6}$$

This means that evaluating the conditional entropy is equivalent to computing the differential entropy of two linked distributions. Note that differential entropy and conditional differential entropy are linked to the mutual information $\mathsf{MI}(\boldsymbol{\theta}, \mathbf{y})$, as

$$\mathsf{MI}(\boldsymbol{\theta}, \mathbf{y}) = h(\boldsymbol{\theta}) - h(\boldsymbol{\theta}|\mathbf{y}). \tag{A.7}$$

Unlike differential entropy and conditional differential entropy, the mutual information is always positive, as $h(\boldsymbol{\theta}) \geq h(\boldsymbol{\theta}|\mathbf{y})$.

Fano bound – The conditional differential entropy provides a lower bound on the general MSE from Eq. A.3. This lower bound is called the Fano bound, and reads

$$\mathbb{E}_{(\boldsymbol{\theta},\mathbf{y})}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\|_2^2\right] \geq \frac{D}{2\pi e} \exp\left(\frac{2}{D} h(\boldsymbol{\theta}|\mathbf{y})\right). \tag{A.8}$$

This entropy-based lower bound on the MSE does not require the choice of an estimator nor its computation for all $\mathbf{y}$ drawn from the joint distribution $\pi(\boldsymbol{\theta}, \mathbf{y})$. It is only based on a summary statistic of the joint distribution $\pi(\boldsymbol{\theta}, \mathbf{y})$, the conditional differential entropy. Besides, this lower bound is quite tight. The inequality turns into an equality in the case where the posterior distribution $\boldsymbol{\theta}$ is Gaussian, and there exists $\lambda > 0$ such that $\mathbf{C} = \lambda \mathbf{I}_D$. Appendix A.A derives this lower bound. In essence, this derivation relies on the fact the Gaussian distribution maximizes the entropy, and on the arithmetic mean geometric mean inequality

$$[\det \mathbf{C}]^{\frac{1}{D}} \leq \frac{\operatorname{Tr} \mathbf{C}}{D} \tag{A.9}$$

Finally, this bound can be shown to be tighter than other lower bounds such as the Bayesian Cramér-Rao bound (Aras et al., 2019).

A.1.3 Selection of the best set of lines as a discrete optimization problem

Unlike the simple cases listed in Table A.1, the distributions on $(\boldsymbol{\theta}, \mathbf{y})$ and $\mathbf{y}$ are not parametric. Therefore, their differential entropies do not have simple closed-form expressions. Both $K, D \lesssim 10$, thus a grid-based quadrature of the differential entropies in Eq. A.6 would be possible. However, for simplicity, we resort to MC estimators $\widehat{h}^{(T_{\mathsf{MC}})}(\boldsymbol{\theta}^*, \mathcal{S})$

$$\widehat{h}^{(T_{\mathsf{MC}})}(\boldsymbol{\theta}^*, \mathcal{S}) = -\frac{1}{T_{\mathsf{MC}}} \sum_{t=1}^{T_{\mathsf{MC}}} \ln \frac{\pi(\boldsymbol{\theta}^{(t)}, \mathbf{y}^{(t)})}{\pi(\mathbf{y}^{(t)})} \tag{A.10}$$

$$= \frac{1}{T_{\mathsf{MC}}} \sum_{t=1}^{T_{\mathsf{MC}}} \ln \pi(\mathbf{y}^{(t)}) - \frac{1}{T_{\mathsf{MC}}} \sum_{t=1}^{T_{\mathsf{MC}}} \ln \pi(\boldsymbol{\theta}^{(t)}, \mathbf{y}^{(t)}), \tag{A.11}$$

with $\mathbf{y}^{(t)} \sim \pi(\mathbf{y}|\boldsymbol{\theta}^*)$ and $\boldsymbol{\theta}^{(t)} \sim \pi(\boldsymbol{\theta}|\mathbf{y}^{(t)})$ for all t. The best set of K lines for a given $\boldsymbol{\theta}^*$ is the solution of the discrete combinatorial optimization problem

$$\mathcal{S}_K^*(\boldsymbol{\theta}^*) \in \underset{\mathcal{S} \in S_K}{\arg\min}\, \widehat{h}^{(T_{\mathsf{MC}})}(\boldsymbol{\theta}^*, \mathcal{S}), \tag{A.12}$$

with $S_K \subset 2^{[\![1,L]\!]}$ the set of sets of K lines.

A.2 Illustration: a Gaussian and linear case

In this section, we consider a multivariate case in which the noise model is Gaussian, the forward model is affine and the prior on $\boldsymbol{\theta}$ is also Gaussian. In this simple case, the MSE can be written in closed form. The MSE is compared with the Fano bound to visualize how tight this bound is.

The considered simple inverse problem consists in estimating a physical parameter vector $\boldsymbol{\theta} \in \mathbb{R}^D$ from an observation vector $\mathbf{y} \in \mathbb{R}^L$. The forward model is set to an affine function $\boldsymbol{\theta} \mapsto \mathbf{A}\boldsymbol{\theta} + \mathbf{b}$, with $\mathbf{A} \in \mathbb{R}^{L \times D}$ and $\mathbf{b} \in \mathbb{R}^L$. The considered noise $\boldsymbol{\varepsilon}$ is Gaussian additive and uncorrelated. We consider the following observation model, that involves

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{b} + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_L). \tag{A.13}$$

The prior is set to a zero-mean Gaussian distribution

$$\boldsymbol{\theta} \sim \mathcal{N}(0, \tau^2 \mathbf{I}_D). \tag{A.14}$$

In this setting, this prior is conjugate. The posterior distribution can thus be written simply by combining Eq. A.13 and Eq. A.14. It is given by

$$\boldsymbol{\theta}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}), \tag{A.15}$$

with

$$\begin{cases} \mathbf{C} = \mathrm{Cov}(\boldsymbol{\theta}|\mathbf{y}) = \left[\frac{1}{\sigma^2}\mathbf{A}^T\mathbf{A} + \frac{1}{\tau^2}\mathbf{I}_D\right]^{-1} = \tau^2 \left[\mathbf{I}_D + \frac{\tau^2}{\sigma^2}\mathbf{A}^T\mathbf{A}\right]^{-1}, \\ \boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}] = \frac{1}{\sigma^2}\mathbf{C}\mathbf{A}^T(\mathbf{y} - \mathbf{b}). \end{cases} \tag{A.16}$$

The posterior expectation $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}]$ is the estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ that minimizes the MSE. With this estimator, the MSE reduces to

$$\mathbb{E}_{(\boldsymbol{\theta},\mathbf{y})}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\|^2\right] = \mathrm{Tr}\,\mathbf{C} = \tau^2 \,\mathrm{Tr}\left(\left[\mathbf{I}_D + \frac{\tau^2}{\sigma^2}\mathbf{A}^T\mathbf{A}\right]^{-1}\right). \tag{A.17}$$

In the following, we derive the Fano lower bounds on the MSE to compare how tight it is.

Fano Bound – The computation of the Fano bound relies on $h(\boldsymbol{\theta}|\mathbf{y}) = h(\boldsymbol{\theta}, \mathbf{y}) - h(\mathbf{y})$. As $(\boldsymbol{\theta}, \mathbf{y})$ and $\mathbf{y}$ are Gaussian random variables,

$$\begin{cases} h(\mathbf{y}) = \frac{1}{2}\ln\left[(2\pi e)^L \det(\sigma^2\mathbf{I}_L + \tau^2\mathbf{A}\mathbf{A}^T)\right], \\ h(\boldsymbol{\theta}, \mathbf{y}) = \frac{1}{2}\ln\left((2\pi e)^{D+L}\tau^{2D}\sigma^{2L}\right), \end{cases} \tag{A.18}$$

and

$$h(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{2}\ln\left[(2\pi e)^D \frac{\tau^{2D}\sigma^{2L}}{\det(\sigma^2\mathbf{I}_L + \tau^2\mathbf{A}\mathbf{A}^T)}\right]. \tag{A.19}$$

The Fano Bound reads

$$\mathbb{E}_{(\boldsymbol{\theta},\mathbf{y})}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\|^2\right] \geq \frac{D}{2\pi e}\exp\left(\frac{2}{D}h(\boldsymbol{\theta}|\mathbf{y})\right) = D\frac{\tau^2\sigma^{2(L/D)}}{\det(\sigma^2\mathbf{I}_L + \tau^2\mathbf{A}\mathbf{A}^T)^{(1/D)}}. \tag{A.20}$$

The Weinstein-Aronszajn identity permits to simplify the comparison with the true MSE. It reads

$$\det(\sigma^2\mathbf{I}_L + \tau^2\mathbf{A}\mathbf{A}^T) = \sigma^{2(L/D)}\det\left(\mathbf{I}_D + \frac{\tau^2}{\sigma^2}\mathbf{A}^T\mathbf{A}\right). \tag{A.21}$$

One can thus rewrite Eq. A.20 to

$$\mathbb{E}_{(\boldsymbol{\theta},\mathbf{y})}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\|^2\right] \geq \tau^2 D \det\left(\left[\mathbf{I}_D + \frac{\tau^2}{\sigma^2}\mathbf{A}^T\mathbf{A}\right]^{-1}\right)^{(1/D)}. \tag{A.22}$$

This bound is very similar to the true MSE. The main difference comes from the arithmetic mean geometric mean inequality.

The Fano bound attains the true MSE if and only if the posterior on $\boldsymbol{\theta}$ is Gaussian, the estimator is the posterior expectation, and the covariance matrix $\mathbf{C} = \lambda \mathbf{I}_D$ for some $\lambda \geq 0$. The first two conditions are verified in this linear and Gaussian simple illustration. For the last one, considering $\mathbf{C}$ from Eq. A.16,

$$\begin{aligned} \exists \lambda \geq 0,\ \mathbf{C} = \lambda \mathbf{I}_D &\Leftrightarrow \exists \lambda_2 \geq 0,\ \mathbf{A}^T\mathbf{A} = \lambda_2 \mathbf{I}_D \\ &\Leftrightarrow \mathbf{A} = 0 \text{ or } \exists \lambda_2 > 0,\ \frac{\mathbf{A}^T}{\sqrt{\lambda_2}} \frac{\mathbf{A}}{\sqrt{\lambda_2}} = \mathbf{I}_D \end{aligned}$$

The first case, $\mathbf{A} = 0$, implies that the observation model does not provide any information on $\boldsymbol{\theta}$, and that the posterior distribution on $\boldsymbol{\theta}$ is the same as the prior distribution. The second case can only be satisfied if $D \leq L$. Indeed, if $D > L$, it is impossible for $\mathbf{A}^T\mathbf{A}$ to be full rank.

Special cases – Table A.2 summarizes the formulae obtained in this case for the true MSE and the Fano bound. The first two formulae illustrate the role of the arithmetic mean geometric mean inequality (Eq. A.9) in the relevance of the Fano bound. The Fano bound equals exactly the true MSE when the prior is dominant, when the matrix $\mathbf{A}$ is square with constant diagonal, or when both the observation and physical parameter vectors are reduced to scalars. The last two cases show that the linear map $\mathbf{A}$ minimizing the true MSE is the largest gradient a. In the general case, the most informative linear map $\mathbf{A}$ has the largest singular values. To obtain large singular values, the lines of $\mathbf{A}$ should have large norms and be as orthogonal as possible.

Table A.2: Comparison of true MSE and Fano bound in Gaussian linear inverse problem

Case	True MSE (Eq. A.17)	Fano bound (Eq. A.22)
General	$\tau^2 \operatorname{Tr}\left(\left[\mathbf{I}_D + \frac{\tau^2}{\sigma^2}\mathbf{A}^T\mathbf{A}\right]^{-1}\right)$	$\tau^2 D \det\left(\left[\mathbf{I}_D + \frac{\tau^2}{\sigma^2}\mathbf{A}^T\mathbf{A}\right]^{-1}\right)^{(1/D)}$
$\sigma \to 0$	$\sigma^2 \operatorname{Tr}\left(\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\right)$	$\sigma^2 D \det\left(\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\right)^{(1/D)}$
$\lVert\mathbf{A}\rVert_F \to 0$ or $\tau \to 0$	$D\tau^2$	$D\tau^2$
$D = L$ and $\mathbf{A} = a\mathbf{I}_D$	$\frac{\tau^2\sigma^2}{a^2\tau^2 + \sigma^2}$	$\frac{\tau^2\sigma^2}{a^2\tau^2 + \sigma^2}$
$D = L = 1$	$\frac{\tau^2\sigma^2}{a^2\tau^2 + \sigma^2}$	$\frac{\tau^2\sigma^2}{a^2\tau^2 + \sigma^2}$

A.3 Conclusion

In this chapter, we presented a variable selection algorithm that permits identifying the lines that lead to the lowest uncertainties in inference. The MSE is used as a quantitative criterion to quantify this uncertainty. As it cannot be evaluated, we resort to comparing lower bounds on the MSE. The Fano bound is a tight lower bound. It is based on the conditional differential entropy of a joint distribution $\pi(\boldsymbol{\theta}, \mathbf{y})$, which can be evaluated efficiently with an MC estimator.

Future work involves adapting the code of the sampling algorithm proposed in Chapter 5 to include a distribution on the observation $\pi(\mathbf{y}|\boldsymbol{\theta}^*)$. Then, it will be applied to compare sets of lines predicted by the Meudon PDR code. The full set of $L = 5\,375$ lines can be considerably reduced to a few hundreds by only considering lines that are observable, i.e., that reach intensities larger than σ_a. However, even with this smaller number of lines, the number of sets to evaluate becomes very large for $K \geq 2$. A heuristic approach may be needed to perform selection. Such methods include e.g., greedy algorithms, stepwise forward selection or backward elimination methods (Shalev-Shwartz and Ben-David, 2014, chapter 25, section 25.1), and meta-heuristics (Dreo et al., 2006).

Appendix A.A Derivation of the multivariate Fano Bound

The scalar case

This subsection's goal is to serve as a simple case to the next one. It is essentially a more detailed version (particularly about hypothesis) than Cover and Thomas (2006, theorem 8.6.6).

Theorem A.A.1 (Estimator MSE and differential entropy in 1D case)**.** *For any scalar random variable $\theta \in \mathbb{R}$ and any estimator $\widehat{\theta}$,*

$$\mathbb{E}\left[(\theta - \widehat{\theta})^2\right] \geq \frac{1}{2\pi e} e^{2h(\theta)}, \tag{A.23}$$

with equality if and only if θ is Gaussian and $\widehat{\theta} = \mathbb{E}[\theta]$

Proof. Let $\widehat{\theta}$ be any estimator of θ. Then

$$\mathbb{E}\left[(\theta - \widehat{\theta})^2\right] \;\geq\; \min_{\widehat{\theta}} \mathbb{E}\left[(\theta - \widehat{\theta})^2\right] \;=\; \mathbb{E}\left[(\theta - \mathbb{E}[\theta])^2\right] \;=\; \mathrm{Var}(\theta). \tag{A.24}$$

Then, to get a relation between the variance of θ and its entropy, one uses the fact that for scalar random variables with given variance, the Gaussian distribution has maximum entropy. Let $\sigma^2 = \mathrm{Var}(\theta)$ and $Z \sim \mathcal{N}(\mu, \sigma^2)$, Then,

$$h(\theta) \leq h(Z) = \frac{1}{2} \ln(2\pi e \sigma^2), \tag{A.25}$$

which gives

$$\mathrm{Var}(\theta) = \sigma^2 \geq \frac{1}{2\pi e} e^{2h(\theta)}, \tag{A.26}$$

with equality if and only if θ is Gaussian.

□

Corollary A.A.1.1. *Given side information $\mathbf{y} \in \mathbb{R}^L$ and estimator $\widehat{\theta}(\mathbf{y})$, it follows that*

$$\mathbb{E}_{(\theta,\mathbf{y})}\left[(\theta - \widehat{\theta}(\mathbf{y}))^2\right] \geq \frac{1}{2\pi e} e^{2h(\theta|\mathbf{y})}, \tag{A.27}$$

with equality if and only if θ is Gaussian and the estimator $\widehat{\theta}(\mathbf{y})$ is the posterior expectation, i.e., $\widehat{\theta}(\mathbf{y}) = \mathbb{E}[\theta|\mathbf{y}]$.

Proof. Let $\widehat{\theta}(\mathbf{y})$ be any estimator of θ from observation $\mathbf{y}$. Then

$$\mathbb{E}_{(\theta,\mathbf{y})}\left[(\theta - \widehat{\theta}(\mathbf{y}))^2\right] \;\geq\; \min_{\widehat{\theta}} \mathbb{E}_{(\theta,\mathbf{y})}\left[(\theta - \widehat{\theta}(\mathbf{y}))^2\right] \;=\; \mathbb{E}_{(\theta,\mathbf{y})}\left[(\theta - \mathbb{E}[\theta|\mathbf{y}])^2\right] \tag{A.28}$$

$$= \mathbb{E}_{\mathbf{y}}\left[\mathrm{Var}(\theta|\mathbf{y})\right] \tag{A.29}$$

Note that $\mathbb{E}_{\mathbf{y}}\left[\mathrm{Var}(\theta|\mathbf{y})\right]$ is the average variance of $\theta|\mathbf{y}$ (when averaging over $\mathbf{y}$). Likewise, by definition, $h(\theta|\mathbf{y})$ is the average entropy of $\theta|\mathbf{y}$ (averaged over $\mathbf{y}$). Therefore, by applying the theorem above, one gets the expected inequality. Like in the theorem above, equality is attained when θ is Gaussian and the estimator $\widehat{\theta}(\mathbf{y})$ is the posterior expectation, ie $\widehat{\theta}(\mathbf{y}) = \mathbb{E}[\theta|\mathbf{y}]$. □

No assumption over the distribution or dimensionality of $\mathbf{y}$ is necessary. This implies that this bound works whether the relation between θ and $y_{n\ell}$ is linear or not.

In the general case, we can't compute $h(\theta|\mathbf{y})$ analytically, and we need to estimate it with an MC estimator. When the pdfs can be evaluated – which is the case in this chapter – then the standard estimator from Eq. A.10 can be used. When the pdfs cannot be evaluated, the entropy can be evaluated with the estimator from Kraskov et al. (2004).

In higher dimensions

Theorem A.A.2 (Estimator MSE and differential entropy in multivariate case)**.** *For any random vector $\boldsymbol{\theta} \in \mathbb{R}^D$ and any estimator $\widehat{\boldsymbol{\theta}}$,*

$$\mathbb{E}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\right] \geq \frac{D}{2\pi e} \exp\left(\frac{2}{D} h(\boldsymbol{\theta})\right), \tag{A.30}$$

with equality if and only if $\boldsymbol{\theta}$ is Gaussian, $\widehat{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta}]$ and the covariance matrix of $\boldsymbol{\theta}$ is proportional to the identity.

Proof. Let $\widehat{\boldsymbol{\theta}}$ be any estimator of $\boldsymbol{\theta}$.

$$\mathbb{E}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\right] \;\geq\; \min_{\widehat{\boldsymbol{\theta}}} \mathbb{E}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\right] \;=\; \operatorname{Tr} \mathbf{C}, \tag{A.31}$$

with $\mathbf{C} \in \mathcal{M}_D(\mathbb{R})$ the covariance matrix of $\boldsymbol{\theta}$. Then, to get a relation between the trace of the covariance matrix of $\boldsymbol{\theta}$ and its entropy, one uses the fact that for a random vector with given covariance matrix, the multivariate Gaussian distribution has maximum entropy. Let $Z \sim \mathcal{N}(0, \mathbf{C})$. Then,

$$h(\boldsymbol{\theta}) \leq h(Z) = \frac{1}{2} \ln(\det(2\pi e \mathbf{C})), \tag{A.32}$$

which yields

$$\det(\mathbf{C}) \geq \frac{1}{(2\pi e)^D} \exp\left(2h(\boldsymbol{\theta})\right). \tag{A.33}$$

Using the arithmetic mean geometric mean inequality, i.e.,

$$[\det \mathbf{C}]^{\frac{1}{D}} \leq \frac{\operatorname{Tr} \mathbf{C}}{D}, \tag{A.34}$$

we get

$$\operatorname{Tr} \mathbf{C} \geq D\,[\det \mathbf{C}]^{\frac{1}{D}} \geq \frac{D}{2\pi e} \exp\left(\frac{2}{D} h(\boldsymbol{\theta})\right), \tag{A.35}$$

with equality iif $\boldsymbol{\theta}$ is Gaussian and C is proportional to the identity.

□

Corollary A.A.2.1. *Given side information $\mathbf{y}$ and estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, it follows that*

$$\mathbb{E}_{(\boldsymbol{\theta},\mathbf{y})}\left[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y})\|^2\right] \geq \frac{D}{2\pi e} \exp\left(\frac{2}{D} h(\boldsymbol{\theta}|\mathbf{y})\right) \tag{A.36}$$

Proof. The proof of this corollary is similar to the previous one: we rewrite the MSE to get an average covariance matrix for $\boldsymbol{\theta}|\mathbf{y}$ (averaged over $\mathbf{y}$), and apply the theorem to get the inequality with $h(\boldsymbol{\theta}|\mathbf{y})$. □

Here again, no assumption is made over the distribution of $\mathbf{y}$.

Acronyms

ABC approximate Bayes computing
AdaGrad adaptive gradient algorithm
AE absolute error
AIC Akaike information criterion
ALMA Atacama large millimeter/submillimeter array
ANN artificial neural network
ASIS ancillarity-sufficiency interweaving strategy
au astronomical unit

BB black body
BIC Bayesian information criterion
BO Bayesian optimization

cdf cumulative density function
cgs centimetre-gram-second
CI credibility interval
CL Cauchy loss function
CLT central limit theorem
CNM cold neutral medium

DGMC distributed genetic MC
DIC deviance information criterion
DMC darting MC

EES equi-energy sampler
EF error factor
elpd expected log-predictive density
EMC evolutionary MC
ESS effective sample size
EUV extreme ultraviolet

FIR far infrared
FUV far ultraviolet

GA genetic algorithm
GD gradient descent
GELU Gaussian error linear unit
GMC giant molecular cloud
GMM Gaussian mixture model
GP Gaussian process

HIM hot ionized medium
HMC Hamiltonian Monte Carlo
HPR high probability region
HST Hubble space telescope

ICA independent component analysis
i.i.d. independent and identically distributed
I-MH independent Metropolis-Hastings
I-MTM independent multiple-try Metropolis
IR infrared
ISM interstellar medium
ISRF interstellar radiation field

JAMS jumping adaptative multimodal sampler
JWST James Webb spatial telescope

KDE kernel density estimation
KL Kullback-Leibler
KS Kolmogorov-Smirnov

L-BFGS limited memory BFGS
LHS latin hypercube sampling
LM Levenberg-Marquardt
LMC large Magellanic cloud
lpd log-predictive density
LTE local thermal equilibrium
LVG large velocity gradient
ly light-year

MAE mean absolute error
MALA Metropolis adjusted Langevin algorithm
MAP maximum a posteriori
MBB modified black body
MC Monte Carlo
MCMC Markov chain Monte Carlo
MH Metropolis-Hastings
ML machine learning
MLE maximum likelihood estimator
MMALA manifold Metropolis adjusted Langevin algorithm
MMSE minimum mean square error
MSE mean squared error
MTM multiple-try Metropolis
MW Milky Way

MYULA Moreau-Yosida unadjusted Langevin algorithm

NIR near infrared

OLHS orthogonal LHS

PAH polycyclic aromatic hydrocarbons
pc parsec
PCA principal component analysis
PCR principal component regression
pdf probability density function
PDR photodissociation region
PGD preconditioned gradient descent
PMALA preconditioned Metropolis adjusted Langevin algorithm
pmf probability mass function

QMC quasi-Monte Carlo

RBF radial basis function
RDMC regeneration darting MC
ReLU rectified error linear unit
RF random forest
RJ Rayleigh-Jeans
RMHMC Riemannian manifold HMC
RMSProp root mean squared propagation
R-SNR reconstruction signal-to-noise ratio
RWMH random walk Metropolis-Hastings

SA simulated annealing
SDE stochastic differential equation
SE squared error
SED spectral energy density
SFR star formation rate
SGD stochastic gradient descent
SG-MCMC stochastic gradient MCMC
SLED spectral line energy distribution
SLHS symmetric LHS
SMC sequential MC
SNR signal-to-noise ratio

UF uncertainty factor
ULA unadjusted Langevin algorithm
UV ultraviolet

VBI variational Bayes inference

WAIC widely applicable information criterion
WBIC widely applicable Bayesian information criterion
WHMC wormhole Hamiltonian Monte Carlo
WIM warm ionized medium
WNM warm neutral medium

YSO young stellar object

References

"Certain authors, speaking of their works, say, "My book", "My commentary", "My history", etc. [...] They would do better to say, "Our book", "Our commentary", "Our history", etc., because there is in them usually more of other people's than their own".

Blaise Pascal, *Pensées*

Abril-Pla, Oriol et al. (2023). "PyMC: a modern, and comprehensive probabilistic programming framework in Python". *PeerJ Comput. Sci.* 9, e1516 (cit. on p. 123).

Acquaviva, Viviana, Eric Gawiser, and Lucia Guaita (2011). "Spectral Energy Distribution Fitting with Markov Chain Monte Carlo: Methodology and Application to z = 3.1 Lyα-emitting Galaxies". *ApJ* 737, p. 47 (cit. on pp. 73, 74).

Agarwal, Shankar, Filipe B. Abdalla, Hume A. Feldman, Ofer Lahav, and Shaun A. Thomas (2012). "PkANN - I. Non-linear matter power spectrum interpolation through artificial neural networks: Matter power spectrum using artificial neural networks". *MNRAS* 424.2, pp. 1409–1418 (cit. on pp. 63, 65).

Ahn, Sungjin, Yutian Chen, and Max Welling (2013). "Distributed and Adaptive Darting Monte Carlo through Regenerations". *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. PMLR, pp. 108–116 (cit. on pp. 58, 130, 131).

Akeret, Joel, Alexandre Refregier, Adam Amara, Sebastian Seehars, and Caspar Hasner (2015). *Approximate Bayesian Computation for Forward Modeling in Cosmology*. (Visited on 07/26/2023). preprint (cit. on p. 70).

Alecian, E. et al. (2008). "Characterization of the magnetic field of the Herbig Be star HD200775". *MNRAS* 385, pp. 391–403 (cit. on p. 144).

Allers, K. N., D. T. Jaffe, J. H. Lacy, B. T. Draine, and M. J. Richter (2005). "H2 Pure Rotational Lines in the Orion Bar". *ApJ* 630, pp. 368–380 (cit. on pp. 25, 167).

ALMA Partnership et al. (2015). "The 2014 ALMA Long Baseline Campaign: First Results from High Angular Resolution Observations toward the HL Tau Region". *ApJ* 808, p. L3 (cit. on p. 17).

Andrae, Rene, Tim Schulze-Hartung, and Peter Melchior (2010). *Dos and don'ts of reduced chi-squared*. (Visited on 02/20/2023). preprint (cit. on pp. 66, 75).

Andrews, Sean M. et al. (2016). "Ringed Substructure and a Gap at 1 au in the Nearest Protoplanetary Disk". *ApJ* 820, p. L40 (cit. on p. 17).

Andricioaei, Ioan, John E. Straub, and Arthur F. Voter (2001). "Smart Darting Monte Carlo". *J. Chem. Phys.* 114.16, pp. 6994–7000 (cit. on p. 58).

Angelopoulos, Anastasios N. and Stephen Bates (2022). *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. URL: http://arxiv.org/abs/2107.07511 (visited on 08/21/2023). preprint (cit. on p. 39).

Aras, Efe, Kuan-Yun Lee, Ashwin Pananjady, and Thomas A. Courtade (2019). "A Family of Bayesian Cramér-Rao Bounds, and Consequences for Log-Concave Priors" (cit. on p. 180).

Ardizzone, Lynton, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe (2019). *Guided Image Generation with Conditional Invertible Neural Networks*. URL: http://arxiv.org/abs/1907.02392 (visited on 09/04/2023). preprint (cit. on p. 32).

Asensio Ramos, Andrés and Moshe Elitzur (2018). "MOLPOP-CEP: an exact, fast code for multi-level systems". *A&A* 616, A131 (cit. on p. 21).

Asmussen, Søren and Peter W. Glynn (2007). *Stochastic Simulation: Algorithms and Analysis*. Red. by B. Rozovskii et al. Vol. 57. Stochastic Modelling and Applied Probability. New York, NY: Springer (cit. on p. 91).

Auld, T., M. Bridges, M. P. Hobson, and S. F. Gull (2007). "Fast cosmological parameter estimation using neural networks". *MNRAS: Letters* 376.1, pp. L11–L15 (cit. on pp. 65, 96, 99).

Bailer-Jones, C. A. L. (2011). "Bayesian inference of stellar parameters and interstellar extinction using parallaxes and multiband photometry". *MNRAS* 411, pp. 435–452 (cit. on pp. 64, 66, 71).

Barnard, John, Robert McCulloch, and Xiao-Li Meng (2000). "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage". *Statistica Sinica* 10.4, pp. 1281–1311 (cit. on p. 71).

Bayarri, M. J. and James O. Berger (2000). "P Values for Composite Null Models". *Journal of the American Statistical Association* 95.452, pp. 1127–1142 (cit. on p. 54).

Beaumont, Mark A, Wenyang Zhang, and David J Balding (2002). "Approximate Bayesian Computation in Population Genetics". *Genetics* 162.4, pp. 2025–2035 (cit. on p. 70).

Beaumont, Mark A. (2010). "Approximate Bayesian Computation in Evolution and Ecology". *Annual Review of Ecology, Evolution, and Systematics* 41, pp. 379–406 (cit. on p. 70).

Beck, Amir (2017). *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics (cit. on pp. 35, 38).

Behrens, Erica et al. (2022). "Tracing Interstellar Heating: An ALCHEMI Measurement of the HCN Isomers in NGC 253". *ApJ* 939, p. 119 (cit. on pp. 62, 66, 71, 73).

Benisty, M. et al. (2013). "Enhanced Hα activity at periastron in the young and massive spectroscopic binary HD 200775". *A&A* 555, A113 (cit. on p. 144).

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Information science and statistics. New York: Springer. 738 pp. (cit. on p. 64).

Blanc, Guillermo A., Lisa Kewley, Frédéric P. A. Vogt, and Michael A. Dopita (2015). "IZI: Inferring the Gas Phase Metallicity (Z) and Ionization Parameter (q) of Ionized Nebulae Using Bayesian Statistics". *ApJ* 798, p. 99 (cit. on pp. 63, 64, 66, 70, 71, 73, 87).

Bohlin, R. C., B. D. Savage, and J. F. Drake (1978). "A survey of interstellar H I from Lalpha absorption measurements. II." *ApJ* 224, pp. 132–142 (cit. on pp. 20, 25).

Bojanov, B. D., H. A. Hakopian, and A. A. Sahakian (1993). *Spline Functions and Multivariate Interpolations*. Dordrecht: Springer Netherlands (cit. on p. 87).

Brewer, Brendon J., Livia B. Pártay, and Gábor Csányi (2011). "Diffusive nested sampling". *Stat Comput* 21.4, pp. 649–656 (cit. on p. 59).

Brinch, C. and M. R. Hogerheijde (2010). "LIME - a flexible, non-LTE line excitation and radiation transfer method for millimeter and far-infrared wavelengths". *A&A* 523, A25 (cit. on p. 21).

Bron, Emeric, Marcelino Agúndez, Javier R. Goicoechea, and José Cernicharo (2018a). *Photoevaporating PDR models with the Hydra PDR Code*. arXiv e-prints. (Visited on 07/18/2023). preprint (cit. on p. 26).

Bron, Emeric et al. (2018b). "Clustering the Orion B giant molecular cloud based on its molecular emission". *A&A* 610, A12 (cit. on p. 1).

Bron, Emeric et al. (2021). "Tracers of the ionization fraction in dense and translucent gas: I. Automated exploitation of massive astrochemical model grids". *A&A* 645, A28 (cit. on pp. 63, 64, 87).

Brooks, K. J. et al. (2003). "The Trumpler 14 photodissociation region in the Carina Nebula". *A&A* 412, pp. 751–765 (cit. on p. 149).

Buchner, Johannes (2016a). "PyMultiNest: Python interface for MultiNest". *Astrophysics Source Code Library*, ascl:1606.005 (cit. on p. 123).
— (2016b). "UltraNest: Pythonic Nested Sampling Development Framework and UltraNest". *Astrophysics Source Code Library*, ascl:1611.001 (cit. on pp. 73, 75).
— (2021). "UltraNest - a robust, general purpose Bayesian inference engine". *The Journal of Open Source Software* 6, p. 3001 (cit. on pp. 73, 75).
— (2023). "Nested Sampling Methods". *Statist. Surv.* 17 (none) (cit. on p. 59).
Burton, Michael G., D. J. Hollenbach, and A. G. G. M. Tielens (1990). "Line Emission from Clumpy Photodissociation Regions". *ApJ* 365, p. 620 (cit. on p. 96).
Cai, Xiaohao, Jason D. McEwen, and Marcelo Pereyra (2022). "Proximal nested sampling for high-dimensional Bayesian model selection". *Stat Comput* 32.5, p. 87 (cit. on p. 59).
Cameron, E. and A. N. Pettitt (2012). "Approximate Bayesian Computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift". *MNRAS* 425, pp. 44–65 (cit. on pp. 70, 73).
Chelouah, R. and P. Siarry (2000). "A Continuous Genetic Algorithm Designed for the Global Optimization of Multimodal Functions". *Journal of Heuristics* 6.2, pp. 191–213 (cit. on p. 56).
Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (cit. on p. 87).
Chevallard, J., S. Charlot, B. Wandelt, and V. Wild (2013). "Insights into the content and spatial distribution of dust from the integrated spectral properties of galaxies". *MNRAS* 432, pp. 2061–2091 (cit. on pp. 66, 73).
Chevallard, Jacopo and Stéphane Charlot (2016). "Modelling and interpreting spectral energy distributions of galaxies with BEAGLE". *MNRAS* 462, pp. 1415–1443 (cit. on pp. 73, 75, 82).
Chevance, M. et al. (2016). "A milestone toward understanding PDR properties in the extreme environment of LMC-30Dor". *A&A* 590, A36 (cit. on pp. 66, 75).
Chib, Siddhartha and Ivan Jeliazkov (2001). "Marginal Likelihood From the Metropolis–Hastings Output". *Journal of the American Statistical Association* 96.453, pp. 270–281 (cit. on p. 50).
Chokshi, A., A. G. G. M. Tielens, M. W. Werner, and M. W. Castelaz (1988). "C II 158 Micron and O i 63 Micron Observations of NGC 7023: A Model for Its Photodisscoiation Region". *ApJ* 334, p. 803 (cit. on p. 144).
Chouzenoux, Emilie, Jean-Christophe Pesquet, and Audrey Repetti (2016). "A block coordinate variable metric forward–backward algorithm". *J Glob Optim* 66.3, pp. 457–485 (cit. on p. 38).
Christensen, Nelson, Renate Meyer, Lloyd Knox, and Ben Luey (2001). "Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements". *Classical and Quantum Gravity* 18, pp. 2677–2688 (cit. on p. 73).
Ciurlo, A., T. Paumard, D. Rouan, and Y. Clénet (2016). "Hot molecular hydrogen in the central parsec of the Galaxy through near-infrared 3D fitting". *A&A* 594, A113 (cit. on pp. 66, 71, 82).
Compiègne, M. et al. (2011). "The global dust SED: tracing the nature and evolution of dust with DustEM". *A&A* 525, A103 (cit. on pp. 22, 23).
Cover, Thomas M and Joy A Thomas (2006). *Elements of Information Theory*. Second Edition. Wiley-Interscience (cit. on p. 184).
Cranmer, Kyle, Johann Brehmer, and Gilles Louppe (2020). "The frontier of simulation-based inference". *PNAS* 117.48, pp. 30055–30062 (cit. on p. 70).
Da Cunha, Elisabete, Stéphane Charlot, and David Elbaz (2008). "A simple model to interpret the ultraviolet, optical and infrared emission from galaxies". *MNRAS* 388.4, pp. 1595–1617 (cit. on p. 73).
Dauphin, Yann N., Harm de Vries, and Yoshua Bengio (2015). "Equilibrated adaptive learning rates for non-convex optimization" (cit. on pp. 38, 39, 46, 175).
De Jong, T., W. Boland, and A. Dalgarno (1980). "Hydrostatic models of molecular clouds." *A&A* 91, pp. 68–84 (cit. on p. 24).

Del Moral, Pierre (2004). *Feynman-Kac Formulae*. Red. by J. Gani, C. C. Heyde, and T. G. Kurtz. Probability and its Applications. New York, NY: Springer (cit. on p. 57).

Del Moral, Pierre, Arnaud Doucet, and Ajay Jasra (2006). "Sequential Monte Carlo Samplers". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.3, pp. 411–436 (cit. on pp. 50, 57, 123, 124).

De Mijolla, D., S. Viti, J. Holdship, I. Manolopoulou, and J. Yates (2019). "Incorporating astrochemistry into molecular line modelling via emulation". *A&A* 630, A117 (cit. on pp. 64, 88).

Dishoeck, Ewine F. van and Edwin A. Bergin (2021). "Astrochemistry and Planet Formation". *ExoFrontiers: Big questions in exoplanetary science*. IOP Publishing (cit. on pp. 16, 17).

Draine, B. T. (1978). "Photoelectric heating of interstellar gas." *ApJS* 36, pp. 595–619 (cit. on p. 20).

Draine, Bruce T. (2011). *Physics of the interstellar and intergalactic medium*. Princeton series in astrophysics. Princeton, N.J: Princeton University Press. 540 pp. (cit. on pp. 13–16, 19, 79).

Dreo, Johann, Patrick Siarry, Alain Petrowski, and Eric Taillard (2006). *Metaheuristics for hard optimization : methods and case studies*. Springer-Verlag. 372 pp. (cit. on pp. 56, 183).

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". *Journal of Machine Learning Research* 12.61, pp. 2121–2159 (cit. on p. 38).

Dullemond, C. P. et al. (2012). "RADMC-3D: A multi-purpose radiative transfer tool". *Astrophysics Source Code Library*, ascl:1202.015 (cit. on p. 21).

Durand, Sylvain, Jalal Fadili, and Mila Nikolova (2010). "Multiplicative Noise Removal Using L1 Fidelity on Frame Coefficients". *J Math Imaging Vis* 36.3, pp. 201–226 (cit. on p. 115).

Durmus, Alain and Éric Moulines (2017). "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". *The Annals of Applied Probability* 27.3, pp. 1551–1587 (cit. on p. 44).

Durmus, Alain, Gareth O. Roberts, Gilles Vilmart, and Konstantinos C. Zygalakis (2017). "Fast Langevin based algorithm for MCMC in high dimensions". *The Annals of Applied Probability* 27.4, pp. 2195–2237 (cit. on p. 45).

Eberhart, R. and J. Kennedy (1995). "A new optimizer using particle swarm theory". *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*. MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, pp. 39–43 (cit. on p. 56).

Einig, L. et al. (2023). "Deep learning denoising by dimension reduction: Application to the ORION-B line cubes". *A&A* (cit. on pp. 77, 79, 98, 175).

Einig, Lucas et al. (n.d.). "Réduction de modèles multi-physiques par réseaux de neurones" (cit. on p. 86).

Fasshauer, Gregory E. (2007). *Meshfree Approximation Methods with Matlab* (cit. on p. 87).

Fendt, William A. and Benjamin D. Wandelt (2007). "Pico: Parameters for the Impatient Cosmologist". *ApJ* 654.1, p. 2 (cit. on p. 65).

Ferland, G J et al. (2017). "THE 2017 RELEASE OF Cloudy", p. 54 (cit. on p. 22).

Feroz, F. and M. P. Hobson (2008). "Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses". *MNRAS* 384, pp. 449–463 (cit. on pp. 59, 75, 123, 124).

Feroz, F., M. P. Hobson, and M. Bridges (2009). "MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics". *MNRAS* 398, pp. 1601–1614 (cit. on pp. 59, 73).

Fitzpatrick, Edward L. and Derck Massa (1990). "An Analysis of the Shapes of Ultraviolet Extinction Curves. III. an Atlas of Ultraviolet Extinction Curves". *ApJS* 72, p. 163 (cit. on pp. 20, 25).

Foreman-Mackey, Daniel, David W. Hogg, Dustin Lang, and Jonathan Goodman (2013). "emcee: The MCMC Hammer". *Publications of the Astronomical Society of the Pacific* 125, p. 306 (cit. on pp. 57, 73, 123).

Friel, Nial and Jason Wyse (2012). "Estimating the evidence - a review". *Statistica Neerlandica* 66.3, pp. 288–308 (cit. on p. 50).

Galliano, F. et al. (2003). "ISM properties in low-metallicity environments. II. The dust spectral energy distribution of NGC 1569". *A&A* 407, pp. 159–176 (cit. on pp. 66, 72).

Galliano, Frédéric (2018). "A dust spectral energy distribution model with hierarchical Bayesian inference - I. Formalism and benchmarking". *MNRAS* 476, pp. 1445–1469 (cit. on pp. 63, 64, 66, 67, 71, 74, 76, 77, 82, 118, 137, 174, 175).

— (2022). "A nearby galaxy perspective on interstellar dust properties and their evolution" (cit. on pp. 14, 15, 174).

Galliano, Frédéric et al. (2021). "A nearby galaxy perspective on dust evolution. Scaling relations and constraints on the dust build-up in galaxies with the DustPedia and DGS samples". *A&A* 649, A18 (cit. on pp. 65–67, 71, 74, 76, 77, 82, 174).

Gaudel, Mathilde et al. (2023). "Gas kinematics around filamentary structures in the Orion B cloud". *A&A* 670, A59 (cit. on p. 79).

Gelman, Andrew (2013). "Two simple examples for understanding posterior p-values whose distributions are far from uniform". *Electron. J. Statist.* 7 (none) (cit. on p. 54).

Gelman, Andrew, Xiao-Li Meng, and Hal Stern (1996). "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies". *Statistica Sinica* 6.4, pp. 733–760 (cit. on pp. 52, 53, 55, 76, 173).

Gelman, Andrew et al. (2015). *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC (cit. on pp. 32, 42, 47, 49–52, 54, 76, 127).

Geman, Stuart and Donald Geman (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6, pp. 721–741 (cit. on pp. 46, 47, 173).

Gilks, Walter R., Gareth O. Roberts, and Sujit K. Sahu (1998). "Adaptive Markov Chain Monte Carlo through Regeneration". *Journal of the American Statistical Association* 93.443, pp. 1045–1054 (cit. on p. 58).

Girolami, Mark and Ben Calderhead (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2, pp. 123–214 (cit. on pp. 45, 46).

Godard, B. et al. (2019). "Models of irradiated molecular shocks". *A&A* 622, A100 (cit. on pp. 22, 157).

Goicoechea, Javier R. et al. (2016). "Compression and ablation of the photo-irradiated molecular cloud the Orion Bar". *Nature* 537, pp. 207–209 (cit. on p. 25).

Goicoechea, Javier R. et al. (2019). "Molecular tracers of radiative feedback in Orion (OMC-1): Widespread CH^+ ($J = 1$–0), CO (10–9), HCN (6–5), and HCO^+ (6–5) emission". *A&A* 622, A91 (cit. on pp. 27, 141, 156).

Gonzalez, Joseph, Yucheng Low, Arthur Gretton, and Carlos Guestrin (2011). "Parallel Gibbs Sampling: From Colored Fields to Thin Junction Trees". *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, pp. 324–332 (cit. on pp. 47, 121, 135, 173).

Goodman, Jonathan and Jonathan Weare (2010). "Ensemble samplers with affine invariance". *CAMCoS* 5.1, pp. 65–80 (cit. on pp. 57, 123, 124).

Gordon, Karl D. et al. (2014). "Dust and Gas in the Magellanic Clouds from the HERITAGE Herschel Key Project. I. Dust Properties and Insights into the Origin of the Submillimeter Excess Emission". *ApJ* 797, p. 85 (cit. on pp. 66, 67).

Graff, Philip, Farhan Feroz, Michael P. Hobson, and Anthony Lasenby (2014). "SKYNET: an efficient and robust neural network training tool for machine learning in astronomy". *MNRAS* 441, pp. 1741–1759 (cit. on p. 64).

Grassi, T. et al. (2011). "ROBO: a model and a code for studying the interstellar medium". *A&A* 533, A123 (cit. on pp. 64, 88).

Grassi, T. et al. (2022). "Reducing the complexity of chemical networks via interpretable autoencoders". *A&A* 668, A139 (cit. on pp. 64, 88).
Gratier, P. et al. (2016). "A New Reference Chemical Composition for TMC-1". *ApJS* 225, p. 25 (cit. on pp. 68, 73, 111).
Gratier, Pierre et al. (2017). "Dissecting the molecular structure of the Orion B cloud: insight from principal component analysis". *A&A* 599, A100 (cit. on p. 1).
Guggenheimer, Heinrich W., Alan S. Edelman, and Charles R. Johnson (1995). "A Simple Estimate of the Condition Number of a Linear System". *The College Mathematics Journal* 26.1, pp. 2–5 (cit. on pp. 36, 38).
Guttman, Irwin (1967). "The Use of the Concept of a Future Observation in Goodness-of-Fit Problems". *Journal of the Royal Statistical Society. Series B (Methodological)* 29.1, pp. 83–100 (cit. on p. 52).
Haber, Seymour (1966). "A Modified Monte-Carlo Quadrature". *Mathematics of Computation* 20.95, pp. 361–368 (cit. on p. 91).
Habing, H. J. (1968). "The interstellar radiation density between 912 A and 2400 A". *Bulletin of the Astronomical Institutes of the Netherlands* 19, p. 421 (cit. on pp. xv, 20).
Hahn, ChangHoon et al. (2019). "Likelihood non-Gaussianity in large-scale structure analyses". *MNRAS* 485.2, pp. 2956–2969 (cit. on p. 70).
Handley, W. J., M. P. Hobson, and A. N. Lasenby (2015). "polychord: nested sampling for cosmology." *MNRAS* 450, pp. L61–L65 (cit. on p. 59).
Hannestad, Steen (1999). "Stochastic optimization methods for extracting cosmological parameters from cosmic microwave background radiation power spectra". *Phys. Rev. D* 61.2, p. 023002 (cit. on p. 72).
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (n.d.). *The Elements of Statistical Learning* (cit. on p. 70).
Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". *Biometrika* 57.1, pp. 97–109 (cit. on p. 43).
Hatta, Yoshiki, Takashi Sekii, Othman Benomar, and Masao Takata (2022). "Bayesian Rotation Inversion of KIC 11145123". *ApJ* 927, p. 40 (cit. on p. 75).
He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (cit. on pp. 88, 105).
Heays, A. N., A. D. Bosman, and E. F. van Dishoeck (2017). "Photodissociation and photoionisation of atoms and molecules of astrophysical interest". *A&A* 602, A105 (cit. on p. 25).
Heitmann, Katrin et al. (2009). "The Coyote Universe. II. Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum". *ApJ* 705.1, p. 156 (cit. on pp. 63, 65).
Hendrycks, Dan and Kevin Gimpel (2023). *Gaussian Error Linear Units (GELUs)*. URL: http://arxiv.org/abs/1606.08415 (visited on 07/13/2023). preprint (cit. on p. 99).
Hensley, Brandon S. and B. T. Draine (2023). "The Astrodust+PAH Model: A Unified Description of the Extinction, Emission, and Polarization from Dust in the Diffuse Interstellar Medium". *ApJ* 948, p. 55 (cit. on p. 22).
Hogerheijde, M. R. and F. F. S. van der Tak (2000). "An accelerated Monte Carlo method to solve two-dimensional radiative transfer and molecular excitation. With applications to axisymmetric models of star formation". *A&A* 362, pp. 697–710 (cit. on p. 21).
Holdship, J., N. Jeffrey, A. Makrymallis, S. Viti, and J. Yates (2018). "Bayesian Inference of the Rates of Surface Reactions in Icy Mantles". *ApJ* 866, p. 116 (cit. on pp. 62, 70, 71, 73, 86).
Holdship, J., S. Viti, T. J. Haworth, and J. D. Ilee (2021). "Chemulator: Fast, accurate thermochemistry for dynamical models through emulation". *A&A* 653, A76 (cit. on pp. 64, 88).
Holdship, J., S. Viti, I. Jiménez-Serra, A. Makrymallis, and F. Priestley (2017). "UCLCHEM: A Gas-grain Chemical Code for Clouds, Cores, and C-Shocks". *The Astronomical Journal* 154, p. 38 (cit. on p. 22).

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". *Neural Networks* 2.5, pp. 359–366 (cit. on p. 88).

Hu, Bo and Kam-Wah Tsui (2010). "Distributed evolutionary Monte Carlo for Bayesian computing". *Computational Statistics & Data Analysis* 54.3, pp. 688–697 (cit. on p. 56).

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger (2017). "Densely Connected Convolutional Networks". *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (cit. on pp. 88, 105).

Ihler, A.T., J.W. Fisher, R.L. Moses, and A.S. Willsky (2005). "Nonparametric belief propagation for self-localization of sensor networks". *IEEE Journal on Selected Areas in Communications* 23.4, pp. 809–819 (cit. on p. 130).

Indriolo, Nick, Thomas R. Geballe, Takeshi Oka, and Benjamin J. McCall (2007). "H+3 in Diffuse Interstellar Clouds: A Tracer for the Cosmic-Ray Ionization Rate". *ApJ* 671, pp. 1736–1747 (cit. on pp. 22, 25, 26).

Ishida, E. E. O. et al. (2015). "COSMOABC: Likelihood-free inference via Population Monte Carlo Approximate Bayesian Computation". *Astronomy and Computing* 13, pp. 1–11 (cit. on p. 70).

Ivezić, Željko, Andrew Connolly, Jacob T. Vanderplas, and Alexander Gray (2020). *Statistics, data mining, and machine learning in astronomy: a practical Python guide for the analysis of survey data*. Updated edition. Princeton series in modern observational astronomy. Princeton: Princeton University Press. 537 pp. (cit. on p. 75).

Jackiewicz, Jason (2020). "Probabilistic Inversions for Time-Distance Helioseismology". *Solar Physics* 295, p. 137 (cit. on p. 73).

Jennings, Elise and Maeve Madigan (2017). *astroABC: An Approximate Bayesian Computation Sequential Monte Carlo sampler for cosmological parameter estimation*. (Visited on 07/26/2023). preprint (cit. on pp. 70, 73).

Jimenez, Raul, Licia Verde, Hiranya Peiris, and Arthur Kosowsky (2004). "Fast cosmological parameter estimation from microwave background temperature and polarization power spectra". *Phys. Rev. D* 70.2, p. 023005 (cit. on p. 64).

Joblin, C. et al. (2018). "Structure of photodissociation fronts in star-forming regions revealed by *Herschel* observations of high-J CO emission lines". *A&A* 615, A129 (cit. on pp. 25, 71, 72, 75, 80, 82, 90, 95, 141, 144, 145, 148, 160, 166–168, 174).

Jóhannesson, G. et al. (2016). "Bayesian Analysis of Cosmic Ray Propagation: Evidence against Homogeneous Diffusion". *ApJ* 824, p. 16 (cit. on pp. 66, 73).

Johnson, Alicia A., Galin L. Jones, and Ronald C. Neath (2013). "Component-Wise Markov Chain Monte Carlo: Uniform and Geometric Ergodicity under Mixing and Composition". *Statist. Sci.* 28.3 (cit. on p. 47).

Jones, A. P. et al. (2013). "The evolution of amorphous hydrocarbons in the ISM: dust modelling from a new vantage point". *A&A* 558, A62 (cit. on p. 22).

Jones, Galin L., Gareth O. Roberts, and Jeffrey S. Rosenthal (2014). "Convergence of Conditional Metropolis-Hastings Samplers". *Advances in Applied Probability* 46.2, pp. 422–445 (cit. on p. 122).

Joseph, V. Roshan (2016). "Space-filling designs for computer experiments: A review". *Quality Engineering* 28.1, pp. 28–35 (cit. on p. 91).

Juvela, M., J. Montillaud, N. Ysard, and T. Lunttila (2013). "The degeneracy between dust colour temperature and spectral index. Comparison of methods for estimating the β(T) relation". *A&A* 556, A63 (cit. on p. 71).

Kamenetzky, J., N. Rangwala, J. Glenn, P. R. Maloney, and A. Conley (2014). "A Survey of the Molecular ISM Properties of Nearby Galaxies Using the Herschel FTS". *ApJ* 795, p. 174 (cit. on pp. 73, 75).

Kao, Yi-Tung and Erwie Zahara (2008). "A hybrid genetic algorithm and particle swarm optimization for multimodal functions". *Applied Soft Computing* 8.2, pp. 849–857 (cit. on p. 56).

Kashyap, Samarth G., Srijan Bharati Das, Shravan M. Hanasoge, Martin F. Woodard, and Jeroen Tromp (2021). "Inferring Solar Differential Rotation through Normal-mode Coupling Using Bayesian Statistics". *ApJS* 253, p. 47 (cit. on p. 73).

Keil, Marcus, Serena Viti, and Jonathan Holdship (2022). "UCLCHEMCMC: An MCMC Inference Tool for Physical Parameters of Molecular Clouds". *ApJ* 927, p. 203 (cit. on pp. 62, 63, 65, 66, 73).

Kelly, Brandon C. (2011). "Comment". *Journal of Computational and Graphical Statistics* 20.3, pp. 584–591 (cit. on pp. 74, 175).

Kelly, Brandon C. et al. (2012). "Dust spectral energy distributions in the era of Herschel and Planck: A hierarchical Bayesian-fitting technique". *ApJ* 752.1, p. 55 (cit. on pp. 67, 71, 74, 77, 115, 175).

Kennedy, J. and R. Eberhart (1995). "Particle swarm optimization". *Proceedings of ICNN'95 - International Conference on Neural Networks*. Proceedings of ICNN'95 - International Conference on Neural Networks. Vol. 4, 1942–1948 vol.4 (cit. on p. 56).

Kingma, Diederik P. and Jimmy Ba (2017). "Adam: A Method for Stochastic Optimization" (cit. on p. 90).

Koenker, Roger (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge: Cambridge University Press (cit. on p. 39).

Konishi, Sadanori and Genshiro Kitagawa (2008). *Information Criteria and Statistical Modeling*. Springer Series in Statistics. New York, NY: Springer (cit. on p. 51).

Kou, S. C., Qing Zhou, and Wing Hung Wong (2006). "Equi-energy sampler with applications in statistical inference and statistical mechanics". *The Annals of Statistics* 34.4, pp. 1581–1619 (cit. on p. 56).

Kramer, C. et al. (2008). "Clumpy photon-dominated regions in Carina. I. [C I] and mid-J CO lines in two 4'×4' fields". *A&A* 477, pp. 547–555 (cit. on p. 149).

Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger (2004). "Estimating mutual information". *Phys. Rev. E* 69.6, p. 066138 (cit. on pp. 137, 184).

Krissian, Karl, Carl-Fredrik Westin, Ron Kikinis, and Kirby G. Vosburgh (2007). "Oriented Speckle Reducing Anisotropic Diffusion". *IEEE Transactions on Image Processing* 16.5, pp. 1412–1424 (cit. on p. 115).

Lan, Shiwei, Jeffrey Streets, and Babak Shahbaba (2014). "Wormhole Hamiltonian Monte Carlo". *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI'14. Québec City, Québec, Canada: AAAI Press, pp. 1953–1959 (cit. on pp. 58, 123, 124, 130, 131).

Le Gall, Jean-François (2022). *Measure Theory, Probability, and Stochastic Processes*. Vol. 295. Graduate Texts in Mathematics. Cham: Springer International Publishing (cit. on p. 30).

Le Petit, F., E. Roueff, and E. Herbst (2004). "H_3^+ and other species in the diffuse cloud towards ζ Persei: A new detailed model". *A&A* 417, pp. 993–1002 (cit. on p. 25).

Le Petit, Franck, Cyrine Nehme, Jacques Le Bourlot, and Evelyne Roueff (2006). "A Model for Atomic and Molecular Interstellar Gas: The Meudon PDR Code". *The Astrophys. J. Supp. Series* 164.2, pp. 506–529 (cit. on pp. 2, 22, 23, 63, 95).

Lebouteiller, Vianney and Lise Ramambason (2022). *Topological models to infer multiphase interstellar medium properties*. (Visited on 10/26/2022). preprint (cit. on pp. 74, 76, 82, 174).

Lee, M.-Y. et al. (2019). "Radiative and mechanical feedback into the molecular gas in the Large Magellanic Cloud - II. 30 Doradus". *A&A* 628, A113 (cit. on pp. 66, 72).

Legin, Ronan, Alexandre Adam, Yashar Hezaveh, and Laurence Perreault-Levasseur (2023). "Beyond Gaussian Noise: A Generalized Approach to Likelihood Analysis with Non-Gaussian Noise". *ApJL* 949.2, p. L41 (cit. on p. 70).

Lemaire, J. L. et al. (1999). "High resolution Fourier transform spectroscopy of H_2 IR emission in NGC 7023". *A&A* 349, pp. 253–258 (cit. on p. 25).

Leshno, Moshe, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken (1993). "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". *Neural Networks* 6.6, pp. 861–867 (cit. on p. 88).

Lewis, Antony and Sarah Bridle (2002). "Cosmological parameters from CMB and other data: A Monte Carlo approach". *Physical Review D* 66, p. 103511 (cit. on p. 73).

Li, Chunyuan, Changyou Chen, David Carlson, and Lawrence Carin (2016). "Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks". *AAAI* 30.1 (cit. on pp. 46, 119, 120, 123, 173).

Liang, Faming and Wing Hung Wong (2000). "Evolutionary Monte Carlo: Applications to Cp model sampling and change point problem", p. 26 (cit. on p. 56).

— (2001). "Real-Parameter Evolutionary Monte Carlo With Applications to Bayesian Mixture Models". *Journal of the American Statistical Association* 96.454, pp. 653–666 (cit. on p. 57).

Liu, Jun S, Faming Liang, and Wing Hung Wong (2000). "The Multiple-Try Method and Local Optimization in Metropolis Sampling". *Journal of the American Statistical Association* 95, p. 15 (cit. on p. 48).

— (2021). "The Multiple-Try Method and Local Optimization in Metropolis Sampling", p. 15 (cit. on pp. 48, 118, 120).

Makrymallis, Antonios and Serena Viti (2014). "Understanding the Formation and Evolution of Interstellar Ices: A Bayesian Approach". *ApJ* 794, p. 45 (cit. on p. 73).

Manrique-Yus, Andrea and Elena Sellentin (2019). "Euclid-era cosmology for everyone: Neural net assisted MCMC sampling for the joint 3x2 likelihood". *MNRAS*, stz3059 (cit. on p. 65).

Marchal, Antoine et al. (2019). "ROHSA: Regularized Optimization for Hyper-Spectral Analysis - Application to phase separation of 21 cm data". *A&A* 626, A101 (cit. on pp. 71, 81, 117, 173, 175).

Marconi, A., L. Testi, A. Natta, and C. M. Walmsley (1998). "Near infrared spectra of the Orion bar". *A&A, v.330, p.696-710 (1998)* 330, p. 696 (cit. on pp. 25, 167).

Maret, Sébastien and Edwin A. Bergin (2015). "Astrochem: Abundances of chemical species in the interstellar medium". *Astrophysics Source Code Library*, ascl:1507.010 (cit. on p. 22).

Martino, Luca (2018). "A review of multiple try MCMC algorithms for signal processing". *Digital Signal Processing* 75, pp. 134–152 (cit. on pp. 48, 120, 122, 173).

Mathis, J. S., P. G. Mezger, and N. Panagia (1983). "Interstellar radiation field and dust temperatures in the diffuse interstellar medium and in giant molecular clouds". *A&A* 128, pp. 212–229 (cit. on pp. 20, 26).

Mathis, J. S., W. Rumpl, and K. H. Nordsieck (1977). "The size distribution of interstellar grains." *ApJ* 217, pp. 425–433 (cit. on p. 25).

McCulloch, Warren S. and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". *Bulletin of Mathematical Biophysics* 5.4, pp. 115–133 (cit. on p. 88).

McElroy, D. et al. (2013). "The UMIST database for astrochemistry 2012". *A&A* 550, A36 (cit. on p. 21).

McEwen, Jason D., Christopher G. R. Wallis, Matthew A. Price, and Matthew M. Docherty (2022). "Machine learning assisted Bayesian model comparison: learnt harmonic mean estimator" (cit. on pp. 50, 75).

McGuire, Brett A. (2018). "2018 Census of Interstellar, Circumstellar, Extragalactic, Protoplanetary Disk, and Exoplanetary Molecules". *ApJS* 239, p. 17 (cit. on p. 17).

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". *Technometrics* 21.2, pp. 239–245 (cit. on p. 91).

Meng, Xiao-Li (1994). "Posterior Predictive p-Values". *The Annals of Statistics* 22.3, pp. 1142–1160 (cit. on p. 52).

Menten, K. M., M. J. Reid, J. Forbrich, and A. Brunthaler (2007). "The distance to the Orion Nebula". *A&A* 474, pp. 515–520 (cit. on pp. 156, 167).

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). "Equation of State Calculations by Fast Computing Machines". *J. Chem. Phys.* 21.6, pp. 1087–1092 (cit. on p. 43).

Miasojedow, Btażej, Eric Moulines, and Matti Vihola (2013). "An Adaptive Parallel Tempering Algorithm". *Journal of Computational and Graphical Statistics* 22.3, pp. 649–664 (cit. on p. 56).

Mizutani, M., T. Onaka, and H. Shibai (2004). "Origin of diffuse C II 158 micron and Si II 35 micron emission in the Carina nebula". *A&A* 423, pp. 579–592 (cit. on p. 149).

Mlikota, Marko and Frank Schorfheide (2022). *Sequential Monte Carlo With Model Tempering*. URL: http://arxiv.org/abs/2202.07070 (visited on 09/07/2023). preprint (cit. on p. 58).

Möller, T. et al. (2013). "Modeling and Analysis Generic Interface for eXternal numerical codes (MAGIX)". *A&A* 549, A21 (cit. on p. 72).

Mootoovaloo, A., A. H. Jaffe, A. F. Heavens, and F. Leclercq (2022). "Kernel-based emulator for the 3D matter power spectrum from CLASS". *Astronomy and Computing* 38, p. 100508 (cit. on pp. 63, 65).

Motulsky, Harvey J. and Ronald E. Brown (2006). "Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate". *BMC Bioinformatics* 7.1, p. 123 (cit. on p. 99).

Nadarajah, Saralees (2005). "A generalized normal distribution". *Journal of Applied Statistics* 32.7, pp. 685–694 (cit. on p. 138).

Nardon, Martina and Paolo Pianca (2009). "Simulation techniques for generalized Gaussian densities". *Journal of Statistical Computation and Simulation* 79.11, pp. 1317–1329 (cit. on p. 139).

Neal, Radford (2011). "MCMC Using Hamiltonian Dynamics". *Handbook of Markov Chain Monte Carlo*. Ed. by Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. Vol. 20116022. Chapman and Hall/CRC (cit. on p. 44).

Nicholson, R. and J. P. Kaipio (2020). "An Additive Approximation to Multiplicative Noise". *J Math Imaging Vis* 62.9, pp. 1227–1237 (cit. on pp. 116, 172).

Nocedal, Jorge and Stephen J. Wright (2006). *Numerical optimization*. 2nd ed. Springer series in operations research. New York: Springer. 664 pp. (cit. on pp. 36, 37, 117).

Nogueira, Fernando (2014–). *Bayesian Optimization: Open source constrained global optimization tool for Python* (cit. on p. 138).

Nwankpa, Chigozie E, Anthony Gachagan, and Stephen Marshall (2021). "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning" (cit. on p. 89).

Oberst, T. E. et al. (2011). "A 205 μm [N II] Map of the Carina Nebula". *ApJ* 739, p. 100 (cit. on p. 149).

Öpik, E. J. (1962). "The lunar atmosphere". *Planetary and Space Science* 9.5, pp. 211–244 (cit. on p. 14).

Ostertagová, Eva (2012). "Modelling using Polynomial Regression". *Procedia Engineering* 48, pp. 500–506 (cit. on p. 105).

Pacifici, Camilla, Stéphane Charlot, Jérémy Blaizot, and Jarle Brinchmann (2012). "Relative merits of different types of rest-frame optical observations to constrain galaxy physical parameters". *MNRAS* 421, pp. 2002–2024 (cit. on p. 73).

Palud, P., P. Chainais, F. Le Petit, P.-A. Thouvenin, and E. Bron (2023a). "Problèmes inverses et test bayésien d'adéquation du modèle". *29° Colloque sur le traitement du signal et des images*. Grenoble: GRETSI - Groupe de Recherche en Traitement du Signal et des Images, p. 705–708 (cit. on pp. 6, 114).

Palud, P., P.-A. Thouvenin, P. Chainais, E. Bron, and F. Le Petit (2023b). "Efficient sampling of non log-concave posterior distributions with mixture of noises". *IEEE Transactions on Signal Processing* 71, pp. 2491–2501 (cit. on pp. 6, 114).

— (in prep[a]). "Bayesian inversion of large interstellar medium observation maps" (cit. on p. 6).

Palud, P. et al. (2022a). "Mélange de bruits et échantillonnage de posterior non log-concave". *28° Colloque sur le traitement du signal et des images*. 001-0176. Nancy: GRETSI - Groupe de Recherche en Traitement du Signal et des Images, p. 705–708 (cit. on pp. 6, 114).

Palud, P. et al. (2022b). "Mixture of noises and sampling of non-log-concave posterior distributions". *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 2031–2035 (cit. on pp. 6, 114).

Palud, P. et al. (2023c). "Neural network-based emulation of interstellar medium models". *A&A (in press)* (cit. on pp. 6, 86).

Palud, Pierre, Pierre-Antoine Thouvenin, Pierre Chainais, Emeric Bron, and Franck Le Petit (in prep[b]). "Entropy-based selection of most informative observables for inference from interstellar medium observations" (cit. on p. 7).

Panter, Benjamin, Alan F. Heavens, and Raul Jimenez (2003). "Star formation and metallicity history of the SDSS galaxy survey: unlocking the fossil record". *MNRAS* 343, pp. 1145–1154 (cit. on p. 72).

Paradis, D. et al. (2010). "Variations of the spectral index of dust emissivity from Hi-GAL observations of the Galactic plane". *A&A* 520, p. L8 (cit. on p. 73).

Paszke, Adam et al. (2017). "Automatic differentiation in pytorch" (cit. on p. 87).

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. (cit. on p. 106).

Patterson, Sam and Yee Whye Teh (2013). "Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex". *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc. (cit. on p. 46).

Paumard, Thibaut, Anna Ciurlo, Mark R. Morris, Tuan Do, and Andrea M. Ghez (2022). "Regularized 3D spectroscopy with CubeFit: Method and application to the Galactic Center circumnuclear disk". *A&A* 664, A97 (cit. on pp. 71, 72, 79, 82, 175).

Paumard, Thibaut, Mark R. Morris, Tuan Do, and Andrea Ghez (2014). "Regularized OSIRIS 3D spectroscopy at the circumnuclear disk ionization front". *The galactic center: Feeding and feedback in a normal galactic nucleus*. Ed. by L. O. Sjouwerman, C. C. Lang, and J. Ott. Vol. 303, pp. 109–113 (cit. on pp. 71, 82).

Pedregosa, F. et al. (2011). "Scikit-learn: Machine learning in Python". *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 92).

Pereyra, Marcelo (2017). "Maximum-a-Posteriori Estimation with Bayesian Confidence Regions". *SIAM J. Imaging Sci.* 10.1, pp. 285–302 (cit. on p. 33).

Pereyra, Marcelo, Jose M. Bioucas-Dias, and Mario A. T. Figueiredo (2015). "Maximum-a-posteriori estimation with unknown regularisation parameters". *2015 23rd European Signal Processing Conference (EUSIPCO)*. 2015 23rd European Signal Processing Conference (EUSIPCO). Nice: IEEE, pp. 230–234 (cit. on p. 137).

Pereyra, Marcelo et al. (2016). "A Survey of Stochastic Simulation and Optimization Methods in Signal Processing". *IEEE Journal of Selected Topics in Signal Processing* 10.2, pp. 224–241 (cit. on pp. 34, 38, 44, 45).

Pérez-Montero, E. (2014). "Deriving model-based Te-consistent chemical abundances in ionized gaseous nebulae". *MNRAS* 441, pp. 2663–2675 (cit. on pp. 63, 65, 66, 73).

Pety, Jérôme et al. (2017). "The anatomy of the Orion B giant molecular cloud: A local template for studies of nearby galaxies". *A&A* 599, A98 (cit. on pp. 1, 27, 76–78, 80, 110, 133, 171, 176, 178).

Pinte, C. et al. (2022). "MCFOST: Radiative transfer code". *Astrophysics Source Code Library*, ascl:2207.023 (cit. on p. 21).

Pompe, Emilia, Chris Holmes, and Krzysztof Latuszynski (2020). "A framework for adaptive MCMC targeting multimodal distributions". *Annals of Statistics* 48, pp. 2930–2952 (cit. on pp. 58, 130).

Raissi, M., P. Perdikaris, and G. E. Karniadakis (2019). "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". *Journal of Computational Physics* 378, pp. 686–707 (cit. on p. 174).

Ramachandra, Nesar, Georgios Valogiannis, Mustapha Ishak, Katrin Heitmann, and LSST Dark Energy Science Collaboration (2021). "Matter power spectrum emulator for f (R) modified gravity cosmologies". *Physical Review D* 103, p. 123525 (cit. on pp. 63, 65).

Ramambason, L. et al. (2022). "Inferring the HII region escape fraction of ionizing photons from infrared emission lines in metal-poor star-forming dwarf galaxies". *A&A* 667, A35 (cit. on pp. 63, 64, 68, 73, 75, 82, 86, 90, 172).

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. Cambridge, Mass: MIT Press. 248 pp. (cit. on p. 87).

Robert, Christian P. and George Casella (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York, NY: Springer New York (cit. on pp. 5, 31, 32, 40–44, 68, 120, 122, 123).

Roberts, G. O. and O. Stramer (2002). "Langevin Diffusions and Metropolis-Hastings Algorithms". *Methodology and Computing in Applied Probability* 4.4, pp. 337–357 (cit. on p. 44).

Roberts, G. O. and R. L. Tweedie (1996). "Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms". *Biometrika* 83.1, pp. 95–110 (cit. on p. 123).

Robin, A. C. et al. (2014). "Constraining the thick disc formation scenario of the Milky Way". *A&A* 569, A13 (cit. on pp. 70, 74).

Robins, James M., Aad van der Vaart, and Valerie Ventura (2000). "Asymptotic Distribution of P Values in Composite Null Models". *Journal of the American Statistical Association* 95.452, pp. 1143–1156 (cit. on p. 54).

Röllig, M. and V. Ossenkopf-Okada (2022). "The KOSMA- τ PDR model: I. Recent updates to the numerical model of photo-dissociated regions". *A&A* 664, A67 (cit. on pp. 22, 26, 96).

Rosenbrock, H. H. (1960). "An Automatic Method for Finding the Greatest or Least Value of a Function". *The Computer Journal* 3.3, pp. 175–184 (cit. on p. 92).

Roueff, Antoine et al. (2021). "C18O, 13CO, and 12CO abundances and excitation temperatures in the Orion B molecular cloud. Analysis of the achievable precision in modeling spectral lines within the approximation of the local thermodynamic equilibrium". *A&A* 645, A26 (cit. on pp. 66, 72).

Roueff, Evelyne and Jacques Le Bourlot (2020). "Sustained oscillations in interstellar chemistry models". *A&A* 643, A121 (cit. on p. 22).

Rousseeuw, Peter J. and Annick M. Leroy (1987). *Robust regression and outlier detection*. Wiley series in probability and mathematical statistics. New York: Wiley. 329 pp. (cit. on p. 99).

Ruaud, Maxime, Valentine Wakelam, and Franck Hersant (2016). "Gas and grain chemical composition in cold cores as predicted by the Nautilus three-phase model". *MNRAS* 459, pp. 3756–3767 (cit. on p. 22).

Rubin, Donald B. (1984). "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician". *The Annals of Statistics* 12.4, pp. 1151–1172 (cit. on p. 52).

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors". *Nature* 323.6088 (6088), pp. 533–536 (cit. on pp. 88, 90).

Saftly, W. et al. (2013). "Using hierarchical octrees in Monte Carlo radiative transfer simulations". *A&A* 554, A10 (cit. on p. 63).

Sarro, L. M., J. Ordieres-Meré, A. Bello-García, A. González-Marcos, and E. Solano (2018). "Estimates of the atmospheric parameters of M-type stars: a machine-learning perspective". *MNRAS* 476, pp. 1120–1139 (cit. on p. 72).

Schilke, P. et al. (2010). "Herschel observations of ortho- and para-oxidaniumyl (H2O+) in spiral arm clouds toward Sagittarius B2(M)". *A&A* 521, p. L11 (cit. on p. 72).

Schwarz, Gideon (1978). "Estimating the Dimension of a Model". *The Annals of Statistics* 6.2, pp. 461–464 (cit. on p. 51).

Serra, Paolo et al. (2011). "CIGALEMC: Galaxy parameter estimation using a Markov chain Monte Carlo approach with CIGALE". *ApJ* 740.1, p. 22 (cit. on p. 73).

Shahriari, Bobak, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas (2016). "Taking the Human Out of the Loop: A Review of Bayesian Optimization". *Proceedings of the IEEE* 104.1, pp. 148–175 (cit. on p. 137).

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. 1st ed. Cambridge University Press (cit. on pp. 35, 85–87, 90, 95, 101, 103, 105, 183).

Sheffer, Y. and M. G. Wolfire (2013). "PDR Model Mapping of Obscured H2 Emission and the Line-of-Sight Structure of M17-SW". *ApJ* 774.1, p. L14 (cit. on pp. 71, 72, 80, 82).

Sheffer, Y., M. G. Wolfire, D. J. Hollenbach, M. J. Kaufman, and M. Cordier (2011). "PDR Model Mapping of Physical Conditions via Spitzer/IRS Spectroscopy of H2: Theoretical Success toward NGC 2023-South". *ApJ* 741.1, p. 45 (cit. on pp. 63, 71, 72, 82).

Simsekli, Umut, Roland Badeau, A Taylan Cemgil, and Gaël Richard (2016). "Stochastic Quasi-Newton Langevin Monte Carlo", p. 10 (cit. on p. 46).

Skilling, John (2004). "Nested Sampling". *AIP Conference Proceedings*. BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Vol. 735. Garching (Germany): AIP, pp. 395–405 (cit. on pp. 50, 59).

— (2006). "Nested sampling for general Bayesian computation". *Bayesian Analysis* 1.4, pp. 833–859 (cit. on pp. 50, 59).

Smirnov-Pinchukov, Grigorii V. et al. (2022). "Machine learning-accelerated chemistry modeling of protoplanetary disks". *A&A* 666, p. L8 (cit. on pp. 63, 64, 87).

Smith, Nathan (2006). "The Structure of the Homunculus. I. Shape and Latitude Dependence from H2 and [Fe II] Velocity Maps of η Carinae". *ApJ* 644, pp. 1151–1163 (cit. on p. 148).

Song, Yang and Stefano Ermon (2020). *Generative Modeling by Estimating Gradients of the Data Distribution*. (Visited on 07/27/2023). preprint (cit. on p. 70).

Speagle, Joshua S (2020). "dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences". *MNRAS* 493.3, pp. 3132–3158 (cit. on pp. 73, 75).

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde (2002). "Bayesian measures of model complexity and fit". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4, pp. 583–639 (cit. on p. 51).

Spurio Mancini, Alessio, Davide Piras, Justin Alsing, Benjamin Joachimi, and Michael P Hobson (2022). "CosmoPower: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys". *MNRAS* 511.2, pp. 1771–1788 (cit. on pp. 63, 65, 103).

Sternberg, Amiel, Franck Le Petit, Evelyne Roueff, and Jacques Le Bourlot (2014). "H I-to-H2 Transitions and H I Column Densities in Galaxy Star-forming Regions". *ApJ* 790, p. 10 (cit. on p. 105).

Sutherland, Ralph, Mike Dopita, Luc Binette, and Brent Groves (2018). "MAPPINGS V: Astrophysical plasma modeling code". *Astrophysics Source Code Library*, ascl:1807.005 (cit. on p. 22).

Tang, Boxin (1993). "Orthogonal Array-Based Latin Hypercubes". *Journal of the American Statistical Association* 88.424, pp. 1392–1397 (cit. on p. 91).

Tegmark, Max et al. (2004). "Cosmological parameters from SDSS and WMAP". *Phys. Rev. D* 69.10, p. 103501 (cit. on p. 73).

Thomas, Adam D. et al. (2018). "Interrogating Seyferts with NebulaBayes: Spatially Probing the Narrow-line Region Radiation Fields and Chemical Abundances". *ApJ* 856, p. 89 (cit. on pp. 70, 71, 73, 87).

Thouvenin, Pierre-Antoine, Audrey Repetti, and Pierre Chainais (2023). *A Distributed Block-Split Gibbs Sampler with Hypergraph Structure for High-Dimensional Inverse Problems*. (Visited on 09/15/2023). preprint (cit. on p. 176).

Thrane, Eric and Colm Talbot (2019). "An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models". *Publications of the Astronomical Society of Australia* 36, e010 (cit. on p. 73).

Tieleman, Tijman and Geoffrey Hinton (2012). "Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude". 4.2, pp. 26–31 (cit. on pp. 38, 90, 118, 119, 173).

Tielens (2005). "The Physics and Chemistry of the Interstellar Medium", p. 510 (cit. on p. 11).

Tielens, A. G. G. M. and D. Hollenbach (1985). "Photodissociation regions. I. Basic model." *ApJ* 291, pp. 722–746 (cit. on p. 167).

Trotta, Roberto (2008). "Bayes in the sky: Bayesian inference and model selection in cosmology". *Contemporary Physics* 49.2, pp. 71–104 (cit. on pp. 75, 113).

Vale Asari, N., G. Stasińska, C. Morisset, and R. Cid Fernandes (2016). "BOND: Bayesian Oxygen and Nitrogen abundance Determinations in giant H II regions using strong and semistrong lines". *MNRAS* 460, pp. 1739–1757 (cit. on pp. 63, 66, 73).

Van der Tak, Floris, John Black, Fredrik Schoeier, David Jansen, and Ewine van Dishoeck (2007). "A computer program for fast non-LTE analysis of interstellar line spectra". *A&A* 468.2, pp. 627–635 (cit. on p. 21).

Van Laarhoven, Peter J. M. and Emile H. L. Aarts (1987). *Simulated Annealing: Theory and Applications*. Dordrecht: Springer Netherlands (cit. on p. 56).

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". *Stat Comput* 27.5, pp. 1413–1432 (cit. on pp. 51, 75).

Veneziani, M. et al. (2013). "A Bayesian Method for the Analysis of the Dust Emission in the Far-infrared and Submillimeter". *ApJ* 772.1, p. 56 (cit. on p. 71).

Vidal, Ana F., Valentin De Bortoli, Marcelo Pereyra, and Alain Durmus (2020). *Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach. Part I: Methodology and Experiments*. URL: http://arxiv.org/abs/1911.11709 (visited on 07/20/2022). preprint (cit. on p. 137).

Villa-Vélez, J. A., V. Buat, P. Theulé, M. Boquien, and D. Burgarella (2021). "Fitting spectral energy distributions of FMOS-COSMOS emission-line galaxies at z $\sim$ 1.6: Star formation rates, dust attenuation, and [OIII]λ5007 emission-line luminosities". *A&A* 654, A153 (cit. on pp. 73–75).

Virtanen, Pauli et al. (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". *Nat Methods* 17.3 (3), pp. 261–272 (cit. on p. 92).

Vono, Maxime, Nicolas Dobigeon, and Pierre Chainais (2019). "Split-and-augmented Gibbs sampler - Application to large-scale inference problems". Version 2. *IEEE Trans. Signal Process.* 67.6, pp. 1648–1661 (cit. on pp. 74, 115).

— (2021). *High-dimensional Gaussian sampling: A review and a unifying approach based on a stochastic proximal point algorithm*. arXiv (cit. on p. 176).

Wakelam, V. et al. (2012). "A KInetic Database for Astrochemistry (KIDA)". *ApJS* 199, p. 21 (cit. on p. 21).

Watanabe, Sumio (2010). *Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory*. URL: http://arxiv.org/abs/1004.2316 (visited on 08/02/2023). preprint (cit. on p. 51).

— (2012). *A Widely Applicable Bayesian Information Criterion*. (Visited on 08/02/2023). preprint (cit. on p. 51).

Welling, Max and Yee Whye Teh (2011). "Bayesian Learning via Stochastic Gradient Langevin Dynamics", p. 8 (cit. on p. 46).

Weyant, Anja, Chad Schafer, and W. Michael Wood-Vasey (2013). "Likelihood-free Cosmological Inference with Type Ia Supernovae: Approximate Bayesian Computation for a Complete Treatment of Uncertainty". *ApJ* 764, p. 116 (cit. on pp. 70, 73).

Whiteoak, Johnathan B. Z. (1994). "High-Resolution Images of the Dust and Ionized Gas Distribution in the Carina Nebula". *ApJ* 429, p. 225 (cit. on p. 148).

Woitke, P., I. Kamp, and W. -F. Thi (2009). "Radiation thermo-chemical models of protoplanetary disks. I. Hydrostatic disk structure and inner rim". *A&A* 501, pp. 383–406 (cit. on p. 22).

Wu, Ronin et al. (2018). "Constraining physical conditions for the PDR of Trumpler 14 in the Carina Nebula". *A&A* 618, A53 (cit. on pp. 25, 63–66, 71, 72, 82, 86, 87, 90, 95, 97, 133, 141, 148–152, 160, 166, 172, 174).

Xifara, T., C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami (2014). "Langevin diffusions and the Metropolis-adjusted Langevin algorithm". *Statistics & Probability Letters* 91, pp. 14–19 (cit. on pp. 45, 46, 118, 119, 173).

Yang, C. et al. (2017). "Molecular gas in the Herschel-selected strongly lensed submillimeter galaxies at z 2-4 as probed by multi-J CO lines". *A&A* 608, A144 (cit. on p. 73).

Ye, Kenny Q, William Li, and Agus Sudjianto (2000). "Algorithmic construction of optimal symmetric Latin hypercube designs". *Journal of Statistical Planning and Inference* 90.1, pp. 145–159 (cit. on p. 91).

Yu, Yaming and Xiao-Li Meng (2011). "To Center or Not to Center: That Is Not the Question—An Ancillarity—Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency". *Journal of Computational and Graphical Statistics* 20.3, pp. 531–570 (cit. on p. 74).

Zucker, Catherine et al. (2018). "Mapping Distances across the Perseus Molecular Cloud Using CO Observations, Stellar Photometry, and Gaia DR2 Parallax Measurements". *ApJ* 869, p. 83 (cit. on p. 73).

Zucker, Catherine et al. (2019). "A Large Catalog of Accurate Distances to Local Molecular Clouds: The Gaia DR2 Edition". *ApJ* 879, p. 125 (cit. on pp. 68, 73).

Zucker, Catherine et al. (2021). "On the Three-dimensional Structure of Local Molecular Clouds". *ApJ* 919, p. 35 (cit. on pp. 73, 75).

Résumé court

Méthodes d'échantillonnage pour l'inférence statistique de problèmes inverses non linéaires : distribution spatiale des propriétés physico-chimiques du milieu interstellaire

Le milieu interstellaire (MIS) est un milieu très diffus qui remplit l'immense volume entre les objets célestes tels que les étoiles et les trous noirs au sein d'une galaxie. L'étude du MIS soulève des questions fondamentales dont la formation d'étoiles. Les étoiles naissent de l'effondrement gravitationnel de parties de régions froides et denses du MIS appelées nuages moléculaires.

Cette thèse analyse des cartes multispectrales de nuages moléculaires dans les domaines infrarouge lointain et radio, obtenues par des télescopes spatiaux ou terrestres. L'attention est portée aux nuages illuminés et chauffés par des étoiles massives voisines émettant des photons UV. La couche de surface de tels nuages, où le champ radiatif UV chauffe et dissocie le gaz moléculaire, est appelée région de photodissociation (PDR). Leur cartes multispectrales contiennent typiquement de 1 à 10 000 pixels, où chaque pixel contient l'intensité intégrée de 5 à 30 raies d'émission. Ces intensités peuvent être comparées avec les prédictions d'un modèle numérique du MIS tel que le code PDR de Meudon, qui calcule ces intensités à partir de paramètres physiques. Cette thèse vise à estimer des cartes de paramètres physiques (tels que la pression thermique ou l'intensité du champ UV incident) à partir d'une carte d'observation et du code PDR de Meudon. Ce problème est une instance d'une classe générale de problèmes inverses.

Une nouvelle méthode d'inférence tenant compte d'autant de sources d'incertitudes que possible est introduite. Une procédure générale est proposée pour construire une approximation de modèles numériques. Elle exploite un réseau de neurones spécifique et surpasse les méthodes d'interpolation en terme de précision, de poids mémoire et de durée d'évaluation. Une régularisation spatiale améliore les estimations. Une approche par échantillonnage est considérée pour fournir des quantification d'incertitudes en plus d'estimateurs ponctuels de cartes de paramètres physiques afin de compenser l'absence vérité terrain, inhérente à l'astrophysique. L'algorithme Monte Carlo Markov chain (MCMC) proposé combine deux échantillonneurs: l'un identifie les minima locaux dans l'espace des paramètres tandis que l'autre les explore efficacement. Finalement, la pertinence du modèle d'observation considéré pour l'inférence est vérifiée. La méthode proposée est appliquée à des données synthétiques pour validation, puis à des observations réelles. Les résultats sont analysés pour fournir des interprétations astrophysiques.

Mots-clés – Problèmes inverses, algorithmes MCMC, milieu interstellaire, nuages moléculaires, régions de photodissociation, apprentissage statistique.

Abstract

Sampling methods for statistical inference of non-linear inverse problems: spatial distribution of physico-chemical properties of the interstellar medium

The interstellar medium (ISM) is a very diffuse medium that fills the extraordinarily large volume between celestial objects such as stars and black holes in a galaxy. The study of the ISM raises fundamental questions including star formation. Stars are born from the gravitational collapse of a part of cold and dense regions of the ISM called molecular clouds.

This thesis analyzes multispectral maps of molecular clouds in the infrared and radio domains, observed by space or ground telescopes. The focus is put on clouds that are illuminated and heated by nearby massive stars emitting UV photons. The surface layer of such clouds, where the UV irradiation heats and dissociates the molecular gas, is called a photodissociation region (PDR). Their multispectral maps typically contain from 1 to 10 000 pixels, where each pixel contains the integrated intensities of 5 to 30 emission lines. These intensities can be compared with the predictions of an ISM numerical model such as the Meudon PDR code that computes intensities from physical parameters. This thesis aims at estimating maps of physical parameters (such as the thermal pressure or the intensity of the incident UV field) from an observation map and the Meudon PDR code. This problem is an instance of a general class of inverse problems.

A new inference method that accounts for as many uncertainty sources as possible is introduced. A general procedure to derive a surrogate approximation of numerical models is proposed. It is based on a specific neural network and outperforms interpolation methods in accuracy, memory weight and evaluation time. A spatial regularization improves estimations. A sampling approach is considered to provide uncertainty quantification along with the estimated physical parameter maps to address the absence of ground truth, inherent to astrophysics. The proposed Monte Carlo Markov Chain (MCMC) algorithm combines two samplers: one identifies local minima in the parameters space while the second efficiently explores them. Finally, the relevance of the observation model considered for inference is assessed. The proposed method is applied to synthetic data for validation and then to real observations. The results are analyzed for astrophysical interpretation.

Keywords – Inverse problems, MCMC algorithms, interstellar medium, molecular clouds, photodissociation regions, machine learning.