PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Statistical study of the non-Gaussian structures of the turbulent ISM in a low observational data regime

Soutenue par

Pablo Richard

Le 14 Octobre 2024

École doctorale n°127

Astronomie et Astrophysique d'Île de France

Spécialité

Astrophysique

Préparée au

Laboratoire de Physique de l'École Normale Supérieure

Composition du jury :

Philippe SALOMÉ Observatoire de Paris	*Président*
Jonathan AUMONT Université Toulouse III Paul Sabatier	*Rapporteur*
Joshua PEEK Space Telescope Science Institute	*Rapporteur*
Susan CLARK Stanford University	*Examinatrice*
Michael EICKENBERG Flatiron Institute	*Examinateur*
Jérôme PETY Institut de Radioastronomie Millimétrique	*Examinateur*
François LEVRIER École Normale Supérieure	*Directeur de thèse*
Erwan ALLYS École Normale Supérieure	*Invité, co-directeur*

ENS | PSL

Résumé

Ce travail cherche à caractériser les structures multi-échelles et non Gaussiennes omniprésentes au sein du milieu interstellaire turbulent (MIS), en mettant néanmoins l'emphase sur le caractère limité du nombre d'observations dont on dispose. Les approches traditionnelles de ce problème, fondées sur des statistiques récapitulatives, n'exploitent pas pleinement la richesse informationelle encodée dans les structures du MIS, tandis que les techniques récentes d'apprentissage automatique, bien qu'extrêmement puissantes, nécessitent d'être entraînées sur des jeux de données adéquats. Si les simulations numériques constituent un outil attrayant pour fournir de tels jeux, la nature multi-physique et complexe de la dynamique du MIS rend sa reproduction *in silico* extrêmement difficile. Ainsi, la modélisation des propriétés non Gaussiennes du MIS ne peut reposer sur des modèles intégralement fondés sur les simulations et doit intégrer une étape d'apprentissage centrée sur les observations. Cependant, cela pose la difficulté de travailler sans supervision et avec une quantité très limitée de données d'entraînement, ce qui favorise fortement l'utilisation de descriptions compressées et de basse variance. Néanmoins, cette étude montre qu'en exploitant les symétries et régularités des processus physiques au travers de statistiques non expansives comme la *scattering transform*, il est possible d'explorer les propriétés non Gaussiennes du MIS à partir d'observations. En particulier, nous complétons le diagnostic standard, fondé sur les statistiques à un point, de l'évolution des nuages moléculaires du stade quiescent au stade actif de formation d'étoiles, par un diagnostic morphologique caractérisant le couplage entre échelles. En outre, nous développons une méthodologie qui permet de comparer, sans supervision, le pouvoir informatif de plusieurs statistiques récapitulatives dans un sens que nous définissons. Cela nous permet d'établir, à partir des observations, une distance morphologique entre les cartes de densité de colonne des nuages moléculaires, reposant sur une description non Gaussienne compacte. Les résultats de ce travail ouvrent de nouvelles perspectives sur le rôle que peuvent jouer les observations dans la caractérisation des structures non Gaussiennes du MIS.

Mots clés : Milieu interstellaire, turbulence, nuages moléculaires, structures non Gaussiennes, géométrie de l'information

Abstract

This work addresses the challenge of characterizing the multi-scale, non-Gaussian structures in the turbulent Interstellar Medium (ISM), especially with limited observational data. Traditional approaches of this problem, based on summary statistics, do not fully exploit the rich information encoded in ISM structures, while recent machine learning techniques, though extremely powerful, require to be trained on appropriate datasets. If numerical simulations constitute an appealing tool to provide such required datasets, the complex multi-physics nature of the ISM makes it extremely challenging to be reproduced *in silico*. Thus, modeling the non-Gaussian properties of the ISM cannot rely solely on simulation-based approaches and must incorporate a learning step grounded in observations. However, this brings the difficulty to work without supervision and with a very limited amount of training data, which strongly favors the use of low variance and compressed descriptions. Still, this study demonstrates that, by leveraging physical symmetries and nonexpansive statistics like the scattering transform, it is possible to explore, from observations, non-Gaussian properties of the ISM. Specifically, we complement the standard one-point based diagnostic of molecular clouds' evolution from quiescent to active star-forming stages with a morphological diagnostic based on scale coupling. Additionally, we develop a methodology that allows to compare, without supervision, the informative power of multiple summary statistics in a sense that we define. This allows us to tailor, from observations, a morphological distance between column density maps of molecular clouds, based on a compressed non-Gaussian description. The results of this work open new perspectives for the role of observations in the characterization of non-Gaussian structures of the ISM.

Keywords : Interstellar medium, turbulence, molecular clouds, non-Gaussian structures, information geometry

Introduction

The Interstellar Medium (ISM) refers to both the baryonic material and radiation that is found in a galaxy, without including the stars. It is at the crossroads of great questions in astrophysics and cosmology. Indeed, it provides the initial conditions for stellar and planetary formation, which occurs within its densest structures, the molecular clouds. It also contributes to the evolution of galaxies over time, both in terms of energy and physicochemical composition. Finally, it presents an unavoidable foreground for the study of cosmological signals.

Its energy density is however close to equipartition between radiative, thermal, kinetic, magnetic and cosmic ray forms. It is thus a complex medium that brings into play multiple fields of knowledge in physics and chemistry. In addition, these processes are nonlinearly coupled and lie in a turbulent cascade, that spans over more than height orders of magnitudes, from thousands of parsecs down to a few astronomical units. As a result of this strong coupling, the interstellar content is organized in remarkably complex and multi-scale structures, that hold information about their physical dynamics. Since the 50s, the understanding of this dynamics has been continuously bewildered by the improvement of observational capabilities: the finer the instrumental resolution, the more detailed structures are unveiled, often unexpectedly.

To explain the formation of these observed structures, numerical simulations constitute a precious tool that reproduces *in silico* a given dynamics. Simulations allow to generate, in a forward manner, the structures resulting from a dynamics, whose ingredients are motivated by our current understanding of the ISM. Consequently, distinguishing between the wide variety of non-Gaussian structures that are observed in the ISM is a crucial step to learn about these physical ingredients and their interplay. This however requires an appropriate statistical description of the structures appearing in the ISM.

To tackle the great challenge of characterizing such non-Gaussian structures, the ISM community has first relied on statistical descriptions that summarize the data into a few quantities. Great care has been dedicated to interpret these descriptors and to make them robust (or at least predictable) under contaminations, both from an analytical perspective, and informed by simulations. However, these data reductions do not fully exploit the complex non-Gaussian information present in the data, and thus characterize only partially the underlying physical processes.

In parallel, the machine learning community has developed tools that are extremely efficient at extracting information in high-dimension, by learning from large datasets. These tools, combined with the controlled world of simulations, have initiated a new paradigm of simulation-based inference. The progress made in this topic has reached a pivotal point where, today, it becomes possible to grasp almost all the information contained in the data generated by a given

simulation, provided these tools are trained in appropriate conditions. Does this mean that, by improving the computational capabilities, we are on the verge of fully characterizing the ISM?

Unfortunately, simulations are only one part of the equation. Observations of the ISM reflect a potentially wide panel of physical conditions, with gravity superseding turbulence here and not there, with magnetic fields being strong enough to influence the dynamics or not, etc. The precise "mixture" of physical processes may vary from one region to the other, and there may be few observed examples associated to a given dynamics. This low observational data regime, combined with the difficulty to reproduce closely the multiple components of the ISM dynamics all-at-once in a simulation, make the task of characterizing ISM non-Gaussianities inherently difficult.

In order to do so, various avenues exist, with a varying simulation-dependence degree, but not one seems to clearly stand out as they all face the difficulty to be confronted to observations. In particular, it is still unclear how to compare, and on which criteria, these characterizations of ISM structures. In this manuscript, I formulate this question in a statistical formalism, aim at providing elements of what could be such a comparison, and give applications mostly in the context of molecular clouds. I dedicate great care to formalize the questions in a rigorous manner, to keep an approach as general as possible as the problems encountered here may find a strong echo in numerous other fields, and to rely as much as possible on existing results in applied mathematics and statistics.

This manuscript is organized as follows:

- the first chapter motivates the study of the ISM, depicts the characteristics of this medium and the inherent difficulties of its study before sketching the main challenges that will be addressed in this thesis.
- The second chapter motivates the statistical approach of the nonlinear dynamics of the ISM and stresses the main difficulties that arise due to the high dimensional non-Gaussianity of its underlying stochastic fields. It then outlines the statistical formulations of various scientific questions relative to the ISM, and the different frameworks usually used to tackle them.
- The third chapter presents statistical tools that can be used to build, from a very limited amount of data, low variance representations of physical processes endowed with enough invariance properties. It then reviews the main statistical diagnostics of turbulence and intermittency, with an emphasis on the coupling between scales, before applying these tools to characterize, from observations, the coupling between scales in column density maps of molecular clouds as they evolve from quiescent to active star forming.
- The fourth chapter addresses the problem of the choice of a statistical description. After presenting key concepts in information theory and standard tools for supervised frameworks, it then motivates the interest for extending such ideas to the much less supervised world of ISM observations, and sets a theoretical ground to perform such a shift.

- The fifth chapter aims at applying in practice, in a low data regime, the theoretical definitions resulting from the previous chapter, in order to compare the ability of various summary statistics to characterize the diversity of structures found in observations of molecular clouds, and from these results, builds a morphological distance to compare observations, simulations and statistical models of these clouds.

Publications related to this thesis

- During this thesis, I have been leading the following paper, submitted to A&A the 13^{th} of July 2024 and that received a positive recommendation for publication after revisions:

  P. Richard, E. Allys, F. Levrier, A. Gusdorf, C. Auclair. "*Molecular clouds: do they deserve a non-Gaussian description?*". (In revision) Astronomy & Astrophysics.
  DOI: 10.48550/arXiv.2407.09832

  This paper constitutes the fifth chapter of this manuscript.

- I also took part in the following paper:

  C. Auclair, E. Allys, F. Boulanger, M. Béthermin, A. Gkogkou, G. Lagache, A. Marchal, M.-A. Miville-Deschênes, B. Régaldo-Saint Blancard, **P. Richard**. "*Separation of dust emission from the cosmic infrared background in Herschel observations with wavelet phase harmonics*". Astronomy & Astrophysics 681 (2024).
  DOI: 10.1051/0004-6361/202346814

- I am involved in the following paper, in preparation:

  P. Lesaffre, J.-B. Durrive, J. Goossaert, S. Poirier, S. Colombi, **P. Richard**, E. Allys, W. Béthune. "*Multiscale Turbulence Synthesis in 2D hydrodynamics*". In prep..

- Finally, I am actively involved in a ongoing work with E. Allys, R. Soletskyi and A. Tsouros that is not yet on the form of a paper but that could be entitled "A Bayesian approach to inverse problems in a low data regime with scattering transform generative models".

Index of main abbreviations and notations

CIB	Cosmic Infrared Background
CMB	Cosmic Microwave Background
CNM	Cold Neutral Medium
CNN	Convolutional Neural Network
DPI	Data Processing Inequality
FBM	Fractional Brownian Motion
HGBS	*Herschel* Gould Belt Survey
IMF	Initial Mass Function
ISM	InterStellar Medium
JS	Jensen-Shannon (divergence)
KL	Kullback-Leibler (divergence)
LOTUS	Law Of The Unconscious Statistician
MC	Molecular Cloud
MHD	MagnetoHydroDynamics
PDF	Probability Distribution Function
RWST	Reduced Wavelet Scattering Transform
SBI	Simulation-Based Inferene
TV	Total Variation (distance)

We will refer to random variables/vectors/fields with capital letters (e.g., X), and to their realizations with lowercase letters (e.g., x). We will also use bold letters to refer to vectors of finite-dimensional vector spaces (e.g.: $\mathbf{r} \in \mathbb{R}^2$). Statistical estimators or estimates of given quantities will be denoted by a hat symbol (e.g.: $\hat{x}$ is an estimate of x). The Fourier transform of a field $x(\mathbf{r})$ will be denoted by a tilde symbol: $\tilde{x}(\mathbf{k})$. The convolution of x and y will be denoted $x \star y$.

Contents

Chapter 1

The Interstellar Medium: context and challenges

"Another One Bites the Dust"

Hit song from the rock band *Queen* (1989) whose guitarist, Brian May, had started a PhD on the dust in our Solar neighborhood. Even though the title of the song is likely to be a coincidence with his work, it has a certain metaphoric echo in astrophysics and especially in observational cosmology.

Objectives

The Interstellar Medium (ISM) is at the crossroads of great scientific questions. It is a multi-phase and multi-scale system that undergoes multiple physical processes in nonlinear interaction. In this chapter, we motivate the study of this complex medium both for astrophysical and cosmological purposes, and sketch the main challenges that will be addressed in this thesis.

Contents

1 ISM great challenges

The Interstellar Medium (ISM) refers to both the baryonic material and radiation that is found in a galaxy, without including the stars. It mostly consists of gas, dust, and cosmic rays lying in a magnetic field, emitting and receiving light. Before going through any further detailed description of the ISM, we start by presenting the great scientific challenges in relation with it. They are mainly twofold: astrophysical and cosmological.

1.1 Astrophysical questions

The ISM is the cradle of stars, that are born in dense collapsing regions of cold *molecular clouds* (MC) (Fig. 1.1 and Fig. 1.10) (Myers et al., 1986; Scoville & Good, 1989). If these clouds certainly set the initial conditions for star formation, predicting quantitative properties about the resulting distribution of stars is still a major open challenge (McKee & Ostriker, 2007). In particular, it is yet unclear if the Initial Mass Function (IMF) (i.e., the mass distribution of newly formed stars) is rather universal or depends on some environmental properties. Furthermore, observations support that only ~ 1 solar mass of the ISM is turned into stars each year in the whole Milky Way (Chomiuk & Povich, 2011; Licquia & Newman, 2015), which is two orders of magnitude below what would be expected from having the remaining gas collapse under its self-gravity without any resisting process. This amount of star formation is also generally overestimated by numerical simulations Hennebelle and Falgarone, 2012.

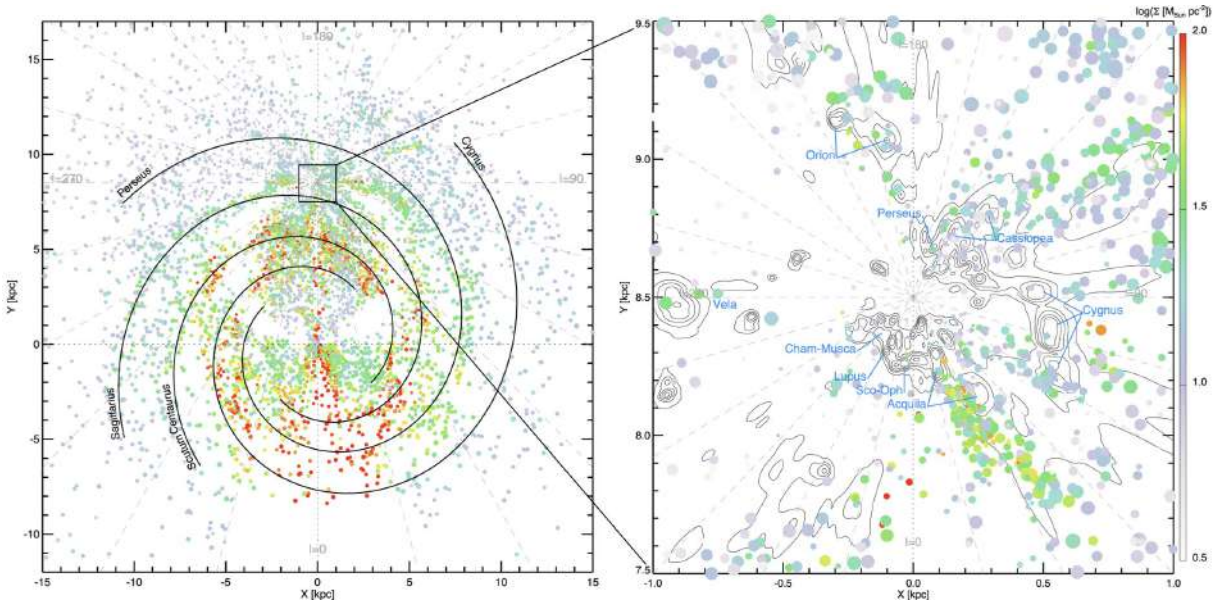


Figure 1.1: Molecular clouds (MCs) in the Milky Way (left) and in the solar neighborhood (right). Each MC is represented by a dot with a radius proportional to its physical radius (typically $\sim 30\,\mathrm{pc}$) and colored according to its H_2 surface density. Figure adapted from M.-A. Miville-Deschênes et al., 2017. The face-on view of the Galaxy sketches the four-spiral-arm model of Vallée, 2008. We display detailed structures of observed and simulated molecular clouds in Fig. 1.10.

One reason for these overestimates is that the relationship between the ISM and stars is far from being unidirectional. In fact, once formed, stars enrich the ISM with heavier elements as molecules and dust, and also energize it in various forms: radiative, kinetic, chemical and cosmic rays. This *stellar feedback* plays a significant contribution in the evolution of the ISM, and in turn in the overall baryonic content of a galaxy (ISM+stars) (Draine, 2010). Star formation is thus not only challenging to understand, but also one of the main driving forces behind the formation and evolution of galaxies (Agertz et al., 2013). Although it represents only $\sim 10\%$ of the mass of the global galactic baryonic content (Draine, 2010), the ISM, through its prominent role in star formation, is therefore a key ingredient in galaxy evolution (Galliano et al., 2018), and its study also impacts our understanding of the evolution of the Universe since the *cosmic dawn* (the period when the first stars appeared, cf. Fig. 1.2) (Barkana & Loeb, 2001).

Figure 1.2: Schematic history of the Universe, from inflation until present. The cosmic structures are an artistic view. The ISM plays a key role in the Universe since the cosmic dawn, marking the end of the dark ages, but also constitutes a foreground to older cosmological signals such as the Cosmic Microwave Background. *Credits*: ESA.

1.2 A foreground for cosmology

As our direct and inescapable surrounding, the ISM of the Milky Way, our own Galaxy, is subject to a specific attention of the scientific community. Indeed, prior to being eventually collected by our instruments, extragalactic signals necessarily travel through part of the Milky Way, and are therefore contaminated by it, through emission and absorption. While the emission from stars can be very bright in certain frequency bands, they remain compact sources whose emission can be removed without altering the large majority of the observed area. On the other hand, the Galactic ISM represents a diffuse component, especially important toward the Galactic disk, at low latitudes. These Galactic foregrounds[1] constitute a major hurdle for cosmology as they contaminate and distort extragalactic signals. We mention below the significant role played by

[1]The ISM of distant galaxies can also be problematic. For instance, besides being a cosmological signal that can be used to characterize the Universe, the Cosmic Infrared Background (CIB) that originates from the integrated thermal dust emission of galaxies (Hauser & Dwek, 2001) is also a foreground to the CMB in total intensity (Planck Collaboration et al., 2014b).

the Galactic ISM in what is one of the most important challenges currently in observational cosmology: measuring the B-mode polarization of the Cosmic Microwave Background (CMB).

The CMB is a signal of utmost importance to constrain our knowledge of the early-Universe. It refers to the electromagnetic radiation that was released right after the recombination of electrons with nuclei approximately 380 000 years after the Big Bang (cf. Fig. 1.2). Indeed, while in its first stages after the Big-Bang, the Universe was too hot to allow atoms to be stable, it eventually cooled along its expansion and reached a sufficiently low temperature to allow for electron-proton recombination ($\sim 3000\,\mathrm{K}$). The cosmic radiation released at this point was searched for by Dicke et al., 1965, but was first unexpectedly discovered by Penzias and Wilson, 1965, as a $\sim 3\,\mathrm{K}$ isotropic radiation at 4080 MHz. Since then, impressive progresses have been made in the characterization of this cosmic signal.

One of the main milestones reached is certainly due to the *Planck* mission that produced the most accurate measurements of the CMB anisotropies both in temperature and linear polarization. These measurements allowed to tighten significantly the constraints on cosmological parameters as illustrated in Fig. 2.9 (Planck Collaboration et al., 2020a), but also had a major impact on many fields of astrophysics[2]. These successful results have been obtained after a component separation step that was one of the greatest challenges of the mission (Delabrouille & Cardoso, 2008; Leach et al., 2008; Planck Collaboration et al., 2020b). Indeed, as shown in Fig. 1.3, the CMB is contaminated at all electromagnetic frequencies, both in intensity (left) and linear polarization (right), mainly by the following (nonexhaustive list of) components of the Galactic ISM:

- the thermal emission of dust grains,
- the synchrotron emission of charged particles (mostly relativistic electrons) gyrating around Galactic magnetic field lines,
- the free-free (or bremsstrahlung) emission of thermal accelerating charged particles, typically when a free electron encounters a proton,
- the emission of the CO molecule (amongst others) through its rotational spectral lines,
- and an anomalous microwave emission expected to originate from fast spinning dust grains with a non-zero electric dipole moment.

It is now considered that temperature fluctuations of the CMB do not really harbor further information about the early Universe than what has been already obtained, mainly because of cosmic variance (Planck Collaboration et al., 2020a). However, beyond this impressive legacy, a new generation of current and future microwave experiments is being designed to improve the constraints on the B-mode polarization of the CMB. These include the Simons Observatory (P. Ade et al., 2019), CMB-S4 (Abazajian et al., 2022) and the LiteBIRD satellite (Hazumi et al.,

[2]The results are so numerous that we directly refer to the following archive: https://www.cosmos.esa.int/web/planck/publications.

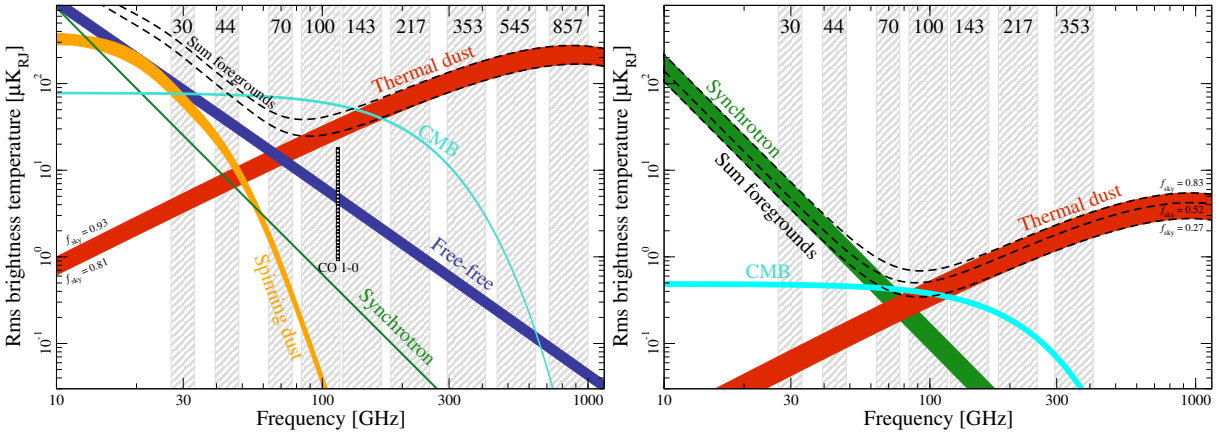


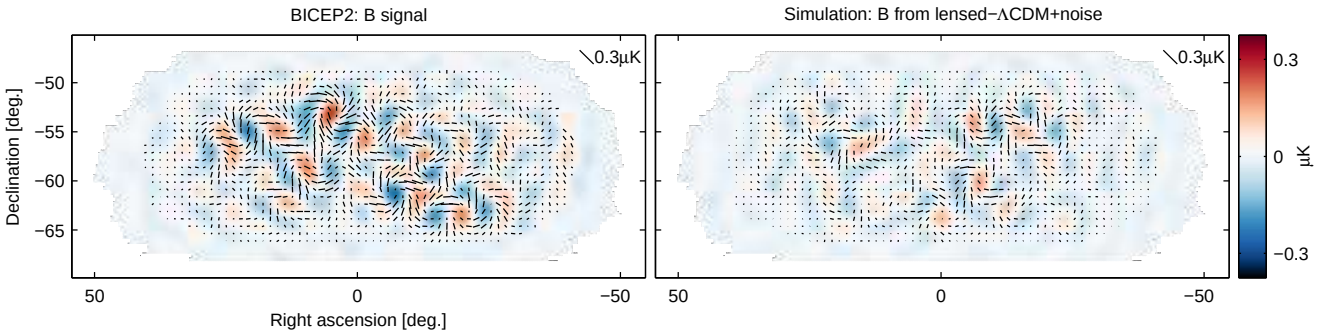
Figure 1.3: Spectral Energy Distributions (SEDs) of the CMB and Galactic ISM foregrounds in total intensity (left) and linear polarization (right). *Credits*: (Planck Collaboration et al., 2020a)

2020). This signal, yet undetected, is a foremost goal of modern cosmology (Kamionkowski & Kovetz, 2016). Indeed, according to the inflation paradigm, the quantum fluctuations of the spacetime fabric that occurred during the inflation era, less than 10^{-32} seconds after the Big-Bang (cf. Fig. 1.2), are expected to have left an imprint in the B-mode polarization of the CMB (Kamionkowski et al., 1997; Seljak & Zaldarriaga, 1997). Detecting these primordial gravitational waves would shed unprecedented light on both the very early Universe but also on the fundamental processes of physics at energy levels way above the capabilities of CERN's Large Hadron Collider (Lyth, 1997).

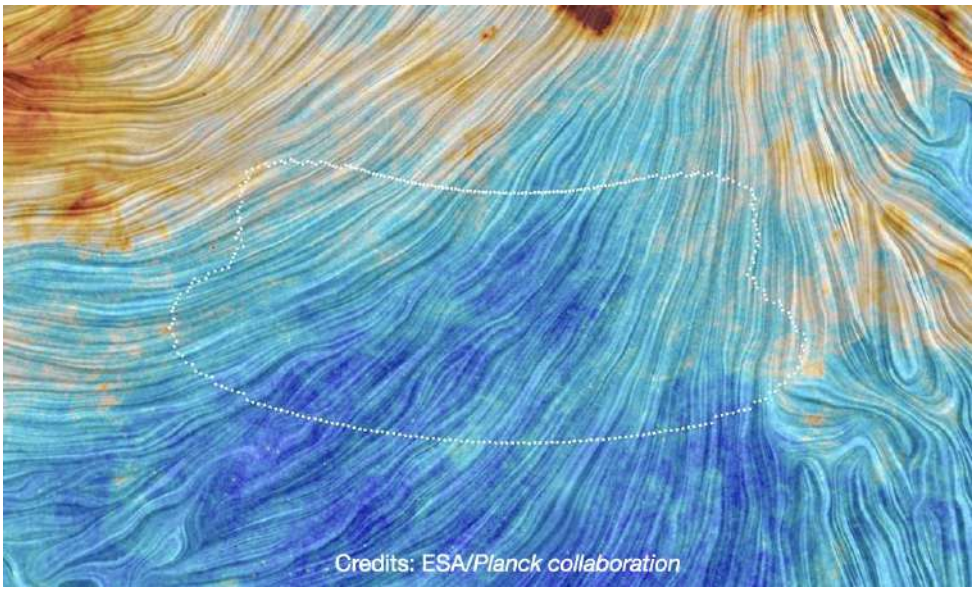
However, the polarized foregrounds due to the Galactic ISM are expected to dominate over primordial B-modes by more than two orders of magnitude (LiteBIRD Collaboration et al., 2023). This brings the difficulty of the component separation task to a next level. As a matter of fact, in 2014, the BICEP2 collaboration has observed B-modes with a power significantly exceeding the one expected for purely lensed ΛCDM B-modes (P. A. Ade et al., 2014) (Fig. 1.4.a). The collaboration attributed this power excess to the presence of primordial B-modes. However, further investigations with the *Planck* team revealed that this excess is much more likely to originate from Galactic dust emission, a key component of the ISM (P. A. R. Ade et al., 2015) (Fig. 1.4.b).

This episode emphasizes the *critical role* that the modeling of ISM foregrounds plays in the quest for cosmological signals, and supports that the intrinsic difficulty of this task should not be underestimated. Indeed, these foregrounds have highly non-Gaussian statistical properties as a result from the nonlinear dynamics of the ISM. These non-Gaussianities are ascertained both in space (Planck Collaboration et al., 2015) and frequency (Hensley & Bull, 2018; Planck Collaboration et al., 2017; Vacher et al., 2023) However, in spite of constituting a clear difficulty in the handling and modeling of these foregrounds, these non-Gaussianities also provide an outstanding lever arm when it comes to separate the CMB signal to its Galactic foregrounds[3].

[3]Indeed, a non-Gaussian process can have a much lower entropy than its Gaussian counterpart. In other words, at fixed variance level, a non-Gaussian process might result (but not necessarily) with regions of likely samples



(a) Figure taken from (P. A. Ade et al., 2014). Left: overall B-modes detected by the BICEP2 collaboration (filtered on a certain angular scale range) and right: B-modes explained by a noisy and lensed ΛCDM signal. The excess of power in the left map was first attributed to an inflationary gravitational wave signal...



(b) ...however, further investigations involving correlation with *Planck* data at higher frequency where thermal emission of dust dominates has shown that this excess is more likely to originate from Galactic dust emission rather than from a cosmological source P. A. R. Ade et al., 2015. This figure reports the BICEP2 field (delimited by the white line) as seen by the *Planck* satellite in the frequency range where dust dominates (cf. Fig. 1.3). The color maps the emission intensity (low in blue) while the drapery pattern traces the orientation of the Galactic magnetic field based on measured dust polarization.

Figure 1.4: The non detection of primordial B-modes due to the underestimation of a Galactic foreground emitted by the dust: a key component of the ISM.

Although modeling Galactic foregrounds is outside the scope of this work, it illustrates the crucial role that understanding the ISM can play, even for seemingly unrelated cosmological questions.

with much smaller measure than its Gaussian counterpart, which means less "uncertainty" about the signal.

2 A multi-phase and multi-scale medium

As discussed above, a detailed understanding of the ISM dynamics constitutes a major goal in astrophysics, which can also impact observational cosmology. These objectives are however highly challenging due to the very complex multi-phase, multi-process, and multi-scale nature of the ISM.

2.1 A multi-phase medium...

The ISM refers to the baryonic content of a galaxy (except stars) and its ambient radiation field. The matter content of the ISM is composed of gas, dust grains (the latter forming only 1% of gas mass), and cosmic rays. In the Milky Way, hydrogen represents 73% of the gas mass and helium 27%, while other elements have a negligible mass contribution. Hydrogen mass is distributed at 60% under its neutral form H I, 23% under its ionized form H II and 17% under its molecular form H_2. The interstellar gas is not in a global thermodynamic equilibrium. Its number density varies from $\sim 4\,10^{-3}\,\mathrm{cm}^{-3}$ to $\sim 10^6\,\mathrm{cm}^{-3}$, that is a range spanning of *more than eight orders of magnitude*, and its temperature ranges from $\sim 10\,\mathrm{K}$ to $10^6 - 10^7\,\mathrm{K}$. In fact, the gas is present in different thermodynamic phases (reported in Tab. 1.1), that are generally classified as follows (Draine, 2010):

- coronal gas, or Hot Ionized Medium (HIM), corresponding to collisionally ionized gas heated to $10^6 - 10^7\,\mathrm{K}$ by supernova shocks. It fills 50% of the ISM volume with regions of typical $10^1 - 10^2\,\mathrm{pc}$ size but is also its most diffuse component ($n_\mathrm{H} \sim 4\,10^{-3}\,\mathrm{cm}^{-3}$), so it accounts only for $\sim 2\%$ of its mass.

- H II gas, corresponding also to hot ($10^4\,\mathrm{K}$) hydrogen, photoionized by ultraviolet radiation emitted by hot stars. It is found in a wide range of densities: $n_\mathrm{H} \sim 0.3 - 10^4\,\mathrm{cm}^{-3}$. Its densest form is found in H II *regions* that originate from dense gas regions photoionized by recently formed stars. Its much more diffuse form, but overall dominating in mass ($\sim 18\%$ of ISM), is called the Warm Ionized Medium (WIM) and fills $\sim 10\%$ of the ISM volume.

- Warm Neutral Medium (WNM), corresponding mostly to hot ($\sim 5000\,\mathrm{K}$) and diffuse $n_\mathrm{H} \sim 0.6\,\mathrm{cm}^{-3}$ H I gas, filling a large fraction of the ISM volume ($\sim 40\%$) and accounting for $\sim 25\%$ of its mass.

- Cold Neutral Medium (CNM), corresponding mostly to cold ($\sim 100\,\mathrm{K}$) and denser H I gas ($n_\mathrm{H} \sim 30\,\mathrm{cm}^{-3}$), filling only $\sim 1\%$ of the ISM volume but still accounting for a significant mass fraction ($\sim 25\%$).

- Diffuse molecular gas, in similar conditions to the CNM ($\sim 50\,\mathrm{K}$, $n_\mathrm{H} \sim 100\,\mathrm{cm}^{-3}$) except that it is sufficiently dense to allow molecular hydrogen H_2 to form. It fills only $\sim 0.1\%$ of the ISM volume.

- Dense molecular gas, corresponding to structures sufficiently dense ($\gtrsim 10^3\,\mathrm{cm}^{-3}$) to become self-gravitating and perhaps eventually form stars. It is slightly colder than the diffuse

gas ($10 - 100\,\mathrm{K}$), fills only $\sim 0.01\%$ of the ISM volume but constitutes with the diffuse molecular gas $\sim 30\%$ of the total ISM mass.

Phase	T (K)	n_H (cm^{-3})	x_e	f_m	f_V	L
HIM	$\gtrsim 10^{5.5}$	$\sim 4\,10^{-3}$	~ 1	2%	$\sim 50\%$	$\sim 100\,\mathrm{pc}$
H II gas	10^4	$0.3 - 10^4$	$\geq 10^{-1}$	18%	$\sim 10\%$	few pc in H II regions
WNM	5000	0.6	$\sim 10^{-2}$	25%	$\sim 40\%$	$\sim 50\,\mathrm{pc}$
CNM	100	30	$\sim 10^{-4}$	25%	$\sim 1\%$	$\sim 10\,\mathrm{pc}$
Diffuse H_2	~ 50	~ 100	$\sim 10^{-4}$	with dense H_2	$\sim 0.1\%$	$\sim 3\,\mathrm{pc}$
Dense H_2	$10 - 100$	$10^3 - 10^6$	$\sim 10^{-7}$	$\sim 30\%$	$\sim 0.01\%$	$\sim 0.1\,\mathrm{pc}$

Table 1.1: Phases of the ISM. x_e denotes the ionization fraction n_e/n_H, f_m and f_V respectively correspond to the mass fraction and volume filling factor. Data combined from (Draine, 2010; Lesaffre, Falgarone, & Hily-Blant, 2024).

The CNM and WNM are two stable phases resulting from the thermal instability of the cold gas (Field, 1965), immersed in the hotter and more diffuse phases filling the galactic volume. The HIM, WNM, CNM and diffuse molecular gas correspond to increasingly cooled gas phases that are in pressure equilibrium $P/k_\mathrm{B} = n_\mathrm{H} T \sim 3000\,\mathrm{K\,cm}^{-3}$. When reported on a temperature-density plot (as shown in Fig. 1.5), these phases form a cooling branch that is the first stage of a matter cycle that also includes a second heating branch to reach hot and dense prestellar conditions. Let us emphasize again that each of these phases is not in global thermodynamic equilibrium: it continuously exchanges matter and energy with the other phases, with stars, and with the intergalactic medium (Draine, 2010).

2.2 ...that undergoes multiple physical processes in interaction...

These multiple gas phases are mixed with dust, magnetic fields, stars and cosmic rays. Below is a list of typical interactions and energy transfers within the ISM:

- Galactic injection: large-scale shear due to differential rotation in spiral galaxies, and also local compression through density waves (McKee & Ostriker, 2007).
- Interactions between clouds allow energy and momentum transfer.
- Conversion of gravitational potential into kinetic and magnetic energy in a self-gravitating cloud.
- Radiative cooling of the gas and dust (cf. cooling branch of Fig. 1.5 and spectra of Fig. 1.6).
- Stellar feedback in several forms:

  ⋆ the ultraviolet radiation of massive stars heats their neighboring gas and dust, dissociates H_2 into hot H I gas creating *PhotoDissociation Regions* (PDRs) (Hollenbach &

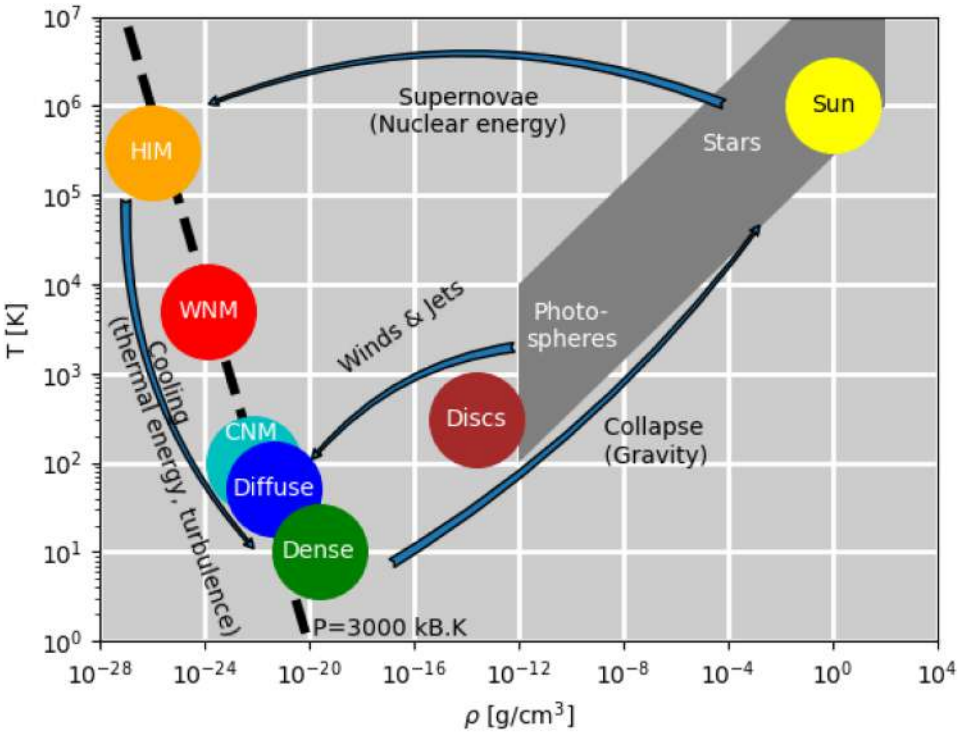


Figure 1.5: ISM baryonic matter cycle in the temperature-density space. The gas phases introduced in Tab. 1.1, from HIM to diffuse molecular gas, lie close to the isobaric locus $P/k_{\rm B} = n_{\rm H}T \sim 3000\,{\rm K\,cm^{-3}}$. Deviations from this branch start to occur with the dense molecular gas that is gravitationally bound. *Credits*: Lesaffre, 2018.

Tielens, 1999) that are further ionized leading to H II regions, but also compresses and "pushes back" (Goicoechea et al., 2016) the cloud (cf. Fig. **??**).

⋆ a supernova releases $\sim 10^{44}$ J in its direct neighborhood, generates shocks that deposit kinetic energy on their progression and accelerate cosmic rays that can in turn transport energy much farther out in the galaxy Dell'Ova, 2021. It is the dominant energy source in most of the diffuse ISM (cf. Fig. 1.5) (Mac Low & Klessen, 2004).

⋆ Stellar winds and bipolar outflows (cf. Fig. 1.5), first studied by Norman and Silk, 1980.

The energy density of the ISM is thus distributed in multiple forms, reported in Tab. 1.2. As emphasized by Draine, 2010, it is remarkable to find all these forms of energy densities matching within less than an order of magnitude. Although we do not have particular reasons to expect the energy density of the CMB to be in accordance with the others, the near-*equipartition* between thermal, turbulent, magnetic, cosmic ray, starlight and far-infrared radiation energy reflects the strong interactions that link these physical processes. For instance, a strong UV radiation from stars heats dust grains to a higher temperature, that in turn reprocess this radiation in the far infrared band.

However, this global balance is not homogeneously verified anywhere and at anytime in the ISM, but rather supports that, if a physical process strongly prevails, locally, over others,

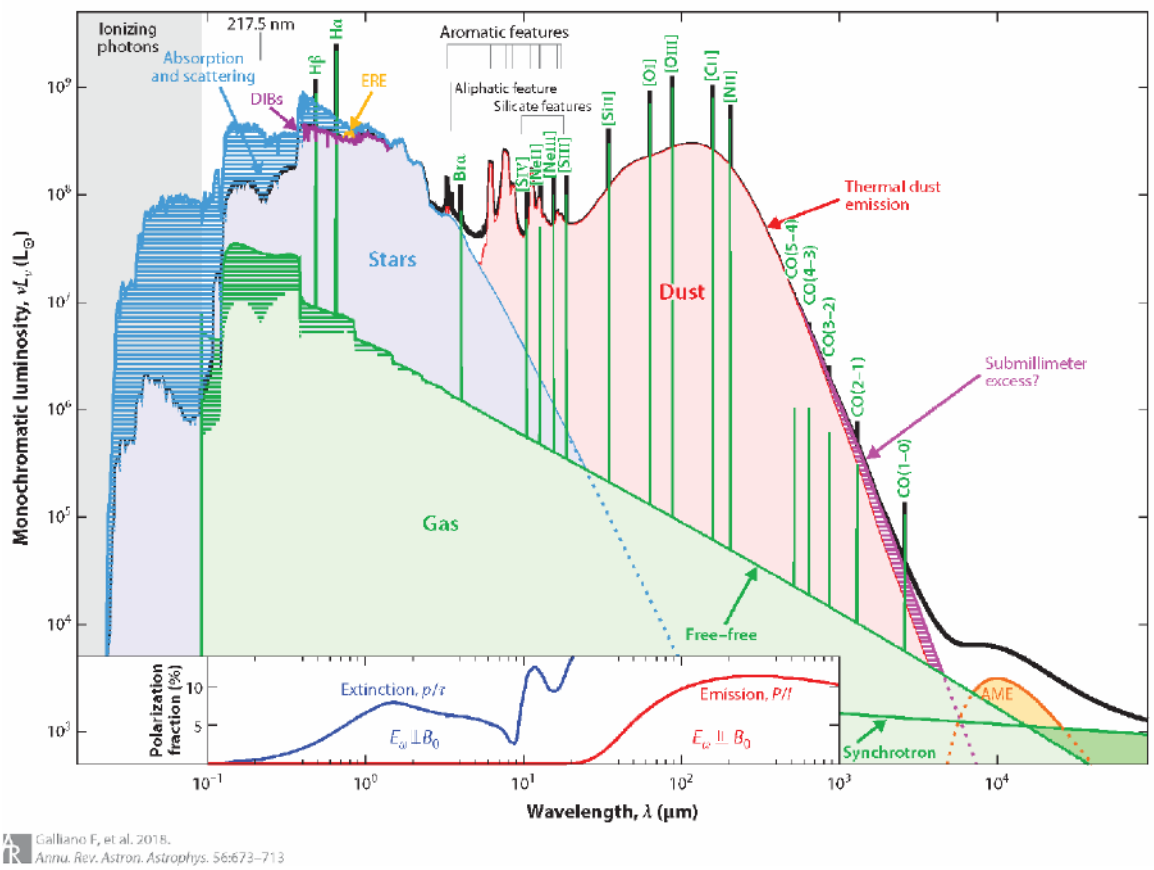


Figure 1.6: Spectral energy distribution of nearby galaxies. Starlight and dust (and Polycyclic Aromatic Hydrocarbons (PAHs)) emission are the main components of the interstellar radiation field. *Credits*: Galliano et al., 2018.

feedback mechanisms will be triggered and tend to mitigate this prevalence, possibly on larger space and time scales.

2.3 ...across multiple scales

The interstellar gas receives energy over a wide range of scales (Elmegreen & Scalo, 2004; Hennebelle & Falgarone, 2012):

- at larges scales ($\sim 1\,\mathrm{kpc}$) *via* the galactic dynamics such as differential rotation,
- at intermediate scales ($10 - 100\,\mathrm{pc}$) *via* supernovæ, winds of massive stars and the over-pressure in H II regions,
- at sub-parsec scales *via* low-mass stellar winds, and where gravity starts dominating,
- at very small scales ($\sim$ au) *via* cosmic-ray streaming.

On the other hand, to build stars from a diffuse cloud, matter density needs to increase by more than 20 orders of magnitude (cf. Fig. 1.5). However, during this contraction, the gravitational collapse faces a resistive pressure from turbulence, rotation and the magnetic field.

Component	Energy density ($\mathrm{eV\,cm^{-3}}$)
Cosmic microwave background ($T_{\mathrm{CMB}} = 2{,}725\,\mathrm{K}$))	0.265
Far-infrared radiation from dust (cf. Fig. 1.6)	0.31
Starlight ($h\nu < 13{,}6\,\mathrm{eV}$) (cf. Fig. 1.6)	0.54
Thermal kinetic energy $(3/2)nk_{\mathrm{B}}T$	0.49
Turbulent kinetic energy $(1/2)\rho v^2$	0.22
Magnetic energy $B^2/8\pi$	0.89
Cosmic rays	1.39

Table 1.2: Main components of the energy density in the local ISM, from Draine, 2010. Thermal kinetic energy is computed for $nT = p/k_{\mathrm{B}} = 3800\,\mathrm{K\,cm^{-3}}$ that corresponds to the peak of the log-normal distribution of thermal pressures in the diffuse CNM obtained by Jenkins and Tripp, 2011. This pressure roughly corresponds to the cooling branch represented by the dashed line in Fig. 1.5.

Therefore, this picture must be completed by efficient dissipation mechanisms of energy and momentum. These include neutral viscosity, ambipolar diffusion, compression, and Ohmic dissipation, that are likely to take place below the milliparsec scale (Hennebelle & Falgarone, 2012).

In the following subsection, we will explain how the energy injected at large scales can be dissipated by these mechanisms occurring at small scales through a turbulent cascade that creates a connection between these scales. Before going into details, the broad picture is the following: the intermittent nature of turbulence transports the energy injected at all scales and focuses it into highly concentrated regions of intense small scale dissipation (Frisch et al., 1978). These dissipative regions are highly structured/coherent and fill a tiny fraction of the volume (Richard, 2022). An illustration of dissipative structures in the (simulated) diffuse ISM is shown in Fig. 1.7. On one side, turbulence fosters star formation as it allows for efficient energy dissipation of the medium by highly concentrating it, but on the other side it provides a coherent kinetic support to the dense regions, that acts as an additive pressure against their gravitational collapse and by so lowers the star formation rate. Hence its role in star formation is not yet clear (Hennebelle & Falgarone, 2012).

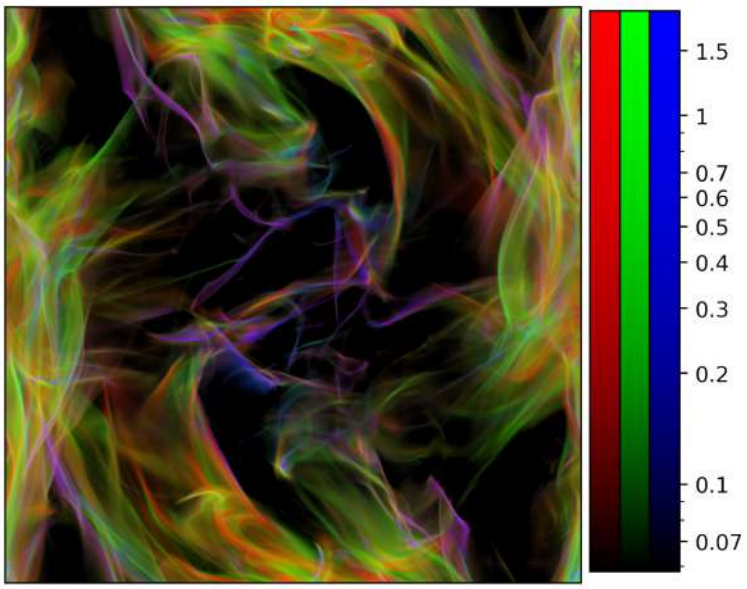


Figure 1.7: Dissipation of MHD turbulence in the diffuse ISM is far from homogeneous: it occurs in localized structures, both in space and time. These bursts of dissipation are referred as *intermittency of turbulence*. Colors account for the type of dissipation: Ohmic $\eta\|\nabla \wedge \mathbf{B}\|^2$ in red, viscous $\rho\nu\|\nabla \wedge \mathbf{u}\|^2$ in green and compressive $\frac{4}{3}\rho\nu|\nabla \cdot \mathbf{u}|^2$ in blue, where η and ν are the resistivity and viscosity. Figure extracted from Lesaffre, Falgarone, and Hily-Blant, 2024 based on simulations of Richard et al., 2022 (ambipolar diffusion is not taken into account).

3 Coupling between scales *via* (notably) turbulence

We have seen that a significant part of the energy in the ISM is injected at large scales, while MHD dissipation occurs at small scales. We now show that the ISM is not just a collection of independent scales, but rather a structured continuum of connected ones (cf. Fig. 1.8 and 1.9), from the large injection scales down to the small scales of dissipation, which creates non-Gaussian structures and dramatically complicates its analysis. We will explain, in the very simplified (but still partially understood) context of 3D incompressible hydrodynamics, why turbulence plays a major role in connecting these scales through a cascade of kinetic energy, before going through the additional difficulties that arise in the compressible, magnetized, and multi-physics ISM.

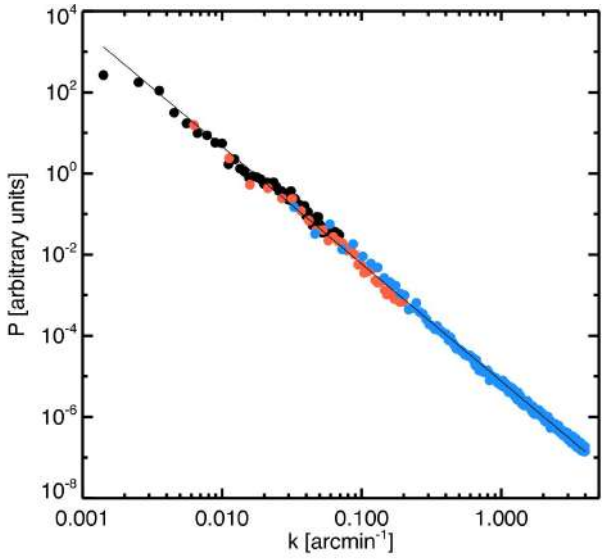


Figure 1.8: Combined power spectrum of dust emission of a diffuse cirrus at high Galactic latitude. Three spectra are used: black is *Planck* radiance, red is WISE and blue is MegaCam. Each spectrum was scaled in order to match the others, and restricted to scales that were corrected for the noise and the beam. The spectrum shows a remarkable power law $PS(k) \propto k^{-2.9\pm0.1}$ over more than three orders of magnitude: scales range from $\sim 50\,\mathrm{pc}$ down to $\sim 0.01\,\mathrm{pc}$. *Credits*: M.-A. Miville-Deschênes et al., 2016.

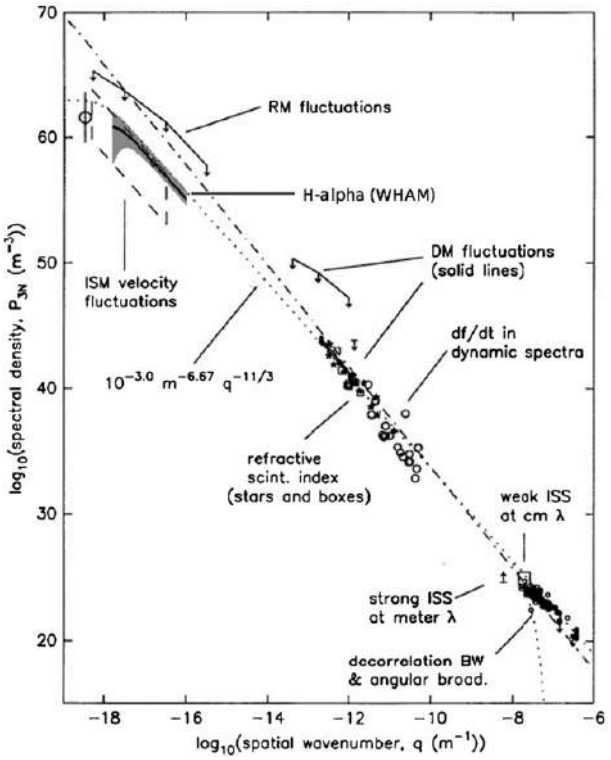


Figure 1.9: The *Big Power Law in the Sky*, from 10^{17} m ~ 3 pc down to 10^{6} m $\sim 3\,10^{-11}$ pc. Power spectrum of the electron density of the ISM and interplanetary ionised medium measured by Armstrong et al., 1995 and then extended at large scales by Chepurnov and Lazarian, 2010. *Credits*: Chepurnov and Lazarian, 2010.

3.1 Insights from the incompressible hydrodynamics

We consider here a 3D flow whose dynamics is given by the incompressible Navier-Stokes equation with a forcing term $\mathbf{f}$ that may depend on time and on the velocity field:

$$\partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p + \nu\Delta\mathbf{u} + \mathbf{f}, \tag{1.1}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{1.2}$$

where $\mathbf{u}$ is the Eulerian velocity field, ρ the fluid density, p its pressure, and ν the kinematic viscosity. We are going to evidence that, when the flow has kinetic energy at large scales (for instance injected by the forcing term $\mathbf{f}$), where the viscosity term $\nu\Delta\mathbf{u}$ dissipates very inefficiently[4], *the nonlinear term* $(\mathbf{u}\cdot\nabla)\mathbf{u}$ transports this energy toward smaller scales, down to scales where the dissipation term prevails over the nonlinear term. This flow regime, referred as *turbulent*, establishes a connection between large injection scales and small dissipation scales, and results in a dramatic increase of the rate at which the flow dissipates.

To do so, we start by deriving a scale-by-scale energy budget of the general Navier-Stokes Eq. 1.1, following the pedagogical approach of Frisch and Kolmogorov, 1995 where the notion of scale is probed by sharp Fourier filtering. This will lead us to the budget of Eq. 1.12 with a term Π_k, called *energy flux*, that encapsulates the kinetic energy transfer at a cut-off wavenumber k. We will then show how this term boils down, in the model of fully developed turbulence, to a strictly positive quantity $\varepsilon > 0$ called *mean rate of energy dissipation*, which is independent of

[4]This linear operator decomposes in Fourier as $-\nu\|\mathbf{k}\|^2\times$, which makes viscous dissipation very inefficient at large scales, where $\|\mathbf{k}\| \to 0$.

the wavenumber k and the viscosity ν, and outline the properties of its associated cascade of energy.

3.1.1 Scale-by-scale energy budget

For a given field $f : \mathbf{r} \in \mathbb{R}^m \mapsto f(\mathbf{r})$, we note $\tilde{f}(\mathbf{k})$ its Fourier transform at wavevector $\mathbf{k}$:

$$\tilde{f}(\mathbf{k}) \equiv \int_{\mathbb{R}^m} f(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}} \, \mathrm{d}\mathbf{r}, \tag{1.3}$$

defined in such a way that:

$$f(\mathbf{r}) = \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} \tilde{f}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}} \, \mathrm{d}\mathbf{k}. \tag{1.4}$$

This allows to define the linear operator $P_k : f \mapsto f_k^<$, that low-pass filters a field at a cut-off scale $2\pi/k$, with $f_k^<$ defined such that:

$$\tilde{f}_k^<(\mathbf{k}) \equiv \tilde{f}_k(\mathbf{k}) \times \mathbf{1}_{\|\mathbf{k}\| \leq k}.$$

We also define $f_k^> \equiv f - f_k^<$, which is the equivalent high-pass filter. For sufficiently regular fields, these operators commute with ∂_t, ∇ and Δ. Therefore, low-pass filtering on space the equations 1.1 and 1.2 at a cut-off scale $2\pi/k$ yields:

$$\partial_t \mathbf{u}_k^< + P_k\big[(\mathbf{u}\cdot\nabla)\mathbf{u}\big] = -\frac{1}{\rho}\nabla p_k^< + \nu\Delta\mathbf{u}_k^< + \mathbf{f}_k^<, \tag{1.5}$$

$$\nabla \cdot \mathbf{u}_k^< = 0. \tag{1.6}$$

Taking the scalar product of these filtered equations with $\mathbf{u}_k^<$ and averaging over space yields the following energy budget of the low-pass filtered velocity $\mathbf{u}_k^<$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\langle\|\mathbf{u}_k^<\|^2\rangle + \langle \mathbf{u}_k^< \cdot P_k\big[(\mathbf{u}\cdot\nabla)\mathbf{u}\big]\rangle = -\langle \mathbf{u}_k^< \cdot \nabla p_k^<\rangle + \nu\langle \mathbf{u}_k^< \cdot \Delta\mathbf{u}_k^<\rangle + \langle \mathbf{u}_k^< \cdot \mathbf{f}_k^<\rangle. \tag{1.7}$$

Then, defining the *cumulative energy*, based on the low-pass filtered velocity field:

$$\mathcal{E}_k \equiv \frac{1}{2}\langle\|\mathbf{u}_k^<\|^2\rangle, \tag{1.8}$$

the *cumulative enstrophy*, based on the low-pass filtered vorticity field $\omega \equiv \nabla \wedge \mathbf{u}$:

$$\Omega_k \equiv \frac{1}{2}\langle\|\omega_k^<\|^2\rangle, \tag{1.9}$$

the *cumulative energy injection*, based on the interaction with the low-pass filtered velocity and low-pass filtered injection field:

$$\mathscr{F}_k \equiv \langle \mathbf{u}_k^< \cdot \mathbf{f}_k^<\rangle, \tag{1.10}$$

and the *energy flux*, resulting from the nonlinear term in Eq. 1.1, based on the total velocity field:

$$\Pi_k \equiv \langle \mathbf{u}_k^< \cdot P_k\big[(\mathbf{u}\cdot\nabla)\mathbf{u}\big]\rangle, \tag{1.11}$$

we obtain[5] the following scale-by-scale energy budget equation:

$$\partial_t \mathcal{E}_k + \Pi_k = -2\nu\Omega_k + \mathscr{F}_k. \tag{1.12}$$

In this equation, Π_k, which is the only term that does not only involve low-pass (large-scales) filtered fields, characterizes the interaction between the different scales, resulting from the non-linearity of Eq. 1.1. Assuming periodic boundary conditions and incompressibility, this term boils[6] down to:

$$\Pi_k = \langle \mathbf{u}_k^< \cdot [(\mathbf{u}_k^< \cdot \nabla)\mathbf{u}_k^>] \rangle + \langle \mathbf{u}_k^< \cdot [(\mathbf{u}_k^> \cdot \nabla)\mathbf{u}_k^>] \rangle \tag{1.16}$$

$$= \langle \mathbf{u}_k^< \cdot [(\mathbf{u} \cdot \nabla)\mathbf{u}_k^>] \rangle, \tag{1.17}$$

and can thus be interpreted as a flux of kinetic energy between the low-pass filtered velocity field $\mathbf{u}_k^<$ and the smaller-scale velocity fluctuations $\mathbf{u}_k^>$, distorted by the nonlinear operator $(\mathbf{u}\cdot\nabla)[\cdot]$.

However, it remains at this stage unclear whether this energy flux can be neglected with respect to the other terms, and if not, what its sign is. Indeed, a positive sign would indicate that energy tends to flow from large to small scales, and a negative sign the opposite. Let us remind that so far, no assumption has been made apart from the incompressible Navier-Stokes dynamics and periodic boundary conditions. In the following, we will show how to compute the energy flux under the assumption that the flow is turbulent.

3.1.2 The fully developed turbulence model

To exhibit the existence of a coupling between scales driven by the term Π_k, we will breakdown this quantity into several pieces. This will allow us to derive in a very straightforward approach a simple form for Π_k, with a clear view on the required assumptions to get there, without going

[5] The term $\langle \mathbf{u}_k^< \cdot \nabla p_k^< \rangle$, written $\langle u_{i,k}^< \partial_i p_k^< \rangle$ with Einstein summation convention, can be decomposed as $\langle \partial_i [u_{i,k}^< p_k^<] \rangle - \langle p_k^< \partial_i u_{i,k}^< \rangle$. The first term vanishes if we assume the fields $\mathbf{u}$ and p to be periodic on the domain on which the spatial averaging is applied. The second term also vanishes, because of the incompressibility assumption that also enforces a divergence free structure to the filtered velocity field $\mathbf{u}_k^<$ (Eq. 1.6).

[6] To analyze the quantity $\Pi_k \equiv \langle \mathbf{u}_k^< \cdot P_k\big[(\mathbf{u}\cdot\nabla)\mathbf{u}\big] \rangle$, let us first note that P_k is self-adjoint, i.e. $\langle f \cdot P_k g \rangle = \langle P_k f \cdot g \rangle$ where $\langle\cdot\rangle$ still denotes spatial averaging, and that it is a projection, i.e. $P_k^2 = P_k$. Hence Π_k becomes:

$$\big\langle \mathbf{u}_k^< \cdot P_k\big[(\mathbf{u}\cdot\nabla)\mathbf{u}\big] \big\rangle = \big\langle P_k\big[\mathbf{u}_k^<\big] \cdot \big((\mathbf{u}\cdot\nabla)\mathbf{u}\big) \big\rangle = \big\langle \mathbf{u}_k^< \cdot \big((\mathbf{u}\cdot\nabla)\mathbf{u}\big) \big\rangle \tag{1.13}$$

Developing the latter expression while decomposing $\mathbf{u}$ as $\mathbf{u}_k^< + \mathbf{u}_k^>$ yields four terms, two of which vanish. Indeed, first:

$$\big\langle \mathbf{u}_k^< \cdot \big((\mathbf{u}_k^> \cdot \nabla)\mathbf{u}_k^<\big) \big\rangle = \langle u_{i,k}^< u_{j,k}^> \partial_j u_{i,k}^< \rangle \tag{1.14}$$

$$= \frac{1}{2}\big(\langle \partial_j [u_{i,k}^{<2} u_{j,k}^>] \rangle - \langle u_{i,k}^{<2} \partial_j u_{j,k}^> \rangle\big), \tag{1.15}$$

The first term in Eq. 1.15 is in the form $\langle \partial_j \dots \rangle$ and therefore vanishes if we assume the field $\mathbf{u}$ to be periodic on the domain on which the spatial averaging is applied. The second term is in the form $\langle \cdots \times \nabla \cdot \mathbf{u}_k^> \rangle$ and also vanishes due to the incompressibility assumption that is also verified by the filtered field $\mathbf{u}_k^>$ (Eq. 1.2-Eq. 1.6). Hence:

$$\big\langle \mathbf{u}_k^< \cdot \big((\mathbf{u}_k^> \cdot \nabla)\mathbf{u}_k^<\big) \big\rangle = 0,$$

and similarly we obtain:

$$\big\langle \mathbf{u}_k^< \cdot \big((\mathbf{u}_k^< \cdot \nabla)\mathbf{u}_k^<\big) \big\rangle = 0,$$

which leads to Eq. 1.16.

too deep in the associated results that will be further discussed in Sec. 2 of Chap. 3. To do so, let us first obtain a global energy budget. Taking $k = \infty$ in Eq. 1.12 yields:

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = -2\nu\Omega + \mathscr{F}, \tag{1.18}$$

with $\mathcal{E} \equiv \frac{1}{2}\langle\|\mathbf{u}\|^2\rangle$, $\Omega \equiv \frac{1}{2}\langle\|\omega\|^2\rangle$ and $\mathscr{F} \equiv \langle\mathbf{u}\cdot\mathbf{f}\rangle$. Now, we will show that, under several assumptions on the flow, and for k in a certain range, called *inertial range*, that will be specified, the energy flux Π_k does not depend on k, nor on ν, and amounts to $\varepsilon \equiv 2\nu\Omega > 0$. To do so let us combine Eq. 1.12 and Eq. 1.18 to decompose Π_k as:

$$\Pi_k = \mathscr{F}_k - \partial_t\mathcal{E}_k - 2\nu\Omega_k \tag{1.19}$$

$$= \mathscr{F} - \mathscr{F} + \mathscr{F}_k - \partial_t\mathcal{E}_k - 2\nu\Omega_k \tag{1.20}$$

$$= 2\nu\Omega + \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} - (\mathscr{F} - \mathscr{F}_k) - \partial_t\mathcal{E}_k - 2\nu\Omega_k. \tag{1.21}$$

Dividing by $2\nu\Omega$, we obtain the exact decomposition:

$$\frac{\Pi_k}{2\nu\Omega} = 1 + \frac{\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} - \partial_t\mathcal{E}_k}{2\nu\Omega} - \frac{\mathscr{F} - \mathscr{F}_k}{2\nu\Omega} - \frac{2\nu\Omega_k}{2\nu\Omega}. \tag{1.22}$$

Now, we will do three assumptions, associated to the so called *fully developed turbulence* flow model, each of which will allow to cancel a specific term in the latter equation. These assumptions are:

- *Stationarity*: we assume[7] that the injection has stationary properties and compensates on average the dissipation $2\nu\Omega$ and that this leads to a flow with stationary averaged properties. Then both $\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t}$ and $\partial_t\mathcal{E}_k$ vanish, which cancels the first term in Eq. 1.22.

- *Injection*: we assume that the force $\mathbf{f}$ injects energy at scales larger than a certain L_0 called *integral scale*. Therefore, considering wavenumbers $k \gg L_0^{-1}$ implies $\mathbf{f}_k^< \simeq \mathbf{f}$ and as a result $\mathscr{F}_k \equiv \langle\mathbf{f}_k^<\cdot\mathbf{u}_k^<\rangle = \langle\mathbf{f}_k^<\cdot\mathbf{u}\rangle \simeq \langle\mathbf{f}\cdot\mathbf{u}\rangle = \mathscr{F}$ which cancels the second term in Eq. 1.22.

- *Anomalous dissipation in the turbulent regime*: experiments and simulations[8] reveal that the viscous dissipation $2\nu\Omega$ is surprisingly not vanishing in the limit $\nu \to 0$, but converges toward a positive value ε (independent of ν), referred to as the *mean rate of energy dissipation* (per unit mass):

$$\lim_{\nu\to 0}[2\nu\Omega] = \varepsilon > [2\nu\Omega]\Big|_{\nu=0} = 0. \tag{1.23}$$

  This result regarding flows that are in the regime $\nu \to 0$ (which corresponds to the infinitely large Reynolds number) is probably one of the most important properties shared by these flows, and confers them with a *turbulent* nature. Indeed, for such a flow, to compensate

[7] This assumption is here to simplify the computation but is actually not necessary to have an energy cascade (cf. e.g., section 11.3.2 (Galtier, 2016)). For instance, in decaying turbulence, one might have $\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} \sim \partial_t\mathcal{E}_k > 0$ but still $|\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} - \partial_t\mathcal{E}_k| \ll 2\nu\Omega$.

[8] Cf. e.g., Fig. 11.6 p. 185 in the textbook (Galtier, 2016).

the vanishing term $\nu \to 0$, its enstrophy Ω must diverge, and therefore its vorticity field $\omega \equiv \nabla \wedge \mathbf{u}$ must diverge too in some regions, which is associated to the development of small-scale fluctuations of the velocity field that feed the curl operator in the vorticity. These regions that arise in a turbulent flow are called *dissipative structures*. An example of such structures is given in Fig. 1.7 (but in the much more complex case of compressible MHD turbulence, which includes other forms of dissipation). On the other hand, for a fixed wavenumber k, the quantity Ω_k remains bounded as $\nu \to 0$, which cancels the third term in Eq. 1.22 for a sufficiently small ν. Indeed:

$$\Omega_k \equiv \langle \|\nabla \wedge \mathbf{u}_k^<\|^2 \rangle \leq k^2 \langle \|\mathbf{u}_k^<\|^2 \rangle \tag{1.24}$$

$$\leq k^2 \langle \|\mathbf{u}\|^2 \rangle \tag{1.25}$$

$$\equiv k^2 \mathcal{E} < \infty. \tag{1.26}$$

The inequality 1.24 is due to bounding properties of the curl operator, and the inequality 1.26 is verified if we assume[9] that the solution to the incompressible Navier-Stokes equation has a bounded energy on any time horizon. Hence, the third term in Eq. 1.22 is controlled as:

$$\frac{2\nu\Omega_k}{2\nu\Omega} \equiv \frac{2\nu\Omega_k}{\varepsilon} \leq \frac{2\nu k^2 \mathcal{E}}{\varepsilon}, \tag{1.27}$$

and therefore, considering wavenumbers $k \ll \sqrt{\varepsilon/\nu\mathcal{E}}$ cancels this term.

In conclusion, if we consider a stationary flow:

- with an injection at scales larger than some integral scale L_0,
- with ν sufficiently small so that the dissipation $2\nu\Omega \equiv \varepsilon$ is strictly positive and does not depend on ν,
- with sufficient kinetic energy $\mathcal{E} \equiv \langle \|\mathbf{u}\|^2 \rangle$ so that the scale $\lambda \equiv \sqrt{10\nu\mathcal{E}/\varepsilon}$, called *Taylor scale*, is negligible compared to the integral scale,

then, for any k such that $L_0^{-1} \ll k \ll \lambda^{-1}$, we finally have:

$$\Pi_k \simeq \varepsilon > 0, \tag{1.28}$$

and the scale-by-scale energy budget:

$$\partial_t \mathcal{E}_k + \Pi_k \simeq \mathscr{F}_k \simeq \varepsilon > 0, \tag{1.29}$$

that can now be interpreted as a cascade of kinetic energy from the large scales of injection toward the small scales of dissipation. The range of scales (or wavenumbers k) that verify this cascading of the energy at rate ϵ, with negligible contribution of the injection and of the

[9] This assumption is not proved for 3D flows and figures among the Millennium Prize Problems. It does not hold in 2D turbulence and other flows having an inverse energy cascade which might have diverging energy. See Frisch and Kolmogorov, 1995 for further details.

dissipation, is called the *inertial range*. We have shown that this range extends at least from L_0 down to λ, but in practice, it extends further to smaller scales[10].

3.2 The much more complex ISM case

In the ISM, however, the fluid motion is very different from the fully developed incompressible hydrodynamical turbulence depicted above (see Elmegreen and Scalo, 2004 for a review). Indeed the ISM turbulence is:

- magnetized: the mean magnetic field $\mathbf{B}_0$ creates anisotropy with difference in the cascade, between scales that are aligned with it and those that are perpendicular. Furthermore, the magnetic flux is frozen with the fluid motion, up to ambipolar diffusion that allows the neutrals to decouple from the ions and constitutes another form of dissipation.
- compressible, with shocks and characterized by high Mach numbers. The gravitational instability enhances gas overdensities and leads to fragmentation (Hennebelle & Chabrier, 2008; Padoan & Nordlund, 2002). These compressive stages are coupled with heating and cooling mechanisms that constitute another form of dissipation.
- with a large diversity of injection scales and mechanisms (discussed in Sec. 2.3).

Yet, turbulence is ubiquitous in the ISM and plays a key role. We do not claim that it dominates the dynamics: other physical processes have a major impact. Their interplay with interstellar turbulence is a highly debated topic that makes it much less understood than the incompressible hydrodynamical turbulence (Hennebelle & Falgarone, 2012). Perhaps surprisingly, numerous statistical diagnostics that are known in some idealized forms of turbulence, such as power law in the density structure (cf. e.g., Fig. 1.8), are observed alike in the ISM (Elmegreen & Scalo, 2004).

4 What tools?

Most of the ISM physical conditions are difficult to reproduce in a laboratory[11]. The study of the ISM thus mainly relies[12] on its observations and on numerical simulations.

4.1 Observations

Each phase of the ISM can be observed with specific tracers, for instance:

- UV, X and radio synchrotron emission for the HIM,

[10] Indeed, assuming a power spectrum $E(k) \sim \varepsilon^{2/3} k^{-5/3}$ in the computation of the cumulative dissipation $2\nu\Omega_k = 2\nu \int_0^k \kappa^2 E(\kappa)\,\mathrm{d}\kappa$ leads to the reformulation of the condition $2\nu\Omega_k/2\nu\Omega \equiv 2\nu\Omega_k/\varepsilon \ll 1 \iff k \ll \left(\frac{\nu^3}{\varepsilon}\right)^{-1/4}$ which gives $\left(\frac{\nu^3}{\varepsilon}\right)^{1/4}$ as a limit scale for the inertial range, called *Kolmogorov dissipation scale*.

[11] The densest regions have a density $n_\mathrm{H} \sim 10^6\,\mathrm{cm}^{-3}$ which remains more tenuous than the most extreme vacuum conditions reproduced in the lab.

[12] Still, let us emphasize that laboratory experiments constitute a precious way to constrain microphysics such as collisions, reaction rates, dust optical properties, chemical reactions and spectroscopic characterization to detect species.

- radio continuum emission and hydrogen optical recombination lines for the WIM,
- 21-cm line absorption and emission, dust emission in the far infrared, CII and OI fine-structure lines and metal absorption lines in the UV/visible, for the atomic gas,
- CO rotational line emission (along with other molecules) in the millimeter range, dust emission in the far infrared and electronic absorption lines of H_2 for molecular phases.

In this work, we will mostly focus on molecular clouds. Since H_2 is a stable molecule, it is therefore difficult to observe in emission (unless it is shock-heated). In order to have emission maps, we will instead rely on the thermal emission of the dust grains, that are intimately coupled with the gas, including in molecular clouds. The emission of arge dust grains[13] peaks in the far infrared (cf. Fig. 1.6) with a specific intensity I_ν that can be modelled as a modified black body:

$$I_\nu = \tau_0(\nu/\nu_0)^\beta B_\nu(T_d), \tag{1.30}$$

with ν the frequency, T_d the grain temperature (typically $T_d \sim 20\,\mathrm{K}$ in the bulk of the ISM, far from any source of UV/visible radiation) and β called *spectral index* (typically $\beta \sim 1.6$ (Planck Collaboration et al., 2014a)). Observing their emission and fitting their parameters allows to estimate their column density assuming a certain choice of grain population (Ysard et al., 2024). This can then be used to trace the H_2 column density assuming a constant dust-to-gas ratio (Ward-Thompson & Whitworth, 2015). An example of such a map is shown for the Orion B giant molecular cloud in Fig. 1.10. In this figure, it is possible to distinguish 2D filamentary structures that are typical morphological features of molecular clouds.

However, relying on observations in order to describe 3D fields and constrain their physical dynamics comes with some difficulties, mostly due to the following limitations:

- observations are integrated along the line of sight, which represents only a partial measurement of the 3D structure,
- absorption measures are very sparse,
- there is no kinematic information in the dust continuum (although CO and H I can give partial information respectively in dense and diffuse regions of molecular clouds),
- instrumental limitations (resolution, sensitivity, noise), and further astrophysical contaminations (e.g., CIB for dust emission in diffuse regions).

4.2 Simulations

On the other hand, numerical simulations are of precious help in order to unveil the dynamics of a complex, multi-physics system such as the ISM. They indeed allow to test *in silico* the

[13] Smaller grains are heated stochastically and emit in narrow bands.

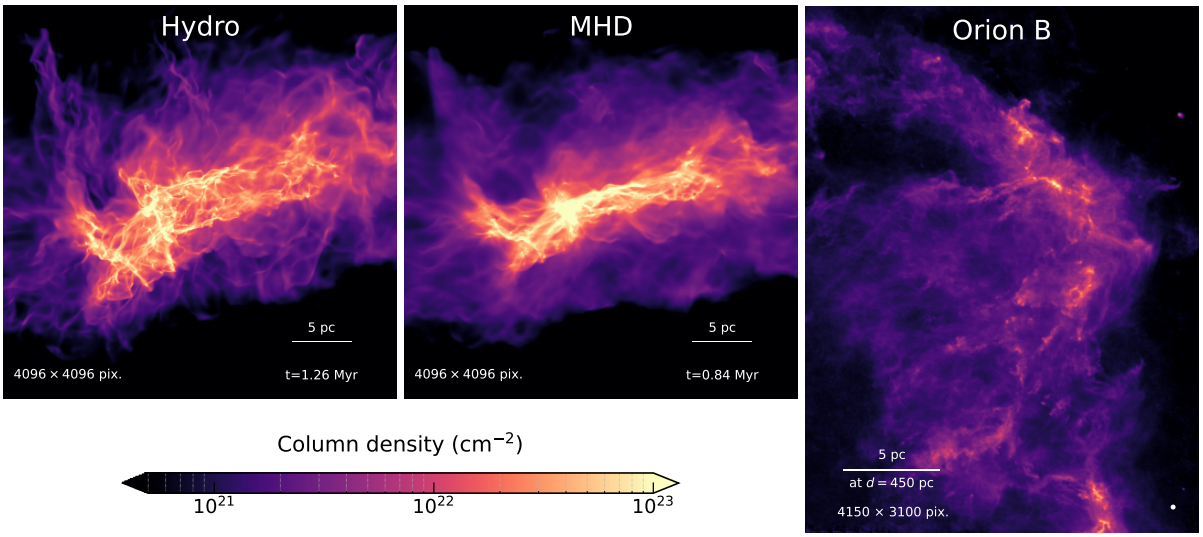


Figure 1.10: H_2 column density maps of a giant molecular cloud. Snapshots from hydrodynamical (left) and magnetohydrodynamical (middle) simulations from Ntormousi and Hennebelle, 2019. These are attempts at reproducing the collapse of a dense molecular cloud under self-gravity, including decaying MHD turbulence but without stellar feedback. See Sec. 8.1.1 of Chap. 5 for a detailed description of this set of simulations. Right: column density map derived from dust emission observation of the *Herschel* satellite André et al., 2010 by Könyves, V. et al., 2020. The diameter of the white dot in the bottom right corner represents five times the map resolution.

influence of such or such physical process and its associated parameters on the observables, in a fully controlled framework.

An example of state-of-the-art numerical simulations that aim at reproducing the star formation process during the gravitational collapse of a giant molecular is shown in Fig. 1.10. Two snapshots are given: one from a fully hydrodynamical dynamics (left), and one including a magnetic field (middle). We see that both simulations reproduce some morphological features, such as filamentary structures, that are present in the column density map derived from dust emission of the Orion B cloud. As these complex morphological structures yielded by observations are difficult to be anticipated when looking at the (relatively) simple compact set of MHD equations, and as they are difficult to be generated by "simplistic" simulations, retrieving such complex structures in a simulation is very encouraging for its underlying physical model. But it also raises important questions discussed in the following section.

5 Problematic of this work

In this chapter, we have shown that the ISM is made of complex multi-scale structures whose study is motivated both by astrophysical and cosmological purposes. However, because of the statistical nature of the ISM dynamics, that will be discussed in Chap. 2, the characterization of these structures must be defined in a statistical framework.

In order to do so, there are on the one hand observations, but which provide a limited

information, both in terms of physical fields that can be measured and in terms of number of data samples. On the other hand, there are numerical simulations, whose growing capabilities continuously extend the set of approachable dynamics. Still, the latter remains far to reach for instance the typical 10^7 Reynolds number of ISM flows, and more importantly, constitutes a very wide range of options to explore in order to unravel the dynamics of the multi-physics ISM. There is therefore the need to identify the subset of simulations that best reproduce observations. Identifying such a subset would then provide physical information about the actual ISM dynamics and statistical models of the ISM data for component separation problems.

However, if the set of simulations used is actually not sufficiently realistic, as it is extremely difficult to precisely reproduce the complex multi-physics of the ISM all-at-once (including the pipeline that maps simulation fields to instrument observable quantities), the latter conclusions might be misleading (cf. Fig. 1.4). Therefore, there is also a vital need to estimate a notion of distance between simulations and observations.

On a more general perspective, there is the need to define a notion of *distance between observations of the ISM and a given statistical model*. In addition to enlighten the comparison between observations and simulations, such a distance would for instance allow to compare two (*a priori* different) observations between them, and also to provide warranties in components separation problems. In this work, we aim at defining and estimating such a statistical distance, in the context of a limited amount of observations.

Chapter 2

Statistical nature of the Interstellar Medium

"It has been realized since the beginning that the problem of turbulence is a statistical problem; that is a problem in which we study instead of the motion of a given system, the distribution of motions in a family of systems..."

Wiener, 1939

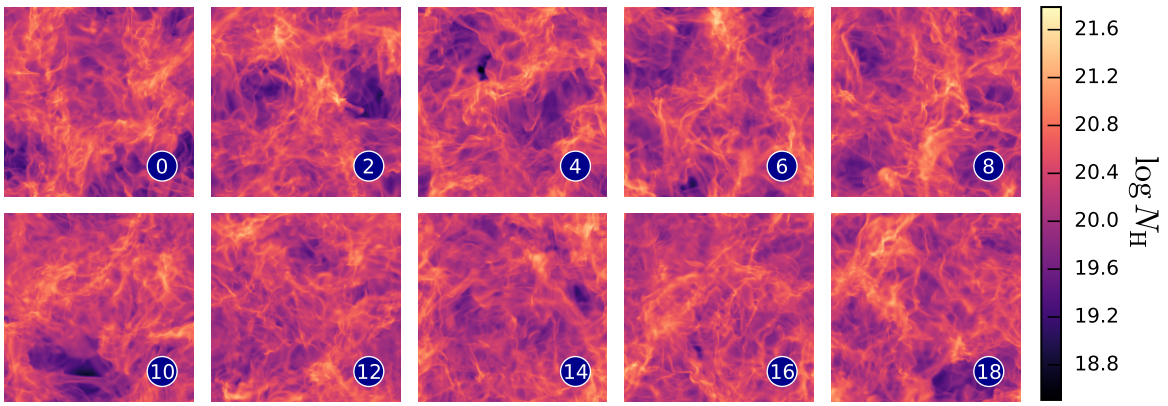
Objectives

We explain the need for statistical approach of nonlinear physical dynamics as those encountered in the ISM, outline the main difficulties that arise due to their high dimensional non-Gaussianity. We then introduce the various goals of the community, discuss how they differ in nature and complexity as well as the different frameworks and approaches usually followed to tackle them.

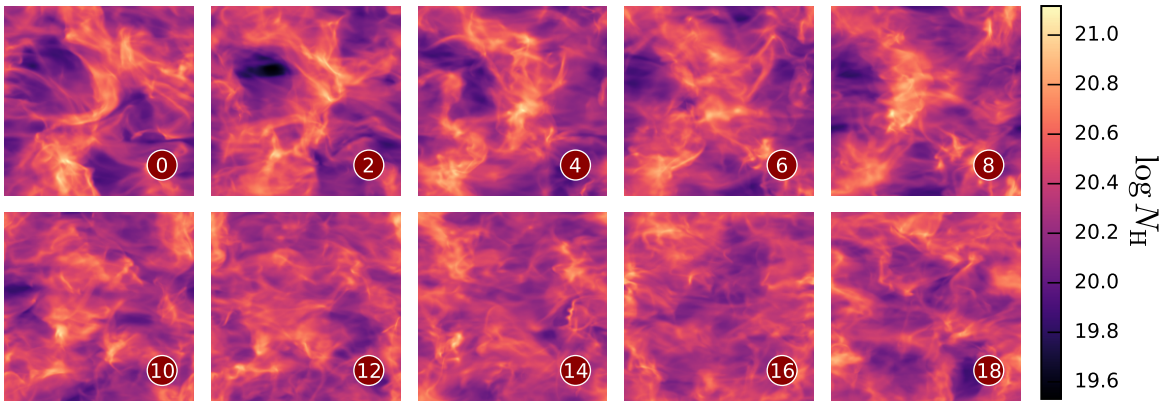
Contents

1 Introduction

The dynamics of the ISM is driven by strongly nonlinear and interacting physical processes such as turbulent MHD, shocks, chemistry, radiative transfer, cosmic rays and gravity (Draine, 2010). This nonlinearity provokes an erratic time evolution: the trajectory followed by a state in the phase space is full of turns and explores nonperiodically a vast range of states (Lorenz, 1963). To illustrate this, we report in Fig. 2.1 snapshots at different time-steps of a stationary turbulent and gravitation-free MHD numerical simulation of diffuse ISM, where the energy dissipated at small scales by viscosity is balanced by a turbulent forcing of the velocity field at large scales (dataset as used in Allys et al., 2019, based on Iffrig and Hennebelle, 2017).



(a) Snapshots from simulation class n°2: non-magnetized with intermediate turbulent forcing.



(b) Snapshots from simulation class n°6: magnetized with high turbulent forcing.

Figure 2.1: Column density maps N_{H} at different time-steps for two different MHD simulations used in Allys et al., 2019 (based on the solver described in Iffrig and Hennebelle, 2017, but with different forcing). Class n°2 has no magnetic field (hydrodynamical case) and intermediate forcing (overall turbulent velocity dispersion $\sigma_{\mathrm{turb}} = 4\,\mathrm{km}\cdot\mathrm{s}^{-1}$), while class n°6 has a low magnetic field ($B_0 = .5\,\mu\mathrm{G}$) and high forcing ($\sigma_{\mathrm{turb}} = 9\,\mathrm{km}\cdot\mathrm{s}^{-1}$).

Let us observe that after a sufficient time, called *turnover time*, the column density field

appears as being fully spatially *reshuffled*. In the case of the simulation class n°2, this time appears to be smaller than the time-step. It is indeed difficult to identify patterns in the snapshots that would give a glimpse on the chronology of the sequence. As it evolves with time, the field is reshaped and takes multiple facets, following an order that is hard to unravel: its trajectory seems indeed erratic. In the case of simulation n°6, while close-by time-steps can be visually related, we nevertheless recover this property for snapshots separated by more than 10 time-steps. To complement[1] this visual observation with a quantitative diagnostic, we report in Fig. 2.2 the spatial correlations

$$\mathrm{Cor}\,(x_0, x_t) \equiv \frac{\langle \bar{x}_0 \cdot \bar{x}_t \rangle}{\sqrt{\langle \bar{x}_0^2 \rangle \langle \bar{x}_t^2 \rangle}}, \tag{2.1}$$

between the first snapshot x_0 of a given class and another snapshot x_t, taken at a later time-step t. In this formula, $\langle\rangle$ denotes spatial averaging and $\bar{x} \equiv x - \langle x \rangle$ probes the fluctuations of the map x around its empirical mean. The fast decay to 0 of the blue curve supports our previous visual analysis.

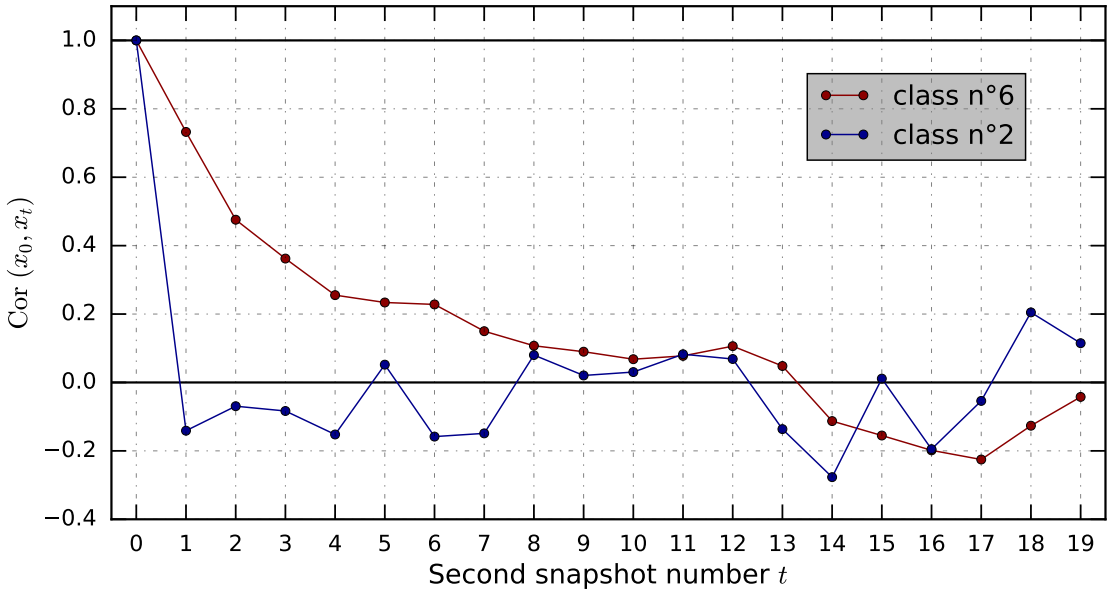


Figure 2.2: Spatial correlations between first image and other images of the same class. As the delay between snapshots increases, their spatial correlation decays to 0, up to chance correlations. The typical decaying time, called *turnover time*, depends on the process.

Hence, this turbulent dynamics gives a typical example of trajectory that runs through various states, in an order difficult to unravel provided that we wait more than a turnover time. This erratic behavior of a trajectory makes turbulence a non-integrable system: in general, to predict the future of a state, there are no shortcuts such as analytical formula, the trajectory

[1] Visual diagnostic is actually very powerful, while such spatial correlation is a strongly partial diagnostic of chronological dependence. For instance an advection dynamics with uniform velocity, which is linear and yields trajectories that simply shift the initial conditions, can lead to quick decorrelation, as seen for class n°2, but remains strongly dependent on the initial conditions and this would be easy to catch visually.

has to be frontally simulated up to the requested instant Tsinober, 2009. Even worse, besides having a complex geometry, this trajectory is extremely dependent on the chosen initial state, conferring a *chaotic* behaviour to the dynamics. This narrows in practice our ability to predict it with precision on a finite horizon called *Lyapunov time* (Bezruchko & Smirnov, 2010). To give an idea, in atmospheric dynamics, this typical duration is estimated to be around two weeks and sets hereby an ineluctable bound to weather forecasting Palmer and Hagedorn, 2006. In the ISM context, interstellar turbulence complicates for instance the task of unraveling the formation history of a quiescent molecular cloud from a given observational snapshot.

On the other hand, besides its unpredictability and seemingly erratic nature, the ISM surprisingly generates highly *organized structures*, with shared high level properties from one observation to another if they share a similar dynamic. This can be seen in Fig. 2.1 as snapshots of a same class have similar morphological features: class n°2 displays strong filamentary structures, in class n°6 elongated structures tend to align horizontally due to the presence of a mean magnetic field. Indeed, while the deterministic details of these textures change from one snapshot to another, it would be easy for a human to infer the originating class of these snapshots based on their appearance, even without supervision. To illustrate what would be a more quantitative approach to such texture comparison task, we report in Fig. 2.3 histograms of each snapshot. This lower dimensional representation of the images is relatively stable from one snapshot to another of the same class, and already allows to separate both classes with a good accuracy. This means that from a single image only, we can retrieve non negligible information about the dynamics from which it originates. There is thus great hope to grasp physical properties of the ISM by characterizing it statistically.

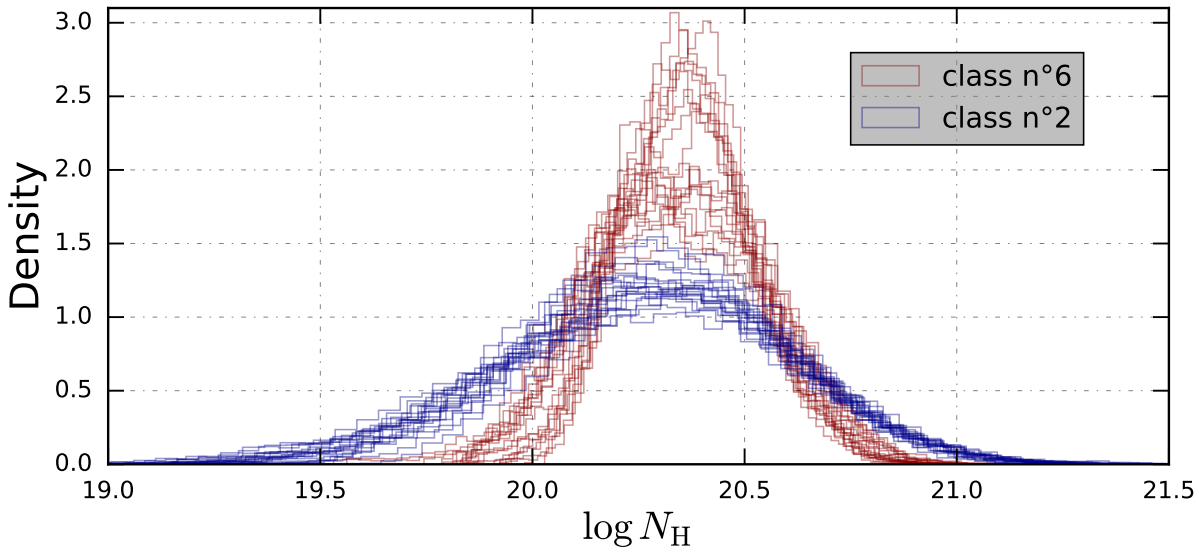


Figure 2.3: Histograms computed for each snapshot. Each class appear well structured in this plot. Despite losing their spatial correlations as shown in Fig. 2.2, time delayed snapshots still share close statistical properties inherited from the simulation's class.

This simplified example of ISM dynamics opens numerous questions that will be the focus

of this chapter:

- What causes the chaotic behavior along time evolution?
- Why is there apparent order in such chaotic systems?
- What statistical framework is suited to characterize this order?
- How are the statistical properties to the physical dynamics?

The investigation of statistical features relevant to characterize such morphological properties yielded in the ISM is left to the following chapter.

2 From physical to stochastic processes

We motivate in this section the relevance of stochastic approach to analyze ISM processes that undergo chaotic dynamics. As chaos can arise in turbulence, and as turbulence is a deterministic process, we therefore justify the following motivation in a fully deterministic context, although stochasticity may occur in the ISM (e.g., through any process involving quantum mechanics, such as radiative transfer) and in our data handling (uncertainty, instrumental measure, numerical error).
First we give a brief flavor of chaos through a toy model. Then we explain how order arises when shifting the focus from a state perspective to a distribution of states. This leads us to the notion of *invariant measure* (or invariant probability distribution). Thereafter, we motivate the relevance of these measures to characterize dynamical processes and finally explain how these measures can be effectively sampled thanks to *ergodicity*.
Along this section, we consider a fixed dynamics, fully deterministic and chaotic, that aims at modeling, from near or afar, a given turbulent process. It can be a mathematical toy model, a physical set of equations with fixed hyper-parameters, its numerically computable counterpart as shown in Fig. 2.1 or even a controlled laboratory experiment.

2.1 Chaos through a deterministic toy model

Chaos: When the present determines the future
but the approximate present does not approximately determine the future.

attributed[2] to Edward Lorenz (2005)

This citation accounts for the first property defining a chaotic dynamics: it is highly sensitive to initial conditions. It is indeed observed in systems like turbulence that an arbitrarily small discrepancy in initial conditions will almost surely change dramatically the future trajectories. This property is popularized as the *butterfly effect*.

[2] http://mpe.dimacs.rutgers.edu/2013/03/17/chaos-in-an-atmosphere-hanging-on-a-wall/

However, this sensitivity effect on initial conditions is not sufficient to account for the erratic behavior that further characterizes a chaotic dynamical systems. Indeed, such sensitivity can be emulated from a linear dynamics of a state $x(t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}x = Ax, \tag{2.2}$$

which evolution boils down to:

$$x(t) = e^{At}x(0). \tag{2.3}$$

Therefore, if a discrepancy $\delta x(0)$ between two initial conditions is an eigenvector of A associated to a strictly positive eigenvalue $\lambda > 0$, this discrepancy will be exponentially amplified and diverge:

$$\delta x(t) = e^{\lambda t}\delta x(0) \underset{t\to\infty}{\to} \infty. \tag{2.4}$$

However, this diverging discrepancy leads to having at least one of the two states that diverges too. Yet, chaotic systems remain bounded. For instance the column density maps shown in Fig.2.1 do not exhibit any diverging trend, even further, all maps of a given a simulation share quite global properties as shown in Fig. 2.3. Therefore, the nonlinearity of these dynamics must dramatically change the exponential divergence of linear dynamical systems depicted here.

To illustrate how these systems can develop an erratic and seemingly unpredictable but bounded time evolution, we consider a famous toy model: the *tent map*. It is defined as the following discrete time dynamical process:

$$x_{t+1} = 1 - |2x_t - 1|, \tag{2.5}$$

starting from an initial condition $x_0 \in [0, 1]$. Because of the absolute value in the time evolution mapping from x_t to x_{t+1}, this toy model has a nonlinear dynamics. This tent shape mapping is shown in Fig. 2.4. Note that it maps $[0, 1]$ to itself, therefore the dynamics remains bounded in this set.

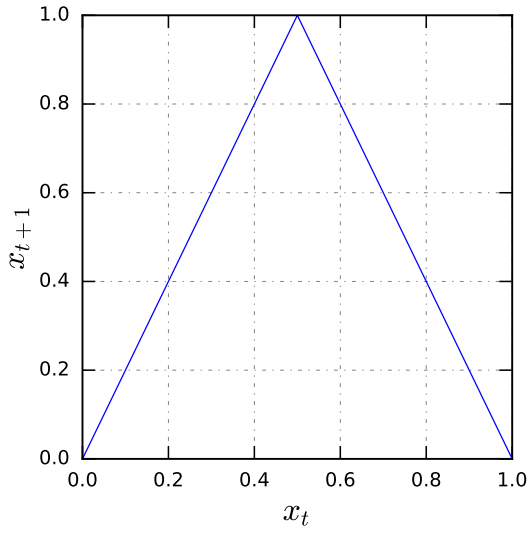


Figure 2.4: The tent map: a toy model of nonlinear dynamical system.

As shown in Fig. 2.5, where we run trajectories from the tent dynamics for 100 time-steps, this relatively simple nonlinearity is already sufficient to generate dramatic disorder: the curves roam in the segment $[0, 1]$ in a nonperiodic and erratic fashion. In fact it can be shown that the n-th bit of the binary decomposition of x_0 will dominate the evolution of the state at step n (see e.g., Chap. 3 of (Frisch & Kolmogorov, 1995)). Therefore, considering an arbitrary initial state x_0 (for instance uniformly sampled in $[0, 1]$), any arbitrary sequence of bits can be expected to occur in its binary decomposition, which in will in turn lead, at proper time, to an associated erratic imprint in the state evolution.

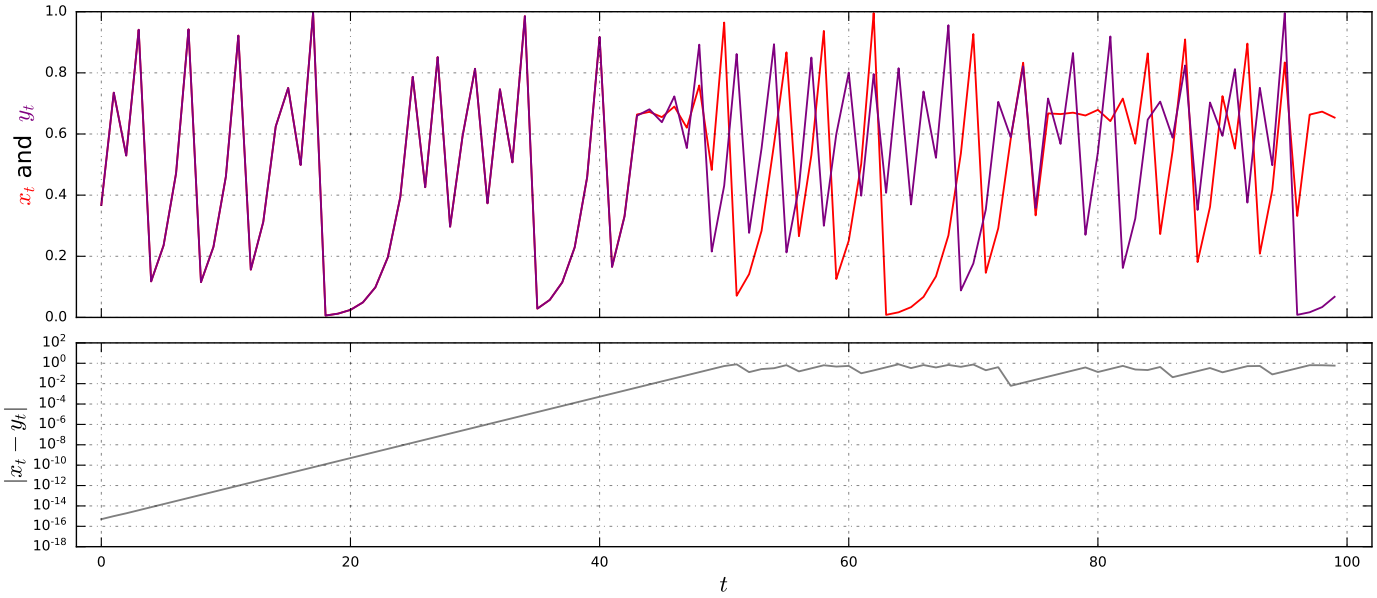


Figure 2.5: Two trajectories x and y following the tent dynamics, and their deviation $|x_t - y_t|$. The dynamics seems very disordered. It is also unpredictable on long term: even though starting from very close initial conditions $|x_0 - y_0| \leq 10^{-15}$, the trajectories deviate exponentially until they saturate and have totally separated after 50 time-steps.

In addition, this importance of *all* the digits of x_0 in the global trajectory renders difficult the task of simulating it on a long time horizon. If one tries to do so in a naive numerical simulation based on standard float formats, the evolution will almost surely converge to 0 in $\sim$ 50 time-steps. To illustrate this high sensitivity on the initial conditions, we run (with a binary-based simulation that overcomes the previous caveat) in Fig. 2.5 two trajectories x_t and y_t with very close initial conditions: $|x_0 - y_0| \leq 10^{-15}$. After 50 time-steps they have totally separated. It is noteworthy that, even though the dynamics at play is nonlinear, we still observe an exponential diverging of the the trajectories in the early times. Indeed, as shown in the plot underneath, at each iteration their discrepancy grows by a factor two (see e.g., (Frisch & Kolmogorov, 1995)), until they cannot diverge anymore due to their bounding property. The logarithm of this dilation factor is referred to as the *Lyapunov exponent* in chaos theory, the inverse of which gives the *Lyapunov time* which sets a typical horizon in forecasting (Bezruchko & Smirnov, 2010). Even more surprising, this exponential diverging period is much longer than the typical time the system takes to have its state spanning accross the phase space $[0, 1]$,

which is of order 1 step here. This diverging period is therefore influenced not by only one local linearization of the dynamics as shown in Eq. 2.4 but undergo a mixture of different (with opposite sign) local expansions, but still maintain an exponential trend.

2.2 Attractor and invariant measure

As chaotic systems display an erratic behavior (Fig. 2.5), it might come as a surprise that they nevertheless exhibit some forms of regularity, as noticeable in the collection of histograms reported in Fig. 2.3. We explain here that these regularities emanate from fixed-point properties of the dynamics, but taken at a distribution of states level instead of a state level.

2.2.1 Fixed-point

A discrete time dynamical system $x_{t+1} = f(x_t)$ consists in iterating the same transformation f. Writing $f^t \equiv f \circ \cdots \circ f$ the composition of f with itself t times with convention $f^0 \equiv \mathrm{id}$, one has

$$x_t = f^t(x_0). \tag{2.6}$$

Not surprisingly, if x is a fixed-point of f, i.e., such that $f(x) = x$, one has $f^t(x) = x$. If f is a contraction mapping on some metric space $(E, \|.\|)$, i.e., there exists $k < 1$ such that

$$\forall x, y \in E \ \|f(x) - f(y)\| \leq k\|x - y\|, \tag{2.7}$$

the Banach-Picard fixed-point theorem Banach, 1922, theorem n°6 ensures[3] that:

- there exists a fixed-point x of f,
- this fixed-point x is unique,
- for all x_0 in E, the dynamical system $f^t(x_0)$ converges to x.

Hence, no matter the initial condition, the dynamics of a contraction system is trapped by an attractive fixed-point on which it progressively collapses.

2.2.2 From point to distributions

Even though insightful, this situation is far from fitting the dynamics of a chaotic system that is exponentially diverging locally, before saturation effect due to boundaries or nonlinearities comes in, in space (for almost every x except null set) and time:

$$\|f^t(x) - f^t(x + \varepsilon)\| \geq e^{\lambda t}\|\varepsilon\|, \tag{2.8}$$

for some Lyapunov exponent $\lambda > 0$.

However, this idea of contraction toward an attractive fixed-state can be extended, for a given dynamical system f, to broader classes of objects, such as attractors (that are sets of states) and

[3] Assuming the metric space $(E, \|\cdot\|)$ is complete (i.e. that contains the limits of its Cauchy sequences) and non-empty.

invariant measures (that are probability distributions over the states). Indeed, if we consider a probability distribution $x \mapsto p(x)$ over the phase space E, and a random variable $X \sim p$ following this distribution, then applying one step of the dynamics f onto X leads to another random variable $f(X)$ which distribution is noted $f_{\#}p$ and reads "the pushforward measure of p by f":

$$X \sim p \quad \Longrightarrow \quad f(X) \sim f_{\#}p. \tag{2.9}$$

Therefore, the state dynamics $f(x)$ induces a dynamics $F(p)$ on the space of distributions:

$$F : p \mapsto F(p) \equiv f_{\#}p. \tag{2.10}$$

In the following we do not make a distinction between these two levels of dynamics. For chaotic systems, the dynamics of a state is almost never trapped toward a fixed-point. However, the dynamics of a measure usually is: as it is pushed forward in time, an initial measure p_0 is likely to be attracted by and converge toward an invariant measure $p^* = F(p^*)$ of the chaotic system:

$$\begin{aligned} X_0 \mapsto X_1 \equiv f(X_0) \quad &\mapsto \dots \quad \mapsto X_t \equiv f^t(X_0) \\ p_0 \mapsto p_1 \equiv f_{\#}p_0 \quad &\mapsto \dots \quad \mapsto p_t \equiv (f^t)_{\#}p_0 \simeq p^* \end{aligned} \tag{2.11}$$

Schematically, the time dynamics progressively favors the states x that have a large basin of attraction (orange state in Fig. 2.6), at the expense of states that have small attraction and are simply depleted by exponential diffusion (blue and violet states in Fig. 2.6).

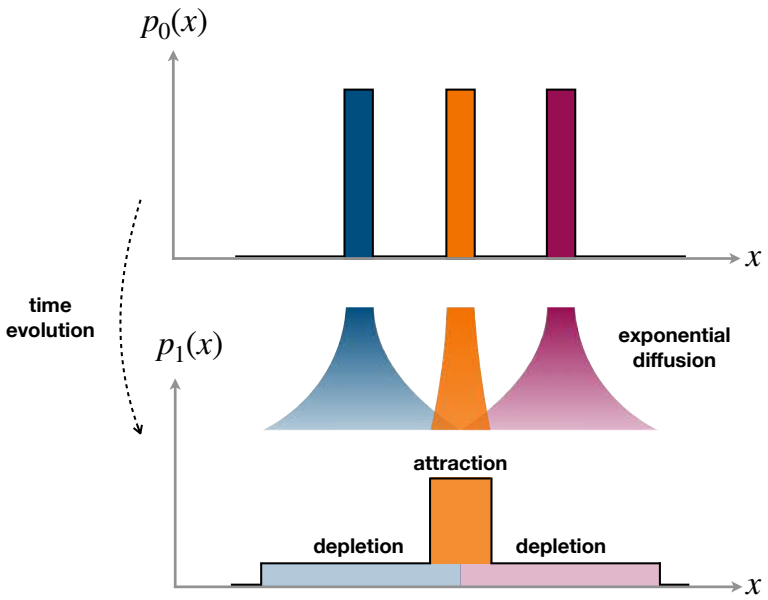


Figure 2.6: Time evolution of a probability distribution of states under chaotic dynamics.

States x with high probability $p^*(x)$ are the ones favored by the dynamics. They can be seen as low energy states, in analogy with statistical physics where these are also the most populated. Hence, if we *access to an invariant measure* of a chaotic system, we can *retrieve information* about its favored and forbidden states, which, if we pursue the analogy with statistical physics, gives clues about its Hamiltonian (Miller, 1990). However this analogy

remains limited and questionable because processes in the ISM are in a large measure out from being in a global thermodynamic equilibrium (Kraichnan & Chen, 1989). As matter of fact, the simulations of diffuse ISM presented in the beginning of this chapter are a typical example of stationary dynamics that is *out of equilibrium*, as a continuous flow of energy is injected in the turbulent cascade that dissipates it at small scales.

The tent map is a very particular chaotic system where the uniform distribution is an invariant measure: it favors no particular state (Frisch & Kolmogorov, 1995). This denotes with turbulent dynamics, that is observed to favor states with highly dissipative structures (cf. Fig. 1.7). The collection of such states have zero Lebesgue measure in the phase space and a fractal structure, in the context of dissipative systems in finite[4] dimension (Ruelle, 1991).

Having a system with an attractive invariant measure implies that, as it evolves, the system not only loses the trace of its initial sample x_0, but it also loses[5] in its distribution of states the footprint of the initial distribution p_0. A clear illustration of this phenomenon is the shape of the histograms reported in Fig. 2.3: they all appear log-normally distributed in N_H while the simulations are initialized with a uniform gas density (Allys et al., 2019). Hence, after one turnover time only (snapshot n°0), the histograms seem to have already erased the initial density footprint and follow some invariant measure. This particular property of chaotic dynamics, in which they tend to project almost any initial distribution onto an invariant measure (that contains significant information about the physical system properties), allows to effectively access these invariant measures by means of numerical simulations. Indeed, by sampling a state x_0 from a chosen initial distribution p_0, and evolving this state through a numerical simulation yields a state that is effectively sampled from an invariant measure of the system (cf. Eq. 2.11).

However the existence and uniqueness of invariant measures for systems like turbulence remain open problems (Ruelle, 1989). Multiple attractors can lead to *intermittent* behavior and dependency of the dynamics on its initial conditions (Frisch & Kolmogorov, 1995). In fact, flows transitioning from laminar to turbulent regime might exhibit hysteretic behavior (Nguyen et al., 2019). However, when considering fully developed turbulence, where the inertial range is large, universal properties seems to hold, suggesting there is *a universality of turbulence*. The extension of this universality to the ISM is evidenced in some situations Elmegreen and Scalo, 2004; Federrath, 2013; Heyer and Brunt, 2004; Padoan et al., 2014 but remains strongly questioned as many other physical processes interact with turbulence. In particular, some of these break the stationarity of the dynamics and might perturb by so the convergence towards an invariant measure. For instance, self-gravity in molecular clouds make the gas collapsing and

[4]For continuous (and differentiable) dynamical systems, the Poincaré–Bendixson theorem ensures that if the phase space is of dimension two, no behavior more complex than periodic orbits occurs. Of course, turbulence is far from having a bidimensional phase space, as discussed more in Sec. 3.1.

[5]Provided that this initial distribution is not a Dirac concentrated on a single state, so that it can diffuse as shown in Fig. 2.6.

by so continuously fills the high gas density tail of the gas density distribution (cf. Fig. 3.2). Extension to such cases is discussed in Sec. 1.2 of the next chapter.

2.3 Ergodicity

If we aim at computing an ensemble average $\mathbb{E}_{X\sim p^*}[\phi(X)]$ of a certain function ϕ, according to the invariant measure p^* of a dynamical system, we can, as explain previously, draw multiple samples $X_1, \dots, X_n$ independently from a distribution p_0 of our choice, evolve them by running n simulations, and finally perform the empirical average $\frac{1}{n}\sum_{i=1}^{n}\phi(X_i)$. Now, a surprising property which is inherent to the definition of a chaotic system, is that we can achieve the same result by *running only one simulation*, and averaging over the time evolution:

$$\text{for almost any } x_0 \quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}\phi(f^t(x_0)) = \mathbb{E}_{X\sim p^*}[\phi(X)], \tag{2.12}$$

the latter formula being in the case of the discrete time dynamics $x_t = f^t(x_0)$. This is referred to as the *ergodic property* of chaotic systems (Halmos, 2017). This has an important implication: we can access to statistical properties of the invariant measure using only one time evolution $\{f^t(x_0)\}_t$ of the system. As we shall see in Sec. 1.2, this property can be extended to other forms of symmetries of the dynamical system. In particular, for spatially ergodic systems, we can replace the time average in Eq. 2.12 by a spatial average $\langle\cdot\rangle_{\mathbf{r}}$ over $x_0(\mathbf{r})$. This means that we can *access to ensemble averages over p^* using only one (large) sample X_0* of the dynamics! We can see in Fig. 2.3 a manifestation of this spatial ergodicity, as the spatial empirical histograms of a same class tend to be very close (in a sense that will be defined later). This gives hope to infer properties of the ISM dynamics using only few observations, with however being aware of the caveats mentioned in the previous paragraph with regard to the universality of the turbulence and its interactions with other processes.

Conclusion

Hence, for the astrophysicist, chaos is much more a gift than a burden as it allows to grasp system properties from very few observations. However, as we shall now see, extracting information from observational data is not straightforward, because of its high dimensionality and non-Gaussianity.

3 Challenges

3.1 The curse of dimensionality

We have explained the interest of studying chaotic dynamical systems through the statistical properties of their invariant measures. We however have little knowledge about these measures. Proving their existence and uniqueness is in general a very hard problem, still open for Navier-Stokes equations for instance. Not surprisingly, we almost never have analytical formula, except for toy dynamical models such as the tent map. Nevertheless, such measures can be approxi-

mated with numerical simulations, by applying iteratively the dynamics on an initial condition as illustrated in Fig. 2.1. It is thus tempting to learn such distributions from the data simulated. Even though very promising, this data-driven approach is not straightforward because of its high dimensionality.

Indeed, for incompressible turbulence for example, the random object under study is the 3D velocity field, while for general MHD, one needs also to consider the density, pressure and magnetic field. These 3D scalar and vector fields are by nature high dimensional objects[6]. For instance, discretizing the 3D fluid density field as a piece-wise flat function onto 100^3 voxels leads to a $100^3 = 10^6$ dimensional random vector. Its associated probability distribution function is thus a $\mathbb{R}^{10^6} \to \mathbb{R}_+$ mapping. Learning such mapping by binning data samples directly in the data space, as illustrated in Fig. 2.7, is doomed to failure since the total number of bins $(L/\varepsilon)^d$ grows exponentially with the dimension d.

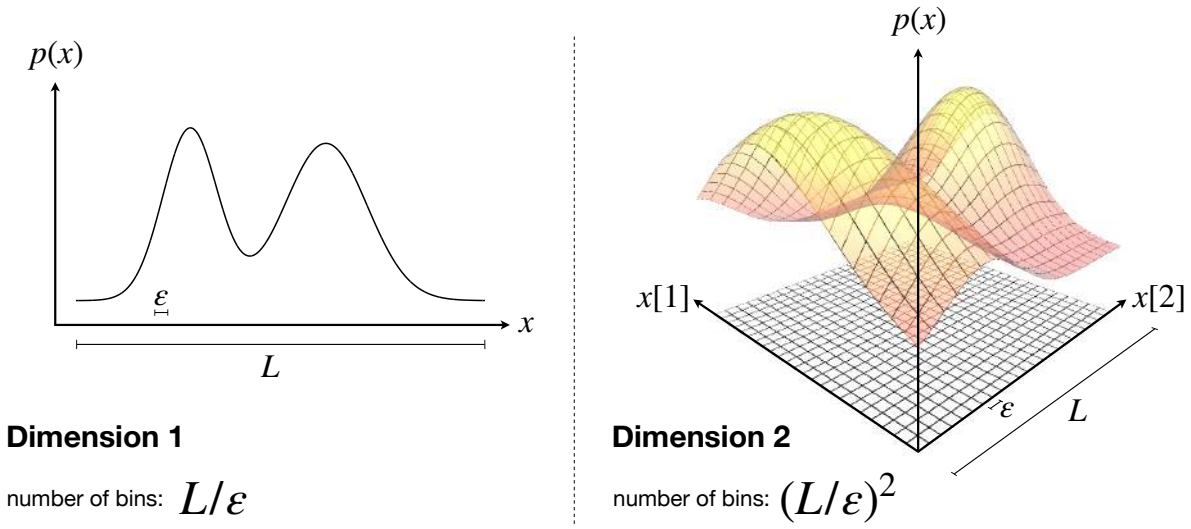


Figure 2.7: The curse of dimensionality: for a given number L/ε of intervals per axis, the total number of bins $(L/\varepsilon)^d$ grows exponentially with the dimension d.

3.2 The curse of multimodality

Despite being defined on high dimensional data space, some processes can be still easy to represent, learn, and sample. This is for instance the case of Gaussian white noise, defined as:

$$p(x_1, \dots, x_d) = \frac{1}{(2\pi)^{d/2}} e^{-d\langle x_i^2 \rangle_i / 2}. \tag{2.13}$$

Indeed, we can see in this case that the density p actually depends on a single scalar compression of the data: $\langle x_i^2 \rangle_i$, that we might call, in Chap. 4, *sufficient statistic*. Furthermore, the mapping $x \mapsto -\log p(x)$ is a *strictly convex* function. This means that the distribution has a unique

[6] Actually infinite dimensional if we forget about quantum formalism. In practice it does not matter as computer encoding of such fields is by far the cruder source of quantification.

local and global maximum, and has a "simple landscape" $x \mapsto p(x)$ in the data space. However, probability distributions generated by nonlinear dynamical systems such as turbulence have a much more complicated landscape, with many modes (i.e. many data elements x that are locally maximizing the density $p(x)$). Indeed, any x that does not represent a physical field will have probability 0. On the other hand, numerous different realisations x are expected to be likely produced by the dynamics. For instance, we have discussed that in the case of turbulence, the intermittency favors realizations that show dissipative structures (Fig. 1.7). To complicate this picture, each mode is not really the "summit" of a strictly concave "bump" of probability. Indeed, if the process has some invariant properties under a class of transformations $\{T_s\}_s$ such as translation or rotation invariance, i.e.,

$$\forall x \ \forall s \ p(T_s x) = p(x),$$

these modes are not isolated points x in the data space, but rather isolated low dimensional submanifolds generated by the equivalent class $\{T_s x\}_s$.

The statistical nature of ISM processes is therefore complex. However, depending on the goal and the framework, statistical tools can overcome, at least partially, this difficulty.

4 From scientific goals to statistical problems

4.1 Emphasis on the distinction between inference and modeling

We have seen that the ISM is made of high-dimensional non-Gaussian processes. This complex nature makes difficult the task of characterizing *ex nihilo* these objects. Indeed, if one wants to understand what makes ISM processes specific, one should be able to characterize the distinction between what is a given ISM process and what is not. Such characterization of a given process (that we assimilate to its probability distribution) p implies that one should be able to test, for any arbitrary process q, whether $p = q$ or not. When p is in high-dimension and is not restricted to belong to specific well-known families such as Gaussian processes, multivariate log-normal processes or Poisson point processes (as it is generally the case with turbulence) this task is very difficult.

However, it can be greatly simplified by restricting the set of test processes q, depending on the objective. For instance, as discussed in Sec. 4.2.2, for parameter estimation task among a given family of processes $\{p_\theta\}_\theta$, it is sufficient to restrict the set of test processes to that specific family: $q \in \{p_\theta\}_\theta$.

Then, an important point is to observe that the subset $\{p_\theta\}_\theta$ has, is in general a not so complicated geometry (cf. for instance Fig. 3.9 of the following chapter). Indeed, provided that the mapping $\theta \mapsto p_\theta$ is regular enough (as it is usually the case with parametric MHD simulations), then $\{p_\theta\}_\theta$ is actually a sub-manifold, with a dimension being at most the one of θ:

$$\dim T_{\theta_0}\{p_\theta\} = \operatorname{rank} \nabla_\theta \, p_\theta\big|_{\theta_0} \leq \dim \theta \ll \dim p_\theta, \tag{2.14}$$

where $T_{\theta_0}\{p_\theta\}$ stands for the tangent subspace of $\{p_\theta\}_\theta$ at $\theta = \theta_0$ and $\nabla_\theta\, p_\theta\big|_{\theta_0}$ for the Jacobian matrix of the mapping $\theta \mapsto p_\theta$ at θ_0. Such tasks, that only require *ăbintus* (from inside) characterization of the ISM submanifold are much simpler, as their scope is much more restricted. One can expect to achieve them with few statistical features, as illustrated in Fig. 2.8. However, these features are in general far from being sufficient when it comes to fully characterize a given ISM process, and even not necessarily the most important ones to consider. In other words, one can achieve very good accuracy results at discriminating ISM processes, without being able to model them.

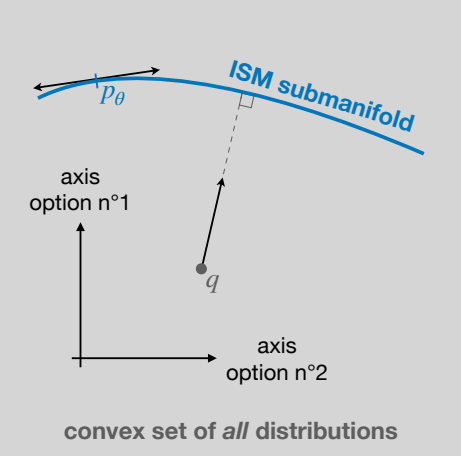


Figure 2.8: Difference between characterizing the submanifold of ISM processes *ex nihilo* (out of nothing) and *ăbintus* (from inside). In the first case, the most relevant axis[1] to choose would be the first option, as the discrepancy between an ISM process and an arbitrary process q is more often dominated by a vertical shift. But this axis, taken alone, is far from being sufficient to have a full characterization. In the second case, the most relevant axis would be the second option, as the discrepancy between two ISM processes is always dominated by an horizontal shift, and such axis is sufficient to identify exactly any process of the ISM submanifold.

[1]An example of such axis in the space of probability distributions is given in Sec. 1.1 of Chap. 3 in the context of moments.

In the following, we give a brief overview of the main inference problems encountered by the ISM community, starting by those where the inference can be postponed after strong dimensionality reduction (typically problems only relative to *ăbintus* characterization), before motivating situations where inference workflow strongly benefits from the deployment of high-dimensional generative models.

4.2 Inference requiring only partial characterization

Many inference problems involving high dimensional stochastic processes actually only necessitate to characterize them partially, as discussed below. This opens the door to great simplifications thanks to dimensionality reduction. These types of problems will be further studied in chapters 3, 4 and 5.

4.2.1 Classification

An extreme example of problem that requires only partial characterization is the classification one. It consists in attributing, to a given data sample x, the likeliest class from which it originates. For instance, Peek and Burkhart, 2019 used a Convolutional Neural Network (CNN) to retrieve the sub or super-Alfvénic nature of a simulation from the morphology of a 2D slice of the density field . Alternatively, Saydjari et al., 2021 used scattering transforms to classify eight different MHD simulations made with both varying Alfvénic and sonic Mach numbers.

In the classification framework, the data is assumed to be sampled from a set of densities $\{p_i\}_{i\in\mathcal{I}}$, where $i \in \mathcal{I}$ designates a class index among the finite set $\mathcal{I}$ of classes. Fig. 2.1 gives an example of such samples regrouped in two classes $\mathcal{I} = \{2, 6\}$, which are differentiated by their values of simulation parameters. We further assume here for simplicity that the data index I, seen as a random variable, is uniformly distributed over the classes, such that the joint density for the data is

$$p_{I,X}(i, x) = p_{X|I=i}(x) \cdot p_I(i) = p_i(x)/\#\mathcal{I},$$

where $\#\mathcal{I}$ is the number of classes. In this framework, the classification task then boils down to building a classifier $x \mapsto c(x) \in \mathcal{I}$ that maximizes the accuracy $\mathcal{A}(c)$ defined as the joint probability over (I, X) of having a correct classification $c(X) = I$:

$$\mathcal{A}(c) \equiv \mathbb{E}_{I,X\sim p_{I,X}}\left[1_{c(X)=I}\right]. \tag{2.15}$$

All the information required to optimally solve this problem holds in the knowledge of the regions $\{\mathcal{R}_j\}_{j\in\mathcal{I}}$ in the data space where each conditional probability density p_j dominates:

$$\mathcal{R}_j \equiv \{x \mid \forall i \in \mathcal{I} \;\; p_i(x) \leq p_j(x)\}.$$

Supplied with this knowledge only, one can build[7] an optimal classifier called *Bayes classifier*, no matter what the specific profiles of the densities inside each $\mathcal{R}_j$. Such classifier only depends on the probability densities through the sign of the quantity $\log p_i/p_j$, which is usually much easier to approximate than p_i. For instance, if we can decompose the probability densities p_i as

$$p_i(x) = r_i(x) \times h(x), \tag{2.18}$$

then the regularity of p_i/p_j will be the one inherited from r_i/r_j, no matter the irregularity level of h. However such decompositions are unlikely to be applicable as is for the processes we are typically interested in. We note that such focus on the probability ratio is not unusual in physics. For instance, in analogy with statistical physics, the quantity p_i/p_j resonates with the Boltzmann factor $e^{-(\varepsilon_i-\varepsilon_j)/k_BT}$, that depends on the energies of the states i and j only through their difference $\varepsilon_i - \varepsilon_j$. It also resonates with the *score function* that will be introduced in the

[7] Indeed, if we choose a classifier c^* such that for all x we have $x \in \mathcal{R}_{c^*(x)}$, then, for any other classifier c we have $p_{c(x)}(x) \leq p_{c^*(x)}(x)$, which translates as:

$$\sum_i p_i(x)1_{c(x)=i} = p_{c(x)}(x) \leq p_{c^*(x)}(x) = \sum_i p_i(x)1_{c^*(x)=i}.$$

Then, let us observe that the accuracy defined in Eq. 2.15 expands as:

$$\mathcal{A}(c) = \int \sum_i p_i(x)1_{c(x)=i}\,\mathrm{d}x/\#\mathcal{I}. \tag{2.16}$$

Using the previous inequality in this integral form of the accuracy yields:

$$0 \leq \mathcal{A}(c) \leq \mathcal{A}(c^*) \leq 1. \tag{2.17}$$

Hence c^* is an optimal classifier, and it can be constructed from the knowledge of $\{\mathcal{R}_j\}_{j\in\mathcal{I}}$ only.

following subsection for continuous parameter inference.

Relying only on the sign of $\log p_i/p_j$, classification tasks thus considerably reduce the level of process characterization required. They however remain difficult from a technical point of view when tackling structured fields as those encountered in the ISM. Indeed, their associated probability densities are still defined over a high dimensional data space (Sec. 3.1), have multiple modes (Sec. 3.2), and are highly intertwined from one density to another so each $\mathcal{R}_j$ has a complex geometry. Furthermore, the densities usually have overlapping supports so the frontier between classes is fuzzy and the maximum accuracy $\mathcal{A}^*$ reachable in theory is then lower than 100%. For instance, if we choose two simulations with very close physical parameters, we expect $\mathcal{A}^*$ not to excess much more than 50%. We give a theoretical value for $\mathcal{A}^*$ in terms of total variation distance between the densities in the binary case ($\#\mathcal{I} = 2$) in Sec. 4.3 of Chap. 4, but in practice, this achievable bound remains unknown for complex processes such as MHD simulations.

Still, in practice, for such ISM fields, classification seems to be a challenge almost fully solved: provided they are fed with a sufficient amount of training data, deep neural networks, such as CNNs or more recently transformers, can for instance define state-of-the-art accuracy levels that are impressively satisfactory. For instance, Peek and Burkhart, 2019 reached more than 98% accuracy for their binary sub/super-Alfvénic classification, while Saydjari et al., 2021 reach 97% on eight different simulation instances. For these examples, this means that the ratio $\mathcal{A}/\mathcal{A}^*$ between the accuracy of their classifier $\mathcal{A}$ and the maximum reachable accuracy $\mathcal{A}^*$ is higher than 97%, since the latter is 100% at most!

While the development of these tools allows for an extremely high discriminative power, we however emphasize that deep networks have a strong data appetite, while being also highly sensitive to data quality. This shifts the main inference bottleneck toward data quality and quantity. For instance, the training set of the network in Peek and Burkhart, 2019 is based, before the data augmentation procedure, on $\sim 10^5$ non overlapping 128×128 patches per class. Other approaches may require less data. For example, by limiting the learning scope to the one of a linear discriminant analysis performed over the Reduced Wavelet Scattering Transform (RWST) features (of dimension 164), Saydjari et al., 2021 achieved their very good classification results with ~ 700 patches (256×256) for each of the eight classes, that is ~ 35 times less total training samples than Peek and Burkhart, 2019 (albeit the two tasks are not directly comparable since they used different simulations, and Peek and Burkhart, 2019 used smaller patches and purposefully increased the difficulty by removing in some way one and two-point information). Still, attaining hundreds of independent cleaned patches rigorously drawn from the same process remains far from being realistic when it comes to observations of ISM components such as molecular clouds. The purpose of this work is to suggest alternative approaches that offer a better balanced trade-off between accuracy (and more generally informativeness) and sensitivity to data quality and quantity.

To do so, we for instance discuss in Sec. 1.2 of Chap. 3 how symmetries of the physical fields can be exploited to partially alleviate the lack of data. We then set a theoretical ground for the general problem of inference with unlabeled data in Chap. 4, before addressing, in Chap. 5, this

issue in a low data regime, that reflects a realistic consideration of the observational framework.

4.2.2 Parameter estimation

Parameter estimation is a direct extension of classification. It is one of the most ubiquitous task in natural sciences, including certainly astrophysics and cosmology. A great instance of such problem is the constraining of cosmological parameters. Fig. 2.9 illustrates how observations from the *Planck* satellite have tightened the constraints of on parameters of the ΛCDM model, and possible extensions.

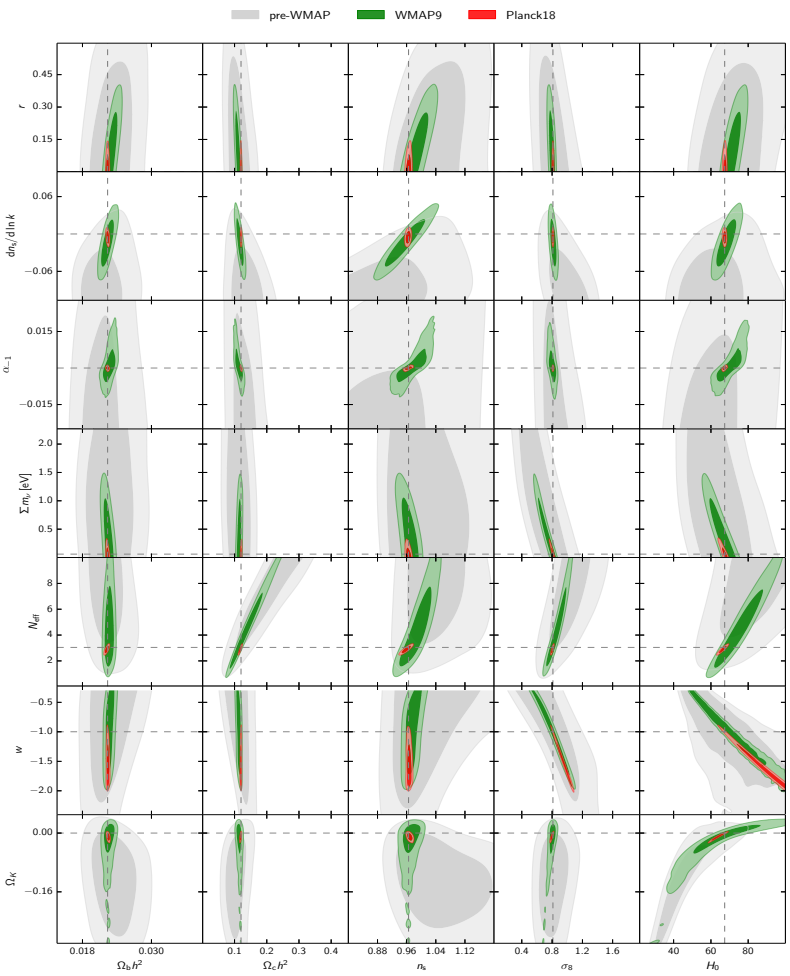


Figure 2.9: Successive constraints on cosmological parameters for various one-parameter extensions to ΛCDM, from pre-WMAP to *Planck* data.

Credits: Figure from *Planck* 2018 results (Planck Collaboration et al., 2020a).

In this type of problems, one observes some data x_0 (possibly a collection of samples), and wants to retrieve some properties θ on the process that generated this data. This is usually modelled by assuming the process to be stochastic, with a density $p : x \mapsto p(x)$ belonging to a certain parameterized family $p \in \{p_\theta\}_\theta$. Once the data x_0 is observed, the general objective called *inference* is then to reveal all the regions in the parameter space that are *compatible with the data* x, as illustrated in Fig. 2.9. The notion of compatibility for a given parameter candidate θ can be simply a sufficient *likelihood* with the data x_0: $p_\theta(x_0) > p_{crit}$, or further combined with an informative prior q on θ: $p_\theta(x) \times q(\theta) > \eta_{crit}$. If q represents a probability distribution for θ, this opens the path to Bayesian framework that allows to link the quantity $p_\theta(x_0) \times q(\theta)$ to the

probability of θ knowing x_0 called *posterior*:

$$p(\theta|x_0) \propto p(x_0|\theta) \times q(\theta) = p_\theta(x_0) \times q(\theta). \tag{2.19}$$

As already explained, numerous ISM components yield high dimensional non-Gaussian processes that make the likelihood $\theta \mapsto p_\theta(x_0)$ often intractable analytically. Instead, the inference procedure is usually operated by making a first overestimated guess of compatible regions by sampling θ according to an initial distribution with a broad support[8], and then tightening this guess by moving the unlikely samples toward a likeliest direction. The main ingredient of this inference procedure is the *score function*:

$$s(\theta_0) \equiv \nabla_\theta \log\, p_\theta(x)\big|_{\theta_0}, \tag{2.20}$$

that captures the dependence between θ and its associated likelihood to the observed data $p_\theta(x)$. This quantity plays a fundamental role in information extraction (Cranmer et al., 2020). It will be further discussed in the context of *sufficient statistics* and *Fisher information* in Sec. 2 of Chap. 3. It is noteworthy that the score is not affected by any irregularity purely based on x. Indeed, as already mentioned in the classification purpose (Eq. 2.18), if the density can be factorized as:

$$p_\theta(x) = r_\theta(x) \times h(x),$$

then:

$$\nabla_\theta \log\, p_\theta(x) = \nabla_\theta \log\, r_\theta(x).$$

In particular, this allows to transform the proportion relation of Eq. 2.19 into an exact equality: $\nabla_\theta \log p(\theta|x_0) = s(\theta) + \nabla_\theta \log q(\theta)$. This property is of further interest if r_θ is (more) regular (than p_θ) on θ and/or x. This resonates with the idea that the submanifold p_θ is of dimension lower than the one of θ (Eq. 2.14 and Fig. 2.8). In particular, as explained in Sec. 1.3 of Chap. 3, if we can write

$$r_\theta(x) = g_\theta(\phi(x)),$$

with $\phi(x)$ of smaller dimension than x, then the compression ϕ, called *summary statistic*, is qualified as *sufficient statistic* for the family $\{p_\theta\}_\theta$, as it does not lose any information on θ.

4.2.3 Statistical properties

As further detailed in Sec. 2 of the following chapter, instead of entirely characterizing the probability density of a given process $x \mapsto p(x)$, accessing to some projections of this distribution is already of great interest. For instance, measuring its mean, variance, power spectrum can help:

- getting acquainted with the process by unraveling some of its basic properties (typical values, power, typical scales),

[8] It should be broader than that of the target distribution. The initial distribution can be for instance q when such prior knowledge is available, or the least informative prior as possible otherwise.

- fostering the interaction between observation and theory (phenomenological laws to be explained, theoretical predictions to be observed),
- anticipating its interaction with other processes which is central in experimental design (signal to noise ratio, range of scales expected to be not contaminated, sensitivity required), and component separation techniques (cf. discussion of the additivity of the power spectrum in Sec. 2.2.1 of the next chapter).

4.3 High dimensional characterization and modeling

Most of the scientific questions in astrophysics and cosmology boil down to binary yes/no test or low dimensional parameter constraints. Indeed, rare are the situations where one actually cares more about a result expressed for a specific observation rather than the implications of this result on the theory. For instance, the CMB map in itself has low interest, but its observation allows to constrain its statistical properties that are in turn used for scientific challenges such as constraining cosmological parameters as shown in Fig. 2.9. This means that an inference end goal can almost always be expressed in low dimension. However, this does not mean that inference can also be carried out in low dimension because of the high dimensional nature of the observed fields. The previous section nevertheless shows that numerous instances of problems actually need only partial characterization of these fields. Yet, there remain situations in which making the effort to directly model the full density $x \mapsto p(x)$ seems to be the most promising strategy. These include:

- the task of estimating the similarity/distance between two processes that do not belong *a priori* to the same submanifold (cf. Fig. 4.1). It is for instance the case when we aim at assessing the overall proximity between some observations and simulations. This task strongly differs from the simpler one of parameter estimation that involves minimizing such distance (or maximizing a likelihood) without requiring to know its absolute value. This task also differs *a priori* from comparing only several moments that are only partial projections of the full densities.
- For inverse problems that involve forward operations in pixel space such as component separation $D = S + C$, having generative models of both distributions p_S and p_D is game changer in inverse problems (Cranmer et al., 2020).

5 Different frameworks

5.1 Nature of the processes : regularity, symmetries and diversity

Despite the fact that the ISM fields previously presented display non trivial and non smooth spatial properties such as power at small scales, intermittency, heavy tail in their one-point statistics, these fields still exhibit multiple assets due to their physical nature that make them much easier to characterize than classical datasets considered by the machine learning community (e.g., human pictures). These assets include:

- symmetries: as discussed in the following chapter, physical processes often have invariant properties which allow to compute empirical averages than benefit from concentration properties. Furthermore, some of these averages can be carried without any loss of information, such as the compression of the one-point distribution functions $\{PDF(X(\mathbf{r}))\}_{\mathbf{r}} \mapsto \langle PDF(X(\mathbf{r}))\rangle_{\mathbf{r}}$ in the case of translation invariance.

- regularity: provided that adequate representations are use, numerous functions defined on a continuous space can be very smooth and therefore compressed into few coefficients. For instance, in a log-log representation, a power spectrum that undergoes a power law turns into a simple straight line. This regularity can be used as a strong lever arm, due to its tight link with approximation and sparsity. See for instance Pr. Mallat's lecture[9] for deeper insights.

- diversity: as shown in Eq. 2.14 and depicted in Fig. 2.8, a physical process usually has few parameters and therefore yield a low dimensional parametric family of associated stochastic processes. If we work with data that is known to be generated by a process belonging to such family, inference problems can be formulated in simple forms that typically boil down to likelihood analysis as discussed in Sec. 4.2.2. However, when working with a diverse collection of processes that exhibit discrepancies over a wide range of dimensions, such as a combination of observations and simulations, inference problems usually have a much higher dimensional nature.

5.2 Available data: quantity and quality

A great difficulty in astrophysics and cosmology is that the observable sky is unique, and its structures are mostly static compared to human timescales. For signals that have large angular features (such as Galactic observations, large scales of the CMB, etc.), we have very few observable examples of these features. Even though they might come to us as big data packets (time series, hyper-spectral observations, extremely high angular resolution), we cannot resample these large scales. Hence for such observation signals, we are in a *low data regime*.

On the other hand, simulations offer, for astrophysicists, a unique way of experimenting, in a controlled manner, the complex ISM dynamics. New samples can be provided upon request. This computationally intensive approach comes however with an economic cost, a certain carbon footprint and archiving/storage problematics. Nevertheless, the deployment of ever growing computing capabilities has really opened a new way to progress in the understanding of these processes, and the scientific community extensively relies on them (Fig. 2.10). High fidelity simulations have outstandingly extended the class of available models that used to be restricted to analytically tractable ones.

[9] https://www.college-de-france.fr/fr/agenda/cours/representations-parcimonieuses/le-triangle-regularite-approximation-parcimonie

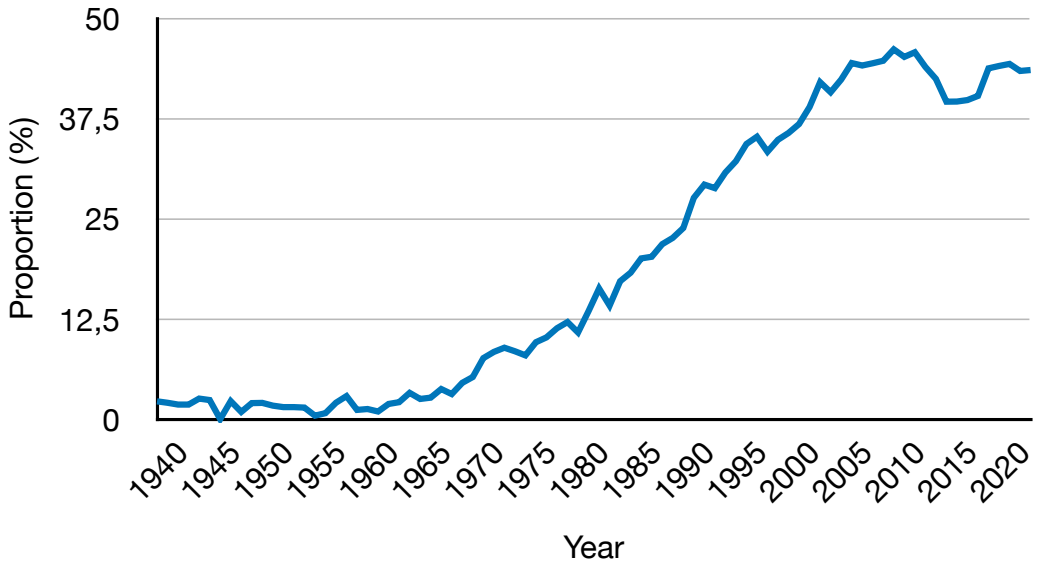


Figure 2.10: Proportion of papers including "numerical simulation" among the set of refereed papers including "turbulence" extracted from the SAO/NASA Astrophysics Data System (ADS).

Most recently, the incorporation of automatic differentiation and stochastic programming into simulation codes have initiated another paradigm shift in simulation-based inference (Cranmer et al., 2020), allowing to probe, with moderate additional computational cost and coding constraints, key ingredients for inference such as the score $\nabla_\theta \log p_\theta(x)$. See for instance the Φ_{Flow} differentiable simulation software Holl and Thuerey, 2024 and Thuerey et al., 2021 textbook.

However, in the context of the ISM, observations reflect a potentially wide panel of physical conditions that are difficult to precisely identify for each observed field, and are therefore very challenging to be reproduced by simulations. In the following chapter we discuss how statistics can be used in order to describe ISM observations and link them with some signatures of physical processes.

Chapter 3

Statistics as a descriptive tool

"The ISM presents astronomy's most visually complex field, and we are often limited by our ability to parse it".

Introduction of the 5^{th} edition of the Interstellar Institute: With Two Eyes,
Institut Pascal, Orsay, France, July 2022.

Objectives

Many tasks such as physical parameter inference or comparison between processes boil down to having suitable description of ISM processes. In this chapter, we show how summary statistics can be used to build such a low variance representation of physical processes endowed with enough invariance properties, and we review the main statistical diagnostics of turbulence and intermittency, with an emphasis on coupling between scales. We apply these tools to characterize, from observations, the evolving coupling between scales in column density maps of molecular clouds as they evolve from quiescent to active star forming.

Contents

1 Reductions: moments and summary statistics approach

As discussed in Sec. 3 of Chap. 2, numerous fields of interest in the ISM yield strong non-Gaussian probability distribution functions in high dimension, making the task of directly modelling them a great challenge. On the other hand, as discussed in Sec. 5.1 of Chap. 2, such processes exhibit strong regularities and multiple sources of invariance. In this section we present how to exploit such properties as a lever arm to construct low variance representations from a few high dimensional observations.

1.1 Moments

1.1.1 Definition and interpretation

In the following, we consider a probability distribution $p : x \in \mathbb{R}^d \mapsto p(x) \in \mathbb{R}_+$ defined on a high-dimensional space ($d \gg 1$), that typically corresponds to the modeling of an ISM process. We write $X \sim p$ an associated random variable, i.e., X has density p. Let us also consider a given function $\phi : x \in \mathbb{R}^d \to \mathbb{R}$ defined in the same space as p, which reduces the high dimensional field x into a scalar $\phi(x)$. Provided that p and ϕ have sufficient regularity, we can consider the following quantity:

$$\mathbb{E}_{X\sim p}\,\phi(X) \equiv \int p(x)\phi(x)\,\mathrm{d}x, \tag{3.1}$$

called a *moment* of the distribution p. It is a deterministic scalar value that depends on p. Indeed, let us emphasize that the mapping $(p, \phi) \mapsto \int p(x)\phi(x)\,\mathrm{d}x$ is well-known to act as a *scalar product* between p and ϕ (both seen as vectors of the space of functions $\{f \mid f : \mathbb{R}^d \to \mathbb{R}\}$). In particular, the moment value $\mathbb{E}_{X\sim p}\,\phi(X)$ is linear in p and can be seen as the coordinate $\langle p|\phi\rangle_x$ of the projection of p along the axis spanned by the vector ϕ. Hence, a certain choice of ϕ supplied with the projection $p \mapsto \mathbb{E}_{X\sim p}\,\phi(X)$, is an example of axis that could be considered in the context of Fig. 2.8. Note that, in the following, ϕ can be multidimensional to account for a collection of moments.

1.1.2 The moment problem

Moments form informative projections that can be used to characterize a process. The problem of retrieving p through the measurements of some prescribed set of moments is referred to as the *moment problem* (Schmüdgen et al., 2017).

In the context of parameter estimation $\{p_\theta\}$, if the projection $\theta \mapsto \mathbb{E}_{X\sim p_\theta}\phi(X)$ is injective, it is possible to entirely characterize θ directly from the moments. Indeed, if $g : \theta \mapsto \mathbb{E}_{X\sim p_\theta}\phi(X)$ is known, it can in practice be inverted by enforcing the so called *moment condition* $g(\theta) = \mu$, using the inverse function:

$$g^{-1} : \mu \mapsto \mathrm{argmin}_\theta \, \|g(\theta) - \mu\|_2^2.$$

This approach was introduced by Hansen, 1982 as the *Generalized Method of Moments*.

For instance, in the unidimensional case ($d = 1$), the knowledge of the standard set of moments

$$\{\mu_n \equiv \mathbb{E}_{X\sim p}\,[X^n] \,|\, n \in \mathbb{N}\},$$

that are associated to the family of functions $\phi_n : x \mapsto x^n$, uniquely determines p if it satisfies the sufficient condition of Carleman (1922) (Akhiezer, 2020):

$$\sum_{n=0}^{+\infty} \mu_{2n}^{-\frac{1}{2n}} = +\infty. \tag{3.2}$$

In this case, p can then be directly retrieved through Carleman's reconstruction formula based on harmonic analysis Akhiezer, 2020.

Moments can thus be of great help to characterize a process, but their analytical computation is often intractable. Indeed, computing them from Eq. 3.1 requires 1) to know the value $p(x)$ for every x and 2) to solve an integral in dimension d. These two limitations render this formula highly inoperable in practice, in particular due to constraints on the data availability. In this chapter, we aim at building a description of p that precisely avoids the very difficult task of implementing a high-dimensional model of p: this leads to the topic of moment estimation. As explained in the following subsection, when p benefits from suitable invariance properties, moments can in particular be estimated from a limited amount of data samples. This trick makes moment estimation an appealing tool to characterize physical processes that are usually endowed with strong invariance properties.

1.2 Symmetries as a lever arm in a low data regime

If we are given n samples $X_1, \ldots, X_n$ of p, we can estimate the moment $\mathbb{E}_{X\sim p}\,\phi(X)$ by the following quantity:

$$\hat{\mu} \equiv \langle \phi(X_i) \rangle_{1\leq i\leq n}. \tag{3.3}$$

Indeed, it is unbiased, by linearity of $\mathbb{E}$. If the samples are furthermore uncorrelated, one has

$$\operatorname{Var} \hat{\mu} = \frac{\operatorname{Var} \phi(X)}{n}. \tag{3.4}$$

This formula shows that to mitigate the variance of this estimator there are two options:

- increase the number n of samples,
- control the variance of $\phi(X)$, for $X \sim p$.

As discussed in Sec. 5.2, the first option is challenging or simply impossible in many cases in observational astrophysics. Hence, in these cases where we face a low data regime, the second option is of great importance.

1.2.1 Reducing the variance

First, let us emphasize that controlling the variance of $\phi(X)$ restricts in practice the usage of the standard moments $\{x \mapsto x^n \,|\, n \in \mathbb{N}\}$, although they may be analytically handy and benefit from deep theoretical understanding as discussed above, to the ones of very low order. Further discussions on this topic in the case of turbulence may be found in Sec. 2.
To control the variance of $\phi(X)$ with $X \sim p$, we can use the invariance properties of p as a lever arm. Indeed, let us assume that p is left unchanged under a class of transformations $\{T_s : x \in \mathbb{R}^d \mapsto T_s x \in \mathbb{R}^d\}_{s\in\mathcal{S}}$ (that is usually endowed with a group structure (Allys, 2017)), i.e.:

$$\forall x \, \forall s \;\; p(x) = p(T_s x).$$

Then, it is advantageous to replace ϕ by

$$\phi_{av}(x) \equiv \langle \phi(T_s x) \rangle_s, \tag{3.5}$$

where $\langle\cdot\rangle_s$ denotes an averaging operator[1] naturally provided over $\mathcal{S}$. Indeed, by invariance of p under T, the random variables $\{T_s X\}_s$ are identically distributed according to p. Hence Eq. 3.5 is analagous to Eq. 3.3 except that it is only based on one input sample X from p. It follows that ϕ and ϕ_{av} share the same moment value:

$$\mathbb{E}_{X\sim p} \, \phi_{av}(X) = \mathbb{E}_{X\sim p} \, \phi(X),$$

but ϕ_{av} benefits from an averaging effect that makes it invariant[2] under $\{T_s\}_s$:

$$\phi_{av}(T_s x) \equiv \langle \phi(T_r T_s x) \rangle_r = \langle \phi(T_r x) \rangle_r \equiv \phi_{av}(x),$$

and reduces its variance:

$$\begin{aligned} \operatorname{Var} \phi_{av}(X) &\equiv \operatorname{Cov}[\langle \phi(T_r X) \rangle_r, \langle \phi(T_s X) \rangle_s] \\ &= \big\langle \operatorname{Cov}[\phi(T_r X), \phi(T_s X)] \big\rangle_{r,s} \leq \operatorname{Var} \phi(X). \end{aligned}$$

Indeed, the equality comes from the bilinearity of the covariance operator and the last inequality comes from the covariance inequality $\operatorname{Cov}[X, Y] \leq \sqrt{\operatorname{Var} X \cdot \operatorname{Var} Y}$, as well as that $\phi(T_r X)$ and $\phi(T_s X)$ are equally distributed (to $\phi(X) \sim \phi_{\#} p$). This covariance inequality however corresponds to the worst case where the variables $\{T_s X\}_s$ are maximally correlated, i.e., all being equal in this case. In practice, the variance of $\phi_{av}(X)$ can be greatly tightened if there is only a limited correlation between these variables. For example, in the extreme case where they are uncorrelated, this translates, if $\mathcal{S}$ is discrete, to:

$$\forall r \neq s \;\; \operatorname{Cov}[\phi(T_r X), \phi(T_s X)] = 0 \quad \Longrightarrow \quad \operatorname{Var} \phi_{av}(X) = \frac{1}{\#\mathcal{S}} \operatorname{Var} \phi(X). \tag{3.6}$$

[1] Such as $\frac{1}{\#\mathcal{S}} \sum_s$ if $\mathcal{S}$ is finite, or $\frac{1}{\operatorname{Vol}(\mathcal{S})} \int \cdot \, \mathrm{d}s$ if it is continuous with finite volume.
[2] Provided that the mapping $T_r \mapsto T_s T_r$ leaves invariant the measure used to define the operator $\langle\cdot\rangle_s$.

This formula is the analogous to Eq. 3.4 with the number of input samples n being effectively superseded by $\#\mathcal{S}$, while only one[3] input sample X is being used here. Hence, the uncorrelatedness between the multiple transformations $\{T_s X\}_s$ of the unique sample X allows to mimic a *data augmentation* effect.
While this uncorrelated case is indeed rarely met, we discuss below how this approach can be adapted to a correlation with a limited spatial range between these variables.

1.2.2 Autocorrelation length: example with translation invariance

Let us give an example for a translation invariant process $X(\mathbf{r}) \sim p$. We assume that the stochastic field $X(\mathbf{r})$ is defined over a bounded space of dimension m: $\mathbf{r} \in [0, L]^m$ (e.g., $m = 2$ for images). In this case, the class of transformations is the group of translations:

$$\{T_\tau : x(\mathbf{r}) \mapsto x(\mathbf{r} - \tau)\}_{\tau \in]-L/2, L/2]^m}.$$

The translations are set to be periodic, so each component of the translated vector $\mathbf{r} - \tau$ has to be taken modulo L. Let us focus on the choice of local operator $\phi : x \mapsto x(\mathbf{0})$, whose moment probes the mean value of the one-point distribution of p (the latter does not depend on the point position since p is translation invariant). Then ϕ_{av} corresponds to the empirical mean of a field $\phi_{av}(x) = \langle x(\mathbf{r}) \rangle_{\mathbf{r}}$ and, using the translation invariance of p, its variance derives as:

$$\begin{aligned}
\operatorname{Var} \phi_{av}(X) &= \big\langle \operatorname{Cov}[\phi(T_{\tau_1} X), \phi(T_{\tau_2} X)] \big\rangle_{\tau_1, \tau_2} \\
&\equiv \big\langle \operatorname{Cov}[X(\tau_1), X(\tau_2)] \big\rangle_{\tau_1, \tau_2} \\
&= \big\langle \operatorname{Cov}[X(\mathbf{0}), X(\tau_2 - \tau_1)] \big\rangle_{\tau_1, \tau_2}.
\end{aligned}$$

To simplify, we assume that the process has an isotropic covariance, so the quantity $\big\langle \operatorname{Cov}[X(\mathbf{0}), X(\tau_2 - \tau_1)] \big\rangle_{\tau_1, \tau_2}$ only depends on $\delta \equiv \|(\tau_2 - \tau_1) \bmod \mathbf{L}\|_2$ the distance between τ_1 and τ_2:

$$\operatorname{Cov}[X(\mathbf{0}), X(\tau_2 - \tau_1)] \equiv \mathcal{C}(\delta).$$

Using the periodic boundary conditions, we then have:

$$\langle \mathcal{C}(\delta) \rangle_{\tau_1, \tau_2 \in]-L/2, L/2]^m} = \langle \mathcal{C}(\delta) \rangle_{\tau_2 - \tau_1 \in]-L/2, L/2]^m}.$$

For simplicity we restrict the latter integral defined on the hypercube $]-L/2, L/2]^m$ to the ball of radius $L/2$ centered at $\mathbf{0}$ and compute it over spheres of radius δ ranging from 0 to $L/2$ with area $S_{m-1}\delta^{m-1}$:

$$\langle \mathcal{C}(\delta) \rangle_{\tau_2 - \tau_1 \in]-L/2, L/2]^m} = \frac{1}{L^m} \int_0^{L/2} \mathcal{C}(\delta) \times S_{m-1} \delta^{m-1} \, \mathrm{d}\delta,$$

[3] Of course we can extend ϕ_{av} to operate with n input samples as $\langle \phi(T_s X_i) \rangle_{s, 1 \le i \le n}$ and obtain $\operatorname{Var} \phi_{av}(X) = \frac{1}{n \times \#\mathcal{S}} \operatorname{Var} \phi(X)$, assuming the samples X_i are *pair-wise independent*.

and finally obtain:

$$\operatorname{Var}\phi_{av}(X) = \frac{S_{m-1}}{L^m}\int_0^{L/2} \mathcal{C}(\delta) \times \delta^{m-1}\,\mathrm{d}\delta. \tag{3.7}$$

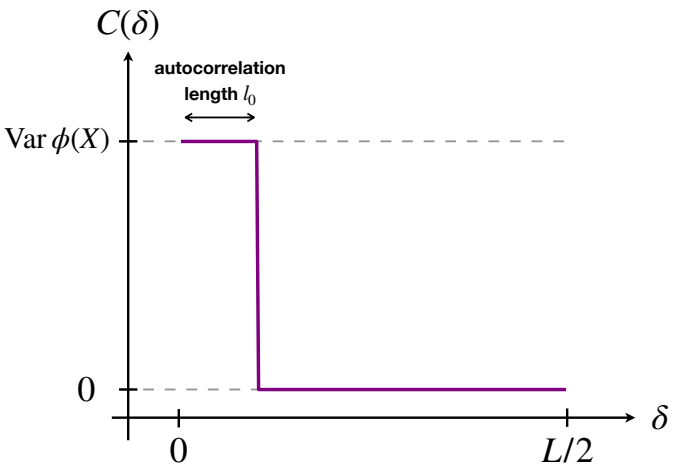
Then, if we take a very simplistic toy model[4] for the autocorrelation function (Fig. 3.1 (a)):

$$\mathcal{C}(\delta) \simeq \begin{cases} \operatorname{Var}\phi(X) & \text{if } 0 \leq \delta \leq l_0 \\ 0 & \text{if } l_0 < \delta \leq L/2 \end{cases}, \tag{3.8}$$

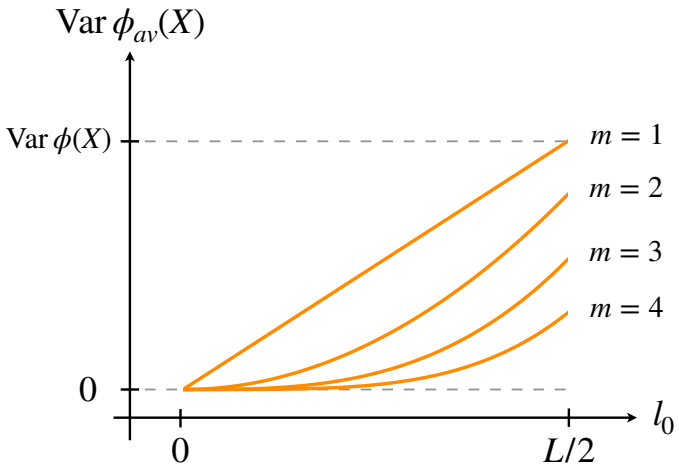
where l_0 plays the role of an *autocorrelation length* (also called *integral length scale* in turbulence), we obtain using Eq. 3.7 the effective variance reduction:

$$\operatorname{Var}\phi_{av}(X) \simeq V_m \left(\frac{l_0}{L}\right)^m \operatorname{Var}\phi(X), \tag{3.9}$$

where $V_m \equiv S_{m-1}/m$ is the volume of the unit ball in dimension m. This example is reported in Fig. 3.1 (b) where the variance of ϕ_{av} is plotted as a function of l_0 for different values of m.



(a) Very simplistic model of Eq. 3.8 for the autocorrelation $\mathcal{C}$, parameterized by l_0.



(b) Dependence of the variance of ϕ_{av} on the autocorrelation length l_0, for various dimensions m, according to Eq. 3.9.

Figure 3.1: (b) Influence of the autocorrelation length l_0 on the variance of ϕ_{av} for various dimensions m, under the simplistic model of Eq. 3.8 reported in (a).

Hence, similarly to the uncorrelated and finite $\mathcal{S}$ case of Eq. 3.6, we obtain a variance reduction that encompasses the ratio between the integrated correlation level $V_m l_0^m$ and the overall volume of $\mathcal{S}$: L^m. For a given ratio l_0/L, the variance reduction is more than[5] exponentially amplified with the space dimension m. In the typical case of images ($m = 2$, $V_m = \pi$), the effective data augmentation factor yielded by translation invariance is quadratic in the ratio L/l_0. In general, l_0 is fixed by the studied physical process. Hence, in the observational context, we might be interested by maximizing the field of view L. However, as explained in the following,

[4] By definition any such model should verify: $|\mathcal{C}(\delta)| \leq \operatorname{Var}\phi(X)$, and $\mathcal{C}(0) = \operatorname{Var}\phi(X)$.

[5] Indeed $V_m \sim m^{-\frac{m+1}{2}}$.

this choice is limited by non stationarity, which is an ubiquitous challenge in the ISM context.

1.2.3 Local symmetry and stationarity length

Beyond translation invariance, physical processes in the ISM may exhibit further symmetries, namely rotation invariance, scale invariance, time invariance (mostly useful in simulations) or space and time reversal. However, these symmetries are in general satisfied only approximately.

To mitigate this effect, we can first remark that the results derived previously do not involve p directly but only $\phi_{\#}p$, the distribution of $\phi(X)$ for $X \sim p$. Hence, we can actually supersede the symmetry condition $p \circ T_s = p$ by the weaker one $\phi_{\#}p \circ T_s = \phi_{\#}p$ without affecting the previous conclusions.

To illustrate this nuance, let us consider that we are interested in a signal S that is contaminated by C such that we observe $X = S + C$. If S is a translation invariant field, but not C, then we do not have in general $p(T_s x) = p(x)$. Nevertheless, if C has negligible power on some scale range, then for any filter ψ whose Fourier support is included in that range we approximately have $\psi \star C = 0$, so $\psi \star X = \psi \star S$. Then, any function of the quantity $\psi \star X$ will lead to a translation invariant distribution:

$$\forall \phi \ \forall \tau \quad \phi(\psi \star T_\tau X) \sim \phi(\psi \star X).$$

For instance, we can take $\phi \equiv |\cdot|^2$ and derive from it the translation invariant representation:

$$\phi_{av}(x) \equiv \left\langle |\psi \star x|^2 \right\rangle_{\mathbf{r}}.$$

A second way to cope with non globally satisfied translation invariance is to reduce the spatial window on which the process is defined. Indeed, in many situations, p might be affected by large scale translations and so failing at satisfying global translation invariance. However, we shall assume that the deformation of p under translation is progressive, i.e., that despite being not entirely constant, the mapping $\tau \mapsto p \circ T_\tau$ is smooth. This means that we can find a certain length L_0, called *stationarity length*, below which p is invariant to a good approximation. Considering the ball $B \equiv \{\tau \mid \|\tau\|_2 \leq L_0/2\}$ of diameter L_0, this translates informally as:

$$\tau \in B \implies p \circ T_\tau \simeq p. \tag{3.10}$$

This ball provides a window $w : x \mapsto \{x(\mathbf{r})\}_{\mathbf{r} \in B}$ from which any performed computation $x \mapsto \phi(w(x))$ allows for a spatial averaging under approximately satisfied translation invariance:

$$\phi_{av}(x) = \langle \phi(w(T_\tau x)) \rangle_{\tau \in B},$$

and therefore benefits from the variance reduction discussed in Eq. 3.9 if the correlation length is smaller than the stationarity length.

Let us emphasize that the smoothness assumption of $\tau \mapsto p \circ T_\tau$ is in no way incompatible

with having p that generates very sharp/edgy-like samples. For instance, an homogeneous Poisson point process generates samples with highly irregular spatial fluctuations, while being translation invariant. As another example, we can observe some ISM regions that contain a great number of sharp filaments and points sources, but look statistically homogeneous.

1.2.4 Summary

In summary, provided that we consider a function ϕ with reasonable[6] variance $\operatorname{Var}\phi(X)$ under p, we can hope to estimate its moment $\mathbb{E}_{X\sim p}\,\phi(X)$ from observations, even in an extremely low data regime. This may happen if p is invariant under a class of transformations $\{x \to T_s x\}_s$ that can be used to build an averaged version ϕ_{av} of ϕ. If p, ϕ and $\{T_s\}_s$ are such that, for a sample $X \sim p$, the augmented set of random variables $\{\phi(T_s X)\}_{s\in\mathcal{S}}$ has a large proportion of pairs with low correlation, which is the case when their autocorrelation length is small compared to the typical radius of $\mathcal{S}$, then ϕ_{av} indeed benefits from a variance reduction (Eq. 3.9). We also discussed that the invariance property of $X \sim p$ can be relaxed at least in two ways: 1) the distribution equality $T_s X \sim X$ can be replaced by $\phi(T_s X) \sim \phi(X)$. 2) Translation invariance (which is of great importance as spatial averaging $\langle\cdot\rangle_{\mathbf{r}}$ is extensively used to reduce the data) may not hold globally, but can be relaxed to a local condition by introducing a stationarity length L_0 of p, which allows for a local averaging that benefits from the effects of translation invariance. The definition of such a stationarity length is discussed in the context of molecular clouds in Chap. 5, but we recognize that it can be very difficult to estimate in practice.

1.3 Summary statistics

As seen previously, the moment $\mathbb{E}_{X\sim p}\,\phi(X)$ of a function $\phi : x \in \mathbb{R}^d \to \mathbb{R}$ under p is a deterministic and reduced description of p. Indeed, it represents the projection $\langle p|\phi\rangle_x$ of p onto the axis spanned by ϕ. We explained how to estimate this integral in practice using only a few samples, which is of great interest when we do not have access to the quantity $p(x)$. We showed in particular how to construct an estimator from an average over these samples (Eq. 3.3), whose variance can be reduced by using the symmetries of p if they exist (Eq. 3.5). It is noteworthy that such estimators access the data only through the preliminary application of ϕ, so the characterization of the process $X \sim p$ is *de facto* forced to take place upon the reduced process $\varphi \equiv \phi(X)$, which is a low dimensional random variable of density $\phi_{\#}p$ (unidimensional as introduced here for simplicity). We may now establish a connection with the *summary statistics* approach, which brings another perspective of dimensionality reduction.

1.3.1 A tool against high dimensional non-Gaussianity

Summary statistics are in essence a dimensionality reduction. Indeed, this approach aims at transforming the random variable X into another one $\phi(X)$, of much lower dimensionality. The mapping ϕ in this case is called *summary statistics*[7]. It is chosen in order to obtain a distribution

[6] What we precisely mean by reasonable might depends on the objective. We give a clear definition in the context of comparing pairs of processes with limited data in Chap. 5.

[7] In general ϕ is multivariate and is therefore considered as a collection of plural statistics.

$\phi_{\#}p$ of $\phi(X)$ simplified compared to the high dimensional distribution p that follows X, in order to allow for practical inference algorithms and models to operate (Cranmer et al., 2020). In general this means having either an analytical model for $\phi_{\#}p$, or a practical way to estimate it with the available data and models[8].

For instance, as it will be further discussed in the following subsection (Sec. 2.1), the one-point distribution of the gas density in numerous forms of turbulence is known to be a log-normal. Hence, for such a turbulent process p, choosing $\phi : x \mapsto x(\mathbf{0})$ (as already done in Sec. 1.2.2), provides an convenient analytical formula for the distribution of $\phi(X)$:

$$\phi(X) \sim \phi_{\#}p = \log \mathcal{N}(\mu, \sigma^2), \tag{3.11}$$

where the parameters μ and σ^2 can be related to physical parameters (cf. Sec. 2.1). The equation 3.11 can then be used as a likelihood for inference purpose.

1.3.2 Symmetries, Gaussianization and concentration

In this quest for simplifying the distribution p of X through summary statistics compression, the use of symmetries is of great help. Indeed, constructing a set of summary statistics as an average over transformations $\phi_{av}(x) \equiv \langle \phi(T_s x) \rangle_s$ of x (as introduced in Sec. 1.2.1) allows the distribution of $\phi_{av}(X)$ to benefit[9] from *Gaussianization* through the central limit theorem:

$$\phi_{av}(X) \sim \phi_{av\#}p \simeq \mathcal{N}(\mu, \sigma^2), \tag{3.12}$$

with μ and σ^2 being respectively the expected value and variance of $\phi_{av}(X)$. This Gaussianization of the compressed data distribution overcomes the curse of non-Gaussianity depicted in Sec. 3.2, but to the prize of a compression which usually comes with a loss of information, that will be addressed in Chap. 4. Concretely, this Gaussianization property is very convenient in practice, and paves the way to a wide range of statistical tools (even in if ϕ is multivariate).

Note that a link can be made between the summary statistics approach, that compresses X into the random variable $\phi(X)$ (we now drop the ϕ_{av} notation), and the moment approach, that computes the deterministic quantity $\mathbb{E}_{X\sim p}\, \phi(X)$ that was introduced as the scalar product $\langle p|\phi \rangle_x$ between $x \mapsto p(x)$ and the mapping $x \mapsto \phi(x)$. Indeed, the center of mass μ of the reduced distribution $\phi_{\#}p$ (introduced in Eq. 3.12) is precisely[10] the moment value $\langle p|\phi \rangle_x$.

[8] For instance if we have a surrogate model with poor quality, we might choose a low dimensional ϕ selecting only features well reproduced by the model to mitigate its bias. Or, when having only few observation data, choosing ϕ such that the distribution of $\phi(X)$ has some regularity (log-concavity or few modes) and is in sufficiently low dimension to be reliably learnt by techniques such as kernel density estimation.

[9] For instance if the variables $\{T_s X\}_s$ are sufficiently uncorrelated, as discussed in Sec. 1.2.2.

[10] Indeed, by the *law of the unconscious statistician* (LOTUS), we have:

$$\mathbb{E}_{X\sim p}\left[\phi(X)\right] = \mathbb{E}_{\varphi\sim\phi_{\#}p}\left[\varphi\right], \tag{3.13}$$

where the left hand side of Eq. 3.13 was introduced in Eq. 3.1 as the low dimensional reduction $\langle p|\phi \rangle_x$ between the two mappings p and ϕ, both acting on the high dimensional space $\mathbb{R}^d$. On the other hand, the right hand side brings into play the reduced distribution $\phi_{\#}p : \mathbb{R} \to \mathbb{R}$ and the identity mapping $\mathrm{id} : \varphi \in \mathbb{R} \mapsto \varphi \in \mathbb{R}$ both

This shows that moment estimation, which aims at reducing p through a deterministic projection $\langle p|\phi\rangle_x$, is a particular case of the summary statistics framework that aims at retrieving the distribution $\phi_{\#}p$ of the compressed process $\phi(X)$. Let us remark that the smaller is $\sigma^2 \equiv \operatorname{Var}\phi(X)$, the better is the moment estimator (Eq. 3.4), and in the mean time, the more concentrated around μ is $\phi_{\#}p$. In the limit $\sigma^2 \to 0$, both approaches are equivalent, in the sense that the results of both are uniquely driven by $\mu = \langle p|\phi\rangle_x$. This situation typically occurs when we consider a translation invariant process X defined on a space of size L that we can arbitrarily increase, while its correlation length l_0 remains fixed and finite. Indeed, thanks to Eq. 3.9, $\sigma^2 \simeq (l_0/L)^m \to_{L\to\infty} 0$ and therefore the distribution $\phi_{\#}p$ concentrates around its expected value μ. This phenomenon is at the heart of the Asymptotic Equipartition Property (Cover & Thomas, 2006) and the Boltzmann Equivalence Principle (Bruna & Mallat, 2019) that are central properties in information theory and statistical physics.

2 Usual statistical diagnostics of turbulence and intermittency

We now review usual summary statistics used by the ISM community. We consider here that the field $X(\mathbf{r})$ is real-valued, defined over a discrete lattice of $\{\mathbf{r}_i\}_{1\leq i\leq d}$ made of d pixels. The translations over this lattice are set to be periodic.

2.1 One-point statistics

The one-point distribution $p_{X(\mathbf{r}_i)}$ of X at point $\mathbf{r}_i$ is defined as:

$$p_{X(\mathbf{r}_i)}(x) \equiv \int_{\mathbb{R}^{d-1}} p(x'_1,\dots,x'_{i-1},x,x'_{i+1},\dots,x'_d)\,\mathrm{d}x'_1\cdots\mathrm{d}x'_{i-1}\mathrm{d}x'_{i+1}\cdots\mathrm{d}x'_d. \tag{3.14}$$

This marginalization induce a strong dimensionality reduction: $p_{X(\mathbf{r}_i)}$ is a univariate distribution. If X is furthermore translation invariant, then its one-point distributions are all equal, usually referred to as *the* one-point distribution, or *Probability Distribution Function* (PDF) (that should not be confused the probability distribution function p of the complete field X). When in the following we mention "the PDF" of a field, we implicitly assume that the field is translation invariant. As already mentioned in Sec. 1.3.1, the PDF of X can be seen as the distribution $\phi_{\#}p$ of the reduced (to one point) random variable $\phi(X)$ through the summary statistic $\phi : x \mapsto x(\mathbf{r}_1)$. The knowledge of this distribution *entirely* captures the one-point properties of X. However, the representation $\{PDF(x)\}_{x\in\mathbb{R}}$ is *a priori* of infinite[11] dimension. Therefore it is often further reduced to quantities such as its expected value, variance, mode(s), median and other quantiles, or binned representation.

belonging to the much smaller vector space of univariate mappings: $\{f \mid f : \mathbb{R} \to \mathbb{R}\}$. In other words:

$$\langle p|\phi\rangle_{x\in\mathbb{R}^d} = \langle \phi_{\#}p|\,\mathrm{id}\rangle_{\varphi\in\mathbb{R}}.$$

[11] Even if the field X is valued in a bounded range $[x_{\min}, x_{\max}]$.

For instance, Vazquez-Semadeni, 1994 suggested that the PDF of the gas density in a supersonic hydrodynamical flow without self-gravity tends to be log-normal as the manifestation of the central limit theorem applied to a multiplicative hierarchical process. This statistical property has been found to hold for multiple forms of turbulent flows (see Padoan et al., 2014 for a review). More precisely, the gas density field ρ normalized[12] by its mean density ρ_0 is such that:

$$\rho/\rho_0 \sim \log \mathcal{N}(-\sigma^2/2, \sigma^2). \tag{3.15}$$

In specific situations, the dispersion σ^2 can be analytically related to some key parameters such as the Mach number $\mathcal{M}_s$, the ratio between compressional and compressional+solenoidal turbulent forcing b, and the ratio between thermal to magnetic[13] pressure β (Padoan & Nordlund, 2011):

$$\sigma^2 = \log\left[1 + b^2\mathcal{M}_s^2\beta/(\beta+1)\right]. \tag{3.16}$$

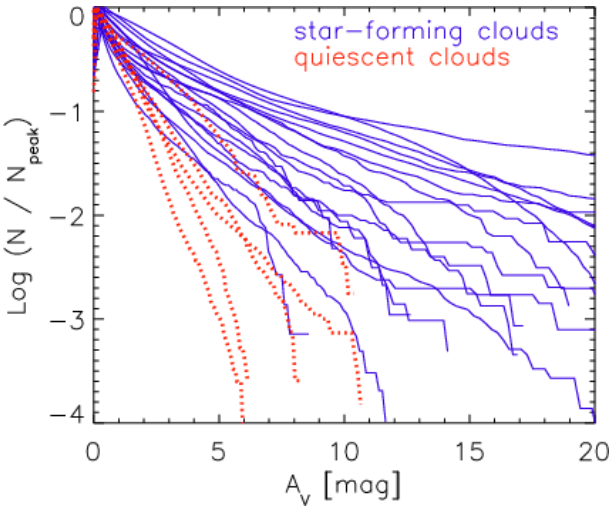
Column density maps usually show a log-normal PDF too. An example from a snapshot of MHD simulation is reported in Fig. 3.5. As shown in Fig. 3.2, this log-normal distribution is further observed in column density maps of quiescent clouds (although often contaminated at low densities by the CIB.). However, for denser molecular clouds that undergo local gravitational collapse, over-densities are amplified and so the gas PDF develops a power law tail[14] (cf. Fig. 3.2). In fact, the PDF of gas density in molecular clouds plays a major role in star formation, see Hennebelle and Falgarone, 2012 for a review.

However, one-point statistics remain a partial diagnostic of the complete process p that is d-dimensional since defined over d pixels. In particular, the PDF (and any further reduced statistic) is invariant[15] under pixel shuffling. Indeed, for any permutation $\tau : i \in [\![1, d]\!] \mapsto \tau(i) \in [\![1, d]\!]$:

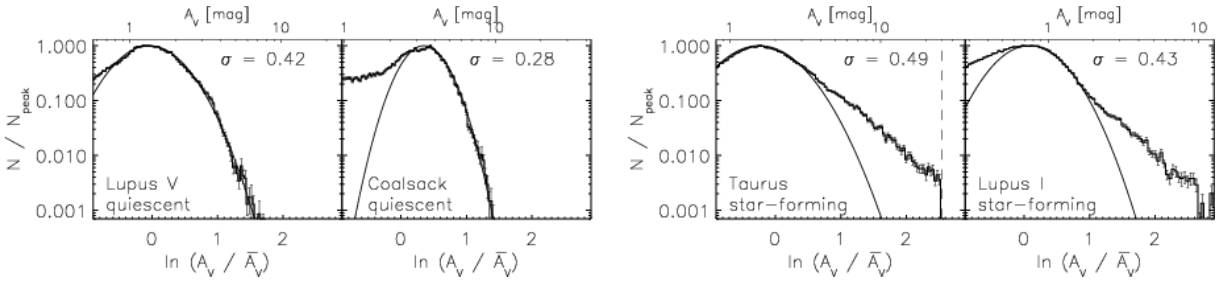
$$PDF(X(\mathbf{r}_i)) = PDF(X(\mathbf{r}_{\tau(i)})), \tag{3.17}$$

and the translation-invariant estimator of one-point distribution function is not modified. Therefore, there exists a wide variety of processes that share the same PDF, but which might have very different morphological properties. We give an illustration of this large inability of the PDF to distinguish between spatial textures in Fig. 3.4, where the empirical histogram of a column density map of a structured MHD simulation is shown to be very close (formal definitions of "closeness" will be given in Chap. 4 and 5) to the one of an unstructured sample of a phase randomized process (close to be Gaussian). To circumvent this degeneracy of the PDF, a natural extension consists in probing the two-point properties of the field.

[12] Note that ρ/ρ_0 is dimensionless and therefore Eq. 3.15 is not inhomogeneous.
[13] The non-magnetized case corresponds to $\beta \to \infty$.
[14] On the other hand, stellar outflows increase the proportion of diffuse gas resulting in deviations from log-normality in the low-density extremity (Appel et al., 2022).
[15] We remind that we assume p to be translation invariant.



(a) High density part of the Cumulative Distribution Functions (CDFs) (actually decreasing, which is the reversed form of the standard definition).



(b) Examples of column density PDFs for quiescent (left) and star forming clouds (right).

Figure 3.2: One-point statistics analysis for column density maps (traced by extinction mapping A_V) of neighboring ($\leq$ 700 pc) molecular clouds. Star-forming clouds tend to develop a heavy tail deviating from the log-normal PDFs observed (except at low-densities) for quiescent clouds (a), resulting in a lower decreasing rate of the (reversed) cumulative distribution function (b). *Credits*: (Kainulainen et al., 2009).

2.2 Two-point statistics

In a similar way than in Eq. 3.14, we can define the two-point distribution function as the joint distribution $(x, y) \mapsto p_{X(\mathbf{r}_i),X(\mathbf{r}_j)}(x, y)$ of the couple $X(\mathbf{r}_i), X(\mathbf{r}_j)$ by marginalizing p over the $d-2$ remaining variables. Again, if we consider a translation invariant process, this joint distribution depends only on $\delta\mathbf{r} \equiv \mathbf{r}_j - \mathbf{r}_i$ (understood modulo the domain size(s) for periodic processes). Then, the collection of two-dimensional distribution functions $\{p_{X(\mathbf{r}_1),X(\mathbf{r}_1+\delta\mathbf{r})}\}_{\delta\mathbf{r}}$ entirely captures the two-point statistics of X. However, even if $\delta\mathbf{r}$ is discretized, this collection is a $\dim \mathbf{r} + 2$-order tensor[16], which is very inconvenient as soon as we work with images, and worsens with hyper-spectral data. It is therefore reduced in several ways.

2.2.1 Basic two-point statistics

The second-order moment $C(\delta\mathbf{r})$, along with its Fourier transform, the power spectrum $PS(\mathbf{k})$, are tools extensively used as a first step to compress the two-points statistics. They consist in:

$$C(\delta\mathbf{r}) \equiv \mathbb{E}\left[X(\mathbf{r}_1) \cdot X(\mathbf{r}_1 + \delta\mathbf{r})\right], \tag{3.18}$$

[16]Indeed, $\dim \delta\mathbf{r}$ + the two values x, y in $p_{X(\mathbf{r}_1),X(\mathbf{r}_1+\delta\mathbf{r})}(x, y)$.

and:

$$PS(\mathbf{k}) \equiv \tilde{C}(\mathbf{k}) \equiv \sum_{\delta\mathbf{r}} C(\delta\mathbf{r}) e^{-2i\pi\mathbf{k}\cdot\delta\mathbf{r}}. \tag{3.19}$$

For a stationary field, it can be shown that the power spectrum, which is defined as a deterministic Fourier transform, relates to the Fourier transform $\tilde{X}$ of a sample X as:

$$PS(\mathbf{k}) = \mathbb{E}\left[|\tilde{X}(\mathbf{k})|^2\right]/d. \tag{3.20}$$

This relation is extensively used in practice as it allows to estimate empirically the power spectrum of a process through the Fourier transform of its sample(s).

Also, the power spectrum is *additive* in the *uncorrelated* case. More precisely, if X and Y are both translation invariant and are such that the $d \times d$ covariance matrix $\mathbb{E}\left[XY^T\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]^T$ is null, then:

$$PS[X+Y] = PS[X] + PS[Y]. \tag{3.21}$$

This property of the power spectrum makes it a very convenient tool for statistical diagnostic in the context of component separation. For instance, let us suppose that we are interested in estimating $PS[X]$ at some wavevector $\mathbf{k}$, from the contaminated observation data $d_0 \equiv x + y$ and the knowledge of the power spectrum $PS[Y](\mathbf{k})$ of the contamination (assumed independent from X). Then, using Eq. 3.20 and Eq. 3.21, we can build the following unbiased estimator for $PS[X](\mathbf{k})$, based on the observation d_0:

$$\widehat{PS}(d_0) \equiv |\tilde{d}_0(\mathbf{k})|^2/d - PS[Y](\mathbf{k}). \tag{3.22}$$

If we are in a regime where $PS[X](\mathbf{k}) \gg PS[Y](\mathbf{k})$ (which can be verified *a posteriori*), then the estimator $\widehat{PS}(d_0)$ is a promising starting point to constrain $PS[X](\mathbf{k})$. Indeed, its variance, noted $\mathrm{Var}\ \widehat{PS}(D)$ with $D \equiv X + Y$, is not impacted by the contamination as it is approximately[17]:

$$\mathrm{Var}\ \widehat{PS}(D) = \mathrm{Var}\left[|\tilde{D}(\mathbf{k})|^2/d\right] \sim PS^2(D)(\mathbf{k}) \simeq PS^2[X](\mathbf{k}), \tag{3.23}$$

which results in a coefficient of variation of the estimator $\widehat{PS}(D)$ (the ratio between its standard deviation and its expected value) of order unity. This coefficient of variation can be further reduced in the isotropic case by extending the estimation to an average over the Fourier modes in the sphere of radius $\|\mathbf{k}\|$. However, the previous derivation conversely shows the *clear difficulty* there is in retrieving $PS[X](\mathbf{k})$ in the regime $PS[X](\mathbf{k}) \ll PS[Y](\mathbf{k})$. Indeed, in this case, the term $PS^2[X](\mathbf{k})$ in the last approximation of Eq. 3.23 is superseded by $PS^2[Y](\mathbf{k})$, and therefore the coefficient of variation of $\widehat{PS}(D)$ is about $PS[Y](\mathbf{k})/PS[X](\mathbf{k}) \gg 1$, meaning that estimating the correct order of magnitude of $PS[X](\mathbf{k})$ is already difficult. Hence, in situations like this, it might be very difficult to mitigate this loss of estimation power due to the contamination.

[17] The first approximation can be for instance motivated in the case where $\tilde{D}(\mathbf{k})$ is a circularly symmetric complex-valued Gaussian random variable, which is typical in practice as the central limit theorem applies to the averaging that occurs in $\tilde{D}(\mathbf{k})$. In this case, the variance of $|\tilde{D}(\mathbf{k})|^2$ is of the same order than the variance of $\tilde{D}(\mathbf{k})$ (that is here $d \times PS[D](\mathbf{k})$) squared.

A promising avenue to tackle this challenge consists in leveraging the potential dependency between the wavevector $\mathbf{k}$ and other scales, either with regard to X (by finding an alternative low-contaminated scale which dependency with $\tilde{X}(\mathbf{k})$ could tighten its constraining of $\tilde{x}(\mathbf{k})$), Y (to lower the uncertainty about $\tilde{y}(\mathbf{k})$), or both. This however requires a way to characterize the statistical dependency between scales. This will be addressed in Sec.2.4.

2.2.2 Self-similarity and the K41 power spectrum

The power spectrum is an ubiquitous tool to characterize physical processes and/or structures. Indeed, for a *self-similar* field, i.e., that is such that there exists an exponent H that verifies, for any $\lambda \in \mathbb{R}_+^*$, the distribution equality:

$$X(\lambda \mathbf{r}) \sim \lambda^H X, \tag{3.24}$$

the power spectrum verifies[18] a power-law, given by:

$$PS(\lambda \mathbf{k}) = \lambda^{-m-2H} PS(\mathbf{k}), \tag{3.27}$$

with m the dimension of the lattice over which $\mathbf{r}$ is described.

This has a great impact in turbulence theory, for which the power-law exponent of the power-spectrum of the velocity field can be prescribed in certain cases. Indeed, for a 3D fluid following the incompressible Navier-Stokes dynamics, the exact Kármán–Howarth-Monin relation (De Karman & Howarth, 1938) (that only requires spatial homogeneity of the flow, but not stationarity, nor isotropy, nor considering a turbulent regime) can be used (cf. e.g., (Frisch & Kolmogorov, 1995)) to relate, by a Fourier transform, the energy flux $k \mapsto \Pi_k$ (introduced in Sec. 2 of Chap. 1) and the *structure function* $\mathbf{l} \mapsto \langle \|\delta \mathbf{u}(\mathbf{l})\|^2 \delta \mathbf{u}(\mathbf{l}) \rangle$ of the increments $\delta \mathbf{u}(\mathbf{l}) \equiv \mathbf{u}(\mathbf{r}+\mathbf{l}) - \mathbf{u}(\mathbf{l})$ of the velocity field $\mathbf{u}$:

$$\Pi_k = -\frac{1}{8\pi^2} \int_{\mathbb{R}^3} \frac{\sin(kl)}{l} \nabla_{\mathbf{l}} \cdot \left(\frac{\mathbf{l}}{l^2} \nabla_{\mathbf{l}} \cdot \langle \|\delta \mathbf{u}(\mathbf{l})\|^2 \delta \mathbf{u}(\mathbf{l}) \rangle \right) \mathrm{d}\mathbf{l}. \tag{3.28}$$

Now, on one hand, if we further assume that the flow undergoes a fully developed turbulence dynamics (which is mostly about assuming an non-zero dissipation rate, cf. Sec. 2 of Chap. 1 for further details), we have shown (Eq. 1.28) that the energy flux does not depend on k in the inertial range. We recall here this equation for convenience:

$$\Pi_k \simeq \varepsilon > 0. \tag{3.29}$$

On the other hand, if we assume that the velocity field in Eq. 3.28 is self-similar with an exponent

[18] Indeed, we can first observe:

$$C(\lambda \delta \mathbf{r}) = \mathbb{E}\left[X(\lambda \mathbf{r}_1) \cdot X(\lambda \mathbf{r}_1 + \lambda \delta \mathbf{r})\right] = \mathbb{E}\left[\lambda^H X(\mathbf{r}_1) \cdot \lambda^H X(\mathbf{r}_1 + \lambda \delta \mathbf{r})\right] = \lambda^{2H} C(\delta \mathbf{r}). \tag{3.25}$$

Then, using the property

$$\tilde{C}(\lambda \delta \mathbf{k}) = \lambda^{-m} \tilde{C}(\delta \mathbf{k}) \tag{3.26}$$

of the m-dimensional Fourier transform leads to Eq. 3.27.

H, as suggested by Kolmogorov in his first paper (Kolmogorov, 1941) of the three that founds the so-called *K41 theory of turbulence*, it can be shown (Frisch & Kolmogorov, 1995) that Eq. 3.28 boils down to:

$$\Pi_k \propto k^{1-3H}. \tag{3.30}$$

Therefore, Eq. 3.29 and Eq. 3.30 are conciliated if and only if $H = 1/3$. Hence, from Eq. 3.27 we see that a solution of the 3D fully developed turbulence, that would have a self-similar velocity increments field, must verify:

$$PS_{3D}[\mathbf{u}](\mathbf{k}) \propto \|\mathbf{k}\|^{-3-2/3} = \|\mathbf{k}\|^{-11/3}. \tag{3.31}$$

This law is usually written in the famous alternative form:

$$E(k) \propto k^{-5/3}, \tag{3.32}$$

where $E(k) \equiv 4\pi k^2 \langle PS_{3D}[\mathbf{u}](\mathbf{k})\rangle_{\|\mathbf{k}\|=k} = \partial_k \mathcal{E}_k$ refers to the overall spectral energy density on the sphere of radius k.

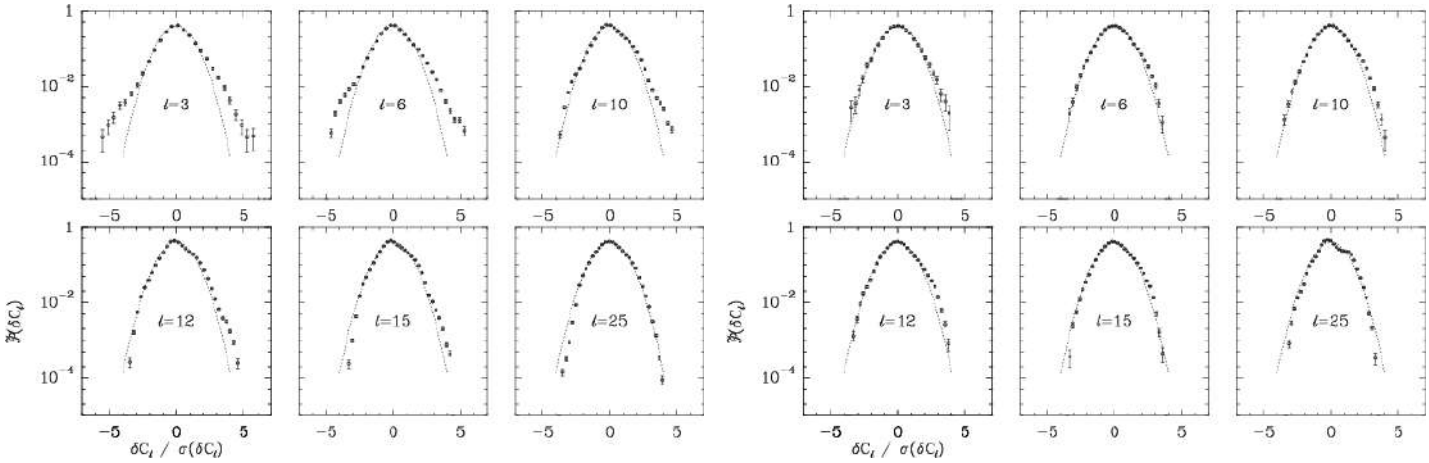
One of the main drawback of this self-similar model of the velocity field is that it leads to a self-similar dissipation rate $\varepsilon \equiv \nu\langle\|\nabla \wedge \mathbf{u}\|^2\rangle$. This, prediction of the K41 theory was the focus of a reproach from Landau. Indeed, dissipation is observed not to be self-similar but rather to occur in highly localized structures both in space and time. This property of the turbulence is referred as *intermittency*. It was first observed in the spatial fluctuations of the amplitude of the small scales structures of velocity field in the hydrodynamical case by Batchelor and Townsend, 1949 and then extend to MHD by Politano and Pouquet, 1995. An example of this dissipative structures in the case of MHD is shown in Fig. 1.7. A standard power spectrum analysis fails at probing this intermittent localization of dissipative structures as the Fourier modes $e^{i\mathbf{k}\cdot\mathbf{r}}$ are non localized in space, in the sense that their local spatial power $|e^{i\mathbf{k}\cdot\mathbf{r}}|^2 = 1$ is uniformly distributed.

2.2.3 Elaborate two-point statistics for intermittency

In the quest for statistical characterization of intermittency, several tools have been developed, for instance:

- PDFs of the velocity increments: $\mapsto PDF(\delta u(l))$. For a fixed lag l, and unit vector $\mathbf{e}$ (usually set to span the line-of-sight in observational context), the velocity increment $\delta u(l)$ is the random variable $\big(\mathbf{u}(\mathbf{r}+l\mathbf{e}) - u(\mathbf{r})\big)\cdot\mathbf{e}$, that does not depend on $\mathbf{r}$. An illustration of such PDFs is given in Fig. 3.3.
- Structure functions: these are moments $S_p(l) \equiv \mathbb{E}\left[|\delta u(l)|^p\right]$ that reduce the latter PDFs. These functions $S_p(l)$ can be further reduced to the scaling exponents ζ_p defined such that $S_p(l) \propto l^{\zeta_p}$ (see e.g., Lesaffre, Falgarone, and Hily-Blant, 2024 for applications in the ISM).
- Local harmonic analysis tools: similarly, PDFs and non linear moments based on wavelet filtered fields have played a significant role in intermittency for turbulence (Farge et al.,

1992), MHD turbulence (Yoshimatsu et al., 2011), ISM (Allys et al., 2019; Robitaille et al., 2019) and wider fields of applications (Bruna et al., 2015).



(a) Polaris Flare. (b) Taurus.

Figure 3.3: PDFs of the centroid (line-of-sight projected) velocity increments with increasing lag l for Polaris Flare (a) and Taurus (b) molecular clouds from Hily-Blant et al., 2008. As the lag decreases, non-Gaussian tails arise, as a signature of intermittency.

These tools operate in quite a similar manner which can be informally summarized as 1) filter the field at small scales, to remove quiescent regions and enhance active regions of intermittency, and 2) analyze the spatial heterogeneity of this filtered field by probing the heavy tail nature[19] of its one-point marginalized version through further compressed quantities, based on moments of non linear operator(s). Let us however emphasize that despite sharing the same core idea, these tools differ dramatically in their reduction step, some of which being compressed and nonexpansive and others, such as structure functions, involving highly expansive operators, which can complicate their estimation.

Finally, by marginalizing the field of regions active at small scales to its one-point statistics (as discussed just above), these tools do not characterize the spatial geometry of these dissipative structures. However, these structures are observed to be highly coherent as they show regular patterns at large scales (cf. e.g., Fig. 1.7). In the following, we illustrate the strong coupling there is between large and small scales yielded by the nonlinear turbulent dynamics, and how this coupling is closely related to the coherent structures observed.

2.3 Exhibiting the coupling between scales

We consider a certain process X that has some coupling between its scales. To exhibit the manifestation of this coupling in the morphology of X, we will progressively remove this coupling, while maintaining the power spectrum of the field. For simplicity, we take X real valued, stationary with 0 mean. To decouple the scales, we confuse their phase coherence by applying a

[19]This property is closely related to the notion of *sparsity* that refers to signals that have a large proportion of values close to 0.

random dephasing linear filter Z to X, computing $Z \star X$ where $\star$ stands for a convolution. The random filter Z is chosen independently from X, with the following Gaussian distribution:

$$Z \equiv \{Z(\mathbf{r}_i)\}_{1 \leq i \leq d} \quad \text{i.i.d.:} \quad Z(\mathbf{r}_i) \sim \mathcal{N}(0, 1/\sqrt{d}). \tag{3.33}$$

We can verify[20] that $Z \star X$ has same mean and power spectrum as X. However, the random filtering obliterates the phase coherence of X. Indeed, writing $\arg[\tilde{X}]$ the field of the phases of $\tilde{X}$, we have:

$$\arg[\tilde{Z} \cdot \tilde{X}] = \arg[\tilde{X}] + \arg[\tilde{Z}] \mod 2\pi. \tag{3.36}$$

Since, for any $\mathbf{k}$, the phase $\arg[\tilde{Z}(\mathbf{k})]$ is uniformly distributed over the possible values for a real-valued field (i.e., $[0, 2\pi[$ if $\mathbf{k} \neq \mathbf{0}$ else $\{0, \pi\}$), the phase $\arg[\tilde{Z} \cdot \tilde{X}(\mathbf{k})]$ is also uniformly distributed:

$$\arg[\tilde{X}(\mathbf{k})] + \arg[\tilde{Z}(\mathbf{k})] \mod 2\pi \sim \arg[\tilde{Z}(\mathbf{k})]. \tag{3.37}$$

Finally, since X and Z are independent, Eq. 3.37 can be extended to an equality between the complete joint distributions:

$$\arg[\tilde{X}] + \arg[\tilde{Z}] \mod 2\pi \sim \arg[\tilde{Z}]. \tag{3.38}$$

Hence, the phase structure of $Z \star X$ fully inherits from the one of Z which has no dependence between scales (except trivial link between $\arg[\tilde{Z}(\mathbf{k})]$ and $\arg[\tilde{Z}(-\mathbf{k})]$ ensuring Z to be real-valued). Note nevertheless that some form of scale coupling might persist between the amplitudes of the scales $\left\{\left|\widetilde{Z \star X}(\mathbf{k})\right|\right\}_{\mathbf{k}}$.

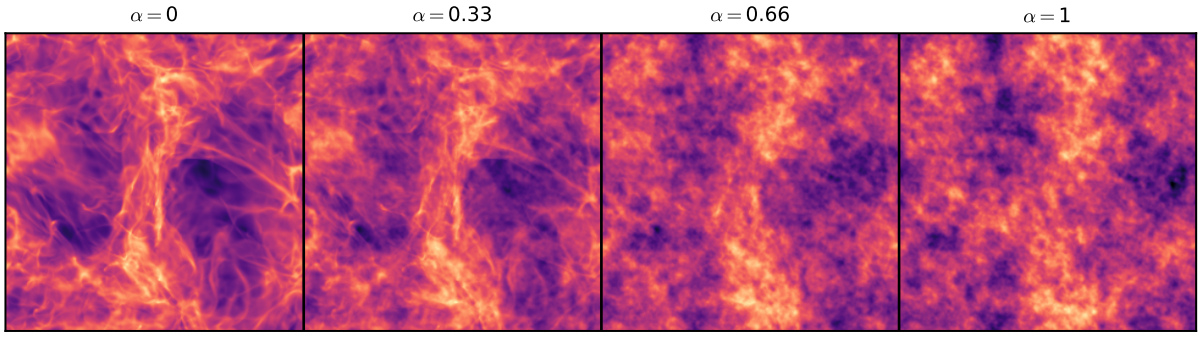


Figure 3.4: Progressive removal of the phase coherence through random dephasing filtering. The left image x corresponds to the result of a MHD simulation (introduced in Sec. 1). The image on the right corresponds to $z \star x$ for z a filter randomly sampled from Eq. 4.44. Intermediate images correspond to $\mathcal{F}^{-1}\left[\mathcal{F}[z]^{\alpha} \times \mathcal{F}[x]\right]$ where $\mathcal{F}$ denotes the 2D discrete Fourier Transform.

[20] Indeed, since X and Z are independent we have, denoting $\{0\}_{1 \leq i \leq d}$ the field composed only of 0:

$$\mathbb{E}[Z \star X] = \mathbb{E}[Z] \star \mathbb{E}[X] = \{0\}_{1 \leq i \leq d} \star \mathbb{E}[X] = \mathbb{E}[X], \tag{3.34}$$

and:

$$\mathbb{E}\left[\left|\widetilde{Z \star X}(\mathbf{k})\right|^2\right] = \mathbb{E}\left[\left|\tilde{Z}(\mathbf{k})\right|^2 \cdot \left|\tilde{X}(\mathbf{k})\right|^2\right] = \mathbb{E}\left[\left|\tilde{Z}(\mathbf{k})\right|^2\right] \cdot \mathbb{E}\left[\left|\tilde{X}(\mathbf{k})\right|^2\right] = \mathbb{E}\left[\left|\tilde{X}(\mathbf{k})\right|^2\right]. \tag{3.35}$$

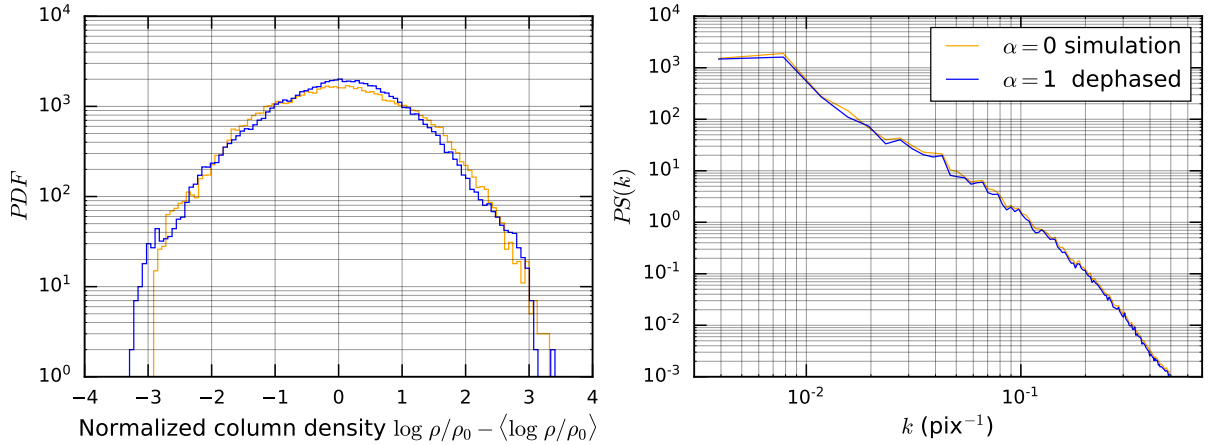


Figure 3.5: Empirical PDFs and power spectra of the simulation ($\alpha = 0$) and the dephased map ($\alpha = 1$) of Fig. 3.4. Note that this phase randomization procedure generates also fluctuations around the power spectrum of the simulation.

We give in Fig. 3.4 an example of this removal of the coupling between scales. This allows to visualize how much this coupling is important to structure the morphology of an image. This further shows the limitations of the PDF and power spectrum to characterize morphology as these descriptors are almost unchanged after removing these structures, as shown in Fig. 3.5.

2.4 Probing the coupling between scales

As motivated previously, there is a great interest in probing the coupling between scales of processes in the ISM. However, such coupling is not straightforward to characterize, since the different scales of a process are linearly *uncorrelated*.

For instance, if we attempt to characterize, through a cross-correlation, the coupling between the scale $\psi_1 \star X$ and $\psi_2 \star X$, where ψ_1 and ψ_2 are two distinct bandpass filters, we have:

$$C(\mathbf{r}) \equiv \mathbb{E}\left[\psi_1 \star X(\mathbf{0}) \cdot (\psi_2 \star X(\mathbf{r}))^*\right], \tag{3.39}$$

which can be written, using Parseval's formula, as:

$$|\tilde{C}(\mathbf{k})| = |\tilde{\psi}_1(\mathbf{k})\tilde{\psi}_2(\mathbf{k})|\tilde{C}_X(\mathbf{k}), \tag{3.40}$$

so we will not retrieve more information than there is in the power spectrum $\tilde{C}_X(\mathbf{k})$ of X. In particular, if the scales probed by the filters ψ_1 and ψ_2 have disjoint Fourier support, this linear correlation is 0. However, uncorrelated does not mean independent. Indeed, let us recall that the previous analysis showed the strong phase dependency there exists between scales.

In order to characterize this appealing dependency, numerous tools have been developed and applied to the ISM purpose. These notably include:

- the three-point correlation function and its Fourier transform: the bispectrum (Burkhart et al., 2009),

- the (Reduced) Wavelet Scattering Transform ((R)WST) (Allys et al., 2019; Lei & Clark, 2023; Saydjari et al., 2021),
- the Wavelet Phase Harmonics (Allys et al., 2020; Auclair et al., 2024; Jeffrey et al., 2022; Regaldo-Saint Blancard et al., 2021),
- and the scattering covariance (or spectra) (Cheng et al., 2024; Mousset et al., 2024).

The bispectrum (and similar types of moments of order 3), involves polynomial expressions of order 3 in the field (i.e., terms like X^3). This can lead to high dependencies of the coefficients on the process (e.g., a slight change of physical parameters induces a significant change on these moments), and also brings the difficulty to estimate these moments out of few samples. On the other hand, the wavelet scattering transform involves nonexpansive operators of the form:

$$||X \star \psi_1| \star \psi_2|, \tag{3.41}$$

which provide them with strong robustness warranties (Bruna & Mallat, 2013), while statistics such as WPH and scattering covariance involve expressions sub-quadratic in the field, of the form:

$$\langle \sigma_1(X \star \psi_1) \cdot \sigma_2(X \star \psi_2) \rangle_{\mathbf{r}}, \tag{3.42}$$

where σ_1 and σ_2 may include non-linear but nonexpansive operations, such as the modulus $|\cdot|$ or the phase harmonic operator (introduced in (Mallat et al., 2020)):

$$[\cdot]^p : x(\mathbf{r}) \mapsto |x(\mathbf{r})| \cdot e^{ip \arg[x(\mathbf{r})]}. \tag{3.43}$$

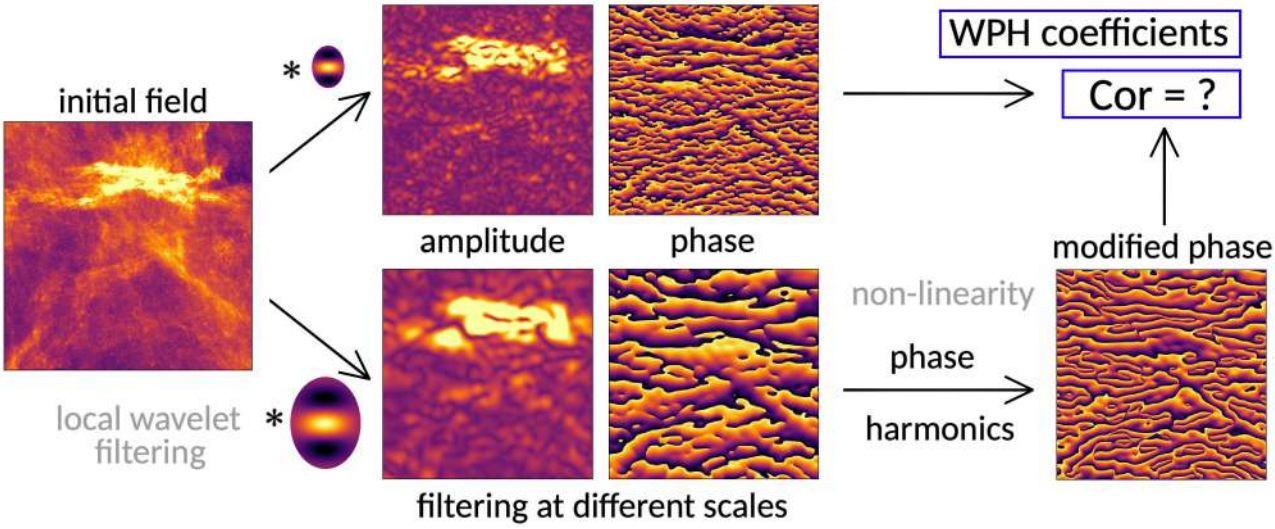


Figure 3.6: Probing the coupling between scales with the Wavelet Phase Harmonics. The phase harmonic operator allows to reduce the wavelength of the oscillations of the field filtered at larges scales. Therefore a non-trivial coupling with the small scales can be measured through a simple cross-correlation. The field corresponds to dust emission of a diffuse cloud called *Spider* (Marchal & Martin, 2023), observed at 250 μm by *Herschel* and cleaned from the CIB (data from Auclair et al., 2024). *Courtesy*: E. Allys.

As explained above, the role of such nonlinear operator is central in order to probe a non-trivial coupling when the scales ψ_1 and ψ_2 have disjoint Fourier support. As illustrated in Fig. 3.6, the phase harmonic operator $[\cdot]^p$ multiplies the wavenumber of the oscillations of $X \star \psi_1$ by a factor p. On the other hand, if the wavelet envelope $|\psi_1(\mathbf{r})|$ is smooth spatially, the modulus operator acts as a non linear smoothing operator that tends to decrease the wavenumber at which the field $X \star \psi_1$ oscillates (cf. e.g., Fig. 2 in (Cheng et al., 2024)).

3 The evolving coupling between scales from quiescent to star forming molecular clouds

We have seen that, as molecular clouds evolve from quiescent to star-forming, they develop a heavy tail in their column density PDF (Fig. 3.2). In this section, we aim at characterizing this evolution from a morphological point of view. It is noteworthy that the quiescent and active star forming regions share rather similar power law in their power spectrum which can therefore not be used to identify a trend in this evolution. To overcome this limitation, we rather aim at probing this morphological evolution from a scale-coupling perspective, leveraging the Wavelet Scattering Transform (WST).

The usual WST consists in two layers of statistics, the first of which depends on a single scale of length $\simeq 2^j$ pixels with orientation θ:

$$S_1(x)[j,\theta] \equiv \langle |x \star \psi_{j,\theta}| \rangle_{\mathbf{u}}.$$

The second layer probes a coupling between two oriented scales (j_1,θ_1), and (j_2,θ_2), with $j_1 < j_2$:

$$S_2(x)[j_1,j_2,\theta_1,\theta_2] \equiv \left\langle \big| |x \star \psi_{j_1,\theta_1}| \star \psi_{j_2,\theta_2} \big| \right\rangle_{\mathbf{u}} / \, S_1(x)[j_1,\theta_1].$$

Note that, with this normalization by the S_1 layer, the S_2 coefficients are both invariant to global additive[21] and multiplicative transformations:

$$\forall \lambda, c \in \mathbb{R}^* \times \mathbb{R} \quad S_2(\lambda \times x + c) = S_2(x). \tag{3.44}$$

Hence, the coupling coefficients S_2 bring, in a large measure, complementary information from the PDF, this latter statistics being very sensitive to the mean and standard deviation of the process under study.

We report in Fig. 3.7 the average

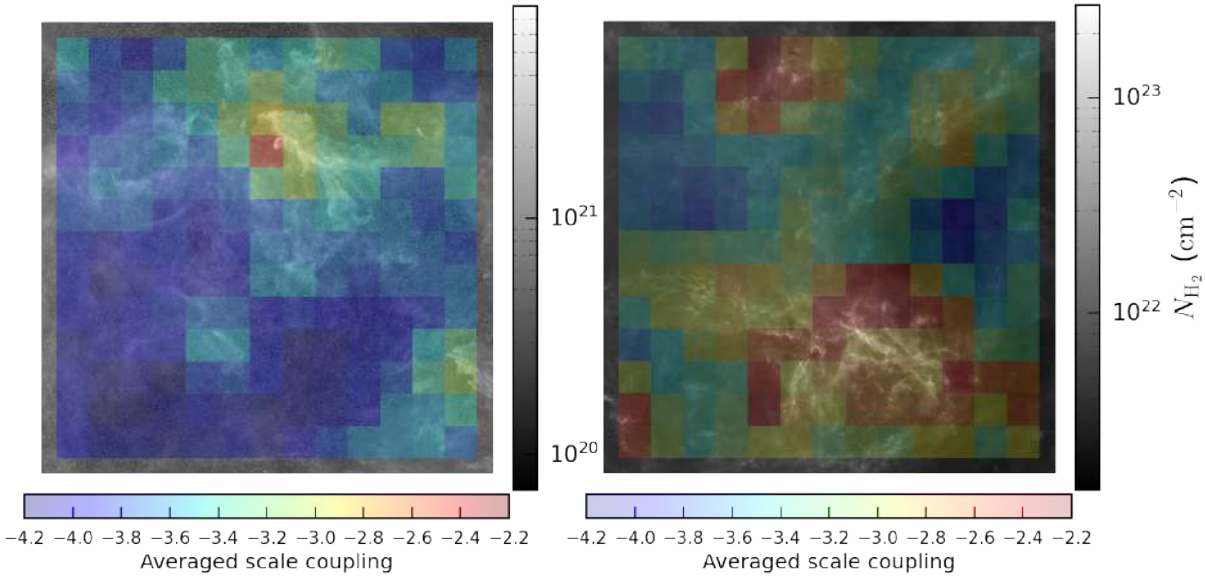
$$\langle \log_2 \, S_2(x) \rangle_{j_1,j_2,\theta_1,\theta_2} \tag{3.45}$$

of this coupling computed on 256×256 patches x of the 3584×3584 column density maps of

[21] Indeed, the filters ψ_j have no power at $\mathbf{k} = \mathbf{0}$.

two molecular clouds Polaris Flare and Aquila (see, in Chap. 5, Sec. 2 for a detailed description of the data and Sec. 4 for the statistics). We observe clear spatial fluctuations of this average coupling inside each molecular cloud, which are organized up to the size of the entire images, while stile varying continuously from one local patch to the other. This suggests, without being a proof, that either the stationarity length L_0 of these clouds is comparable or smaller than the size of the entire images studied here, either that the correlation length l_0 is larger than the size of the local patches used, and in reality both limitations may apply. This further suggests that particular care should be taken when estimating statistical properties over wide regions such as an entire molecular cloud.

Despite being invariant under the global additive and multiplicative column density enhancements (Eq. 3.44), this averaged scale coupling seems to be correlated with the regions of high column density. This correlation is not due to a mathematical relation that would hold for any process, but rather bolsters that the physical dynamics of molecular clouds constrain their densest regions to be also the ones associated to the strongest coupling between scales[22].



(a) Polaris Flare. (b) Aquila.

Figure 3.7: Averaged coupling between scales $\langle \log_2 S_2(x) \rangle_{j_1,j_2,\theta_1,\theta_2}$ for different patches x of size 256×256 in a quiescent molecular cloud (left) and actively star forming cloud (right).

Beyond this correlation of fully averaged $S2$ coefficients with the presence of dense regions in a given patch x, further constraints can be highlighted. To exhibit these constraints, we use another compressed form of the S_2 coefficients. To construct them, we first compress the S_2 coefficients into the isotropic levels of coupling:

$$\langle \log_2 S_2(x) \rangle_{\theta_1,\theta_2}[j_1, j_2] \tag{3.46}$$

between the non-oriented scales $(2^{j_1}, 2^{j_2})$ (still with $j_1 < j_2$). We then further compress them

[22] At least as probed by the scattering transform on the column density field.

by only keeping their $\delta = j_2 - j_1$ dependency, i.e., their dependency on the ratio between the scales, by considering the following average that we call *Scale Coupling Spectrum* (SCS):

$$SCS(\delta) \equiv \big\langle \langle \log_2 S_2(x) \rangle_{\theta_1,\theta_2,j_2-j_1=\delta}. \tag{3.47}$$

This spectrum probes a coupling between scales that only depends on their discrepancy $j_2 - j_1$ and benefits from the following symmetries (up to the usual limitations, as e.g., the discretization of the scales amplitudes and orientations):

- translation invariant,
- rotation invariant,
- scale invariant[23],
- multiplicative and additive invariant (cf. Eq. 3.44),
- (and also parity $\mathbf{r} \mapsto -\mathbf{r}$ invariant).

When computing this spectrum on 256×256 patches x extracted from column density maps of nearby molecular clouds ($d \leq \sim 700\,\mathrm{pc}$), we observe the following regularities on the distribution of coupling spectra:

- each spectrum is in good approximation affine:

$$SCS(x)[\delta] \simeq a_x + b_x \times \delta. \tag{3.48}$$

  The quantity a_x is related to the *averaged scale coupling* $\langle \log_2 S_2(x) \rangle_{j_1,j_2,\theta_1,\theta_2}$ introduced in Eq. 3.45 and reported in Fig. 3.7. We call b_x the *scale coupling rate* as it probes the rate at which the scale coupling evolves with respect to the scale discrepancy.

- Across the distribution of observed patches $\{x\}$, the averaged scale coupling and the scale coupling rate are positively correlated. This relation is partly due to mathematical definition of these quantities, but is also due to the ISM dynamics, as it does not hold when taking a wide set of everyday textures (referred to as DTD for the Describable Texture Dataset[24] in the following (Cimpoi et al., 2014)).

- As already depicted in Fig. 3.7, the averaged scale coupling of a patch x is positively correlated with the presence of dense structures in this patch. In conclusion: the denser the region, the higher its averaged scale coupling, and the higher its scale coupling rate (cf. Fig. 3.8 for a schematic summary). To quantify the latter notion of dense regions in a patch x, we suggest[25] the following quantification $\eta(x)$ of its empirical *PDF tail heaviness*:

$$\eta(x) \equiv \log \left[q_{85\%}(x) / q_{50\%}(x) \right], \tag{3.49}$$

[23] Up to a dilation factor that is negligible with respect to the range of scales probed $j_{\max} - j_{\min}$.
[24] Further details may be found in Sec. 2 of Chap. 5 and at https://www.robots.ox.ac.uk/~vgg/data/dtd/
[25] This choice is motivated by the simple evolution that undergoes the peaked (reversed) cumulative distribution function from quiescent to active molecular clouds reported in Fig. 3.2.a.

where $q_\alpha(x)$ designates the α-quantile of the distribution of values in patch x.

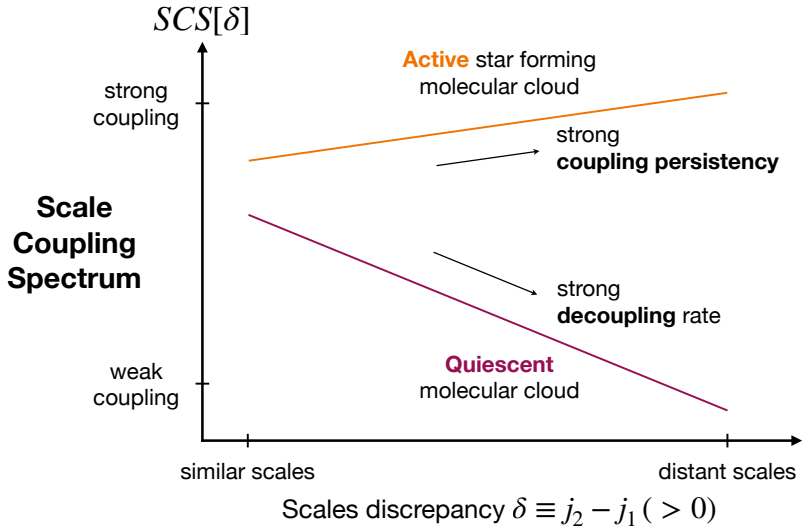


Figure 3.8: Schematic view of the evolving coupling between scales of column density maps from quiescent to active regions of molecular clouds.

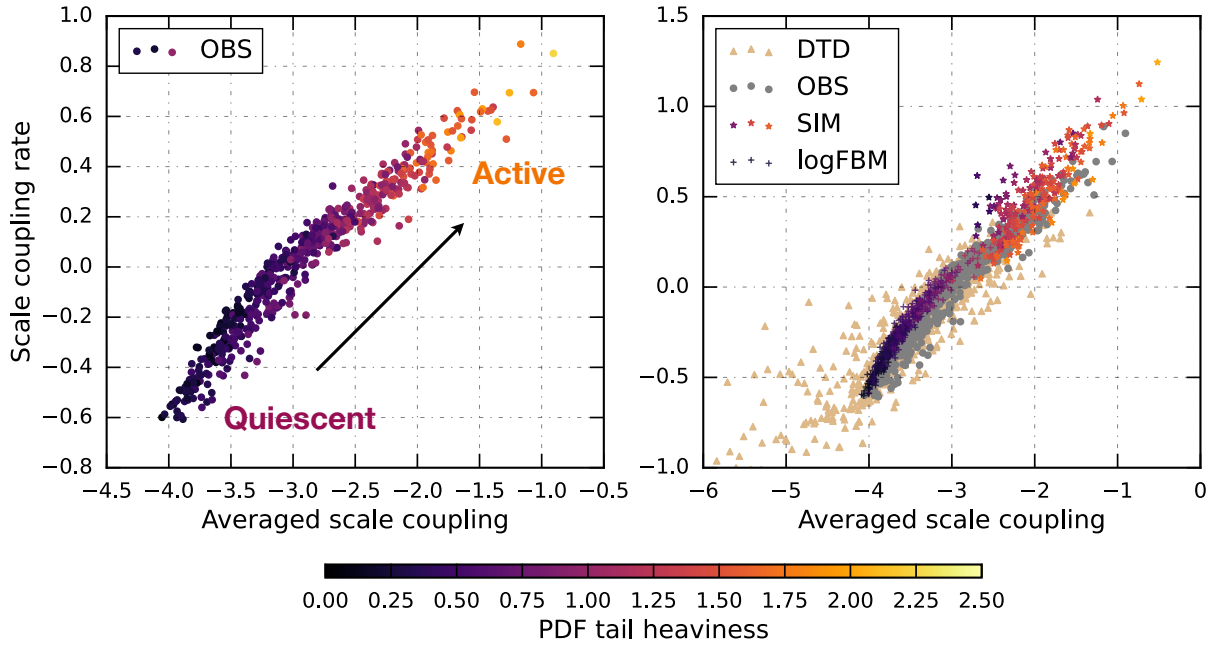


Figure 3.9: Properties of the scale coupling of column density maps of observations (left) and other datasets (right). DTD refers to a wide collections of textures, SIM to a set of simulations of dense star forming molecular cloud and logFBM to statistical models. Each point in this plot represent a 256×256 patch x, and is colored according to the heaviness of its PDF tail as probed by $\eta(x)$ (cf. Eq. 3.49).

We report these correlations for the set of observations in Fig. 3.9 (left). The patches associated to quiescent regions according to the criteria η land in the bottom left corner of this plot, while dense active regions to the top right, and intermediate observations lie in a tight submanifold. The latter is reported in gray on the right plot. When adding a wide variety of textures (DTD, camel triangles), we observe that the "mathematically admissible" space is wider than the one filled by the observations. This bolsters the idea presented in the previous chapter that ISM processes are constrained by a dynamics with a moderate number of parameters which makes them lie in a specific low-dimension submanifold (cf. Eq. 2.14).

We also report in this plot the results of a set of simulations of a dense star forming molecular cloud (SIM) and a statistical model (logFBM) used in the literature to reproduce some properties of diffuse regions of molecular clouds (these two datasets are described respectively in Sec. 8.1.1 and 8.1.2 of Chap. 5). The dense simulations lie in the area where we identified the dense active observations, while conversely, the logFBM processes lie in the area corresponding to diffuse patches. At first sight, this supports the analysis performed on the submanifold of observations in which we identified the corresponding quiescent and dense areas. However, this plot remains a very reductive 2D projection of various sets of highly non-Gaussian processes. Therefore, there is absolutely no warranty that processes that do end up at the same location in this plot are actually similar (cf. schematic view in Fig. 4.1). As explained in Sec. 4.1 of the previous chapter, this is all the more likely to happen that we make use of different datasets, with no warranty on their actual proximity. And as matter of fact, it is well known by the community that logFBM models are far from matching the rich statistical nature of molecular clouds (cf. e.g., Fig. 3.4 and 5.1). This thus brings the question: what statistics should be used in order to account for the actual similarity between observations, simulations (and other statistical models)?

Chapter 4

How to quantify information without supervision?

"The purpose of generality here is not to solve immediate practical problems, but rather to capture the logical essence of an important concept (sufficient statistic), and in particular to disentangle that concept from such ideas as Euclidean space, dimensionality, partial differentiation, and the distinction between continuous and discrete distributions, which seem to us extraneous."

Halmos & Savage in Halmos and Savage, 1949

Objectives

Summary statistics are a major tool used to grasp information from high dimensional processes. However, the choice of these descriptors highly depends on the processes considered. Fisher analysis is a tool extensively used to assess the amount of information captured by a given set of summary statistics for a given family of processes. It however operates in a supervised framework. We motivate in this chapter the interest for extending such ideas to the much less supervised world of ISM observations, and set a theoretical ground to perform such a shift.

Contents

1 Introduction

The multiple sets of summary statistics presented in the previous chapter offer a wide range of options to reduce and represent the input data, and their choice plays a pivotal role in inference. This open numerous questions. For a given inference task, how to choose these summary statistics? Does this choice actually depend on the task, or are there intrinsically better descriptors than others? Is it really necessary to make a choice in descriptors, or can they simply be all jointly used in a cooperative manner? Could we further tighten the inference results by designing new sets of summary statistics, and if so what should lead this development?

The topic relative to these questions is named *feature selection* in data science. To begin answering these questions, we present in this chapter the main tools brought by statistics and information theory. The large majority of these tools have been established by Sir Ronald Fisher in the beginning of the 20^{th} century and are nowadays routinely used in applied science, but also keep playing a major role in the theoretical development of data science. These tools, that we present in a first part of this chapter, are however defined in the framework of *parametric families of distributions* $\{p_\theta\}_\theta$ where the parameters θ define a natural geometry or similarity between the processes p_θ, and are therefore not directly applicable in cases where the structure of such parametrization is yet to be unveiled or simply inexistent. Therefore, in the second part of this chapter (starting at Sec. 3), we use information geometry to extend these ideas in a framework where the parameters θ are reduced to simple indexes i that contain no information about the similarity between the processes $\{p_i\}_i$.

2 What statistics for a supervised task?

2.1 Sufficient statistics

When we consider a parametric family of distributions $\{p_\theta\}_\theta$, if we find a summary statistic ϕ such that the conditional distribution of X given $\phi(X)$ does not depend on θ, then knowing the data x brings no additive information about θ than knowing the reduction $\phi(x)$. In other words, to constrain θ, it is *sufficient* to have $\phi(x)$ instead of x. This leads to the following definition of *sufficient summary statistics*:

$$\phi \text{ is sufficient for } \{p_\theta\}_\theta \quad \iff \quad \forall \theta_1, \theta_2 \quad p_{\theta_1}(x|\phi(x)) = p_{\theta_2}(x|\phi(x)), \tag{4.1}$$

where the conditional probability $p_\theta(x|\phi(x))$ is defined as:

$$p_\theta(x|\phi(x)) \equiv \frac{\mathbb{P}[(X = x) \wedge (\phi(X) = \phi(x)) \,|\, \theta]}{\mathbb{P}[\phi(X) = \phi(x)) \,|\, \theta]} = \frac{p_\theta(x)}{\phi_\# p_\theta(\phi(x))}. \tag{4.2}$$

Note that the definition of sufficiency (Eq. 4.1) is given for a certain family $\{p_\theta\}_\theta$. It is possible to find a summary statistic sufficient for a first family but not for another. As matter of fact,

we shall see below (Eq. 4.8) that the empirical mean and variance are sufficient statistics for a family of uncorrelated Gaussian processes but are well-known to be insufficient for more complex processes as those encountered in the ISM, as shown in the following chapter.

2.1.1 Fisher-Neyman factorization criterion

Sufficiency can be characterized by a convenient *factorization criterion* that was first studied by Fisher, 1925 and progressively generalized by Neyman, 1936 and then Halmos and Savage, 1949. This criterion claims that a summary statistic ϕ is sufficient for $\{p_\theta\}_\theta$ if and only if there exist nonnegative functions g_θ and h such that:

$$p_\theta(x) = g_\theta(\phi(x)) \times h(x). \tag{4.3}$$

This characterization evidences that the dependency on θ of p_θ is entirely captured by g_θ, that relies on x only through $\phi(x)$. Since the quantities involved are nonnegative, this factorization can be rephrased by taking the logarithm and then deriving with respect to θ:

$$\nabla_\theta \log\, p_\theta = \nabla_\theta \log\, [g_\theta \circ \phi]. \tag{4.4}$$

The left hand side of this equation involves the *score function* $\nabla_\theta \log p_\theta$ of p_θ. Thus the Fisher-Neyman factorization theorem tells us that the score of p_θ depends on θ only through sufficient statistics. In particular, $\phi : x \mapsto \{\nabla_\theta \log p_\theta(x)\}_\theta$ is a sufficient set of statistics, but of infinite dimension. However, if $\nabla_\theta \log\, p_\theta(x)$ varies few with respect to θ, one might expect to compress this set for a fixed x. This will be further detailed in the context of Fisher information in Sec. 2.3 (cf. notably Eq. 4.15). The utility of the score function has been briefly motivated in Sec. 4.2 for parameter estimation and find great applications in the context of high dimensional modeling and inference.

We can use the Fisher-Neyman characterization to show the following (surprising) result: for the family of d-dimensional stationary Gaussian processes $\{p_{\mu,\sigma}\}$ given by:

$$p_{\mu,\sigma} = \mathcal{N}(\mu, \sigma^2 \mathbb{I}_d), \tag{4.5}$$

the simple summary statistics that are the empirical mean $x \mapsto \langle x_r \rangle_r$ and the empirical variance $x \mapsto \langle x_r^2 \rangle_r - \langle x_r \rangle_r^2$ (where $\langle \cdot \rangle_r$ stands for $(1/d)\sum_{r=1}^d$) are sufficient. Indeed, $p_{\mu,\sigma}(x)$ can be expressed as a function of x only through $\langle x_r \rangle_r$ and $\langle x_r^2 \rangle_r$:

$$p_{\mu,\sigma}(x) \propto \exp\left[\frac{1}{2}(x-\mu)^T \sigma^{-2} \mathbb{I}^{-1} (x-\mu)\right] \tag{4.6}$$

$$\propto \exp\left[\frac{1}{2\sigma^2}(x^T x - 2\mu^T x)\right] \tag{4.7}$$

$$= \exp\left[\frac{d}{2\sigma^2}(\langle x_r^2 \rangle_r - 2\mu \langle x_r \rangle_r)\right], \tag{4.8}$$

therefore the latter expression matches the Fisher-Neyman factorization criterion (Eq. 4.3) (with e.g., $h(x) \equiv 1$).

This impressive reduction of the dimensionality without loss of information makes sufficient statistics a very appealing tool. In a more general perspective, if the processes of the family $\{p_\theta\}_\theta$ share some symmetries, it is possible to construct sufficient statistics ϕ of dimension strictly smaller than the input data x (Diaconis, 1988). However, the Pitman–Koopman–Darmois theorem (Darmois, 1935; Koopman, 1936; Pitman, 1936) gives us that, for a family of distributions $\{p_\theta\}_\theta$ with θ of finite dimension, the existence of a set of sufficient statistics, whose dimension remains finite when taking the sample size to infinity (this excludes for instance the trivial statistic: $x \mapsto x$), is possible if and only if the family is an *exponential family*, i.e., g_θ is of the form:

$$g_\theta(x) = \exp\left[\theta^T \phi(x) + c(\theta)\right]. \tag{4.9}$$

This excludes a dramatic set of families. Informally, this reduces the scope to parametric processes that have an "Hamiltonian" $\mathcal{H}_\theta(x) \equiv -\log p_\theta(x)$ where the variables θ and x are separated. For ISM processes, we should therefore not expect the existence of sets of low dimensional summary statistics that would be exactly sufficient. Nevertheless, by relaxing the latter constraint to the one of *approximate sufficiency*, we can aim at finding statistics that verify a trade-off between their dimensionality and sufficiency level. This however requires to quantify, still with respect to a given family, the notion of insufficiency.

2.2 From sufficiency to informativeness

A convenient way to measure the *insufficiency* of ϕ relatively to a parameterized family $\{p_\theta\}_\theta$ (and for a given[1] distribution $\mathbb{P}_\theta$ on θ) is to resort to the mutual information $I(\theta, \phi(X))$. Indeed, this quantity probes the reduction in the uncertainty on θ yielded by the observation of $\phi(X)$ (Cover & Thomas, 2006):

$$I(\theta, \phi(X)) = H(\theta) - H(\theta|\phi(X)) \geq 0, \tag{4.10}$$

where the notion of uncertainty is quantified by the entropy, defined for $Z \sim p_Z$ as:

$$H(Z) \equiv \mathbb{E}_{Z\sim p_Z}\left[-\log p_Z(Z)\right], \tag{4.11}$$

and the conditional entropy:

$$H(Z|Y) \equiv \mathbb{E}_{Y,Z\sim p_{Z|Y}}\left[-\log p_{Z|Y}(Z|Y)\right]. \tag{4.12}$$

[1]Indeed, to define the *mutual information* between θ and a statistic of X, we need to define a joint probability distribution $\mathbb{P}_{\theta,X}$. In our case of parametric distributions $\{p_\theta\}_\theta$, we have the natural definition for the conditional probability $\mathbb{P}_{X|\theta}(x) \equiv p_\theta(x)$. Then, adding a marginal distribution $\mathbb{P}_\theta$, called *prior* in the Bayesian framework, allows to fully specify the joint distribution: $\mathbb{P}_{\theta,X}(\vartheta, x) = p_\theta(x) \cdot \mathbb{P}_\theta(\vartheta)$.

Hence, the higher $I(\theta, \phi(X))$, the tighter $\phi(X)$ constrains θ. Indeed, if $\phi(X)$ entirely prescribes θ, then $H(\theta|\phi(X)) = 0$, and $I(\theta, \phi(x))$ is maximized. On the other hand, if $\phi(X)$ gives no information on θ, then $H(\theta|\phi(X))$ boils down to $H(\theta)$, and we have no reduction in the uncertainty on θ.

The *Data Processing Inequality* (DPI) (also called *Information Monotonicity*) ensures that

$$0 \leq I(\theta, \phi(X)) \leq I(\theta, X). \tag{4.13}$$

Hence, not surprisingly, no summary statistics are more informative than the non reductive operation $X \mapsto X$. But this property is further shared to any statistic that is sufficient:

$$\phi \text{ is sufficient for } \{p_\theta\}_\theta \quad \Longrightarrow \quad I(\theta, \phi(X)) = I(\theta, X). \tag{4.14}$$

Conversely, if ϕ holds the equality in the DPI for any distribution $\mathbb{P}_\theta$, then it is sufficient for $\{p_\theta\}_\theta$.

Hence, the mutual information $I(\theta, \phi(X))$ allows to go beyond the binary sufficient/not sufficient outcome yielded by the definition in Eq. 4.1, and quantifies how far from sufficient ϕ is. As for the definition of sufficiency, this quantity is relative to a parametric family $\{p_\theta\}_\theta$, but further requires a distribution $\mathbb{P}_\theta$ over θ.

2.3 Fisher information and Fisher analysis

Mutual information is however hard to estimate directly in practice, as it involves an integral over the joint distribution of $\phi(X)$ and θ and more importantly as it requires the knowledge of the absolute value (in the sense non relative) of the conditional probability $p_{\theta|\phi(X)}$. Instead, the *Fisher Information* constitutes an alternative measure of information which is local (in θ) and relative (on the probability value). It is thus extensively used in applied statistics, the ISM not being an exception, see for instance Allys et al., 2020; Hothi et al., 2024; Park et al., 2023.

For a given θ_0, this information refers[2] to as the curvature[3] in θ of $-\log \phi_\# p_\theta(\phi(X))$ averaged over $X \sim p_{\theta_0}$ (Lehmann & Casella, 2006):

$$J_{\phi(X)}(\theta_0) = -\mathbb{E}_{X \sim p_{\theta_0}} \left[\partial^2_{\theta^2} \log \phi_\# p_\theta(\phi(X)) \big|_{\theta_0} \right]. \tag{4.16}$$

Hence, the higher $J_{\phi(X)}$, the more sensitive the distribution $\phi_\# p_\theta$ becomes with θ, the tighter the constraint on θ can be derived from $\phi(X)$. This link is for instance exhibited by the Cramér–Rao bound: if $\phi(X)$ is designed to be an unbiased estimator of θ, its variance is lower bounded by

[2]The actual definition is based on the variance of the score statistics $\nabla_\theta \log \phi_\# p_\theta\big(\phi(X)\big)$ of the summary statistic ϕ (Cover & Thomas, 2006):

$$J_{\phi(X)}(\theta_0) \equiv \mathrm{Var}_{X \sim p_{\theta_0}} \left[\nabla_\theta \log \phi_\# p_\theta\big(\phi(X)\big) \big|_{\theta_0} \right]. \tag{4.15}$$

[3]For simplicity we only develop the unidimensional case for θ here.

the reciprocal of the Fisher information

$$\operatorname{Var} \phi(X) \geq \frac{1}{J_{\phi(X)}(\theta)}. \tag{4.17}$$

As the mutual information $I(\theta, \phi(X))$, the Fisher information $J_{\phi(X)}(\theta)$ verifies a chain rule of the form presented in Eq. 4.10 which allows to derive another Data Processing Inequality based on this alternative definition of information (Zamir, 1998):

$$J_{\phi(X)}(\theta_0) \leq J_X(\theta_0), \tag{4.18}$$

with equality if (but not only if (Pollard, 2013)) ϕ is sufficient for the family:

$$\phi \text{ is sufficient for } \{p_\theta\}_\theta \quad \Longrightarrow \quad J_{\phi(X)}(\theta_0) = J_X(\theta_0). \tag{4.19}$$

It is also noteworthy that the Fisher information defines a metric *within* the submanifold of distributions on which it is computed, and by so allows to define a distance between two distributions of this submanifold called the *Fisher-Rao distance* (see (Nielsen, 2013) for an insightful introduction). Let us however emphasize that this distance is *degenerate* as soon as one distribution leaves the submanifold. As we shall see in the following section, this might be problematic when comparing simulations with observations (cf. Fig. 4.1).

Finally, to compare the informativeness of various sets of summary statistics, for a given dataset, and by means of a Fisher analysis, requires the data $\{x_i\}_i$ to be sampled from processes lying in a smooth parametric family $\{p_\theta\}_\theta$:

$$X_i \sim p_{\theta_i}, \quad p_{\theta_i} \in \{p_\theta\}_\theta,$$

and it requires to have the parameters jointly provided with the samples, i.e. to have $\{(\theta_i, x_i)\}_i$. Hence, Fisher analysis cannot be applied in numerous ISM situations where we are left only with a collection of observations $\{x_i\}_i$.

3 Motivating the unsupervised approach

Supervised frameworks refer to situations where we collect data samples x simultaneously with some parameters or labels θ revealing some additional[4] information to the samples x, on the processes that generated these samples. This is well suited to situations where the data generator can be actively piloted (e.g., numerical simulations (Federrath et al., 2010; Hothi et al., 2024; Peek & Burkhart, 2019; Saydjari et al., 2021), or laboratory experiments). The scientific community is starving for such controlled frameworks as they allow to progressively dissect the role of the parameters into reproducible experiments. Supervised frameworks can also be considered in less actively controlled situations. For instance when the data can be classified

[4]Note that we exclude in this definition *self-supervised learning* where the data itself constitute the labels (Gidaris et al., 2019; Liu et al., 2021).

a posteriori by humans with a simple/consensus criterion (Peek & White, 2021), or classified according to a complementary tracer. As an example, Lei and Clark, 2023 learned to predict the CNM fraction with H I emission morphology, from a training set of emission observations where direct absorption measurements allow to access the CNM fraction.

However, numerous challenges that involve ISM observations still cannot be cast in such supervised frameworks. For instance, in the next chapter we aim at exploiting the morphological properties of observed molecular clouds in order to compare them. Indeed, as motivated hereafter, we do not want to learn entirely these properties relying on an prior model (such as simulations), so their exploitation should include[5] an unsupervised learning step on observations. We present below the limits of fully simulation-based inference in some ISM related challenges and motivate the consideration of unsupervised observation-based approaches.

Limitations of the supervised approach

Numerical simulations constitute a controlled framework which is appealing to explore and model physical processes. These models can then be used in *Simulation-Based Inference* (SBI) to extract information from observations (Cranmer et al., 2020). However, there are parts of the ISM that are so challenging that relying on a fully simulation-based supervision can be misleading. For instance, the scenario of star formation in molecular clouds involves a complex interplay of turbulence, magnetic fields, chemistry, radiation, gravity and feedback. The highly non-linear combination of these ingredients makes it very difficult to understand the precise role they each play (Falgarone et al., 2004). Any omission or modeling error in one of these ingredients is likely to lead to a non realistic simulation. In other words, the submanifold of probabilistic processes yielded by a parameterized simulation is likely not to overlap the set of observations (here each observation is seen as a realization of a stochastic process associated to the combination of physical processes that generated it, cf. Sec 2 of Chap. 2), as illustrated in Fig. 4.1. This unknown but often significant discrepancy between observations and simulations might be problematic for at least two approaches:

- estimating the distance between observations and simulations with standard information criteria such as the likelihood, the Akaike or Bayesian Information Criteria (see (Stoica & Selen, 2004) for an insightful review). Indeed, these measures of closeness operate by building a likelihood based only on simulations, evaluated on the observations. Doing so implicitly assumes that the observed process belongs to the same parametric family as the simulations, and is blind to the so-called *model error* (Anderson & Burnham, 2002).

- characterizing the set of observations with a set of summary statistics designed under the supervision of simulations (i.e. to supersede the geometry of the submanifold of observation processes by the one learnt on simulations). Indeed, the informativeness of summary

[5] We could still imagine some supervised tasks in this setup, such as observing a joint quantity like the distribution of stars formed. But if the inference procedure does not include an unsupervised assimilation of the morphological properties, these will bring no additional information (to the quantity used for the supervision).

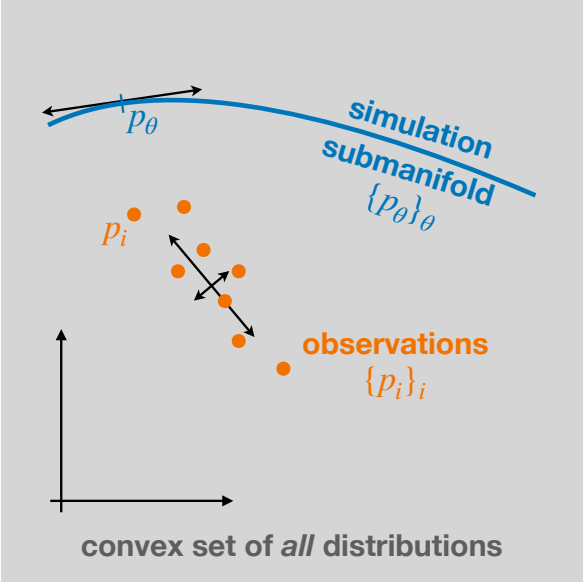


Figure 4.1: For complex processes such as molecular cloud dynamics, the submanifold of simulations is likely to differ from the set of observed processes (Falgarone et al., 2004). The supervision of simulations, if used alone, can lead to misleading characterization of the observations (Nunes & Balding, 2010) or estimation of distance to simulations (Anderson & Burnham, 2002).

statistics is highly dataset dependent (Blum et al., 2013; Nunes & Balding, 2010). Further quantitative results about this statement are given on ISM data in the next chapter.

Hence, for these two tasks, there is an interest to complement the simulation-supervised framework with a step of direct learning from the unsupervised observations.

Consequences

We consider an unlabelled collection of data $\{x_i\}_i$, sampled from various unknown processes $X_i \sim p_i$, and we aim at comparing these processes (i.e. to know whether there are pairs or clusters of processes that are similar). Here i stands for a simple data index, containing no physical meaning. This brings to two major issues:

- the notions of insufficiency and informativeness introduced previously are ill-defined as the index i cannot play the same role as θ. For instance i does not identify a process: it may be possible to have $i \neq j$ but simultaneously $p_i = p_j$. More generally the indexes (i, j) do not provide any information about the actual similarity between p_i and p_j, which is precisely what we aim at estimating.

- we have to estimate the similarities between processes from only one training sample x_i per process p_i. Indeed, contrary to supervised frameworks where samples x_i and x_j that have same labels $\theta_i = \theta_j$ can be regrouped in the same class, associated to $p_{\theta_i} = p_{\theta_j}$, we have no way to regroup the available data into classes.

The first issue will be addressed in this chapter, while the second one will be the focus of the next chapter. However, as may be already anticipated by the reader, the problem presented here is ill-posed without further assumptions. Indeed, p_i and p_j are each seen only through one sample so they are both extremely poorly constrained, resulting in general in any similarity level possible. Even though this issue will not be the focus of this chapter, we briefly clarify here the assumptions that will be made in the following chapter. They consist in adding further

constraints (or prior information) on the unknown processes $\{p_i\}_i$ that we are interested in. For instance, if we assume that p_i is stationary, with a correlation length small compared to the size of the data element x_i, this means that from the single observation x_i, some statistical properties of p_i can be constrained, as discussed in Sec. 1.2 of Chap. 3. With such assumptions, non-trivial bounds can be derived on the the similarity between the processes, even when working *a priori* in this *single data regime*, as shown in the next chapter in the context of molecular cloud observations.

In what follows, we will see how the notions of insufficiency and informativeness of a set of summary statistics can be extended to the case of a nonparametric family of distributions $\{p_i\}_i$. These will lead to information characterizations that involve manipulations of the densities $x \mapsto p_i(x)$. We assume here that this densities are given and we will tackle the challenge of estimating this characterizations from a very low amount of data in the next chapter.

4 From parameter-based information to dissimilarity contraction between pairs of processes

In order to measure the informative level of a summary statistic ϕ with respect to a nonparametric family (i.e. raw collection) of processes $\{p_i\}_i$, we proceed as follows:

- first we show that the sufficiency of ϕ for a whole family $\{p_i\}_i$ can be characterized from pairwise sufficiency, which means the sufficiency of ϕ for each binary family $\{p_i, p_j\}$.
- Then we show that, in the context of a binary family $\{p_i, p_j\}$, the informativeness of ϕ as probed by the Data Processing Inequality (DPI) (Eq. 4.13) boils down to a measure of the contraction, induced by ϕ, of the dissimilarity between the two processes, with the dissimilarity being quantified by the so called *Jensen-Shannon divergence*.
- Then we show that this characterization of informativeness through dissimilarity contraction can be extended to (but not further) a broad class of measures of dissimilarity called the *f-divergences*. This class includes in particular the Kullback-Leibler divergence and the total variation distance.
- We finally motivate the relevance of the total variation to perform such a measure of informativeness by linking it with the optimal accuracy of the $\{p_i, p_j\}$ classification problem.

4.1 From family sufficiency to pairwise sufficiency

The sufficiency of ϕ for the family $\{p_i\}_i$ is equivalent[6] to the combination of the pairwise sufficiency of ϕ for each binary family $\{p_i, p_j\}$ (Halmos & Savage, 1949). Therefore, the informative level of ϕ for a given dataset of processes $\{p_i\}_i$ can be broken down into smaller pieces. But how to characterize pairwise sufficiency, or more importantly pairwise insufficiency?

[6] If the family of stochastic processes considered is not a dominated set of measures, case that we exclude all along this document, the reciprocal implication does not hold: pairwise sufficiency does not necessarily imply sufficiency. See Halmos and Savage, 1949 for a counter-example.

4.2 From pairwise sufficiency to dissimilarity contraction

We aim at characterizing the deviation to sufficiency of ϕ in the case of a family made of two distributions $\{p_1, p_2\}$.

4.2.1 DPI for mutual information as a contraction of Jensen-Shannon divergence

We showed that the DPI (Eq. 4.13) is a way to characterize the insufficiency of ϕ through the reduction it induces in mutual information $I(\theta, \phi(X)) \leq I(\theta, X)$. Let us investigate the form that takes the mutual information $I(\theta, X)$ in the particular case of a binary family of processes $\{p_i\}_{i\in\{1,2\}}$ with a uniform distribution on the random index that we denote here θ:

$$p_\theta(1) = p_\theta(2) = 1/2.$$

We thus have:

$$p_{X|\theta=i}(x) \equiv p_i(x), \tag{4.20}$$

$$p_{\theta,X}(i, x) = p_{X|\theta=i}(x) \cdot p_\theta(i) = p_i(x) \cdot 1/2, \tag{4.21}$$

and p_X, the density of the random variable X marginalized over θ, is the mixture of p_1 and p_2:

$$p_X(x) = \sum_{i\in\{1,2\}} p_\theta(i) \cdot p_{X|\theta=i}(x) = \big(p_1(x) + p_2(x)\big)/2. \tag{4.22}$$

With this structure of joint distribution $p_{\theta,X}$, we can expand the mutual information computation (defined in Eq. 4.10) and observe that it boils down to a measure of dissimilarity between the two distributions of the family $\{p_1, p_2\}$:

$$\begin{aligned}
I(\theta, X) &= H(X) - H(X|\theta) && (4.23)\\
&= -\int p_X(x) \cdot \log\, p_X(x)\, \mathrm{d}x \quad + \sum_{i\in\{1,2\}} \int p_{\theta,X}(i, x) \cdot \log\, p_{X|\theta=i}(x)\, \mathrm{d}x && (4.24)\\
&= -\int \frac{p_1(x) + p_2(x)}{2} \cdot \log\, \frac{p_1(x) + p_2(x)}{2}\, \mathrm{d}x \quad + \frac{1}{2} \sum_{i\in\{1,2\}} \int p_i(x) \cdot \log\, p_i(x)\, \mathrm{d}x && (4.25)\\
&= \frac{1}{2} \sum_{i\in\{1,2\}} \int p_i(x) \cdot \log\, \frac{p_i(x)}{(p_1(x) + p_2(x))/2}\, \mathrm{d}x && (4.26)\\
&\equiv D_{JS}(p_1||p_2). && (4.27)
\end{aligned}$$

In this computation, we identify D_{JS} as the *Jensen-Shannon* (JS) divergence. It is a measure of dissimilarity between the two probability distributions p_1 and p_2. It is symmetric, vanishes if and only if $p_1 = p_2$. Its square root defines a metric (Endres & Schindelin, 2003).

Now, applying the same computations but replacing X by $\phi(X)$ yields:

$$I(\theta, \phi(X)) = D_{JS}(\phi_\# p_1||\phi_\# p_2). \tag{4.28}$$

Therefore, the mutual information $I(\theta, \phi(X))$ boils down to a *measure of dissimilarity between the reduced distributions* $\phi_{\#}p_1$ and $\phi_{\#}p_2$. The closer these distributions get, the less informative is ϕ. With this novel perspective on the mutual information for this binary case, the Data Processing Inequality (DPI) (Eq. 4.13) becomes (cf. next subsection for references on these results):

$$D_{JS}(\phi_{\#}p_1||\phi_{\#}p_2) \leq D_{JS}(p_1||p_2), \tag{4.29}$$

and:

$$\phi \text{ is sufficient for } \{p_1, p_2\} \quad \Longleftrightarrow \quad D_{JS}(\phi_{\#}p_1||\phi_{\#}p_2) = D_{JS}(p_1||p_2). \tag{4.30}$$

Thus, a way to quantify the informative level of a set of summary statistics ϕ for a nonparametric family of processes $\{p_i\}_i$ is to evaluate, for every pair $\{p_i, p_j\}$, what we can call the *contraction of dissimilarity* as probed by the Jensen-Shannon divergence:

$$\eta_{JS} \equiv D_{JS}(\phi_{\#}p_1||\phi_{\#}p_2) \,/\, D_{JS}(p_1||p_2) \tag{4.31}$$

induced by the compression of $X \sim p_i$ to $\phi(X) \sim \phi_{\#}p_i$. Note that this contraction lies in fact in $[0, 1]$ and cannot be increased by an artificial expansion of ϕ. Indeed, Eq. 4.30 directly implies[7] that any invertible transformation ψ applied after ϕ does not affect the Jensen-Shannon divergence:

$$D_{JS}\big((\psi \circ \phi)_{\#}p_1||(\psi \circ \phi)_{\#}p_2\big) = D_{JS}(\phi_{\#}p_1||\phi_{\#}p_2). \tag{4.33}$$

4.2.2 Extension to f-divergences

The Jensen-Shannon divergence is not the only divergence whose contraction can be used to measure informativeness. Indeed, it is a particular instance of a broader class of dissimilarity measurements called *f-divergences* that share interesting properties for information geometry. These divergences, introduced by Ali and Silvey, 1966; Csiszár, 1964; Rényi, 1961, are of the form:

$$D_f(p_1||p_2) \equiv \int f\left(\frac{p_1(x)}{p_2(x)}\right) p_2(x)\,\mathrm{d}x, \tag{4.34}$$

where the generator $f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ is a *convex* function. The Jensen-Shannon divergence corresponds to the particular choice $f : t \mapsto (t \log t - (t+1)\log[(t+1)/2])/2$. Other choices for f generate well-known divergences. Among these, we might mention the *Kullback-Leibler* (KL) divergence Kullback and Leibler, 1951, generated by $f : t \mapsto t \log t$:

$$D_{KL}(p_1||p_2) \equiv \int p_1(x)\, \log \frac{p_1(x)}{p_2(x)}\,\mathrm{d}x \quad \geq 0, \tag{4.35}$$

[7] Indeed, if ψ is invertible then it is sufficient for $\{\phi_{\#}p_1, \phi_{\#}p_2\}$, so by data processing equality:

$$D_{JS}\big((\psi \circ \phi)_{\#}p_1||(\psi \circ \phi)_{\#}p_2\big) = D_{JS}(\psi_{\#}(\phi_{\#}p_1)||\psi_{\#}(\phi_{\#}p_2)) = D_{JS}(\phi_{\#}p_1||\phi_{\#}p_2). \tag{4.32}$$

and the *Total Variation* (TV) distance, based on the L^1-norm between p_1 and p_2, yielded by $f : t \mapsto |t-1|/2$:

$$D_{TV}(p_1||p_2) \equiv \frac{1}{2}\int |p_1(x) - p_2(x)|\,\mathrm{d}x \quad \in [0,1]. \tag{4.36}$$

These f-divergences are measures of dissimilarity that comply with and therefore generalize the DPI of Eq. 4.29, and its equality case holds if[8] and only if[9] the statistics are sufficient (generalizing by so Eq. 4.30) (Liese & Vajda, 2006). Conversely, the f-divergences are the only[10] divergences that satisfy the DPI (Pardo & Vajda, 1997).

Hence, these f-divergences allow to broaden the measures of informativeness we might consider. In particular, we introduce the f-informativeness of a set of summary statistics ϕ with respect to a pair $\{p_1, p_2\}$ of processes based on the contraction coefficient η_f (Sason, 2019):

$$\eta_f[\phi, \{p_1, p_2\}] \equiv D_f(\phi_\# p_1||\phi_\# p_2) \,/\, D_f(p_1||p_2) \quad \in [0,1]. \tag{4.37}$$

As explained below, the TV-based measure of information η_{TV} is of particular interest since it represents the optimal accuracy that a classifier could reach.

4.3 Total Variation contraction coefficient: a measure of optimal accuracy reduction

The total variation distance is a particularly relevant choice of f-divergence since it is linked to the optimal accuracy reachable for the classification task (defined in Sec. 4.2.1 of Chap. 2), in the binary case: $i \in \{1,2\}$. To exhibit this link, we assume as previously that the two classes

[8] The proof for the implication *sufficient* $\implies$ *equality in the DPI* holds in few lines with the tools introduced so far, so we detail it for the interested reader. Since ϕ is assumed sufficient, there exist g_1, g_2 and h nonnegative mappings verifying Fisher-Neyman's factorization:

$$p_{1,2}(x) = g_{1,2}\big(\phi(x)\big) \times h(x).$$

For simplicity, we further assume that $p_{1,2}(x) > 0$. Then, applying the divergence D_f defined in Eq. 4.34, the factorization cancels h so the result depends only on the random variable $\varphi \equiv \phi(X) \sim \phi_\# p_2$:

$$D_f(p_1||p_2) = \int f\left(\frac{g_1(\phi(x))}{g_2(\phi(x))}\right) p_2(x)\,\mathrm{d}x \equiv \mathbb{E}_{X\sim p_2}\, f\left(\frac{g_1(\phi(X))}{g_2(\phi(X))}\right) = \mathbb{E}_{\varphi\sim\phi_\# p_2}\, f\left(\frac{g_1(\varphi)}{g_2(\varphi)}\right),$$

with the latter equality stemming from the LOTUS formula. Finally, for φ in the support of $\phi_\# p_2$ we have:

$$\frac{\phi_\# p_1(\varphi)}{\phi_\# p_2(\varphi)} = \frac{\int \delta_{\phi(x)-\varphi} g_1(\varphi) h(x)\,\mathrm{d}x}{\int \delta_{\phi(x)-\varphi} g_2(\varphi) h(x)\,\mathrm{d}x} = \frac{g_1(\varphi)}{g_2(\varphi)},$$

where δ denotes the Dirac delta distribution in dimension $\dim\phi$, so we obtain:

$$\mathbb{E}_{\varphi\sim\phi_\# p_2}\, f\left(\frac{g_1(\varphi)}{g_2(\varphi)}\right) = \mathbb{E}_{\varphi\sim\phi_\# p_2}\, f\left(\frac{\phi_\# p_1(\varphi)}{\phi_\# p_2(\varphi)}\right) \equiv D_f(\phi_\# p_1||\phi_\# p_2),$$

in order to finally verify $D_f(p_1||p_2) = D_f(\phi_\# p_1||\phi_\# p_2)$.

[9] Assuming that f is strictly convex.

[10] Among the class of *decomposable* divergences (i.e., that can be written as $D(p_1||p_2) = \int d\big(p_1(x), p_2(x)\big)\,\mathrm{d}x$, or $\sum_{x\in X(\Omega)} d\big(p_1(x), p_2(x)\big)$ in the finite case $\#X(\Omega) < \infty$, for some function d), except the case $\#X(\Omega) = 2$. Otherwise, there are non decomposable divergences, or divergences acting over probabilities defined on a binary alphabet that verify the DPI without belonging to f-divergences (Polyanskiy & Verdú, 2010).

have the same probability so the data is sampled as: $p_{\theta,X}(i,x) = p_{X|\theta=i}(x) \cdot p_\theta(i) = p_i(x)/2$.

As introduced in Sec. 4.2.1 of Chap. 2, the goal is to build a classifier $c : x \mapsto c(x) \in \{1,2\}$ that maximizes the accuracy $\mathcal{A}(c)$:

$$\mathcal{A}(c) \equiv \mathbb{E}_{\theta,X\sim p_{\theta,X}}\left[1_{c(X)=\theta}\right] = \int \sum_{i=1}^{2} p_i(x) 1_{c(x)=i}\, \mathrm{d}x/2. \tag{4.38}$$

We have shown (Eq. 2.17) that a classifier c^* defined such that

$$p_{c^*(x)}(x) = \max_{i\in\{1,2\}} p_i(x), \tag{4.39}$$

yields the optimal accuracy:

$$\text{for any classifier } c, \quad \mathcal{A}(c) \leq \mathcal{A}(c^*) \equiv \mathcal{A}^*. \tag{4.40}$$

In the literature, c^* is referred as *a Bayes classifier*[11] and $1 - \mathcal{A}^*$ as *the Bayes error rate*. Let us use one such classifier, c^*, verifying Eq. 4.39 to compute $\mathcal{A}^*$. For any x we have:

$$\sum_{i=1}^{2} p_i(x) 1_{c^*(x)=i} = \max\left\{p_1(x), p_2(x)\right\}, \tag{4.41}$$

and observing the general identity:

$$\max\left\{p_1(x), p_2(x)\right\} = \frac{1}{2}\Big[|p_1(x) - p_2(x)| + p_1(x) + p_2(x)\Big], \tag{4.42}$$

we obtain, using Eq. 4.38, 4.41 and 4.42, $\mathcal{A}^* \equiv \mathcal{A}(c^*) = \int \left(|p_1(x) - p_2(x)| + p_1(x) + p_2(x)\right) \mathrm{d}x/4$, and finally:

$$\mathcal{A}^* = \frac{D_{TV}(p_1||p_2) + 1}{2} \quad \in [50\%, 100\%]. \tag{4.43}$$

Hence, if p_1 and p_2 have disjoint supports, then the accuracy is maximal: $\mathcal{A}^* = 100\%$, and so is the TV distance: $D_{TV}(p_1||p_2) = 1$. Inversely, when the TV distance is minimal: $D_{TV}(p_1||p_2) = 0$, then $p_1 = p_2$ so the classes are identical and an optimal strategy cannot beat the 50% accuracy imposed by the random class selection $p_\theta(i) = 1/2$ of the data generator.

The total variation distance thus gives a direct measure of the best discrimination performance between two classes that can be carried out from a sample. If we are able to learn perfectly the distributions p_1 and p_2, then the simple classification procedure $c(x) = 1$ if $p_1(x) \geq p_2(x)$ else $c(x) = 2$ achieves this optimal performance, and no other procedure can beat it in average. Hence, for a given set of summary statistics ϕ, the contraction coefficient $\eta_{TV} \equiv D_{TV}(\phi_\# p_1||\phi_\# p_2)/D_{TV}(p_1||p_2)$ translates how much of this theoretical maximum discrimination power we lose by compressing X into $\phi(X)$.

[11] In the general case where the class index distribution p_θ is not especially uniform, c^* should be such that $p_{c^*(x)}(x) = \max_{i\in\mathcal{I}}\left[p_i(x) \cdot p_\theta(i)\right]$, which modifies the computation of $\mathcal{A}^*$ derived here.

5 Application: what statistics to discriminate between flat log-FBMs?

We now apply the total variation distance to show how Gaussian statistics (defined in Eq. 4.46) can fail at discriminating flat log-FBM (Fractional Brownian Motion) processes that are (optimally) separated by log-Gaussian statistics (also defined in Eq. 4.46).

Log-FBMs are statistical models widely used by the ISM community to reproduce high dimensional fields such as images or 3D cubes that display a log-normal PDF and show spatial correlations in the form of a power law. These models consist in exponentiating a Gaussian process Z, with Z having a power-law power spectrum. More details and references about these are given in Sec. 8.1.2 of the following chapter.

We restrain the scope of this application to the family of flat log-FBMs, i.e., generated by exponentiating a stationary Gaussian random field that has no spatial correlations, therefore having a flat power spectrum. In fact, we consider the following parametric family of processes:

$$X \sim \log \mathcal{N}(\mu, \sigma^2 \mathbb{I}_d), \quad \text{with } d \gg 1, \tag{4.44}$$

defined in dimension d. So that $X \equiv \{X(r)\}_{1 \leq r \leq d}$ mimics a high dimensional field such as an image, we take $d = 32 \times 32 = 1024$. We choose not to consider spatial correlations so that we can build a vector of sufficient summary statistics $\phi_{\text{log-Gaussian}} : \mathbb{R}^d \to \mathbb{R}^2$ of dimension only 2, defined as:

$$\phi_{\text{log-Gaussian}} \equiv \phi_{\text{Gaussian}} \circ \log \quad \text{with} \quad \phi_{\text{Gaussian}} : X \mapsto \begin{pmatrix} \langle X_r \rangle_r \\ \langle (X_r)^2 \rangle_r - \langle X_r \rangle_r^2 \end{pmatrix}, \tag{4.45}$$

that is sufficient[12] for the family of flat log-FBMs. This high compression allows to visualize the separation frontier between pairs (p_1, p_2) of high dimensional log-FBMs without loss of information (cf. Fig. 4.2) and to estimate numerically very efficiently their actual total variation distance through the following data processing equality:

$$D_{TV}(p_1 || p_2) = D_{TV}(\phi_{\text{log-Gaussian}\#} p_1 || \phi_{\text{log-Gaussian}\#} p_2), \tag{4.50}$$

[12] Indeed, the mapping $x \mapsto \log x$ is invertible, so sufficient and so verifies the data processing equality, for $p_{1,2} = \log \mathcal{N}(\mu_{1,2}, \sigma^2_{1,2} \mathbb{I}_d)$:

$$D_{TV}(p_1 || p_2) = D_{TV}(\log_\# p_1 || \log_\# p_2) \tag{4.46}$$

$$\equiv D_{TV}\Big(\mathcal{N}(\mu_1, \sigma_1^2 \mathbb{I}_d) \,\Big|\Big|\, \mathcal{N}(\mu_2, \sigma_2^2 \mathbb{I}_d)\Big). \tag{4.47}$$

Since, we have shown in Sec. 2.1.1 that ϕ_{Gaussian} is sufficient for the family $\{\mathcal{N}(\mu, \sigma^2 \mathbb{I}_d)\}_{\mu,\sigma}$, we obtain, pursuing Eq. 4.47:

$$D_{TV}(p_1 || p_2) = D_{TV}(\phi_{\text{Gaussian}\#} \log_\# p_1 || \phi_{\text{Gaussian}\#} \log_\# p_2) \tag{4.48}$$

$$\equiv D_{TV}(\phi_{\text{log-Gaussian}\#} p_1 || \phi_{\text{log-Gaussian}\#} p_2). \tag{4.49}$$

that simplifies the d-dimensional integral into a 2-dimensional one.

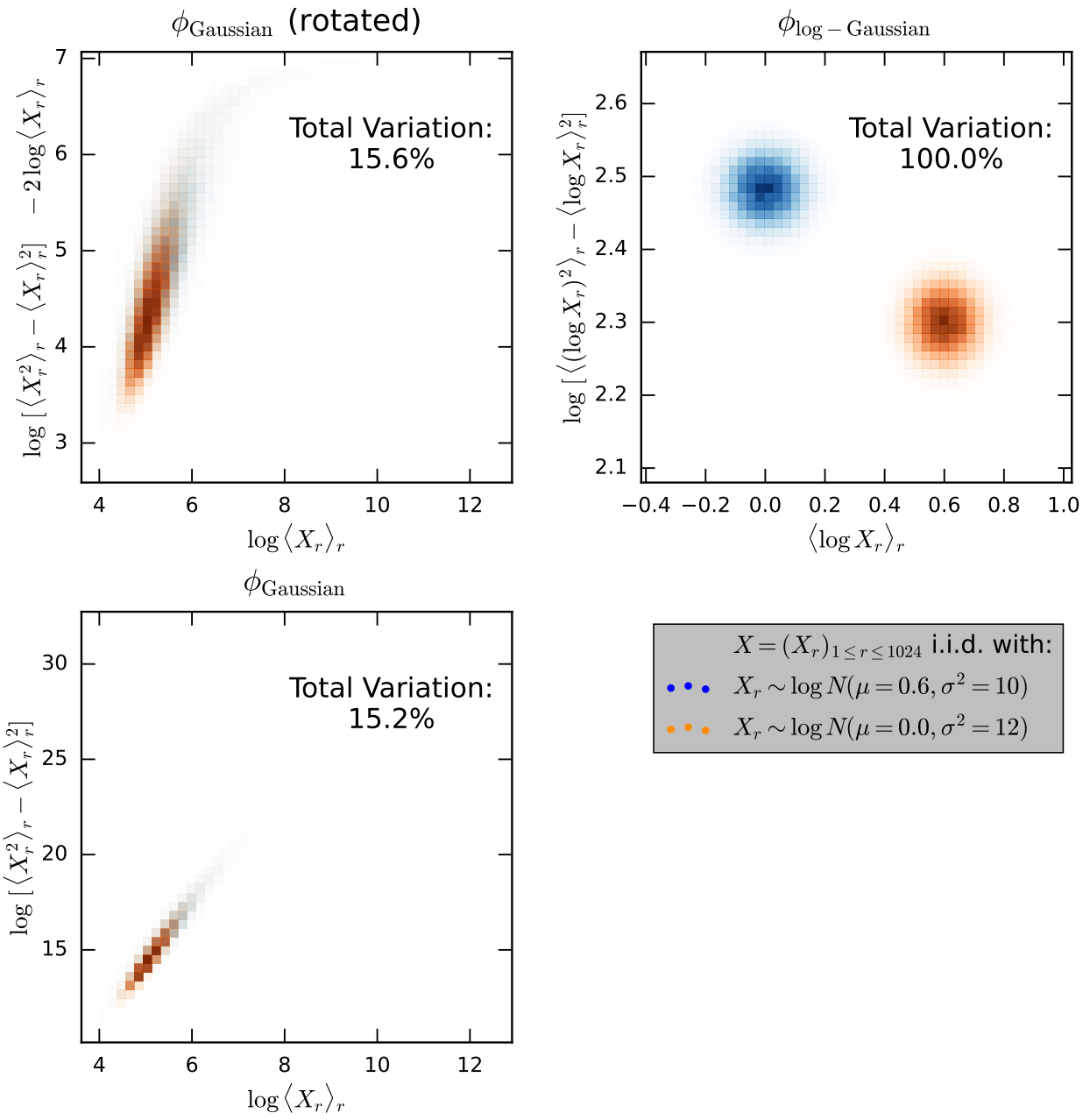


Figure 4.2: Gaussian (left) and log-Gaussian (right) statistics to discriminate between two flat log-FBMs. Gaussian statistics mix the two distributions while log-Gaussian statistics exhibit a clear separation that allows to classify between the processes with $\sim 100\%$ accuracy despite its reduction of the data from $d = 1024$ to $d = 2$ (reduction that is proved to drop no information for classification purpose). Note that some axes in these plots might differ from the definition given in Eq. 4.45, but are associated to invertible transformations applied after the summary reduction and therefore do not affect the results up to numerical estimation errors (cf. Eq. 4.32 that can be generalized to any f-divergence).

Yet, the statistics ϕ_{Gaussian} are not sufficient over the global family $\{\log \mathcal{N}(\mu, \sigma^2 \mathbb{I}_d)\}_{\mu,\sigma}$. Indeed, as illustrated in Fig. 4.2, we can exhibit a pair $\{p_1, p_2\}$ of flat log-FBMs for which

the total variation distance $D_{TV}(\phi_{\text{Gaussian}\#p_1}||\phi_{\text{Gaussian}\#p_2}) \simeq 16\%$ estimated numerically is much smaller than the 100% one estimated for $\phi_{\text{log-Gaussian}}$. This translates in the figure as a strong overlapping between the distributions $\phi_{\text{Gaussian}\#p_1}$ and $\phi_{\text{Gaussian}\#p_2}$. If we were to classify between p_1 and p_2 by accessing x only through $\phi_{\text{Gaussian}}(x)$, that is its empirical mean and variance, the optimal accuracy that we could reach would be $(0.16+1)/2 \simeq 57\%$ (Eq. 4.43), while the frontier between these distributions is clear in the $\phi_{\text{log-Gaussian}}$ space that would lead to an accuracy of $\sim 100\%$.

We extend this insufficiency analysis of the Gaussian statistics to the following pairs of flat log-FBMs:

$$p_1[\delta] = \log \mathcal{N}(\mu = 0, \quad \Sigma = 10 \times (1+\delta)), \tag{4.51}$$

$$p_2[\delta] = \log \mathcal{N}(\mu = 3 \times \delta, \quad \Sigma = 10), \tag{4.52}$$

with $\delta \in \{0, 0.02, 0.05, 0.1, 0.2, 0.5\}$.

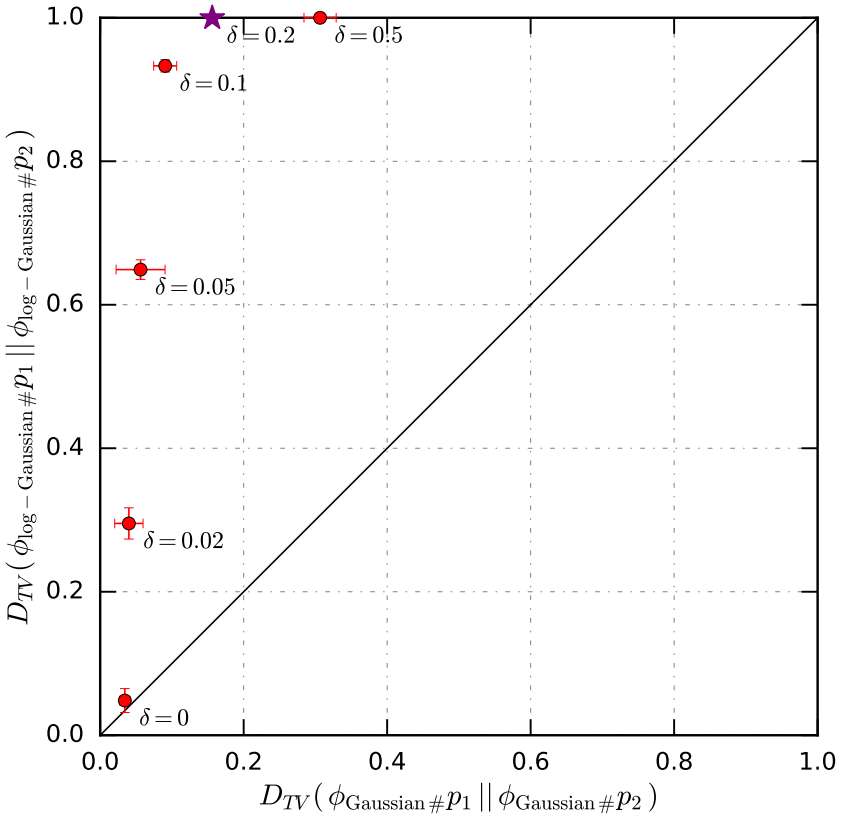


Figure 4.3: Gaussian *vs* log-Gaussian statistics to discriminate between two flat log-FBMs. The purple star corresponds to the pair of distributions reported in Fig. 4.2. Numerical set up: Total Variation (TV) distance is estimated by drawing 10^5 samples (each one having a dimension 1024) per distribution that are then binned into 50×50 bins as shown in Fig. 4.2. We run this estimation 10 times to estimate a standard deviation. The latter is small so magnified by 10 and then reported as an errorbar. We observe a small bias of our TV estimation at low TV values since the point associated to the pair $\delta = 0$ should be located at $(0, 0)$ in this plot. This bias is likely due to the finite number of samples used to estimate the binned distributions.

We confront in Fig. 4.3 the TV distance obtained by ϕ_{Gaussian} and by $\phi_{\text{log-Gaussian}}$ for these pairs. The off-diagonal position of a point in this plot indicates the existence of a pair $\{p_1, p_2\}$ for which, at least one of two sets of summary statistics, is detected to be insufficient. Here, we clearly see that ϕ_{Gaussian} looses discriminative power for all the pairs considered.

Conclusions

In this chapter we have motivated and set the theoretical ground to quantify informativeness of summary statistics in an unsupervised regime. We suggest to decompose the analysis of the associated collection of processes $\{p_i\}_i$ to multiple pair-wise discrimination tasks $\{p_i, p_j\}$. To do so, we introduced the contraction coefficient $\eta_{TV}[\phi, \{p_1, p_2\}]$. However, such measurements require to estimate $D_{TV}(p_1||p_2)$ and $D_{TV}(\phi_{\#}p_1||\phi_{\#}p_2)$, while as a consequence of lack of supervision, we often have to work in a very low data regime where we have few samples x_i per process p_i. In the next chapter, we show how to estimate informativeness in practice, in such a regime and apply this methodology to characterize the morphology of molecular clouds from an observational base.

Chapter 5

Comparing Molecular Clouds' morphology without supervision

Objectives

We analyse the statistical properties of Molecular Clouds using column density maps from *Herschel* observations. In a first part we question what summary statistics are the most informative to characterize this unlabeled set of observed images. To answer this, we use the approach described in the previous chapter for the unsupervised case. However, due to this lack of supervision, we have to work *a priori* with only one sample per process. This additional difficulty leads us to develop a simplified approach based on further assumptions in statistical space, notably ergodicity. In a second part we use the summary statistics identified in the first step to build a morphological distance to compare observations, simulations and statistical models of molecular clouds.

Contents

Contextualizing the paper in this manuscript

This chapter is based on a paper I led as a first author, submitted to A&A the 13^{th} of July 2024 entitled *Molecular clouds: do they deserve a non-Gaussian description?* It received a positive recommendation for publication after revisions that are currently being carried out.

In this paper, we address the problem of characterizing the morphology of column density maps of molecular clouds without relying on supervised knowledge. In order to do so, we aim at comparing the informative level between various sets of summary statistics. In the previous chapter, we suggest a proper definition for this informative level of a given summary statistics ϕ, based on the contraction coefficients $\{\eta_{TV}[\phi, \{p_i, p_j\}]\}_{i,j}$ that measure how the dissimilarity between two processes is reduced once these are compressed through ϕ. However, these coefficients are defined as the ratios of $D_{TV}(\phi_\# p_i || \phi_\# p_j)$ to $D_{TV}(p_i||p_j)$ that require *a priori* high dimensional integrals to compute (the first being nevertheless reduced to $\dim \phi$) and that are inaccessible with the few data we rely on.

Instead, we suggest not to compute $D_{TV}(p_i||p_j)$ and to focus only on $D_{TV}(\phi_\# p_i||\phi_\# p_j)$ which is simpler to probe. This quantity alone does not allow to estimate the absolute informativeness level of a set of summary statistics ϕ, but allows a relative comparison of these levels between various sets of summary statistics, as $D_{TV}(p_i||p_j)$ is a common denominator to all these levels. Nevertheless, the quantity $D_{TV}(\phi_\# p_i||\phi_\# p_j)$ remains very challenging to estimate in a low data regime. Instead, we reduce this measure of dissimilarity between $\phi_\# p_i$ and $\phi_\# p_j$ into a simplified statistical test of compatibility between these two reduced processes, motivated by the unimodal form that these should take due to the averaging properties of ϕ.

For this reduced test to detect an incompatibility between a given pair of actually different non reduced processes $\{p_i, p_j\}$, the summary statistics ϕ must satisfy a trade-off between being informative and having low variance. This therefore motivates the title of this paper, since highly informative but expansive and not compressed descriptions are not warranted to perform well in order to compare the morphology of molecular clouds in the low data regime set by observations.

Molecular clouds: do they deserve a non-Gaussian description?

Abstract Molecular clouds show complex structures reflecting their non-linear dynamics. Many studies, investigating the bridge between their morphology and physical properties, have shown the interest provided by non-Gaussian higher-order statistics to grasp physical information. Yet, as this bridge is usually characterized in the supervised world of simulations, transferring it onto observations can be hazardous, especially when the discrepancy between simulations and observations remains unknown. In this paper, we aim at identifying relevant summary statistics directly from the observation data. To do so, we develop a test that compares the informative power of two sets of summary statistics for a given dataset. Contrary to supervised approaches, this test does not require the knowledge of any data label or parameter, but focuses instead on comparing the degeneracy levels of these descriptors, relying on a notion of statistical compatibility. We apply this test to column density maps of 14 nearby molecular clouds observed by Herschel, and iteratively compare different sets of usual summary statistics. We show that a standard Gaussian description of these clouds is highly degenerate but can be substantially improved when being estimated on the logarithm of the maps. This illustrates that low-order statistics, properly used, remain a very powerful tool. We then further show that such descriptions still exhibit a small quantity of degeneracies, some of which are lifted by the higher order statistics provided by reduced wavelet scattering transforms. This property of observations quantitatively differs from state-of-the-art simulations of dense molecular cloud collapse and is not reproduced by logfBm models. Finally we show how the summary statistics identified can be cooperatively used to build a morphological distance, which is evaluated visually, and gives very satisfactory results.

1 Introduction

Molecular clouds (MCs) play a key role in star formation. Their highly non-linear dynamics tends to couple spatial scales over a wide range, to create dense filamentary structures, in which clumps form that will eventually lead to prestellar cores (McKee & Ostriker, 2007). However, the precise role of each physical ingredient (such as turbulence, magnetic fields, and gravity, to name but a few) in this dynamics still remains to be fully understood.

A key step towards such an understanding certainly resides in our ability to decipher, in the morphology of these structured clouds, some signature of the physical processes at play. For instance, the power-law shape of the power spectrum in emission maps may trace properties of the turbulence (Federrath and Klessen, 2013; M.-A. Miville-Deschênes et al., 2007), or the shape of the probability distribution function (PDF) of column density maps (Appel et al., 2022; Federrath & Klessen, 2013; Hennebelle & Chabrier, 2008; Kainulainen et al., 2009; Schneider et al., 2022; Vázquez-Semadeni et al., 1997) may trace the impact of gravitational collapse or stellar feedback as it transitions from a log-normal shape to the development of a heavy tail.

However, the strong non-linear interplay between the physical processes at work leads to the emergence of non-Gaussian features in interstellar structures, such as filaments. This has motivated numerous studies to capture physical information beyond one-point and two-point

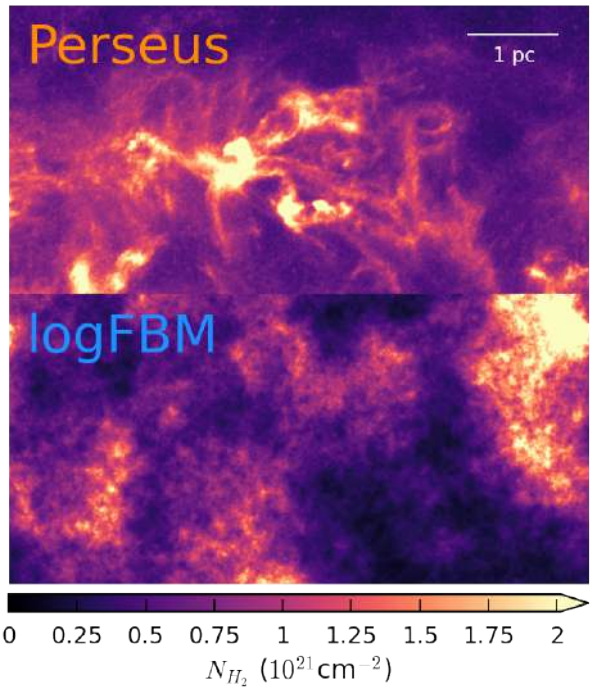


Figure 5.1: Column density map of a region in Perseus *(top row)* and sample of a log-Fractional Brownian Motion field *(bottom row)*. Each image has 256 × 512 pixels but is tiled into 2 complementary patches of size 256 × 256 on which the statistics shown in Fig. 5.2 are computed. While these one-point and two-point statistics (some being non-Gaussian) are clearly compatible, these two images have manifestly different morphologies.

statistics. Many approaches have been investigated including, for instance, diagnostics of the phase coherence of the Fourier modes (Burkhart & Lazarian, 2016; Levrier et al., 2006), bispectrum (Burkhart et al., 2009), structure functions (Heyer & Brunt, 2004), dendograms (Goodman et al., 2009), multiscale segmentation (Robitaille et al., 2019), scattering transforms (Allys et al., 2019; Regaldo-Saint Blancard et al., 2020; Saydjari et al., 2021), and neural networks (Peek & Burkhart, 2019; Zavagno et al., 2023).

If higher-order statistics represent an appealing way of refining the geometrical description of complex ISM structures, they are nevertheless subject to the following caveats:

1. They should not screen out the potential of low-order statistics, but should come as complementary diagnostics (Burkhart & Lazarian, 2016). Low-order descriptors should not be underestimated: it is not because the processes under study are highly non-Gaussian that low-order statistics are only marginally informative. Yet, in numerous studies, such easy-access information is not considered, or even intentionally discarded[1], to emphasize the contribution of high-order statistics.

2. Refining the statistical description with higher-order terms complicates the connection between the statistical descriptions and the physical processes and associated parameters. Such a connection is thus usually learned in the supervised world of simulations, but transferring it to observations can be hazardous, especially when their distance to simulations remains unknown (Falgarone et al., 2004; Peek & Burkhart, 2019).

In this paper, we aim at identifying what sets of summary statistics are relevant to characterize the diversity of observed molecular clouds, without relying on simulations nor prior

[1]The mean and variance of the data may be normalized, the power spectra flattened, and histograms (non-linearly) equalized.

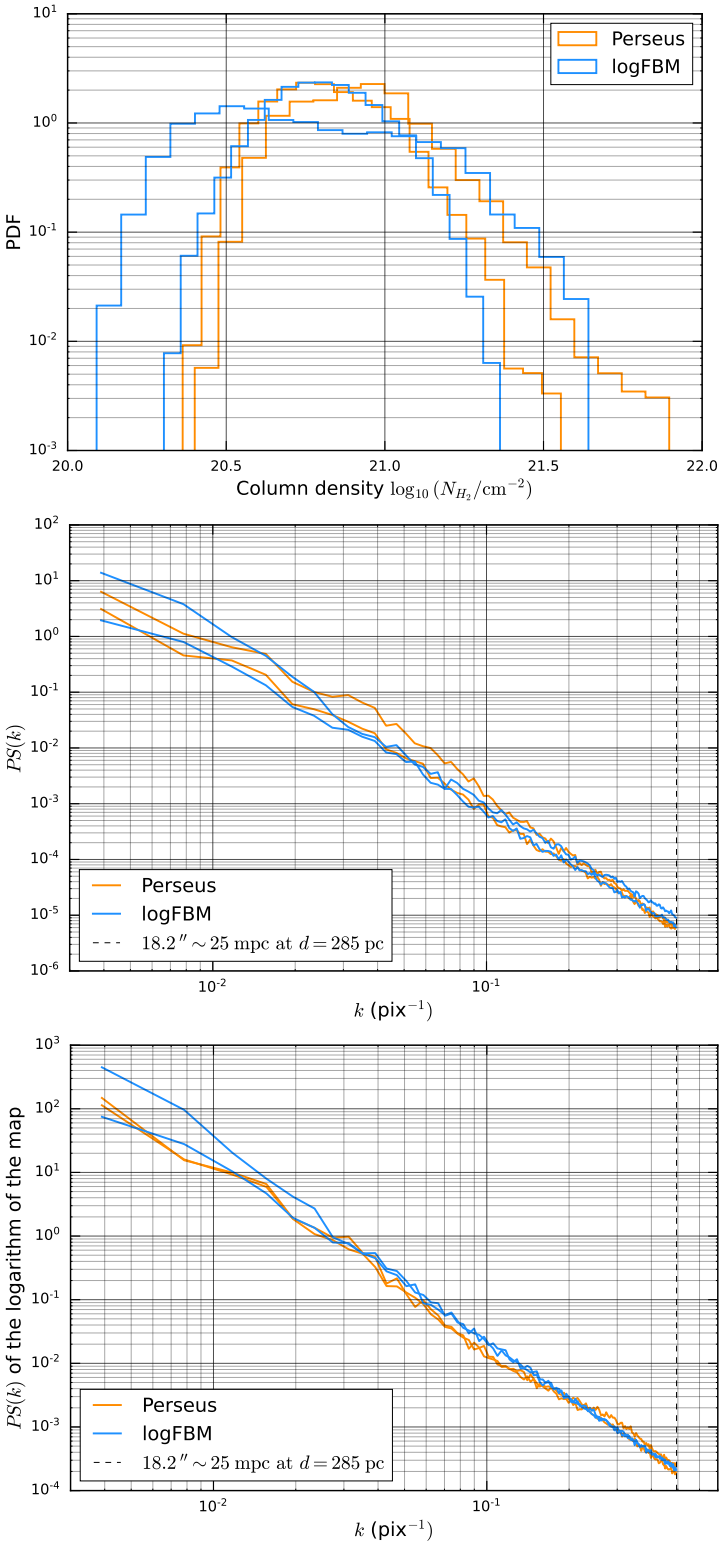


Figure 5.2: Statistical properties of the maps from Fig. 5.1: PDFs *(top)*, power spectra *(center)*, and power spectra of the logarithms of the maps *(bottom)*. The dashed vertical line in the power spectra plots is located at the $18.2''$ resolution of the column density map from the observational dataset. Power spectra are apodized as explained in Sec. 4. We see that the logFBM process is compatible with this observed portion of MC for all these statistics (some of theme being non-Gaussian), while having a manifestly different morphology, as revealed in Fig. 5.1. All these statistics are thus degenerate for such a comparison.

knowledge. In such an unsupervised framework, various observations cannot be grouped *a priori* in a similar class, so one often has to work in a very low data regime, which restricts dramatically the range of tools that can be used. This unsupervised and prior-knowledge free framework excludes features that require significant tuning to extract information, such as neural networks. This priority shift, from theoretical informativeness to actual information retrievable in this framework, might reshuffle the cards among the various sets of summary statistics to consider. This leads us to wonder to what extent non-Gaussian statistics can actually bring meaningful contributions in a fully observation-driven and unsupervised framework.

To answer this question, we rely on a dataset of column density maps constructed from a survey of nearby molecular clouds. Our approach is to identify the amount of degeneracy that a set of summary statistics can have on this dataset, i.e., finding a pair of maps in this dataset that are compatible according to these statistics, but actually have a different morphology. A simplified illustration of such a degeneracy is given in Figs. 5.1 and 5.2, where we compare an observed column density map with a realization of a field whose logarithm is sampled from a Fractional Brownian Motion, that is a specific type of Gaussian process with power-law type constrained power spectrum. In this example, the PDFs of these maps, their power spectra, and the power spectra of their logarithms, all seem to be compatible, even though the maps clearly have different structures. This illustration underlines the limitation of those statistics, but only for this specific comparison, where we compare one observation to a specific known model. However, in an unsupervised framework, when working solely with unlabeled observations, we will exhibit these limitations by introducing additional statistics that could be used to lift these degeneracies.

In this work, we use such a degeneracy diagnostic on the observed molecular clouds dataset to progressively build a robust set of summary statistics. We start with basic statistics, such as the power spectrum and one-point statistics. We evidence strong limitations that we investigate and substantially mitigate by taking the logarithm of the maps beforehand. We evidence further but moderate limitations of this improved set, leveraging higher-order statistics through the reduced wavelet scattering transform (RWST). This study, initially carried out only on the observational dataset, is then extended to logFBM data, as well as to a set of magnetohydrodynamics (MHD) simulations intended to reproduce observations of dense star-forming molecular clouds. We conclude by discussing the relevance and limitations of the final set of statistics we have constructed, and show how it can be used to assess the distance, in a statistical sense, between different maps of the interstellar medium.

The structure of the paper is as follows:

- In section 2, we present the set of MCs considered in this work and the corresponding column density maps, as well as the numerical simulations and the logFBM models.
- We present in section 3 the diagnostic of statistical compatibility that we will rely on, and our general methodology.
- We present in section 4 the sets of statistics that will be confronted.

- We apply this methodology in section 5 to confront these sets of statistics on observations, and from these results we design an informative and low-dimensional set of summary statistics $\phi_{\rm final}$.
- We finally define, in section 6, a distance based on $\phi_{\rm final}$ that allows us to compare datasets, such as observations and simulations.

2 Data

In this section, we present the main dataset that we use for this study: an ensemble of ∼550 molecular hydrogen (H_2) column density maps from a survey of nearby molecular clouds (MCs). Three other datasets will be used in this paper:

- a set of ∼230 total gas column density maps built from magnetohydrodynamical (MHD) numerical simulations of dense molecular clouds, classified into three subsets with varying values of the magnetic field, from null (*hydro*), to medium (*MHD*), and high (*MHD high B*). These simulations are state-of-the art attempts to reproduce the physics at play in the early stages of star formation, including self-gravity and decaying turbulence, and are therefore well adapted to compare to our observations of the HGBS clouds.
- A set of ∼500 synthetic maps sampled from logFBM models, whose parameters reproduce the diversity of the observational dataset. This type of purely synthetic fields has already been extensively used to model the interstellar medium (see, e.g., Brunt & Heyer, 2002; Elmegreen, 2002; Levrier et al., 2018; M.-A. Miville-Deschênes et al., 2007).
- A set of ∼1000 images from a large collection of everyday textures, the Describable Texture Dataset (DTD, Cimpoi et al., 2014). We use these to emphasize the specificity of ISM fields in the context of image texture analysis.

More details about these sets are given in appendix 8.1. All these maps have a size of 512 × 512 pixels.

2.1 Observations: column density maps from the HGBS

We focus our study on a set of MCs targeted by the Gould Belt Survey (HGBS) (André et al., 2010) with the *Herschel* Space Observatory (Pilbratt et al., 2010), whose footprints on the sky are shown in Fig. 5.3. Some of these clouds are also pointed out in the face-on view of Fig. 1.1.

The HGBS combines two main criteria we require for this work:

- it sampled numerous MCs with a diversity of physical and environmental conditions, from diffuse and quiescent regions with no sign of star-formation activity such as the Polaris Flare (André et al., 2010; Heithausen & Thaddeus, 1990; M. .-. Miville-Deschênes et al., 2010) to very dense and active ones such as Orion B (Schneider et al., 2013) or the Aquila Rift cloud (Könyves et al., 2015). Examples of molecular column density maps from the

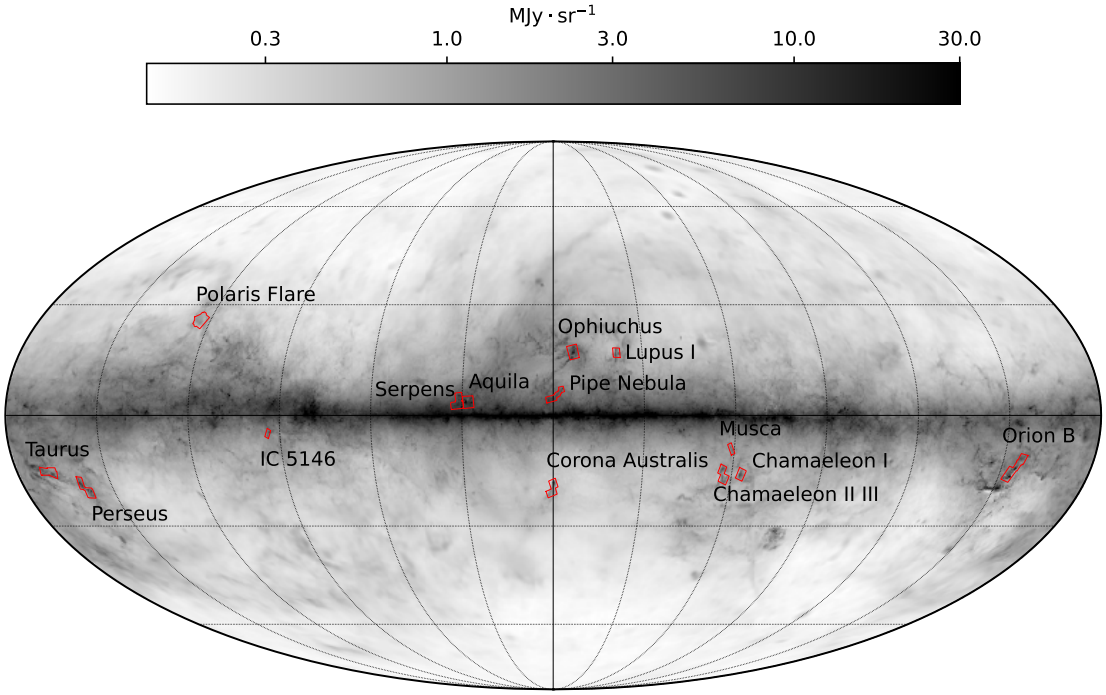


Figure 5.3: Footprints of the *Herschel* Gould Belt Survey (HGBS) fields used in this study, overlaid on the total thermal dust intensity at 353 GHz from the GNILC (Remazeilles et al., 2011) variable resolution data of *Planck* (Planck Collaboration et al., 2020c). Some of these clouds are also pointed out in the face-on view of Fig. 1.1.

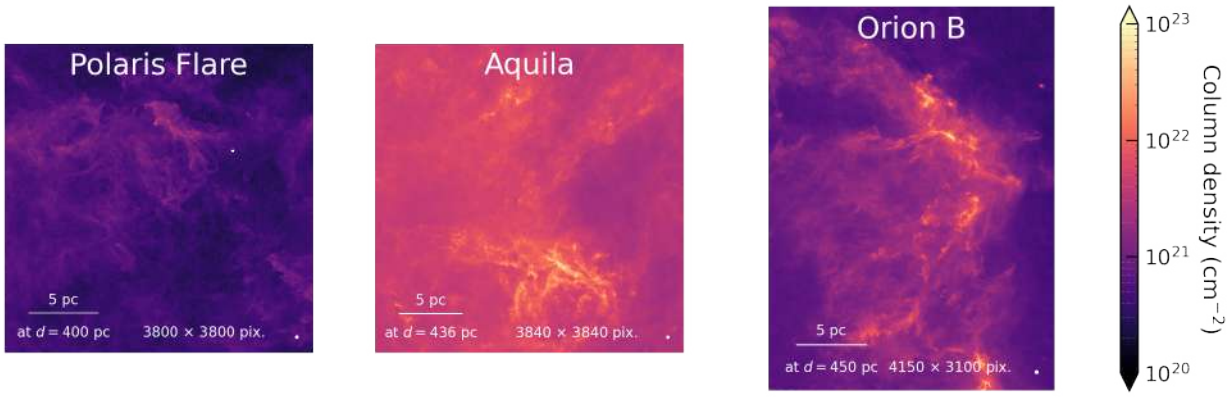


Figure 5.4: Column density maps of three regions from the HGBS. The spatial resolution of the maps is 18″, sampled with a pixel size of 3″. A white circle with diameter $5 \times 18''$ is shown at the bottom right corner of each map. Estimates of spatial scales, also shown on the maps, are based on reported distances (see Table 5.1). The Polaris Flare is an example of a diffuse and quiescent cloud, while Aquila and Orion B are dense and very active star-forming regions.

HGBS are shown in Fig. 5.4. We however emphasize that this survey is limited to local clouds (distances $d \leq 500$ pc) and does not cover the whole range of conditions expected in Galactic molecular clouds[2].

- It imaged a broad range of scales, from the full cloud size ($\sim$ 10 pc, corresponding to a few degrees for these nearby clouds) down to the $\sim$ 0.1 pc scales of filaments (André et al., 2014; Arzoumanian et al., 2011), which are spatially resolved for the nearest clouds[3], thus covering more than 2 orders of magnitude in spatial scales. This allows us to perform an in-depth morphological analysis, based on a local description of multi-scale interactions.

We consider the high resolution (18″) column density maps of 14 regions, produced with the procedure described in Appendix A of Palmeirim et al., 2013 and publicly available from the *Herschel* Gould Belt Survey Archive[4]. This 18″ angular resolution corresponds to a 12 mpc spatial resolution for the nearest clouds, such as Ophiuchus and Taurus ($d \sim 140$ pc), and up to 40 mpc for the most distant clouds, such as Orion B ($d \sim 450$ pc). The main properties of these clouds are given in Table 5.1.

2.2 Subsampling and tiling

The presence of physical processes such as gravity in the dynamics of MCs implies spatial variations of their statistical properties, which prevents us from modeling them as stationary stochastic processes. For instance, they usually have strong local overdensities, while being surrounded by diffuse borders. This non-homogeneity makes it difficult to compare such objects as a whole, as well as to properly estimate their statistical properties, when seen as realizations of a random process, as these properties will for instance depend heavily on the identification of the cloud boundary, or on the precise definition of what a cloud is. Furthermore, variance estimates for random processes generally rely on a homogeneity assumption to estimate the intrinsic variability of a process from its spatial variations. More broadly, the estimation of the statistical moments and properties of a process is usually based on spatial averaging, which assumes the homogeneity of the sample.

To avoid this problem, we choose to restrict the comparison to local patches of these MCs, assuming statistical homogeneity within each individual patch. This requires the identification of a characteristic stationarity length over which it can be assumed that the statistical properties of the cloud do not vary significantly. Different patches extracted from the same cloud may have different statistical properties, which will be representative of the non-homogeneity of the cloud as a whole. This encourages us to use small patches. However, an additional difficulty is that these patches must be large enough to make statistical estimates. In particular, variance estimates require sampling beyond the correlation length of the process under study.

In this paper, we have chosen to cut patches with angular sizes ranging from 0.85° to 2.55°, that correspond respectively to 3 pc and 9 pc for a cloud at a typical distance of 200 pc. We

[2] A further study could target more distant, more massive star forming regions with interferometric observations. We postpone such a study to a future paper.

[3] The typical resolutions range from 10″ at 70 μm to 36″ at 500 μm.

[4] http://www.herschel.fr/cea/gouldbelt/en/Phocea/Vie_des_labos/Ast/\ast_visu.php?id_ast=66

believe this represents a good compromise with regard to the trade-off expressed above. However, we are aware that there may not be an ideal solution, especially as we have no precise estimate of neither the stationarity length nor the correlation length. This means that we cannot rule out some confusion between local intrinsic variability and large-scale variations in the statistical properties of the clouds. However, we believe that, for lack of a better solution, this does not undermine the relevance of this study.

More precisely, we make multiple versions of the original (18″ resolution, 3″pixel size) column density maps at different pixel sizes (6″, 9″, 12″, and 18″). The down-sampling is done using a bivariate spline approximation of order 3 of the original maps. Then, we cut from these maps patches of size in pixels 512×512, with a step size of 256 pixels, such that two neighboring patches will have 50% of their pixels in common. In the following, these patches will be themselves tiled into 4 sub patches of size 256×256 to perform a variability estimation. The choice of 6″ for the finest sampling allows us to exploit the 18″ resolution of these maps with mitigated sampling artifacts. The 9″, 12″, and 18″pixel sizes respectively correspond to a relative shrinking of the 6″ pixel maps with ratios of 1.5, 2, and 3, which allows to accommodate for the significant and quite uncertain range of distances of the different MCs in the HGBS. This entire procedure, which is illustrated in Fig. 5.5, and summarized in Table 5.1, leads to a total of 551 patches with size 512×512 pixels, each of which is then subdivided into four 256×256 sub-patches.

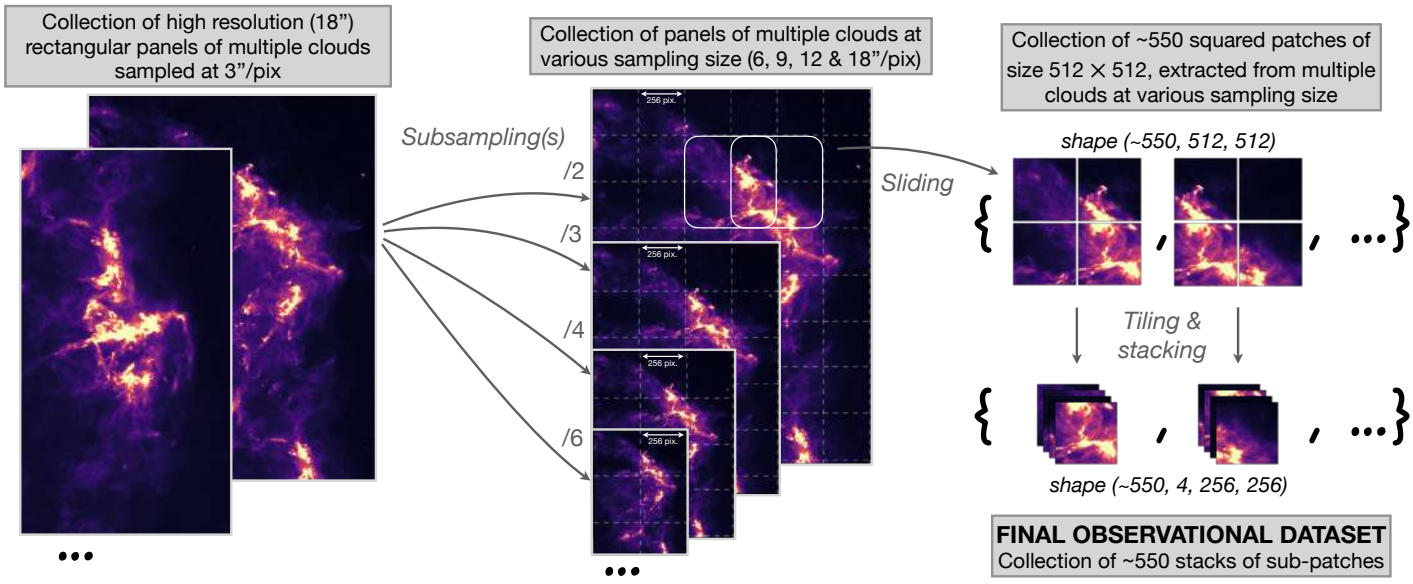


Figure 5.5: Illustration of the pre-processing of observational data. Details are given in Sec. 2.2.

Region	Dist.	Coord.		# patches with resolution					References
	d (pc)	l (°)	b (°)	Tot.	6”	9”	12”	18”	
Aquila	436	28	4	57	36	16	4	1	Ortiz-León et al., 2018
Chamaeleon I	210	296	-16	28	20	6	2	0	Zucker et al., 2020
Chamaeleon II III	160	302	-16	43	32	8	3	0	Zucker et al., 2020
Corona Australis	160	0	-19	44	34	7	3	0	Zucker et al., 2020
IC 5146	750	94	-5	4	4	0	0	0	Zucker et al., 2020
Lupus I	182	339	17	17	12	4	1	0	Galli et al., 2013
Musca	200	301	-9	6	6	0	0	0	Knude and Hog, 1998
Ophiuchus	139	353	17	61	42	12	6	1	Mamajek, 2008
Orion B	450	205	-14	88	61	19	7	1	Schlafly et al., 2014
									Zucker et al., 2020
Perseus	285	159	-20	44	32	10	2	0	Zucker et al., 2020
Pipe Nebula	180	359	6	23	18	5	0	0	Zucker et al., 2020
Polaris Flare	400	123	26	50	36	9	4	1	Schlafly et al., 2014
									Zucker et al., 2020
Serpens	436	31	4	46	33	9	4	0	Ortiz-León et al., 2018
Taurus	145	170	-16	40	30	6	4	0	Yan et al., 2019

Table 5.1: Summary of the properties of the different MCs studied in this paper, as well as their division into patches at different sampling. References in the rightmost column are for the distance estimates, which can be quite uncertain (uncertainties are given in the references quoted).

3 Quantifying informative power of summary statistics on an unlabeled dataset

3.1 General methodology

We want to quantify the amount of information we can exploit from a given set of summary statistics on a given dataset. However, because we work with unlabeled data, we cannot use supervised frameworks, such as Fisher analysis, that have a label-based approach to quantify information. For instance, when working with simulations, each sampled data x_i can be labeled by its corresponding physical parameters θ_i. In our case, we have to deal with an ensemble of unlabeled maps, for which we have *a priori* no notion of distance between pairs.

In this paper, we choose to rely on a notion of compatibility between patches, that can be estimated for any set of summary statistics. This approach allows to be quantitative, even in this unsupervised setting. Nevertheless, without supervision, it remains difficult to interpret a compatibility result between two patches for a given set of summary statistics: do the two patches actually have very similar properties, or do they have very distinct properties whose

differences are not effectively caught by the statistics?

This difficulty can be partially overcome by using complementary sets of statistics. Indeed, if two patches are distinguished by a first set of statistics, this is sufficient to assess that they are in practice different, and therefore to highlight the degeneracies of another set of statistics. It is this comparative approach, which is all the more relevant when the panel of statistics compared is comprehensive, that we use in this paper.

The informative power of a set of statistics strongly depends on the family of processes studied. For instance, we know that the empirical mean and power spectrum are sufficient statistics for stationary Gaussian fields (Cover & Thomas, 2006), which is not the case for physical processes in general. In this paper, we will apply our approach to the dataset of observations defined above, as well as to numerical simulations and synthetic logFBM models. In addition, we note that this diagnostic can also be used to compare maps from two different datasets.

In the rest of this section, we introduce the compatibility criterion, and explain how we extend it from a pair level to a dataset level. The set of statistics used in this paper will be presented in the following section.

3.2 Statistical compatibility for a pair of patches

We want to measure a notion of compatibility, according to a given set of summary statistics ϕ, between the two processes that generated the patches (x_i, x_j). To estimate this ϕ-compatibility, we need to make a number of simplifications, given the low-data regime, which will bring us back to a simplified case of statistical hypothesis testing. To do so, we tile each patch x_i into 4 sub patches $\{x_i^{(l)}\}_{1\leq l\leq 4}$, as illustrated in Fig. 5.5. We then compute the statistics at this sub patch level: $\{\phi(x_i^{(l)})\}_{1\leq l\leq 4}$ and we assume that these random variables can be considered as independent samples of the same distribution, that we furthermore model as multivariate normal distribution[5] of mean μ_i and variance Σ_i:

$$\phi(x_i^{(l)}) \sim \mathcal{N}(\mu_i, \Sigma_i). \tag{5.1}$$

Under these assumptions, the problem of measuring the compatibility between the distributions of $\phi(x_i^{(l)})$ and $\phi(x_j^{(l)})$ boils down to testing the compatibility between the two normal distributions, i.e., to test the hypothesis: $\mu_i = \mu_j$ and $\Sigma_i = \Sigma_j$. In our case, however, we have to estimate this compatibility from very few samples, most of the time fewer than the dimension of the vector of statistics ϕ. We thus choose to focus only on testing if the means μ_i and μ_j are statistically compatible, but not Σ_i and Σ_j, a problem known as the *multivariate two-sample mean test*. The most widespread test statistic for this problem is the Hotelling's two-sample T^2-statistic, a multivariate extension of Student's t-test (Hotelling, 1931). However, this test statistic requires to invert an estimation S of the full covariance matrix $\Sigma_i + \Sigma_j$, which is usually

[5]This assumption is not far from being true when the law of large numbers can be applied to the distribution of $\phi(x)$. This is for instance the case when $\phi(x)$ is defined as an average over the image pixels of a certain local distortion (such as filtering, possibly non linear) $\phi_{loc}(x)$ of x: $\phi(x) = \langle \phi_{loc}(x)[\mathbf{u}] \rangle_{\mathbf{u}}$ and when the process $\phi_{loc}(x)$ is stationary with a correlation length that is small compared to the image length.

intractable in our low data regime.

To overcome this, Srivastava and Du, 2008 proposed a test statistic based only on the diagonal D_S of the covariance estimator S, and on the trace of the square of its associated correlation matrix $R \equiv D_S^{-1/2} S D_S^{-1/2}$:

$$d^2_\phi(x_i, x_j) \equiv \alpha\Big[(\hat{\mu}_i - \hat{\mu}_j)^T D_S^{-1}(\hat{\mu}_i - \hat{\mu}_j) - \beta\Big], \tag{5.2}$$

where $\hat{\mu}_i$ is the estimator of μ_i obtained through an average over the 4 sub-patches $x_i^{(l)}$, α is an overall factor to normalize the variance of d^2_ϕ, and β is a debiasing term. We emphasize that, through a dependence on $tr(R^2)$, the α factor accounts, at least partially, for the correlation structure of ϕ. More details about these terms are given in appendix 8.3.

Under the assumption that $\Sigma_i = \Sigma_j$, the $d^2_\phi(x_i, x_j)$ test statistic has a variance of order unity. Thus, when it is much larger than 1, μ_i & μ_j cannot be considered compatible:

$$d^2_\phi(x_i, x_j) \gg 1 \implies \mu_i \,\&\, \mu_j \text{ incompatible}, \tag{5.3}$$

whereas when it is of the order of 1 or less, it is not possible to detect a discrepancy between the two means μ_i and μ_j with the available amount of data:

$$d^2_\phi(x_i, x_j) \lessapprox 1 \implies \mu_i \,\&\, \mu_j \text{ not incompatible based on the available amount of data.} \tag{5.4}$$

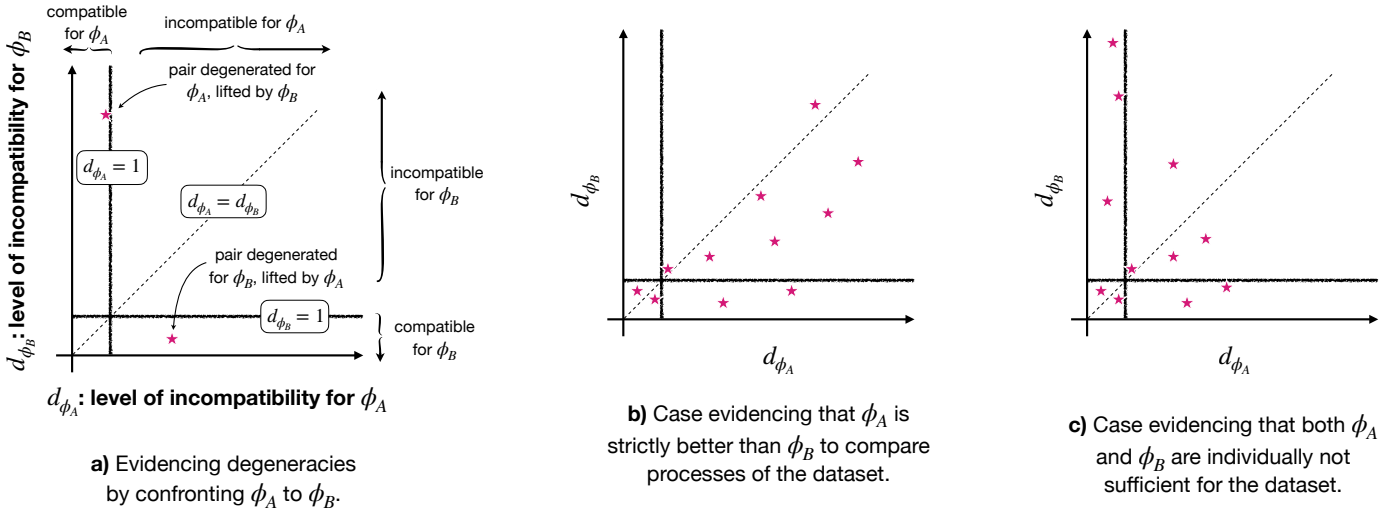
3.3 Comparing summary statistics on a dataset

By extending the ϕ-compatibility test introduced above, we set up a comparison between two sets of summary statistics ϕ_A and ϕ_B on a given dataset. This comparison consists in studying whether each set of summary statistics has degeneracies that the other set can lift. These degeneracies are evidenced by the presence in the dataset of pairs of maps that are clearly incompatible for ϕ_A, but for which ϕ_B detects no incompatibility, or *vice versa*. To be relevant, this comparison requires that the dataset contain maps that actually have different properties, which will be the case for the data sets studied later.

In practice, we suggest the following algorithm:

- Step 1: for every patch x_i of the dataset, compute the statistics ϕ_A and ϕ_B of its corresponding sub-patches $\{x_i^{(l)}\}_l$.
- Step 2: for every pair $\{x_i, x_j\}$ with $i \neq j$, compute $d^2_{\phi_A}(x_i, x_j)$ and $d^2_{\phi_B}(x_i, x_j)$ from the quantities derived in step 1.
- Step 3: place every pair as a point on a 2D scatter plot showing $d^2_{\phi_A}(x_i, x_j)$ against $d^2_{\phi_B}(x_i, x_j)$, such as Fig. 5.6.

The resulting scatter plot can then be used to detect whether there are some pairs of processes in the dataset that are degenerate for ϕ_A but lifted by ϕ_B (or *vice versa*), as explained in Fig. 5.6. In such plots, the presence of points in the bottom right region, i.e., where $d_{\phi_A} \gg 1$ and $d_{\phi_B} \lessapprox 1$,



a) Evidencing degeneracies by confronting ϕ_A to ϕ_B.

b) Case evidencing that ϕ_A is strictly better than ϕ_B to compare processes of the dataset.

c) Case evidencing that both ϕ_A and ϕ_B are individually not sufficient for the dataset.

Figure 5.6: Illustration of the proposed test to confront two sets of summary statistics, ϕ_A *vs* ϕ_B, on their degeneracy level for a given dataset. Each star represents a pair of patches. Panel a: the presence of stars in the bottom right region, i.e., where $d_{\phi_A} \gg 1$ and $d_{\phi_B} \lessapprox 1$, reveals that some pairs of this dataset are identified by ϕ_A as incompatible but not by ϕ_B: such pairs thus evidence degeneracies of ϕ_B lifted by ϕ_A. Conversely, the presence of stars in the top left region evidences degeneracies of ϕ_A lifted by ϕ_B. Hence, if all the points land mainly in the sub-diagonal part (panel b), this evidences that ϕ_A is better suited than ϕ_B to compare the pairs of this dataset. If on the contrary the points are spread both in the upper left and bottom right regions of the plot (panel c), this shows that both ϕ_A and ϕ_B are individually not sufficient to describe the processes of this dataset.

reveals that some pairs of this dataset are identified by ϕ_A as incompatible but not by ϕ_B: such pairs thus evidence degeneracies of ϕ_B lifted by ϕ_A (panel a). Conversely, the presence of points in the top left region evidences degeneracies of ϕ_A lifted by ϕ_B. Hence, if all the points land mainly in the sub-diagonal part (panel b), this evidences that ϕ_A is better suited than ϕ_B to compare the pairs of this dataset. If on the contrary the points are spread both in the upper left and bottom right regions of the plot (panel c), this shows that both ϕ_A and ϕ_B are individually not sufficient to describe the processes of this dataset.

4 Summary statistics

We present below the different summary statistics that are used in this paper. In the following, x is an image, $x(\mathbf{u})$ is the value of the image at pixel $\mathbf{u}$, $\tilde{x}(\mathbf{k})$ is the discrete Fourier transform of image x evaluated at wave-vector $\mathbf{k}$, $\star$ stands for the convolution operator, and $\langle\rangle_{\mathbf{u}}$ for the averaging over pixels. When $\phi(x)$ is multivariate, $\phi(x)[i]$ is the value of its i-th dimension. $\bar{x}$ designates the following normalization of x: $\bar{x} \equiv x/\mathrm{std}(x)$ where $\mathrm{std}(x) \equiv \sqrt{\langle [x - \langle x \rangle_{\mathbf{u}}]^2 \rangle_{\mathbf{u}}}$.

Note that for some usual statistics, we consider their logarithms instead of their raw values. For instance we use $\log \langle x \rangle_{\mathbf{u}}$ instead of $\langle x \rangle_{\mathbf{u}}$. This is possible when we work with positive-valued statistics. We made this choice to better fit the Gaussianity assumption given Eq. (5.1). In

addition, we found that logarithms expressed more discriminatory power with our diagnostic. See appendix 8.4 for more details. If no precision is made, log designates the logarithm in base 10.

4.1 One-point based statistics

We list below the one-point statistics used in this paper. Even though some of these statistics can probe non Gaussian information such as sparsity, they are all point wise statistics. This means in particular that they cannot capture spatial arrangement in the maps.

- The mean:

$$\boxed{\phi_{\text{mean}}(x) \equiv \log \langle x \rangle_{\mathbf{u}}.} \tag{5.5}$$

- The variance:

$$\boxed{\phi_{\text{var}}(x) \equiv \log \left\langle \left[x - \langle x \rangle_{\mathbf{u}} \right]^2 \right\rangle_{\mathbf{u}}.} \tag{5.6}$$

- The mean of the logarithm:

$$\boxed{\phi_{\text{mean of log}}(x) \equiv \langle \log x \rangle_{\mathbf{u}}.} \tag{5.7}$$

- The variance of the logarithm:

$$\boxed{\phi_{\text{var of log}}(x) \equiv \log \left\langle \left[\log x - \langle \log x \rangle_{\mathbf{u}} \right]^2 \right\rangle_{\mathbf{u}}.} \tag{5.8}$$

- The quantile functions[6] (QF) normalized by the median:

$$\boxed{\phi_{\text{QF}}(x)[i] \equiv \log \left[q_{\alpha_i}(x) / q_{1/2}(x) \right],} \tag{5.9}$$

  where $q_\alpha(x)$ designates the α-quantile of the distribution of values $\{x(\mathbf{u})\}_{\mathbf{u}}$. Hence $q_{1/2}$ stands for the median operator. We consider 10 quantiles $\{\alpha_i\}_i$ such that $1 - \alpha_i$ are logarithmically spaced between 10^{-4} and 0.4. Low quantiles are not considered as, for observational data, they are contaminated notably by the noise and the Cosmic Infrared Background (CIB) (Auclair et al., 2024; Ossenkopf-Okada et al., 2016). This logarithmic binning of the high column density values is motivated by the log-normal to power-law behavior of the tail of MCs' PDFs. See for instance Pouteau et al., 2023 and references therein.

4.2 Two-point based statistics

A very popular way to describe spatial properties of a process is through the power spectrum (PS). It is defined as the Fourier transform of the auto-correlation function, and describes the

[6] We consider quantile functions rather than probability distribution functions, as the difficulty of defining a unique binning range makes comparisons less efficient. Indeed, to define a range in terms of quantiles allows for a natural adaptation to each MCs' regions.

energy distribution scale by scale in the process studied. In this paper, we consider the isotropic PS, defined as:

$$PS[k] \equiv \langle \mathbb{E}[|\tilde{x}(\mathbf{k})|^2] \rangle_{\|\mathbf{k}\|=k}.$$

A power-law behavior $PS[k] \propto k^{-\beta}$ is expected for fields arising in turbulent MHD (Schekochihin, 2022). Thus we adopt a log-log representation, to linearize it and for stability purposes (Bruna & Mallat, 2013):

$$\boxed{\phi_{\mathrm{PS}}(x)[i] \equiv \log \left\langle |x_{\mathrm{apo}} \star \psi_i|^2 \right\rangle_{\mathbf{u}},} \tag{5.10}$$

where x_{apo} is apodized to mitigate non periodic boundary conditions (PBC) as explained in appendix 8.2. We use 6 band-pass filters $\{\psi_i\}_{1 \leq i \leq 6}$ defined as:

$$\tilde{\psi}_i[\mathbf{k}] = 1_{\|\mathbf{k}\| \in [k_i, k_{i+1}[},$$

where k_i is logarithmically spaced between $k_{\min} = 1/256\,\mathrm{pix}^{-1}$ and $k_{\max} = 1/4\,\mathrm{pix}^{-1}$. For the finest maps with pixel size $6''$, this $k_{\max}$ corresponds to a smallest angular scale of $24''$, which remains above the $18''$ resolution of the observations.

Because some MCs tend to have log-normal behavior in their one-point statistics, we also consider PS statistics of the logarithm of the maps:

$$\boxed{\phi_{\text{PS of log}}(x)[i] \equiv \log \left\langle |(\log x)_{\mathrm{apo}} \star \psi_i|^2 \right\rangle_{\mathbf{u}}.} \tag{5.11}$$

4.3 Scattering statistics

The Wavelet Scattering Transform (WST) is a set of non-Gaussian descriptors with a hierarchical and multi-scale structure (Bruna & Mallat, 2013). It has been shown to be very efficient at describing astrophysical fields (see, for instance, Allys et al., 2019; Saydjari et al., 2021). The usual WST consists in two layers of statistics, the first of which depends on a single scale of length $\simeq 2^j$ pixels with orientation θ:

$$S_1(x)[j, \theta] \equiv \langle |x \star \psi_{j,\theta}| \rangle_{\mathbf{u}}.$$

The second layer probes a coupling between two oriented scales (j_1, θ_1), and (j_2, θ_2), with $j_1 < j_2$:

$$S_2(x)[j_1, j_2, \theta_1, \theta_2] \equiv \left\langle \left| |x \star \psi_{j_1,\theta_1}| \star \psi_{j_2,\theta_2} \right| \right\rangle_{\mathbf{u}} / \, S_1(x)[j_1, \theta_1].$$

Then, observing that many processes of interest in astrophysics and cosmology exhibit strong regularities in their angular dependencies, Allys et al., 2019 proposed the Reduced WST (RWST): an angular compression of WST statistics for 2D data. We are going to use three of the main descriptors they introduced (we refer to the above reference for more details):

- S_1^{Iso}, a scale-by-scale isotropic descriptor, that we normalize by estimating it on $\bar{x}$ instead of x:

$$\boxed{S_1^{Iso}(x)[j] \equiv \left\langle \log_2 \langle |\bar{x} \star \psi_{j,\theta}| \rangle_{\mathbf{u}} \right\rangle_{\theta}.} \tag{5.12}$$

Note that it computes a L^1 norm of the filtered field, in contrast with the power spectrum that probes a L^2 norm.

- S_2^{Iso1} and S_2^{Iso2}, that measure an isotropic coupling between scales and are defined by fitting the following model:

$$\log_2 S_2(x)[j_1, j_2, \theta_1, \theta_2] \simeq S_2^{Iso1}(x)[j_1, j_2] + S_2^{Iso2}(x)[j_1, j_2] \cos\left[2(\theta_2 - \theta_1)\right]. \quad (5.13)$$

They characterize, respectively, a coupling between non-oriented scales and the relative coupling between parallel and perpendicular scales.

In this paper, we will furthermore compress both S_2^{Iso1} and S_2^{Iso2} coefficients that depend on two scales $(2^{j_1}, 2^{j_2})$ by only keeping their $j_2 - j_1$ dependency, i.e., a dependency on their ratio, by considering the following average:

$$\boxed{\langle S_2^{Iso1} \rangle_{j_2-j_1}(x)[\delta] \equiv \langle S_2^{Iso1}(x)[j_1, j_2] \rangle_{j_2-j_1=\delta},} \quad (5.14)$$

and

$$\boxed{\langle S_2^{Iso2} \rangle_{j_2-j_1}(x)[\delta] \equiv \langle S_2^{Iso2}(x)[j_1, j_2] \rangle_{j_2-j_1=\delta}.} \quad (5.15)$$

In practice, we consider 4 scales between $j_{\min} = 2$ and $j_{\max} = 5$, which leads to the 3 possible values $1 \leq \delta \leq 3$ for our reduced S_2 coefficients. We also divide the half-plane $[0, \pi[$ in 8 different θ angles. For the finest maps of pixel size $6''$, this $j_{\min}$ corresponds to a scale of $24''$ that remains above the $18''$ resolution of the observations.

For RWST computations, convolutions are done with local wavelets. This allows us to mitigate non PBC after convolution by suitably cropping the corrupted edges, without using apodization. These computations are done with the `pywst`[7] Python package (Regaldo-Saint Blancard et al., 2020).

4.4 Overview

The statistics introduced above are compiled in Table 5.2. In the following, in addition to the sets of summary statistics defined above, we will also consider some aggregated sets of summary statistics that are made from groups of these building blocks. For instance, we will refer to ϕ_{Gaussian} statistics to mean the joint set of $\{\phi_{\text{mean}}, \phi_{\text{var}}, \phi_{\text{PS}}\}$. These groups are defined in Table 5.3.

[7] https://github.com/bregaldo/pywst

ϕ	Dim.	Non-Gaussian	Scale description	Scale interactions
mean	1	×	×	×
var	1	×	×	×
mean of log	1	log-Gauss.	×	×
var of log	1	log-Gauss.	×	×
QF	10	✓	×	×
PS	6	×	✓	×
PS of log	6	log-Gauss.	✓	×
S_1^{Iso}	4	✓	✓	×
$\langle S_2^{Iso1} \rangle_{j_2-j_1}$	3	✓	×	✓
$\langle S_2^{Iso2} \rangle_{j_2-j_1}$	3	✓	×	✓

Table 5.2: Overview of the sets of summary statistics, their dimensions and properties (whether they are sensitive to non-Gaussianity, provide a multiscale description, and probe couplings between scales). "log-Gauss." refers to statistics that are not sensitive to non-log-Gaussianity, but these can be sensitive to some extent to non-Gaussianity.

Set of stats.	Composed of	Dim.
Gaussian	mean + PS	7
log-Gaussian	mean of log + PS of log	7
RWST	$S_1^{Iso} + \langle S_2^{Iso1} \rangle_{j_2-j_1} + \langle S_2^{Iso2} \rangle_{j_2-j_1}$	10
final	log-Gaussian, RWST	17

Table 5.3: Aggregated sets of summary statistics used in this work.

5 Towards a low-degeneracy set of statistics

We want to construct a set of informative summary statistics for the observational dataset. To do so, we will follow a bottom-up approach, starting with usual low-order statistics, and increasingly trying to improve on them, by exhibiting and lifting potential degeneracies. The underlying idea is that low-order statistics are good candidates to concentrate most of the informative power into a few coefficients. Such a concentration property is of interest, especially as our low data regime prevents us from learning many features. Another benefit of this approach is to first promote simple statistics, and use more elaborate ones only if they add a significant contribution.

5.1 Molecular clouds have Gaussian degeneracies

To begin with, we investigate whether Gaussian statistics (i.e., the set made of mean and PS statistics) are degenerate for the observational data. To point out such potential degeneracies, we confront these Gaussian statistics, starting with a low-order non-Gaussian set of statistics,

the quantile function QF, using the methodology introduced in Sec. 3.3. As shown in Fig. 5.7, this first confrontation on observational data already underlines a strong degeneracy level for both Gaussian and QF features, showing that neither statistics is by itself sufficient, according to the compatibility diagnostic we introduced. In particular, the quantile statistics extract a significant amount of information that cannot be efficiently captured by Gaussian statistics.

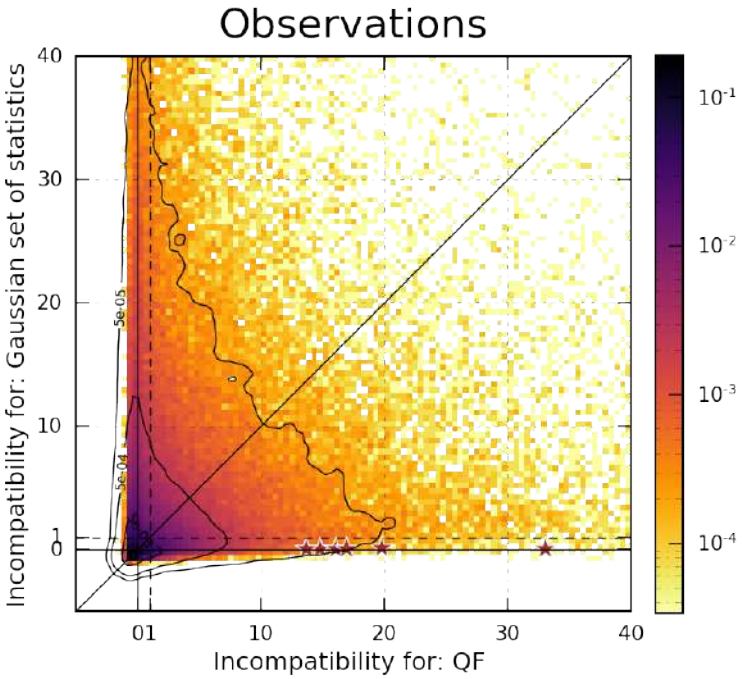


Figure 5.7: Confronting Gaussian statistics with QF statistics on observational data, based on the test presented in Fig. 5.6. Each set of statistics has strong degeneracies lifted by the other set. To investigate the Gaussian confusions, we pick six pairs of 512 × 512 patches, whose locations on the scatter plot are given by the red stars. These pairs are shown in Fig. 5.8.

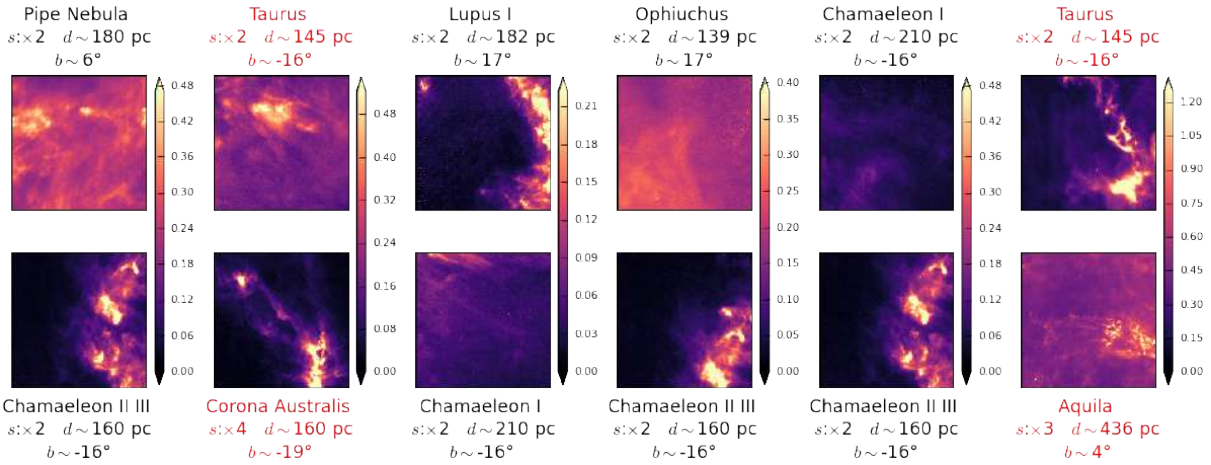


Figure 5.8: Examples of Gaussian confusions. Six pairs of 512 × 512 patches are chosen, whose locations on the scatter plot of Fig. 5.7 are given there by the red stars. The column density maps are shown in units of $10^{20}\mathrm{cm}^{-2}$. For each patch, we report: s the sub-sampling factor from the original $3'$/pix map, d and b the approximated distance and Galactic latitude of the cloud. The pixel size (in mpc) of a patch is thus proportional to $s \times d$. If a pair has patches (i, j) with incompatible pixel sizes, that we define according to the following criterion $\max\{\frac{s_i d_i}{s_j d_j}, \frac{s_j d_j}{s_i d_i}\} \geq 3/2$, we color the labels in red.

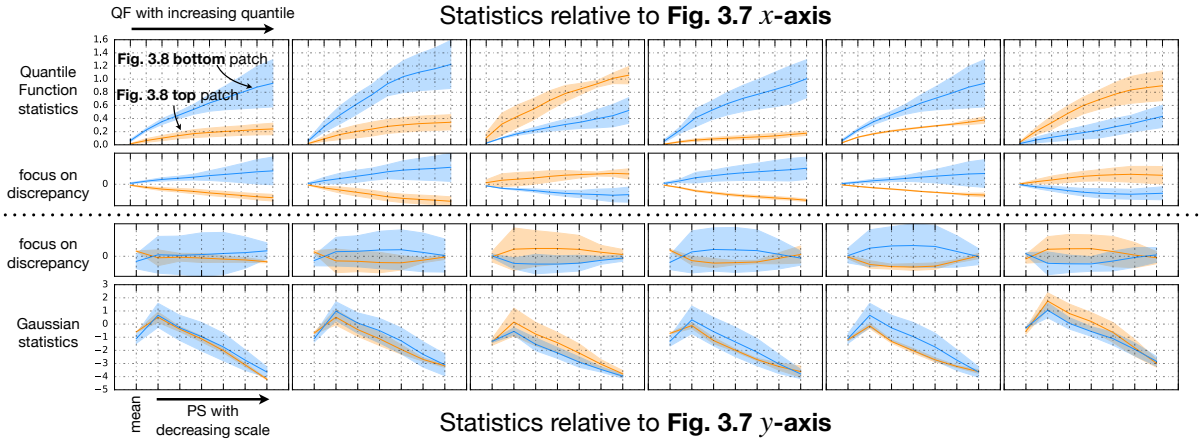


Figure 5.9: Statistics for the examples of Gaussian confusions shown in Fig. 5.8. In each row, the orange filled line (resp. band) corresponds to the mean (resp. std) of the statistics computed over the four 256 × 256 sub-patches of the top patch of each pair of Fig. 5.8, and the corresponding blue lines and areas refer to the bottom patch of the pair. The top row corresponds to the statistics used in the x-axis of the scatter plot of Fig. 5.7, i.e., QF statistics, plotted with 10 increasing quantile values, while the bottom row corresponds to the y-axis, i.e., Gaussian statistics, plotted starting with mean and followed by the binned PS with six decreasing scales. To better highlight the discrepancies between the two patches of a given pair, we report in second and third rows the offsets of the orange and blue filled lines with respect to their common mean.

In addition to this dataset-wide diagnostic, we display in Fig. 5.8 six randomly picked pairs of 512 × 512 patches degenerate for Gaussian statistics but with increasing QF incompatibility. The locations of these pairs on the scatter plot of Fig. 5.7 are given there by the red stars. For these six pairs of patches, we show in Fig. 5.9 the two sets of statistics used in the diagnostic (QF statistics and Gaussian statistics), where the orange filled lines (resp. bands) correspond to the mean (resp. standard deviation) of the statistics computed over the four 256 × 256 sub-patches of the top patch of each pair of Fig. 5.8, and the corresponding blue lines and areas refer to the bottom patch of the pair. To better highlight the discrepancies between the two patches of a given pair, the offsets of the orange and blue filled lines with respect to their common mean are also shown in Fig. 5.9. These results illustrate the ability of QF statistics to distinguish between images with the same Gaussian statistics, which confirms that our compatibility diagnostic works as expected. We emphasize here that the ϕ_{QF} statistics are complementary to the one-point properties probed by Gaussian statistics.

This behavior is not particularly surprising as one-point properties of column density maps of MCs are expected to be at least log-normal, if not with a power-law tail. Hence, estimating such properties directly from the logarithm of the maps, as does ϕ_{QF} but not ϕ_{Gaussian}, might enhance the discriminative power. To test this idea, we study in the first column of Fig. 5.10 the degeneracy level of the set of mean and variance statistics estimated respectively on the raw maps and their logarithm. More precisely, top row plots confront QF statistics to $\{\phi_{\mathrm{mean}}, \phi_{\mathrm{var}}\}$, while bottom row plots confront QF statistics to $\{\phi_{\mathrm{mean\ of\ log}}, \phi_{\mathrm{var\ of\ log}}\}$. We see in this figure that the pairs of observations are clearly better discriminated when using the mean and variance

computed on the logarithm of the maps (d) rather than directly (a). This effect is very well reproduced by logFBM data (middle column), but does not hold in general, as shown with the Describable Textures Dataset (DTD) (right column).

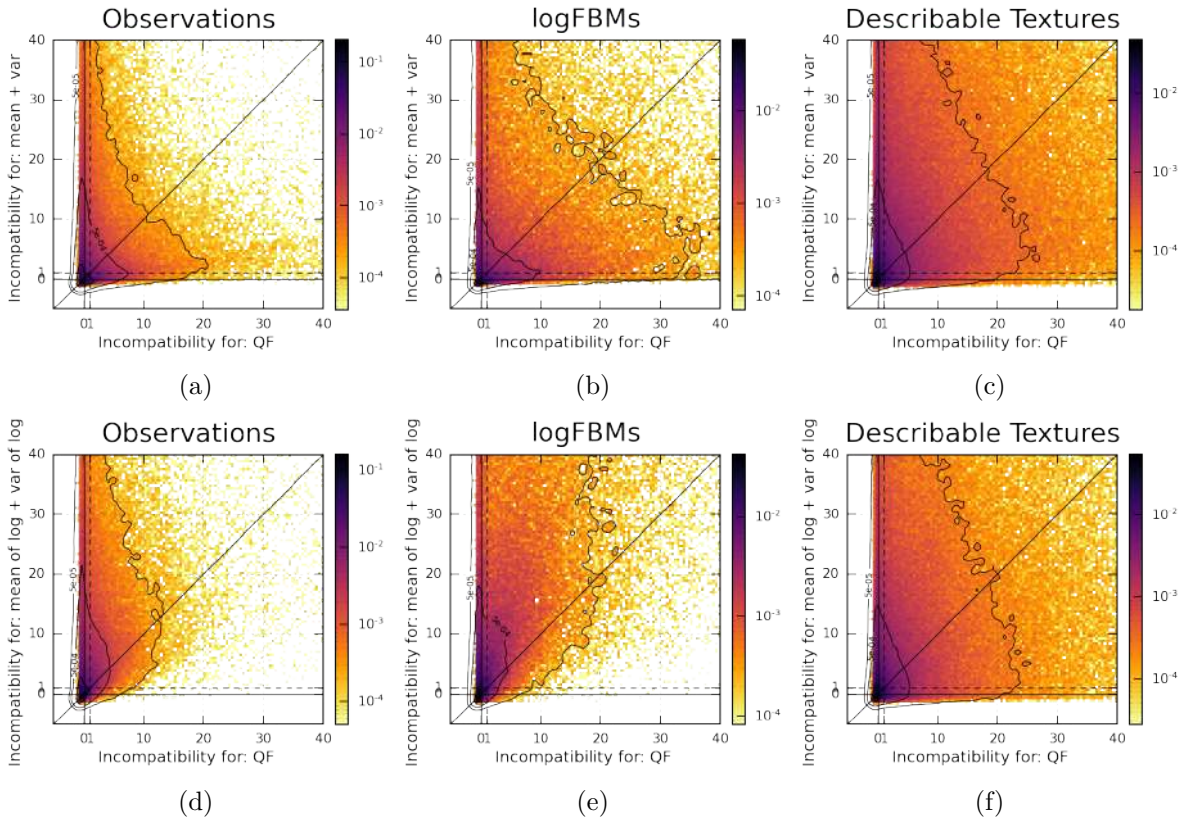


Figure 5.10: Top row plots confront QF statistics to $\{\phi_{\text{mean}}, \phi_{\text{var}}\}$, while bottom row plots confront QF statistics to $\{\phi_{\text{mean of log}}, \phi_{\text{var of log}}\}$. These results evidence that taking the logarithm of the map enhances the discriminative power of the mean and variance statistics on both observation and logFBM data (left and middle columns) but not on DTD (right column).

These results suggest that the specific Gaussian degeneracies evidenced in Fig. 5.7 are mainly explained by the inefficiency of Gaussian one-point statistics $\{\phi_{\text{mean}}, \phi_{\text{var}}\}$ to characterize one-point properties of processes that have a log-normal (or heavier tail) nature.

Surprisingly, this analysis also shows that, for datasets such as observations or logFBMs, probing the PDF properties through the prism of a mean and a variance is more discriminative than probing its shape. Here, in Fig. 5.10, in both d) and e), our suitably constructed set $\{\phi_{\text{mean of log}}, \phi_{\text{var of log}}\}$, of dimension 2, performs almost always a better discrimination than the set ϕ_{QF}, of dimension 10. This emphasizes the importance of suitably constructed low dimensional descriptions in such analysis. The performative power of such low-dimensional features is usually due to the exploitation of underlying regularities of the studied processes that allow these compression, without much loss of information. For deeper insights on the connection between regularity, approximation and sparsity, we recommend Pr. Mallat's lecture[8]. However,

[8] https://www.college-de-france.fr/fr/agenda/cours/representations-parcimonieuses/le-triangle-regularite-approximation-parcimonie

when dealing with datasets made of a wide diversity of irregular processes, such as everyday life textures like DTD, it is hard to identify a compression that does not lead, for some pairs, to poorer performances (c, f).

5.2 Molecular clouds have log-Gaussian degeneracies

The ability of Gaussian one-point statistics to grasp efficiently one-point properties of observations from the logarithm of the column density maps, shown in Fig. 5.10, suggests to shift towards log-Gaussian statistics, i.e., mean and PS estimated on the logarithms of the maps. We thus investigate now whether we may point out some degeneracies of these statistics on observational data. However, if we confront this set with QF statistics, as we did previously for Gaussian statistics, we do not expect to lift significant log-Gaussian degeneracies. Instead, we suggest to search for such degeneracies using a set of higher order statistics: the RWST. To make sure that, on other datasets, the potential degeneracies lifted by these higher order statistics could not be lifted by one-point statistics, we will confront the log-Gaussian set to the RWST one.

We show in Fig. 5.11 that log-Gaussian statistics have degeneracies on observational data lifted by RWST. We show six examples of degenerate pairs in Fig. 5.12 and the statistics of these patches in Fig. 5.13. In most of these pairs, the S_1^{Iso} coefficients heavily contribute to lift the log-Gaussian degeneracy. In the first and last pairs, $\langle S_2^{Iso2}\rangle_{j_2-j_1}[\delta]$ and respectively $\langle S_2^{Iso1}\rangle_{j_2-j_1}[\delta]$ are also very discriminative.

When applying the same analysis on logFBM data (Fig. 5.14.a), this finding however does not hold. The RWST diagnostic brings no additional information that was not already probed by log-Gaussian statistics.

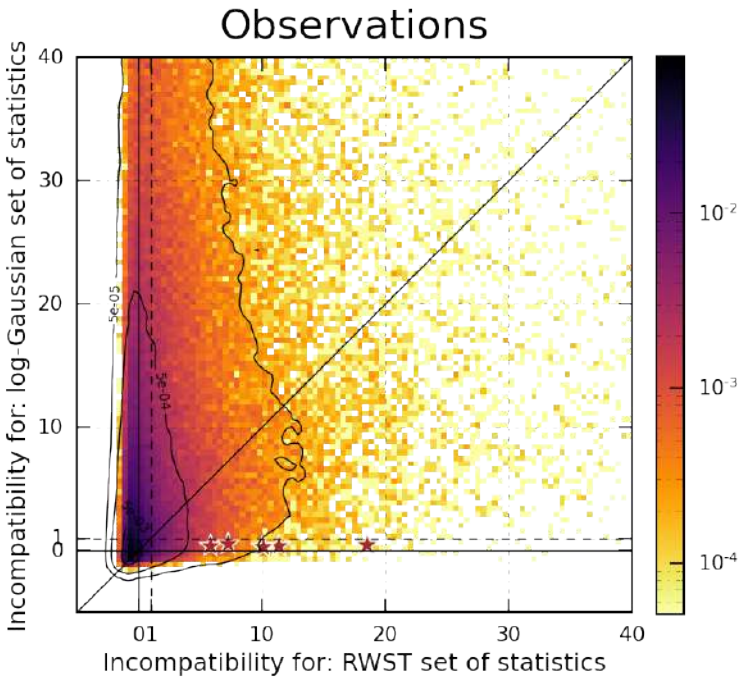


Figure 5.11: Same as Fig. 5.7, but using the log-Gaussian and RWST sets of statistics on the observational dataset. To investigate the degeneracies of the log-Gaussian statistics, six pairs of patches are selected, corresponding to the six red stars on this plot.

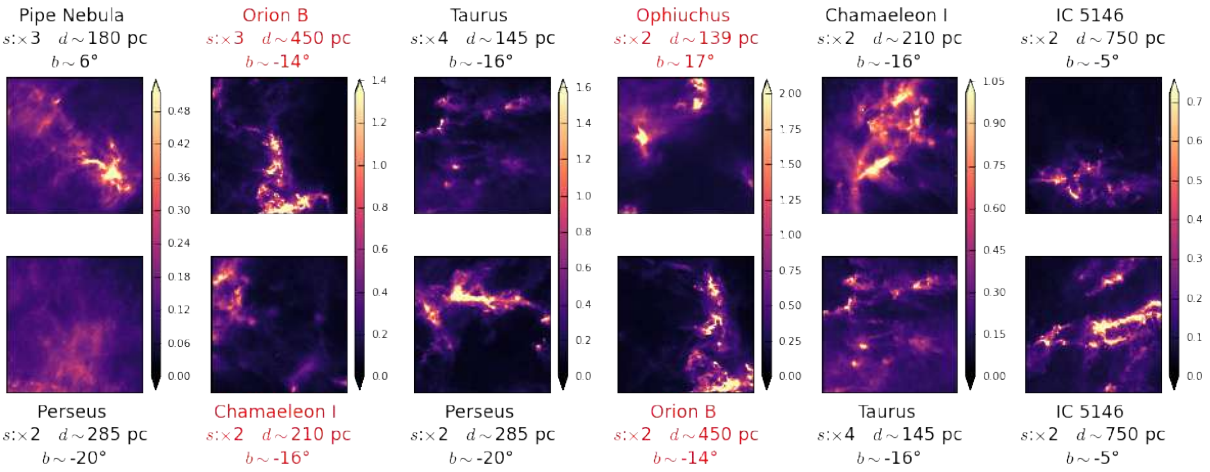


Figure 5.12: Examples of log-Gaussian degeneracies. Six pairs of 512 × 512 patches are chosen, whose locations on the scatter plot of Fig. 5.11 are given there by the red stars. The column density maps are shown in units of $10^{20}\mathrm{cm}^{-2}$. For each patch, we report: s the sub-sampling factor from the original 3″/pix map, d and b the approximated distance and Galactic latitude of the cloud. If a pair has patches (i, j) with incompatible pixel sizes according to the following criterion $\max\{\frac{s_i d_i}{s_j d_j}, \frac{s_j d_j}{s_i d_i}\} \geq 3/2$, we color the labels in red.

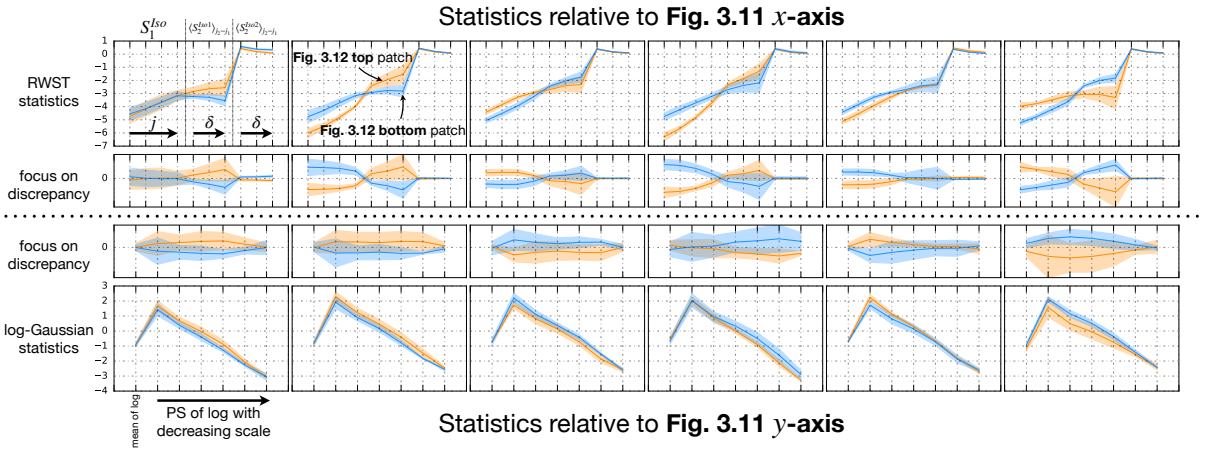
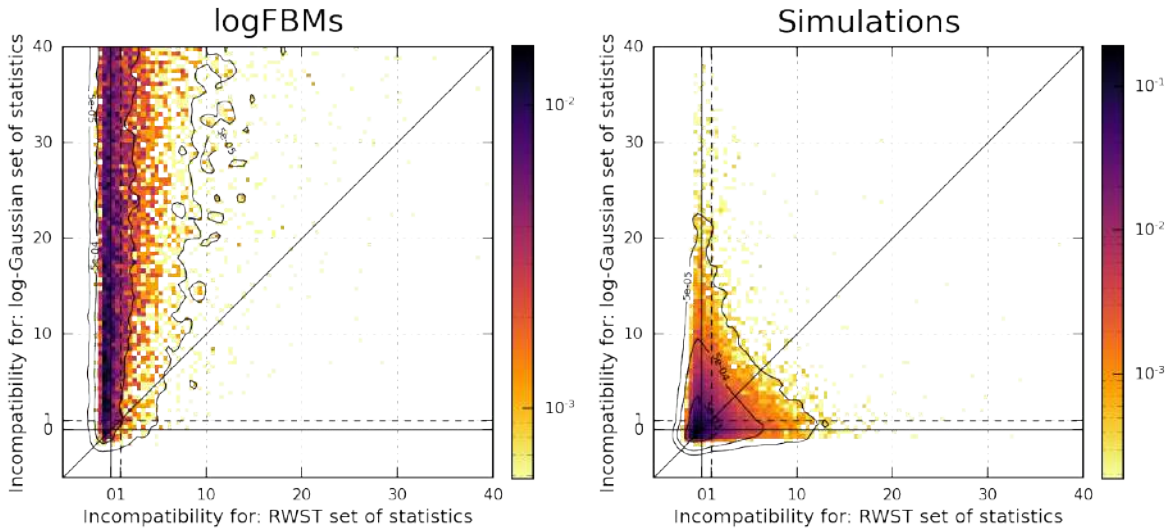


Figure 5.13: Statistics for the examples of log-Gaussian confusions shown in Fig. 5.12. In each row, the orange filled line (resp. band) corresponds to the mean (resp. std) of the statistics computed over the four 256 × 256 sub-patches of the top patch of each pair of Fig. 5.12, and the corresponding blue lines and areas refer to the bottom patch of the pair. The top row corresponds to the RWST, starting with $S_1^{Iso}[j]$ coefficients with four increasing scales j, then $\langle S_2^{Iso1}\rangle_{j_2-j_1}[\delta]$ coefficients with three increasing scale ratios δ and finally $\langle S_2^{Iso2}\rangle_{j_2-j_1}[\delta]$ coefficients with the same three scale ratios. The bottom row represents in the following order: mean of log followed by PS of log with 6 decreasing scales. The second and third rows show the offsets of these statistics with respect to the mean of the two.

Indeed, as expected for such a dataset, log-Gaussian statistics are sufficient (Cover & Thomas, 2006). Finally, simulations yield yet another result (Fig. 5.14.b). Although this latter case leads to the same qualitative conclusions as observations, namely underlining the insufficiency of log-Gaussian statistics, it differs from it quantitatively. This supports the caveat of simulation-based



(a) (b)

Figure 5.14: log-Gaussian statistics have degeneracies lifted by RWST for simulations in (b), but not for logFBM processes in (a). This is expected because log-Gaussian statistics are sufficient for families made of logFBM processes (cf. e.g., Eq. 4.8).

inference according to which observations and simulations of such processes lie in sub-manifolds having different geometries, and motivates the observation-based approach of this paper.

5.3 Final set of statistics

In this paper, we choose to limit our study to the log-Gaussian + RWST set, which we call ϕ_{final}, that seems well suited to describe this observational dataset. To ensure this suitability any further, we could continue to study the degeneracies of ϕ_{final} by confronting it with other sets of statistics. In any case, it is difficult to guarantee with certainty at any stage that we have not missed any remaining degeneracy of the resulting set that would be obtained by such a construction, even if each confrontation with a complementary set reinforces our confidence in its completeness.

Once a set of statistics has been fixed, however, it is also possible to use a complementary diagnostic to study it, which consists in checking the visual proximity of maps identified as close. If they are clearly visually different, this indicates complementary limitations of the set of statistics used. This is done with ϕ_{final} in the following section, where we introduce a distance to evaluate a notion of proximity rather than compatibility between pairs of maps.

Finally, we want to emphasize that low-order statistics generally constitute a precious and easily-accessed source of information. They should not be underestimated, and attempting to tailor them to the type of data considered can be fruitful, as shown here when shifting from a Gaussian to a log-Gaussian description.

6 Comparing pairs and datasets

The previous section confronted multiple sets of statistics to assess their information content through their level of degeneracy. In this section, we fix the set of statistics: $\phi = \phi_{\text{final}}$, composed of 17 coefficients (7 log-Gaussian descriptors and 10 RWST statistics), and use it to define a distance between maps. We illustrate this distance by exhibiting closest pairs of images in a dataset, as well as between different datasets such as observations and simulations.

6.1 Defining a morphological distance

Our objective is to define a distance between two patches based on ϕ_{final}. One of our requirements is to enable a comparison between distance values for different pairs, so that these pairs can be ordered according to the morphological proximity of their patches. This requirement prevents us from using the statistical compatibility diagnostic introduced earlier. Indeed, since it is weighted by the local variance of each patch's statistics, this can lead, for example, to some patches in a dataset being compatible with almost all the others simply because their spatial variance is very high. This could also encourage a simulation to approach an observation, following this criterion, by arbitrarily increasing its variance $\operatorname{Var}\phi(x_{SIM})$ without focusing on reducing the discrepancy of its average statistical properties $\hat{\phi}(x_{SIM}) - \hat{\phi}(x_{OBS})$. This property was purposely used to act as a penalization when confronting different sets of statistics in the previous section, but is no longer desired for pairs' ordering, once ϕ is fixed.

To build a distance that avoids this drawback, we choose instead to normalize it by the variability of the statistics evaluated over the entire observational dataset. We thus use the following distance:

$$d^2_{\mathcal{D}}(x_i, x_j) \equiv (\hat{\mu}_i - \hat{\mu}_j)^T (\operatorname{diag} M_{\mathcal{D}})^{-1} (\hat{\mu}_i - \hat{\mu}_j), \tag{5.16}$$

which is normalized by the spanning of the estimated values $\hat{\mu}_i$ of ϕ_{final} over all maps of a given dataset $\mathcal{D}$:

$$M_{\mathcal{D}} \equiv \left\langle (\hat{\mu}_i - \langle \hat{\mu}_j \rangle_j)(\hat{\mu}_i - \langle \hat{\mu}_j \rangle_j)^T \right\rangle_i, \tag{5.17}$$

where the brackets indicate an average over $\mathcal{D}$. Note, however, that it is difficult to interpret the value of $d^2_{\mathcal{D}}$ in absolute terms. Indeed, the $M_{\mathcal{D}}$ term does not describe a typical variance for a given process, but describes the variety of morphologies encountered in the entire dataset, that can be wide, as for the MC data investigated here. Unlike statistical compatibility diagnostics, the $d^2_{\mathcal{D}}$ distance is therefore modified by the addition or removal of maps in the $\mathcal{D}$ dataset, and can be affected by the presence of outliers.

In the following, we work with different datasets. For instance, we aim at comparing the minimum distance between observations and simulations to the typical distance of the closest pairs of observations. To do so, we use in this paper a common metric M_{obs} computed on the observation dataset, as well as the associated d^2_{obs} distance.

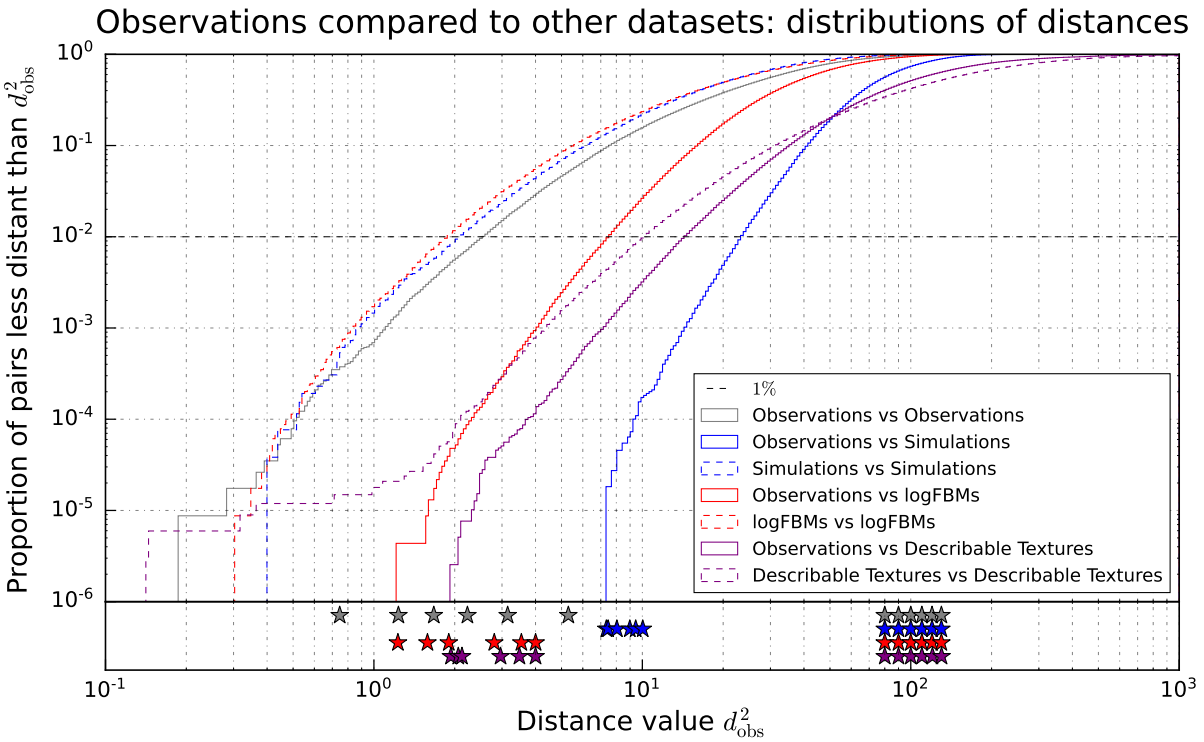


Figure 5.15: Cumulative distributions of $d^2_{\rm obs}$ distances between pairs extracted from multiple datasets. The different curves correspond to different choices of datasets from which the two patches of a given pair are extracted. This distance is based on $\phi_{\rm final}$, a set of 17 coefficients (7 log-Gaussian descriptors and 10 RWST statistics). The same metric $M_{\rm obs}$ is used for all distances and is defined based on the observational dataset. The stars in the bottom sub-panel correspond to "close" and "distant" pairs shown in Fig. 5.16. An horizontal line at the value 1% is drawn to represent the proportion of pairs of observations that are neighboring patches in the sky. For these pairs, the distance is underestimated and thus the gray curve below this line is not meaningful. Same applies for pairs of simulations (blue dashed curve).

6.2 Closest pairs

We use the $d^2_{\rm obs}$ distance to identify the closest pairs of patches that can be found between two datasets. To do so, we report in Fig. 5.15 the cumulative distributions of distances associated to pairs of observation patches (in gray), as well as to pairs made of one observation patch and one logFBM, DTD, and simulation patch (in red, purple, and blue, respectively). For each case, we pick up six of the typical closest pairs[9], as well as six relatively distant pairs, that we show in Fig. 5.16. In dashed lines with the same colors, we also report the cumulative distributions of distances associated to pairs of patches from a common dataset (logFBM, DTD[10], and simulation patches, respectively).

Closest (OBS, OBS) pairs and (OBS, logFBM) pairs are visually rather similar, while distant

[9]We do not investigate the $\sim 1\%$ closest pairs of observations, since the comparison between observations is slightly biased with respect to the other ones. The reason is that, in that case only, we have pairs made of non independent patches, such as neighboring patches in the sky or two slightly different scaled versions of the same region. This corresponds to approximately 1% of the pairs, that we therefore exclude.

[10]Note that the ~ 10 closest pairs of DTD images, located at $d^2_{\rm obs} \lessapprox 1$ are artifacts in this dataset corresponding in practice to almost identical images.

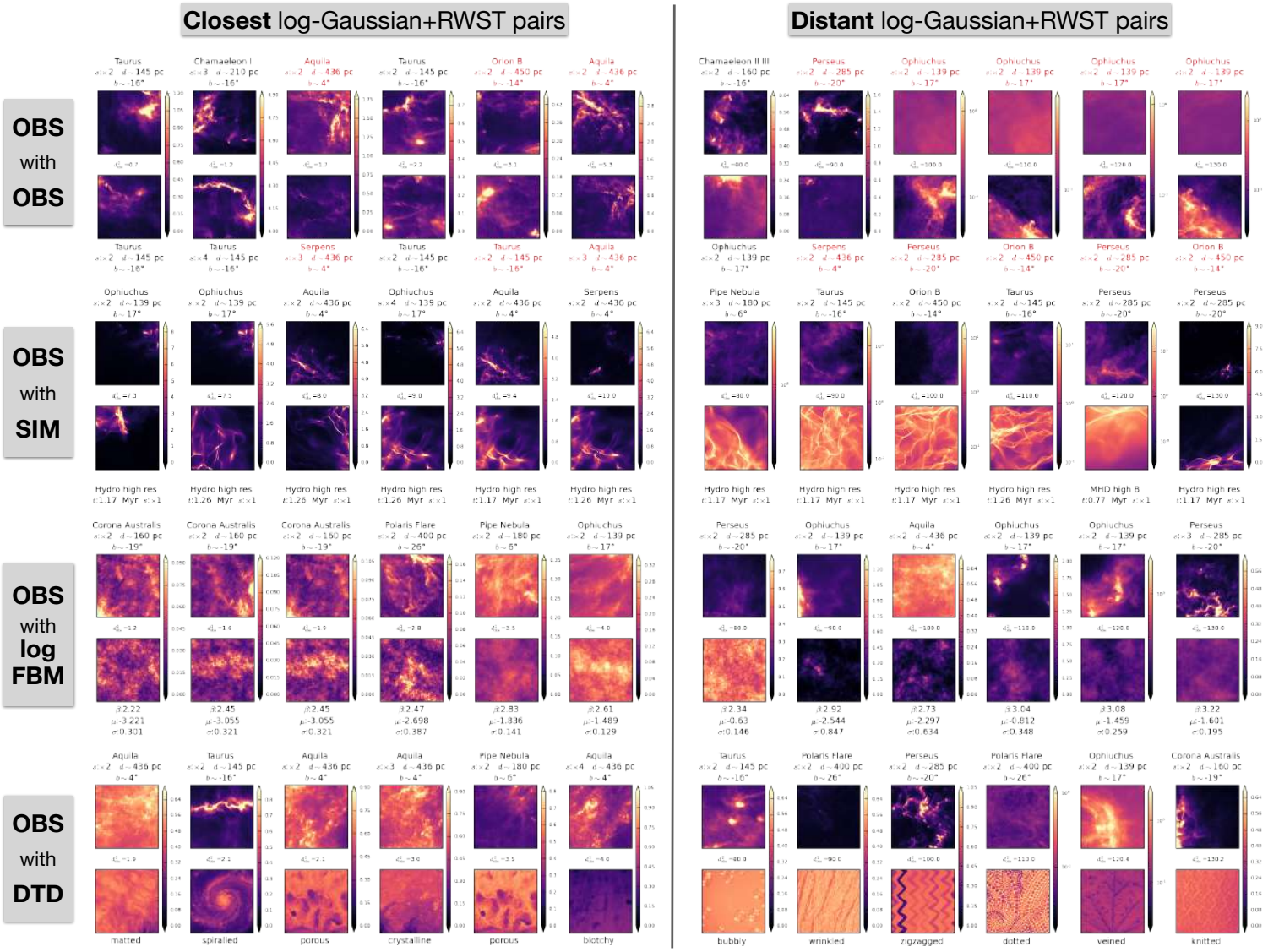


Figure 5.16: Closest (left) and more distant (right) pairs extracted from distributions of distances reported in Fig. 5.15. We see that the closest (OBS, SIM) pairs are much more distant ($d^2_{\text{obs}} \sim 7$) than the closests (OBS, OBS) pairs ($d^2_{\text{obs}} \sim .7$). See Sec. 6.3 for more detailed interpretation on these results. Some logFBM models end up quite close to the most diffuse regions observed: Polaris Flare and Corona Australis. Such regions are highly contaminated by CIB emission. Many (OBS, DTD) pairs are found to be close whereas the textures look very different. This shows that the set of summary statistics developed in this paper is tuned for ISM observations but is far from being sufficient for any kind of data. This also illustrates that MCs have much more regularity in terms of morphology than DTD textures.

pairs look very different. This relative agreement between morphological proximity as probed by our statistical distance as well as by human vision is encouraging, because it should be satisfied by an ideal distance. However, it remains far from being an exhaustive diagnostic.

The closest (OBS, logFBM) pairs have a distance $d^2_{\text{obs}} \sim 2$, which marks a high agreement. Indeed, only $\sim 0.5\%$ of (OBS, OBS) or (logFBM, logFBM) pairs exhibit smaller distance values, and we have already mentioned that the closest 1% of (OBS, OBS) pairs are very similar by construction. Comparatively, simulations are more distant to observations: the closest pairs between observations and simulations have a distance $d^2_{\text{obs}} \sim 8$, four times larger than the closest (OBS, logFBM) pairs, when already $\sim 10\%$ of pairs of observations exhibit smaller distances (see Sec. 6.3 for more details on this interpretation). As expected, these closest (OBS, SIM) pairs are visually less similar than the closest (OBS, OBS) or (OBS, logFBM) pairs (left panels of Fig. 5.16), but are still not so different, for instance with respect to the more distant pairs, that can be seen in the right panels of Fig. 5.16.

The closest (OBS, logFBM) samples obtained here are very diffuse regions of MCs such as Corona Australis, Polaris Flare, and Ophiuchus. This is not surprising because for MCs, diffuse regions are closer to logFBM models than dense regions, whose PDFs are known to deviate from log-normality. However, these diffuse regions are still supposed to exhibit coherent structures that should induce deviations from logFBM models. Here, such deviations are found to be small. We note, however, that this proximity only means that certain diffuse regions are close to logFBM processes relative to the total variability of the observational dataset, and not necessarily that they are well described by such models, or could not be distinguished by the previous compatibility diagnostic. Moreover, this proximity to logFBM models is also partly due to contamination by the CIB, that has non negligible power at such low levels of column density, and is expected to Gaussianize the data, including their RWST statistics (Auclair et al., 2024). Incidentally, we remind that the observational dataset has some artifacts that make it deviate from an "ideal" MC column density dataset, even if we try to limit as much as possible their effects (noise, finite resolution), as discussed in Sec. 4.

On the contrary, the observations that are closest to the simulations of dense MCs correspond, unsurprisingly, to dense regions of MCs such as Ophiuchus, Aquila, and Serpens. Note that Ophiuchus exhibits patches that are close to these dense simulations, but also at least one patch that is close to a logFBM model, underlining the spatial variability of molecular cloud morphologies. This illustrates the difficulty to treat a MC as a single entity, and emphasizes the relevance of our local approach.

The comparison between observations and DTD shows the limitations of our distance diagnostic. Indeed, the closest (OBS, DTD) pairs are found at a distance $d^2_{\text{obs}} \sim 2$ that is the typical distance between close (OBS, OBS) pairs or close (OBS, logFBM) pairs, although they are visually very different. This illustrates that, for a diagnostic based on a low-dimensional set of statistics, it is difficult to probe a distance over a family of processes that has such a wide variety of textures as DTD. On the contrary, because the simulations that are closest to observations are more distant ($d^2_{\text{obs}} \sim 8$), this suggests that the set of simulations does not intersect the set of observations and that ϕ_{final} is able to pinpoint this discrepancy, as discussed in the

following subsection.

Finally, it is quite impressive to see that we can build a distance diagnostic from a representation of dimension 17 only that still manages to identify morphological similarity between maps quite satisfactorily. This illustrates the possibility of constructing a highly informative but low-dimensional description tailored to a family of processes from an ensemble of compressed sets of usual statistics. It should be stressed, however, that this study remains partial, notably because it is based mainly on the observation of a few close pairs in our dataset. Yet, the confrontation diagnostics studied in Sec. 5 showed that degenerate counterexamples remain largely in the minority. In addition, we lack solid baselines since it is inherently difficult to quantify visual impressions of morphological proximity, although some work has been done in this direction (Peek & White, 2021).

6.3 Interpreting the minimal distance between observations and simulations

A last question we tackle is whether the relatively high value of the minimal distance $d^2_{\rm obs}$ that we get between observations and simulations indeed indicates distinct statistical properties of these two sets. More precisely, we question whether the set of observations' moments $\{\bar{\mu}_i\}_{i\in\rm obs}$ overlaps that of simulations $\{\bar{\mu}_j\}_{j\in\rm sim}$, where $\bar{\mu}_i \equiv \mathbb{E}[\phi_{\rm final}(x_i)]$ stands for the expected value over a given process i. In practice, we aim at retrieving the minimal distance

$$\bar{d}^2_{ij} \equiv (\bar{\mu}_i - \bar{\mu}_j)^T (\mathrm{diag}\, M_{\rm obs})^{-1} (\bar{\mu}_i - \bar{\mu}_j) \tag{5.18}$$

between $\bar{\mu}_i$ and $\bar{\mu}_j$ over (OBS, SIM) pairs (i, j). However, the distance $d^2_{\rm obs}$ introduced previously is a statistical estimator, so that in general

$$d^2_{ij} \neq \bar{d}^2_{ij},$$

and in particular: $\mathbb{E}[d^2_{ij}] > \bar{d}^2_{ij}$. Indeed, the variance of the $\hat{\mu}_i$ estimator biases d^2_{ij} with respect to $\bar{d}^2_{ij}$:

$$\mathbb{E}[d^2_{ij}] \equiv \bar{d}^2_{ij} + b_{ij}. \tag{5.19}$$

This non-negative bias, which boils down to:

$$b_{ij} = \mathrm{tr}\{\mathrm{cov}\left[(\mathrm{diag}\, M_{\rm obs})^{-1/2} (\hat{\mu}_i - \hat{\mu}_j)\right]\}, \tag{5.20}$$

prevents us from interpreting directly the value of d^2_{ij} as $\bar{d}^2_{ij}$. Furthermore, its value increases with the amplitude of the fluctuations of $\hat{\mu}_i - \hat{\mu}_j$, and can thus change depending on the pair (i, j) considered. This dependency on the pair also prevents us, without further check, from actually comparing $\bar{d}^2$ between different pairs based on the estimations d^2.

To probe and interpret the minimal distance between observations and simulations, we propose the following strategy:

- find an observation i and a simulation j that minimize the biased estimation d^2_{ij}. Such pairs are already reported in Fig. 5.16 and are good candidates to minimize $\bar{d}^2_{ij}$. The

goal then becomes to compare $\bar{d}^2_{ij}$ to the minimal value $\bar{d}^2_{ii\prime}$, for $i\prime$ another observation, independent from i. But $\bar{d}^2$ is unknown. Instead:

- find an observation $i\prime$, that minimizes the biased distance $d^2_{ii\prime}$, while checking it does not overlap i in the sky to assume them independent.

- Since $\bar{d}^2_{ij} - \bar{d}^2_{ii\prime} = \mathbb{E}[d^2_{ij}] - \mathbb{E}[d^2_{ii\prime}] + b_{ii\prime} - b_{ij}$, the discrepancy between $\bar{d}^2_{ij} - \bar{d}^2_{ii\prime}$ can be estimated based on the measured discrepancy $d^2_{ij} - d^2_{ii\prime}$ up to the unknown bias shift $b_{ii\prime} - b_{ij}$.

- If, in addition the observation patch $x_{i\prime}$ is such that spatial fluctuations of $\phi_{\rm final}$ over its sub-patches are at least of the order of those estimated from the sub-patches of the simulation patch x_j, then based on Eq. 5.20, we have $b_{ii\prime} \gtrapprox b_{ij}$. This allows then to use $d^2_{ij} - d^2_{ii\prime}$ as an estimated lower bound for $\bar{d}^2_{ij} - \bar{d}^2_{ii\prime}$.

We report in Fig. 5.17 such independent[11] pairs $(i, i\prime)$. In the first row, we fix i associated to the patch of Ophiuchus x_i that is the closest observation to simulations ($d^2_{ij} = 7$, Fig. 5.16). For that choice of patch, the reported distances $d^2_{ii\prime}$ of neighboring observations $i\prime$ are rather similar to d^2_{ij}. However, in the second row, where i is associated to the patch of Aquila that is the second closest observation to simulations (as shown in Fig. 5.16), we see other observations $i\prime$, such as Serpens or Orion B, that are closer than the closest simulation: $d^2_{ii\prime} = 3 < d^2_{ij} = 8$. We checked that the fluctuations of $i\prime$ are comparable to the one of j so that $b_{ii\prime} \gtrapprox b_{ij}$. This implies then $\bar{d}^2_{ij} - \bar{d}^2_{ii\prime} \gtrapprox 5$ while $\bar{d}^2_{ii\prime} \leq d^2_{ii\prime} = 3$. Hence, $\bar{d}^2_{ij}/\bar{d}^2_{ii\prime} \gtrapprox 8/3$. This second case shows that one of the closest observation/simulation pair is definitely further apart than this specific observation is with other observations, at least according to $\phi_{\rm final}$. This example also supports that the potential bias on the $d^2_{\rm obs}$ estimators remains moderate for the study of (OBS, SIM) pairs ($\bar{d}^2_{ij} \gtrapprox 5$ and $d^2_{ij} = 8$ implies $b_{ij} \lessapprox 3$).

We conclude that the minimum (OBS, SIM) distance value, obtained on all possible pairs between those datasets, evidences a meaningful but moderated distinction between these datasets. Note that the $\phi_{\rm final}$ statistics set may still have some degeneracies, particularly for the (OBS, SIM) comparison, and that accounting for these should likely deepen this gap. We however believe that current analysis illustrates the usefulness of such a distance, and leave a more detailed study for future work.

[11] Indeed, only one of these pairs is made of patches retrieved from the same MC (Ophiuchus), but these patches do not overlap.

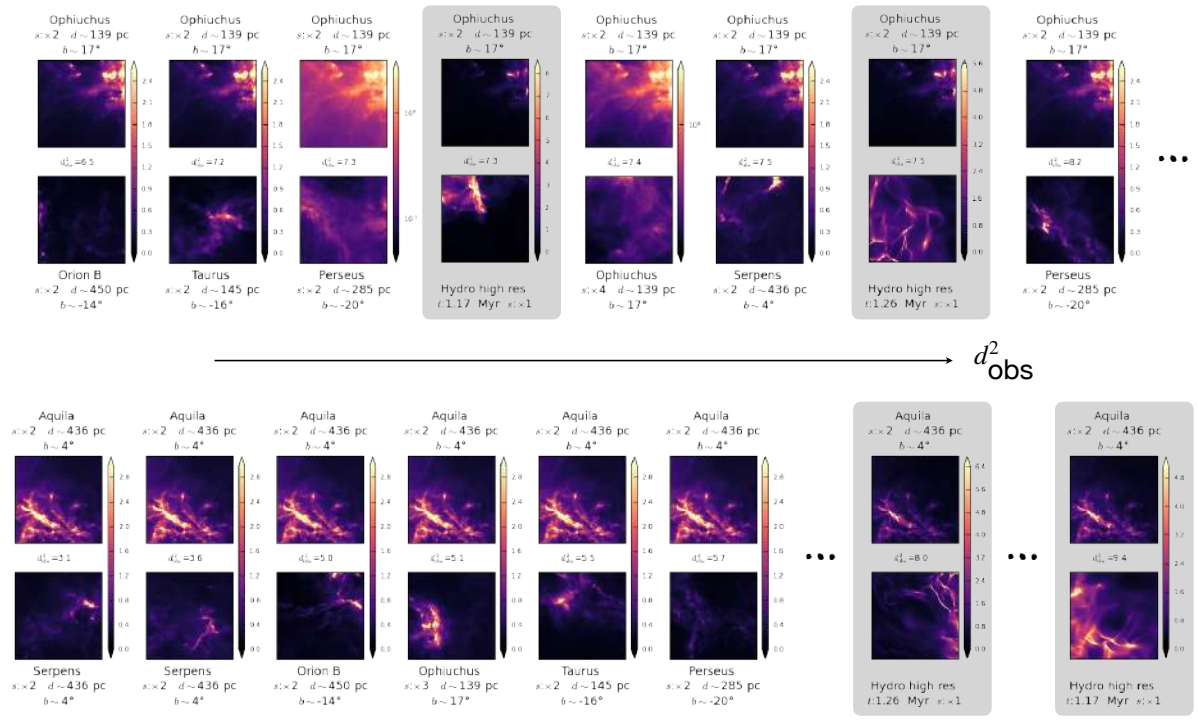


Figure 5.17: Comparing closest (OBS, SIM) pairs (i, j) (gray boxes) to neighboring (OBS, OBS) pairs $(i, i\prime)$. Top row focuses on the closest observation to simulations, that is a patch of Ophiuchus, while bottom row focuses on the second closest patch, that is in Aquila cloud. In the latter case, neighboring (but independent) observations of Aquila are closer than the closest simulations. The color-bars can change from a pair to another.

7 Conclusions

In this paper, we aim at studying the diversity of morphologies of observed molecular clouds. To do so we construct a set of $\sim$ 500 patches of size 512×512 pixels, extracted at different resolutions from column density maps of 14 nearby clouds derived from the HGBS dust emission observations. We compute several sets of statistics (mean, variance, quantile function, power spectrum and scattering transform) from these maps, and we compare their informative power. To do so, we introduce a new methodology (Fig. 5.6) that allows to confront, without any supervision, two sets of summary statistics on their respective abilities to detect statistical incompatibility between pairs of processes in a given dataset.

Applying this methodology to this set of observations, we find that Gaussian statistics have degeneracies for the observational dataset, some of which can be lifted by one-point statistics (Fig. 5.7). We then show that even if log-Gaussian statistics are much less degenerate, a compressed set of scattering statistics still succeeds to demonstrate further degeneracies (Fig. 5.11). This confirms that the diversity of morphologies arising in these observed clouds cannot be sufficiently described by Gaussian nor log-Gaussian statistics. This means that using such descriptions, typically to compare numerical simulations with observational data, can lead to misleading projections: in addition to missing potential absolute discrepancies, observations with

different properties can still be matched to the same simulation.

We apply the same diagnostic to simulations and to the set of logFBMs (Fig. 5.14), and find deviations from observations' geometry. In particular, the outcome of this diagnostic strikingly supports the sufficiency of log-Gaussian statistics to discriminate between logFBM patches. Regarding observations and simulations, there remain deviations between their respective geometries, even though great care has been taken in this work to design robust and low dimensional sets of summary statistics. This supports the difficulty of transferring simulation-based priors[12] to observations, especially high-order information content, and also buttresses the supervision-detached approach developed in this paper, along with the choice to work with compressed and robust summary statistics.

From these results, we introduce a morphological distance d^2_{obs} based on a set of summary statistics ϕ_{final} composed of 7 log-Gaussian and 10 RWST coefficients. The similarity probed by this distance is in agreement with visual impression when comparing observations with themselves, with logFBMs and with simulations (Fig. 5.16). It remains however insufficient to operate on datasets made of a wider diversity of textures such as DTD.

This work opens multiple perspectives:

- the methodology we developed to confront summary statistics requires very few assumptions: it can operate in an unsupervised and very low data regime. Hence such methodology can easily be assimilated by the ISM community and applied on a wider set of statistics and physical tracers of data (velocities, polarization, temperature).
- The low dimensional and analytical morphological embedding ϕ_{final} developed here allows for a directly interpretable comparison, for instance between different observed clouds, between observations and simulations or statistical models, but it also paves the way for the use of more sophisticated unsupervised learning techniques.
- Leveraging saliency maps $\nabla_{\text{pixels}} d^2_{\text{obs}}(x_i, x_j)$, the distance d^2_{obs} can be used to highlight the main areas responsible for morphological discrepancies between two patches (x_i, x_j), broadening the scope of the work initiated by Peek and Burkhart, 2019 to the unsupervised world of observations.
- The confrontation methodology can be used to make a feature selection algorithm, in the spirit of the FRAME model developed by Zhu et al., 1998, but designed to optimize the comparison task of a non-parametric collection of processes $\{p_i\}_i$, instead of modeling a single process.

The following improvements could also be of great benefit:

- reduce the uncertainty in the compatibility diagnostic d^2_ϕ due to the precision matrix estimation. A promising idea is to use a maximum entropy model conditioned on the data on which to perform the precision estimation.

[12] Such as the outcome of a Fisher analysis, or a trained neural network.

- Reduce the overall dependency of the distance $d^2_{\mathcal{D}}$ on the dataset $\mathcal{D}$, and in particular on its outliers. This can be done by building a localized metric, based on local estimations of the geometry, restricting $\mathcal{D}$ to $\mathcal{D}_{\text{loc}}[i]$, the neighbors of a given process i.

Acknowledgements. This research has made use of data from the *Herschel* Gould Belt survey (HGBS) project[13]. The HGBS is a *Herschel* Key Programme jointly carried out by SPIRE Specialist Astronomy Group 3 (SAG 3), scientists of several institutes in the PACS Consortium (CEA Saclay, INAF-IFSI Rome and INAF-Arcetri, KU Leuven, MPIA Heidelberg), and scientists of the *Herschel* Science Center (HSC).

This work reused datasets available on the Galactica simulations database[14].

This work reused the Describable Textures Dataset[15].

8 Appendices

8.1 Other datasets

8.1.1 Numerical simulations

The set of numerical simulations used in this paper as an example of state-of-the-art attempts to reproduce the physics of the ISM *in silico* is taken from the ORION project of the Galactica database[16]. Using the adaptive mesh refinement (AMR) code RAMSES (Fromang et al., 2006; Teyssier, 2002), the data simulates the collapse of a dense molecular cloud under self-gravity, including decaying MHD turbulence but without stellar feedback. The focus of these simulations is to study the early stages of the star formation process in a molecular cloud ($10^5\,\mathrm{M}_\odot$ in a 66 pc cubic box), and so fits in well with the HGBS observational data used in this paper.

Three classes of simulations are considered, one without magnetic field ("Hydro high res"), one with a typical magnetic field of order $5-10\,\mu$G ("MHD"), and one with a higher value of the magnetic field of order $10-25\,\mu$G ("MHD high B"). With the adaptive mesh refinement scheme, the spatial resolution of the models can reach down to 1 mpc ($\sim$ 200 au), and even 100 au for the "Hydro high res" model. For more details about these simulations, we refer the reader to Ntormousi and Hennebelle, 2019.

For each of these models, we take two snapshots, integrate along the three axes, and crop the resulting column density images to keep the central $33\,\mathrm{pc}\times 33\,\mathrm{pc}$ field, on a regular 4096×4096 grid. Examples of such images are shown in Fig. 5.18. As we did for observations, we cut these panels into 512×512 patches, keeping only the ones where the effective resolution is fine enough, to avoid artifacts due to the AMR scheme affecting our morphological analysis. The pixels in the resulting patches have sizes 8 mpc. This is in accordance with the typical spatial

[13] http://gouldbelt-herschel.cea.fr
[14] http://www.galactica-simulations.eu
[15] https://www.robots.ox.ac.uk/~vgg/data/dtd/
[16] http://www.galactica-simulations.eu/db/STAR_FORM/ORION/

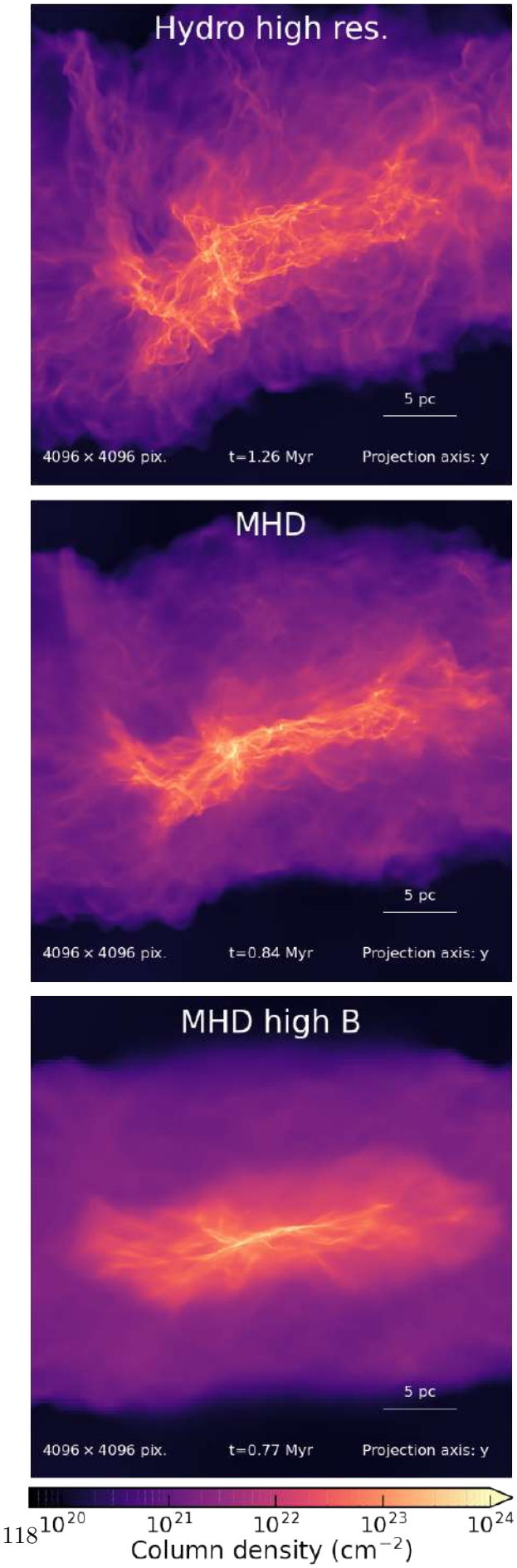


Figure 5.18: Overview of the simulations. One snapshot is shown for each model ("Hydro high res", "MHD", and "MHD high B"), and the maps show the central 33 pc × 33 pc field for gas column density integrated along the y axis.

sampling found in our observational dataset. Indeed, this corresponds to the finest pixels (6 ″) of intermediate-distance clouds $d \sim 300\,\mathrm{pc}$, but also to the closest clouds $d \sim 150\,\mathrm{pc}$ sampled at 12 ″.

8.1.2 logFBM models

To perform our analysis of observational data, we also consider purely synthetic, parametric models of column density maps, derived from exponentiations of fractional Brownian motions (fBm). Such fields have been previously studied by the ISM community (Brunt & Heyer, 2002; Elmegreen, 2002; Levrier et al., 2018; M.-A. Miville-Deschênes et al., 2007). It is indeed a convenient class of models that allows to simultaneously reproduce exactly the log-normal one-point properties (encountered in quiescent regions) and power-law power spectra to a good approximation. An example is given in Fig. 5.1.

In this paper, we consider the following model:

$$X \equiv e^{\sigma F_\beta \star Z + \mu}, \text{ with } Z \sim \mathcal{N}(\mathbf{0}, I), \tag{5.21}$$

parametrized by μ, σ and β, corresponding respectively to the mean, standard deviation and spectral index of the fBm. The latter is generated by sampling a 512×512 Gaussian white noise map Z, which is then filtered with F_β, that is defined in Fourier space as:

$$\tilde{F}_\beta(\mathbf{k}) \equiv \begin{cases} \propto \|\mathbf{k}\|^{-\beta/2} & \text{if } \mathbf{k} \neq \mathbf{0}, \\ 1 & \text{if } \mathbf{k} = \mathbf{0}, \end{cases} \tag{5.22}$$

such that $\sigma F_\beta \star Z + \mu$ is a real valued Gaussian process with mean μ, variance σ^2 and β-decaying power spectrum power-law.

The parameters (μ, σ, β) are fitted from the observations. More precisely, for each 512×512 column density patch that has no negative pixel (that is about 87% of the 550 observational patches), we estimate the mean μ, variance σ^2 and spectral index β on the logarithm of the column density patch, expressed in $10^{22}\,\mathrm{cm}^{-2}$ units. This procedure leads to a distribution of $\sim$ 480 estimated parameters triplets (μ, σ, β) that is reported in Fig. 5.19. For each triplet, we then sample one 512×512 logFBM patch. Contrary to the other patches considered in this paper (observations, simulations, everyday textures), these samples have periodic boundary conditions, by construction. However, their 256×256 sub-patches, that are the only ones on which the various summary statistics are applied, do not, just as the rest of the data.

We dub these synthetic maps logFBM models, to recall that the statistics of their logarithms are those of the Gaussian, fBm random fields. In the literature, emission maps are often produced by integrating a 3D generated field. In this paper, for simplicity and computational efficiency, we directly generate 2D fields. Even though integrated 3D fields could be expected to better model MCs' column density maps, they do not differ much in morphology from 2D models and both are far from reproducing MCs' morphology. Thus, the 2D logFBM models used here are suited to the scope of this paper.

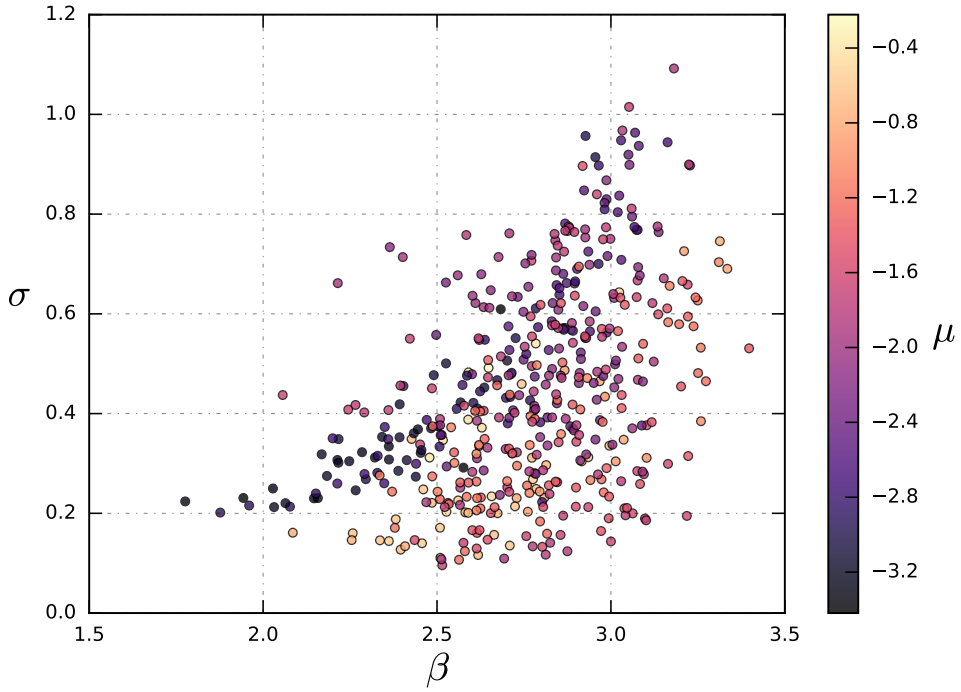


Figure 5.19: log-Gaussian parameters fitted on the observations used to sample logFBM data.

8.1.3 Describable Textures Dataset (DTD)

To better understand the peculiarities of MCs' statistical properties, we compare them to a very wide set of everyday textures: the Describable Textures Dataset (DTD) (Cimpoi et al., 2014). This dataset is in open access[17]. In this paper, we keep only the images with smaller edge size larger than 512 pixels. We then crop each one of these $\sim$ 1050 resulting pictures around its center into a 512×512 patch. Then each patch is converted to gray-scale by summing its three RGB channels. These maps have integer values that range from 0 to $3 \times 255 = 765$ that we finally divide by a factor 10^3 to match the typical values of the other datasets.

8.2 Apodization

The 256×256 sub-patches on which we perform 2D Fourier transform do not have periodic boundary conditions (PBC). Thus, before applying such transform during power spectrum estimation on a given sub-patch x, we first apodize it as follows:

$$x \mapsto rw \cdot (x - \langle x \rangle_{\mathbf{u}}),$$

where $w\cdot$ denotes the pixel-wise multiplication by the window w, reported in Fig. 5.20 and r is a scalar factor depending on the input map x such that the output apodized map has same variance as the input. The resulting map does not share the same mean with the input but this property is not regarded by the Fourier-based statistics used here.

[17] https://www.robots.ox.ac.uk/~vgg/data/dtd/

Figure 5.20: 256×256 apodization window w used. The orange line is a 1D slice of the middle of the window.

8.3 Srivastava & Du test statistic

The test statistic $d^2_\phi(x_i, x_j)$ we introduce in Eq. 5.2 is derived from Srivastava and Du, 2008. We apply this test on the two collections of summary statistics $\{\phi(x_i^{(l)})\}_{1\leq l\leq N_i}$ and $\{\phi(x_j^{(l)})\}_{1\leq l\leq N_j}$ computed on the sub-patches (indexed by l) of the patch x_i and the patch x_j. In our case, $N_i = N_j = 4$. We recall Eq. 5.2 here:

$$d^2_\phi(x_i, x_j) \equiv \alpha\Big[(\hat{\mu}_i - \hat{\mu}_j)^T D_S^{-1}(\hat{\mu}_i - \hat{\mu}_j) - \beta\Big]. \tag{5.23}$$

This equation is based on the following quantities:

$$\hat{\mu}_i \equiv \langle \phi(x_i^{(l)}) \rangle_{1\leq l\leq N_i},$$

$$p = \dim\, \phi, \text{ and } n = N_i + N_j - 2,$$

$$\begin{cases} S_i \equiv \langle [\phi(x_i^{(l)}) - \hat{\mu}_i]\,[\phi(x_i^{(l)}) - \hat{\mu}_i]^T \rangle_{1\leq l\leq N_i}, \\ S \equiv \frac{1}{n}(N_i S_i + N_j S_j), \\ D_S \equiv \operatorname{diag} S, \end{cases}$$

$$\begin{cases} R \equiv D_S^{-1/2} S D_S^{-1/2}, \\ c_{p,n} := 1 + trR^2/p^{3/2}, \\ \alpha \equiv \frac{N_i N_j}{N_i+N_j} \frac{1}{\sqrt{2(trR^2 - p^2/n)c_{p,n}}}, \\ \beta \equiv \frac{N_i+N_j}{N_i N_j} \frac{np}{n-2}. \end{cases}$$

8.4 Why taking the logarithm of some standard statistics?

In theory, because the mapping $\psi \mapsto \log \psi$ is invertible, the same amount of information should be contained in $\phi(x)$ and in $\log \phi(x)$ (when the latter is well defined). However, for the compatibility diagnostic we chose, these two options might perform differently. Such issues are not specific to this test and also occur with standard diagnostics such as Fisher analysis Park et al.,

2023.

First, this non-linear mapping can change the fulfillment level of the test assumptions. In the Srivastava & Du test used here, the two input distributions that are to be compared are assumed to be normally distributed and have equal variance. In many cases, typically when $\phi(x) > 0$, taking the logarithm of $\phi(x)$ decreases the deviation to normality and decreases the relative deviations between the variances of $\phi(x_i)$ and $\phi(x_j)$.

Second, this could lead to dramatic improvement in incompatibility detection, as illustrated in Fig. 5.21 for $\phi = \phi_{\text{var}}$ (provided that the assumptions of the test are not too discarded in both cases for these results to be appropriately interpreted).

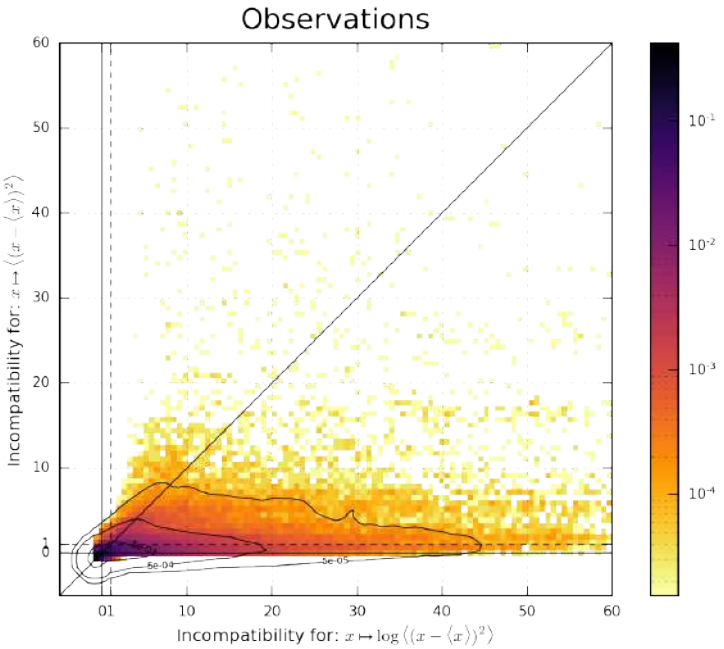


Figure 5.21: Effect of the log rescaling after computing variance statistics on observation data. Provided that the assumptions of the test are not too discarded in both cases for these results to be appropriately interpreted on both axes, this shows that $\left\langle \left[x - \langle x \rangle_{\mathbf{u}}\right]^2 \right\rangle_{\mathbf{u}}$ is strictly more degenerate than $\log \left\langle \left[x - \langle x \rangle_{\mathbf{u}}\right]^2 \right\rangle_{\mathbf{u}}$.

Conclusions & perspectives

In this work, I addressed the problem of characterizing the multi-scale non-Gaussian structures observed in the turbulent ISM. I motivated the importance of developing observational-based characterizations, which however brings the difficulty of operating in a low observational data regime. This regime strongly favors the use of low variance and compressed descriptions.

Still, I showed that, even from very few observational samples, a significant amount of information can be grasped by leveraging the symmetries and regularities of the physical fields. In particular, the compressed set of nonexpansive statistics that are the scattering transform allows for a characterization of some non-Gaussian properties by probing a coupling between scales. I showed that the evolution from the quiescent to active star forming stage in observations of nearby molecular clouds is accompanied with an increasing coupling between distant scales of the gas column density. This morphological diagnostic complements the usual one based on one-point statistics.

Then, being aware that many types of summary statistics are used in the ISM context, and that a single specific one is likely not to be sufficient to account for the ISM complexity, I established a methodology to combine different statistics in order to build an informed description for ISM purpose. I developed this method in order to operate in an unsupervised and low data regime, typically to use it as a tool for estimating the model error of simulations, and also to compare observations between them in a simulation-detached manner. In this unsupervised regime, standard parametric-based definitions of information do not apply. I instead suggested to use a measure of how much dissimilarity between pairs of processes is lost once these are reduced through the summary statistics. Finally, I developed a methodology that allows to estimate this information loss, not in an absolute way, but in a form that can be compared between different sets of summary statistics.

As a result of this work, I built a morphological distance $d_\phi(x, y)$ between two images x and y based on a compressed set of summary statistics ϕ that includes non-Gaussian scale-coupling information. This distance is not based on a prior model and only assumes ergodicity.

Perspectives

This works opens numerous perspectives. To start with, the distance $d_\phi(x, y)$ can be derived as $\nabla_\phi d_\phi(x, y)$ in order to retrieve and interpret the relative contribution of each statistical component in what makes the morphological discrepancy $d_\phi(x, y)$ between the maps x and y. Pushing this derivation further to $\nabla_{x,y} d_\phi(x, y)$, yields saliency maps that enhance in each image the spatial structures that account the most for the morphological discrepancy. Last but not least, "integrating" this saliency map by performing a gradient descent $u(t+1) = u(t) - \alpha(t)\nabla_u d_\phi(u(t), y)$ starting from $u(0) = x$ leads to a spatial "projection" $p_{\phi(y)}[x]$ of x onto the images with texture close to the one of y. The spatial structures of this projection $p_{\phi(y)}[x]$ can then be deterministically compared to the ones of x as shown in Fig. 5.22. Both these approaches, based on the introduction of a proper comprehensive distance between maps, that can be built in a low data regime, offer a new opportunity to interpret the morphological difference between x and y.

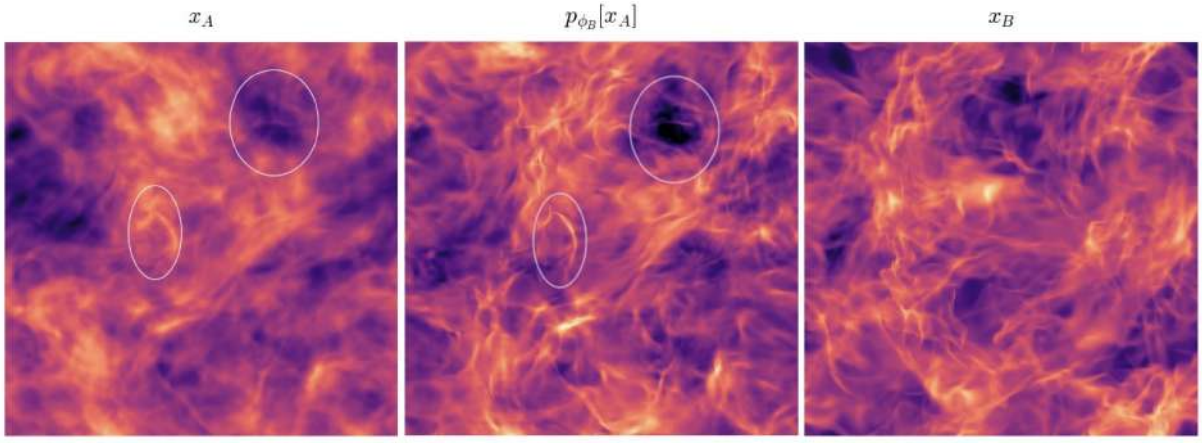


Figure 5.22: Gradient descent projection $p_{\phi_B}[x_A]$ of image x_A onto the image space having a WPH texture $\phi_B \equiv \phi(x_B)$. The maps x_A and $p_{\phi_B}[x_A]$ have different textures but are very close in terms of deterministic structures locations. White circles show how initial structures such as filaments and voids in x_A have been magnified to resemble texture ϕ_B but without involving a significant transport of structures in pixel space. Therefore, the deterministic difference $p_{\phi_B}[x_A] - x_A$ can be used as a good proxy for the morphological difference between x_A and x_B.

Next, to deal with the global non stationarity of molecular clouds, their comparison with the distance $d_\phi(x, y)$ was restricted to pairs of patches of limited size. While it is still not completely clear how to properly estimate the stationarity length related to the size of these patches, such approach could allow for a comparison between molecular clouds. To compare such clouds in their spatial entirety, a first approach could consist in focusing on the distribution of distances $d_\phi(x, y)$ obtained between pairs of local patches extracted from these clouds, as shown in Fig. 5.23. We emphasize, however, that there is no single method for making such a global comparison from (morphological) distances between local patches, and that any such comparison must be guided by the astrophysical scientific goals one wishes to achieve.

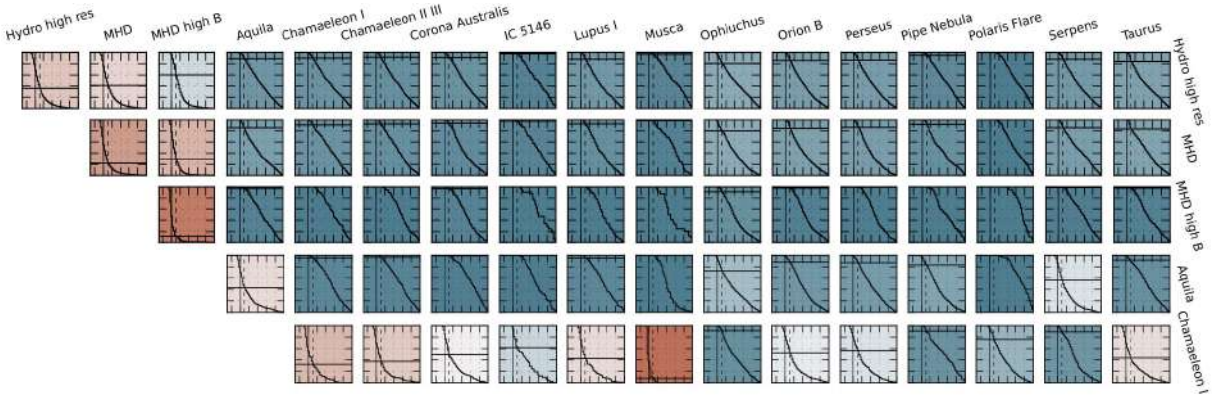


Figure 5.23: Similarity matrix between observations and simulations of molecular clouds (red means similar, blue means dissimilar). The similarity criterion between two molecular clouds is based on a quantile value (horizontal line) of the inverse cumulative distribution (decreasing curve) of distances $d_\phi(x, y)$ obtained between pairs of local patches extracted from these clouds. Note that according to such criteria, a strongly non stationary molecular could can be very dissimilar to itself.

Another point that deserves further work is the mitigation of the uncertainty in the estimation of the precision matrices involved in the computation of $d_\phi(x, y)$ and that limits the discriminative power of this distance. Improvements could be obtained in the case of a comparison between an observation with a simulation, where the low data regime is actually asymmetric since the simulation can be sampled on demand. But care should be made not to fall in the caveat of fully asymmetric techniques such as simulation-based likelihood where only the statistical properties of simulations are used to build the statistical metric. Another promising avenue is to use macrocanonical maximum entropy generative models, conditioned on the empirical moments of the data on which to perform the precision estimations, in order to improve statistical estimates through bootstrap-related approaches[18].

On a more general perspective, the use of generative models, even though they do not bring additional information to the prior data and knowledge that is provided in input, can be of great help for inference algorithms that require a large amount of training data to operate. In the context where there is a very small amount of prior knowledge and data samples to build a generative model, maximum entropy models, that can rely on a predetermined reduced set of non linear but low order empirical moments that concentrate efficiently, such as moments of scattering transform, provide outstanding results. These models are able to faithfully reproduce a wide variety of textures, even when estimated from a single example and without prior knowledge, where neural based methods (such as diffusion models based on neural score matching) fail in this context Allys et al., 2020; Cheng et al., 2024. During my PhD, I took part in various projects that leveraged such generative models for inference purpose, with a special focus on the component separation problem.

[18]See for example Allys et al., 2020, where generative models built from scattering statistics manage to accurately reproduce the sample variance of several summary statistics, including of the scattering statistics themselves.

In (Auclair et al., 2024), we showed that, formulating a $d = s + c$ component separation problem in a maximum entropy framework based on scattering transform constraints, allows to retrieve, using only one sample of the contamination c, non-Gaussian properties of the signal s of interest, down to scales where the contamination dominates in power by more than one order of magnitude. In this paper, this approach allowed to separate the Galactic dust emission from the cosmic infrared background, for infrared observations made by the *Herschel* satellite. Still, this maximum entropy formulation leads to a biased estimation of the statistical properties of the signal of interest at scales where it is sufficiently dominated by the contamination (as entropy is favored when the signal has a large variance), and does not allow to retrieve uncertainties on this bias.

In a ongoing work with E. Allys, R. Soletskyi and A. Tsouros, we suggest to replace this maximum entropy formulation, where the constraints are formulated on the data once it has been contaminated, to a parametric approach where the process of interest is supposed to belong to a certain parametric family, reducing the component separation to estimating the parameters values of this process. Formulated as is, the problem can benefit from the promising simulation-based inference tools, where the "simulation" in question corresponds, for a given parameter of the family, to draw a sample from the associated model of the signal of interest and then apply on it a random contamination. A great challenge of this problem is to find an adequate parametric family of models for the signal of interest. As a response, we suggest to use scattering maximum entropy models, that constitute a promising avenue to parameterize the space of textures. However, this yields the difficulty of having to define a relevant prior in that space of models, as well as adapt existing algorithms to what remains a high-dimensional problem for Bayesian parameter inference.

Last but not least, there is an intensifying demand for sampling generative models in astrophysics, with growing dimensionality of the data (e.g., fields defined on the sphere, hyper-spectral observations, 3D simulation cubes). Therefore, improving the computational efficiency of these samplers is a key objective. Yet, in their macrocanonical form, maximum entropy models designed for the non-Gaussian and multi-scale processes that we found in the ISM are extremely challenging to sample. Indeed, standard methods such as Gibbs sampling are affected by a critical slowing down due to the long-range interactions of these processes. Instead, approximations of microcanonical models based on gradient descent sampling constitute an operational method. Still, significant improvements on the computational efficiency of such models are likely to be made if we take advantage of the hierarchical multi-scale structure of the data. I am therefore interested into various developments of scale-by-scale synthesis algorithms which find echo in the simulation community (Lesaffre, Durrive, et al., 2024).

The intense development over the past decades of the computational capabilities and algorithms in data science are continuously opening unprecedented means to exploit data. Combining these tools with the ever growing simulations, and with the observations that often unveil more complexity as instrumental capabilities are improved too, promises an exciting and enlightening future for astrophysics and cosmology.

Bibliography

Abazajian, K., Addison, G. E., Adshead, P., Ahmed, Z., Akerib, D., Ali, A., Allen, S. W., Alonso, D., Alvarez, M., Amin, M. A., et al. (2022). Cmb-s4: Forecasting constraints on primordial gravitational waves. *The Astrophysical Journal*, *926*(1), 54 (cit. on p. 4).

Ade, P. A. R., Aghanim, N., Ahmed, Z., Aikin, R. W., Alexander, K. D., Arnaud, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barkats, D., Barreiro, R. B., Bartlett, J. G., Bartolo, N., Battaner, E., Benabed, et al. (2015). Joint analysis of bicep2/keck array and planck data. *Phys. Rev. Lett.*, *114*, 101301 (cit. on pp. 5, 6).

Ade, P., Aguirre, J., Ahmed, Z., Aiola, S., Ali, A., Alonso, D., Alvarez, M. A., Arnold, K., Ashton, P., Austermann, J., et al. (2019). The simons observatory: Science goals and forecasts. *Journal of Cosmology and Astroparticle Physics*, *2019*(02), 056 (cit. on p. 4).

Ade, P. A., Aikin, R. W., Barkats, D., Benton, S., Bischoff, C. A., Bock, J., Brevik, J., Buder, I., Bullock, E., Dowell, C., et al. (2014). Detection of b-mode polarization at degree angular scales by bicep2. *Physical Review Letters*, *112*(24), 241101 (cit. on pp. 5, 6).

Agertz, O., Kravtsov, A. V., Leitner, S. N., & Gnedin, N. Y. (2013). Toward a Complete Accounting of Energy and Momentum from Stellar Feedback in Galaxy Formation Simulations. *The Astrophysical Journal*, *770*(1), Article 25, 25 (cit. on p. 3).

Akhiezer, N. I. (2020). *The classical moment problem and some related questions in analysis*. SIAM. (Cit. on p. 45).

Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, *28*(1), 131–142 (cit. on p. 77).

Allys, E. (2017). *Au-delà des modèles standards en cosmologie* [Doctoral dissertation, Université Pierre et Marie Curie - Paris VI]. (Cit. on p. 46).

Allys, E., Levrier, F., Zhang, S., Colling, C., Regaldo-Saint Blancard, B., Boulanger, F., Hennebelle, P., & Mallat, S. (2019). The rwst, a comprehensive statistical description of the non-gaussian structures in the ism. *Astronomy & Astrophysics*, *629*, A115 (cit. on pp. 23, 31, 58, 61, 88, 100).

Allys, E., Marchand, T., Cardoso, J.-F., Villaescusa-Navarro, F., Ho, S., & Mallat, S. (2020). New interpretable statistics for large-scale structure analysis and generation. *Physical Review D*, *102*(10), 103506 (cit. on pp. 61, 71, 125).

Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of wildlife management*, 912–918 (cit. on pp. 73, 74).

André, P., Men'shchikov, A., Bontemps, S., Könyves, V., Motte, F., Schneider, N., Didelon, P., Minier, V., Saraceno, P., Ward-Thompson, D., et al. (2010). From filamentary clouds to prestellar cores to the stellar imf: Initial highlights from the herschel gould belt survey. *Astronomy & Astrophysics*, *518*, L102 (cit. on pp. 20, 91).

André, P., Di Francesco, J., Ward-Thompson, D., Inutsuka, S.-i., Pudritz, R. E., & Pineda, J. (2014). From filamentary networks to dense cores in molecular clouds: Toward a new paradigm for star formation. *Protostars and Planets VI*, *27* (cit. on p. 93).

Appel, S. M., Burkhart, B., Semenov, V. A., Federrath, C., & Rosen, A. L. (2022). The effects of magnetic fields and outflow feedback on the shape and evolution of the density probability distribution function in turbulent star-forming clouds. *The Astrophysical Journal*, *927*(1), 75 (cit. on pp. 53, 87).

Armstrong, J., Rickett, B., & Spangler, S. (1995). Electron density power spectrum in the local interstellar medium. *Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 443, no. 1, p. 209-221*, *443*, 209–221 (cit. on p. 13).

Arzoumanian, D., André, P., Didelon, P., Könyves, V., Schneider, N., Men'shchikov, A., Sousbie, T., Zavagno, A. e., Bontemps, S., Di Francesco, J., et al. (2011). Characterizing interstellar filaments with herschel in ic 5146. *Astronomy & Astrophysics*, *529*, L6 (cit. on p. 93).

Auclair, C., Allys, E., Boulanger, F., Béthermin, M., Gkogkou, A., Lagache, G., Marchal, A., Miville-Deschênes, M.-A., Régaldo-Saint Blancard, B., & Richard, P. (2024). Separation of dust emission from the cosmic infrared background in herschel observations with wavelet phase harmonics. *Astronomy & Astrophysics*, *681*, A1 (cit. on pp. 61, 99, 112, 126).

Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta mathematicae*, *3*(1), 133–181 (cit. on p. 29).

Barkana, R., & Loeb, A. (2001). In the beginning: The first sources of light and the reionization of the universe. *Physics Reports*, *349*(2), 125–238 (cit. on p. 3).

Batchelor, G. K., & Townsend, A. A. (1949). The nature of turbulent motion at large wave-numbers. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *199*(1057), 238–255 (cit. on p. 57).

Bezruchko, B. P., & Smirnov, D. A. (2010). *Extracting knowledge from time series: An introduction to nonlinear empirical modeling*. Springer Science & Business Media. (Cit. on pp. 25, 28).

Blum, M., Nunes, M., Prangle, D., & Sisson, S. (2013). A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, *28*(2), 189–208 (cit. on p. 74).

Bruna, J., & Mallat, S. (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1872–1886 (cit. on pp. 61, 100).

Bruna, J., & Mallat, S. (2019). Multiscale sparse microcanonical models. *Mathematical Statistics and Learning*, *1*(3), 257–315 (cit. on p. 52).

Bruna, J., Mallat, S., Bacry, E., & Muzy, J.-F. (2015). Intermittent process analysis with scattering moments (cit. on p. 58).

Brunt, C. M., & Heyer, M. H. (2002). Interstellar turbulence. i. retrieval of velocity field statistics. *The Astrophysical Journal*, *566*(1), 276 (cit. on pp. 91, 119).

Burkhart, B., Falceta-Gonçalves, D., Kowal, G., & Lazarian, A. (2009). Density studies of mhd interstellar turbulence: Statistical moments, correlations and bispectrum. *The Astrophysical Journal*, *693*(1), 250 (cit. on pp. 60, 88).

Burkhart, B., & Lazarian, A. (2016). The phase coherence of interstellar density fluctuations. *The Astrophysical Journal*, *827*(1), 26 (cit. on p. 88).

Cheng, S., Morel, R., Allys, E., Ménard, B., & Mallat, S. (2024). Scattering spectra models for physics. *PNAS Nexus*, *3*(4), pgae103 (cit. on pp. 61, 62, 125).

Chepurnov, A., & Lazarian, A. (2010). Extending the big power law in the sky with turbulence spectra from wisconsin hα mapper data. *The Astrophysical Journal*, *710*(1), 853 (cit. on p. 13).

Chomiuk, L., & Povich, M. S. (2011). Toward a unification of star formation rate determinations in the milky way and other galaxies. *The Astronomical Journal*, *142*(6), 197 (cit. on p. 2).

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613 (cit. on pp. 64, 91, 120).

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory (wiley series in telecommunications and signal processing)*. Wiley-Interscience. (Cit. on pp. 52, 70, 71, 96, 107).

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062 (cit. on pp. 39, 40, 42, 51, 73).

Csiszár, I. (1964). Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, *8*, 85–108 (cit. on p. 77).

Darmois, G. (1935). Sur les lois de probabilitéa estimation exhaustive. *CR Acad. Sci. Paris*, *260*(1265), 85 (cit. on p. 70).

De Karman, T., & Howarth, L. (1938). On the statistical theory of isotropic turbulence. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, *164*(917), 192–215 (cit. on p. 56).

Delabrouille, J., & Cardoso, J.-F. (2008). Diffuse source separation in cmb observations. In *Data analysis in cosmology* (pp. 159–205). Springer. (Cit. on p. 4).

Dell'Ova, P. (2021, December). *Exploring the interstellar contents of a TeVatron : molecules, dust and star formation in the evolved supernova remnant IC443* (Publication No. 2021UPSLO005) [Theses]. Université Paris sciences et lettres. (Cit. on p. 9).

Diaconis, P. (1988). Sufficiency as statistical symmetry. *Proceedings of the AMS Centennial Symposium*, 15–26 (cit. on p. 70).

Dicke, R. H., Peebles, P. J. E., Roll, P. G., & Wilkinson, D. T. (1965). Cosmic black-body radiation. *Astrophysical Journal, vol. 142, p. 414-419*, *142*, 414–419 (cit. on p. 4).

Draine, B. T. (2010). *Physics of the interstellar and intergalactic medium* (Vol. 19). Princeton University Press. (Cit. on pp. 3, 7–9, 11, 23).

Elmegreen, B. G. (2002). A fractal origin for the mass spectrum of interstellar clouds. ii. cloud models and power-law slopes. *The Astrophysical Journal*, *564*(2), 773 (cit. on pp. 91, 119).

Elmegreen, B. G., & Scalo, J. (2004). Interstellar turbulence i: Observations and processes. *Annu. Rev. Astron. Astrophys.*, *42*(1), 211–273 (cit. on pp. 10, 18, 31).

Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, *49*(7), 1858–1860 (cit. on p. 76).

Falgarone, E., Hily-Blant, P., & Levrier, F. (2004). Structure of molecular clouds. *Astrophysics and Space Science*, *292*(1), 89–101 (cit. on pp. 73, 74, 88).

Farge, M., et al. (1992). Wavelet transforms and their applications to turbulence. *Annual review of fluid mechanics*, *24*(1), 395–458 (cit. on p. 57).

Federrath, C. (2013). On the universality of supersonic turbulence. *Monthly Notices of the Royal Astronomical Society*, *436*(2), 1245–1257 (cit. on p. 31).

Federrath, C., & Klessen, R. S. (2013). On the star formation efficiency of turbulent magnetized clouds. *The Astrophysical Journal*, *763*(1), 51 (cit. on p. 87).

Federrath, C., Roman-Duval, J., Klessen, R. S., Schmidt, W., & Mac Low, M.-M. (2010). Comparing the statistics of interstellar turbulence in simulations and observations-solenoidal versus compressive turbulence forcing. *Astronomy & Astrophysics*, *512*, A81 (cit. on p. 72).

Field, G. B. (1965). Thermal instability. *Astrophysical Journal, vol. 142, p. 531*, *142*, 531 (cit. on p. 8).

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical proceedings of the Cambridge philosophical society*, *22*(5), 700–725 (cit. on p. 69).

Frisch, U., & Kolmogorov, A. N. (1995). *Turbulence: The legacy of an kolmogorov*. Cambridge university press. (Cit. on pp. 13, 17, 28, 31, 56, 57).

Frisch, U., Sulem, P.-L., & Nelkin, M. (1978). A simple dynamical model of intermittent fully developed turbulence. *Journal of Fluid Mechanics*, *87*(4), 719–736 (cit. on p. 11).

Fromang, S., Hennebelle, P., & Teyssier, R. (2006). A high order Godunov scheme with constrained transport and adaptive mesh refinement for astrophysical magnetohydrodynamics. *Astronomy & Astrophysics*, *457*(2), 371–384 (cit. on p. 117).

Galli, P. A. B., Bertout, C., Teixeira, R., & Ducourant, C. (2013). A kinematic study and membership analysis of the Lupus star-forming region. *Astronomy & Astrophysics*, *558*, Article A77, A77 (cit. on p. 95).

Galliano, F., Galametz, M., & Jones, A. P. (2018). The interstellar dust properties of nearby galaxies. *Annual Review of Astronomy and Astrophysics*, *56*(1), 673–713 (cit. on pp. 3, 10).

Galtier, S. (2016). *Introduction to modern magnetohydrodynamics*. Cambridge University Press. (Cit. on p. 16).

Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., & Cord, M. (2019). Boosting few-shot visual learning with self-supervision. *Proceedings of the IEEE/CVF international conference on computer vision*, 8059–8068 (cit. on p. 72).

Goicoechea, J. R., Pety, J., Cuadrado, S., Cernicharo, J., Chapillon, E., Fuente, A., Gerin, M., Joblin, C., Marcelino, N., & Pilleri, P. (2016). Compression and ablation of the photo-irradiated molecular cloud the orion bar. *Nature*, *537*(7619), 207–209 (cit. on p. 9).

Goodman, A. A., Rosolowsky, E. W., Borkin, M. A., Foster, J. B., Halle, M., Kauffmann, J., & Pineda, J. E. (2009). A role for self-gravity at multiple length scales in the process of star formation. *Nature*, *457*(7225), 63–66 (cit. on p. 88).

Halmos, P. R. (2017). *Lectures on ergodic theory*. Courier Dover Publications. (Cit. on p. 32).

Halmos, P. R., & Savage, L. J. (1949). Application of the radon-nikodym theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, *20*(2), 225–241 (cit. on pp. 67, 69, 75).

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, 1029–1054 (cit. on p. 44).

Hauser, M. G., & Dwek, E. (2001). The cosmic infrared background: Measurements and implications. *Annual Review of Astronomy and Astrophysics*, *39*(1), 249–307 (cit. on p. 3).

Hazumi, M., Ade, P. A., Adler, A., Allys, E., Arnold, K., Auguste, D., Aumont, J., Aurlien, R., Austermann, J., Baccigalupi, C., et al. (2020). Litebird satellite: Jaxa's new strategic l-class mission for all-sky surveys of cosmic microwave background polarization. *Space Telescopes and Instrumentation 2020: Optical, Infrared, and Millimeter Wave*, *11443*, 431–450 (cit. on p. 4).

Heithausen, A., & Thaddeus, P. (1990). The Polaris Flare: Extensive Molecular Gas near the North Celestial Pole. *Astrophysical Journal, Part 2-Letters*, *353*, L49 (cit. on p. 91).

Hennebelle, P., & Chabrier, G. (2008). Analytical theory for the initial mass function: Co clumps and prestellar cores. *The Astrophysical Journal*, *684*(1), 395 (cit. on pp. 18, 87).

Hennebelle, P., & Falgarone, E. (2012). Turbulent molecular clouds. *The Astronomy and Astrophysics Review*, *20*, 1–58 (cit. on pp. 2, 10, 11, 18, 53).

Hensley, B. S., & Bull, P. (2018). Mitigating complex dust foregrounds in future cosmic microwave background polarization experiments. *The Astrophysical Journal*, *853*(2), 127 (cit. on p. 5).

Heyer, M. H., & Brunt, C. M. (2004). The universality of turbulence in galactic molecular clouds. *The Astrophysical Journal*, *615*(1), L45 (cit. on pp. 31, 88).

Hily-Blant, P., Falgarone, E., & Pety, J. (2008). Dissipative structures of diffuse molecular gas-iii. small-scale intermittency of intense velocity-shears. *Astronomy & Astrophysics*, *481*(2), 367–380 (cit. on p. 58).

Holl, P., & Thuerey, N. (2024). Φ_{Flow} (PhiFlow): Differentiable simulations for pytorch, tensorflow and jax. *International Conference on Machine Learning* (cit. on p. 42).

Hollenbach, D. J., & Tielens, A. (1999). Photodissociation regions in the interstellar medium of galaxies. *Reviews of Modern Physics*, *71*(1), 173 (cit. on p. 8).

Hotelling, H. (1931). The generalization of student's ratio (cit. on p. 96).

Hothi, I., Allys, E., Semelin, B., & Boulanger, F. (2024). Wavelet-based statistics for enhanced 21cm eor parameter constraints. *Astronomy & Astrophysics*, *686*, A212 (cit. on pp. 71, 72).

Iffrig, O., & Hennebelle, P. (2017). Structure distribution and turbulence in self-consistently supernova-driven ism of multiphase magnetized galactic discs. *Astronomy & Astrophysics*, *604*, A70 (cit. on p. 23).

Jeffrey, N., Boulanger, F., Wandelt, B. D., Regaldo-Saint Blancard, B., Allys, E., & Levrier, F. (2022). Single frequency cmb b-mode inference with realistic foregrounds from a single training image. *Monthly Notices of the Royal Astronomical Society: Letters*, *510*(1), L1–L6 (cit. on p. 61).

Jenkins, E. B., & Tripp, T. M. (2011). The distribution of thermal pressures in the diffuse, cold neutral medium of our galaxy. ii. an expanded survey of interstellar ci fine-structure excitations. *The Astrophysical Journal*, *734*(1), 65 (cit. on p. 11).

Kainulainen, J., Beuther, H., Henning, T., & Plume, R. (2009). Probing the evolution of molecular cloud structure-from quiescence to birth. *Astronomy & Astrophysics*, *508*(3), L35–L38 (cit. on pp. 54, 87).

Kamionkowski, M., Kosowsky, A., & Stebbins, A. (1997). A probe of primordial gravity waves and vorticity. *Physical Review Letters*, *78*(11), 2058 (cit. on p. 5).

Kamionkowski, M., & Kovetz, E. D. (2016). The quest for b modes from inflationary gravitational waves. *Annual Review of Astronomy and Astrophysics*, *54*(Volume 54, 2016), 227–269 (cit. on p. 5).

Knude, J., & Hog, E. (1998). Interstellar reddening from the HIPPARCOS and TYCHO catalogues. I. Distances to nearby molecular clouds and star forming regions. *Astronomy & Astrophysics*, *338*, 897–904 (cit. on p. 95).

Kolmogorov, A. N. (1941). The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. *Dokl. Akad. Nauk SSSR*, *30* (cit. on p. 57).

Könyves, V., André, P., Men'Shchikov, A., Palmeirim, P., Arzoumanian, D., Schneider, N., Roy, A., Didelon, P., Maury, A., Shimajiri, Y., et al. (2015). A census of dense cores in the aquila cloud complex: Spire/pacs observations from the herschel gould belt survey. *Astronomy & Astrophysics*, *584*, A91 (cit. on p. 91).

Könyves, V., André, Ph., Arzoumanian, D., Schneider, N., Men'shchikov, A., Bontemps, S., Ladjelate, B., Didelon, P., Pezzuto, S., Benedettini, M., Bracco, A., Di Francesco, J., Goodwin, S., Rygl, K. L. J., Shimajiri, Y., Spinoglio, L., Ward-Thompson, D., & White, G. J. (2020). Properties of the dense core population in orion b as seen by the herschel gould belt survey. *Astronomy & Astrophysics*, *635*, A34 (cit. on p. 20).

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, *39*(3), 399–409 (cit. on p. 70).

Kraichnan, R. H., & Chen, S. (1989). Is there a statistical mechanics of turbulence? *Physica D: Nonlinear Phenomena*, *37*(1-3), 160–172 (cit. on p. 31).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86 (cit. on p. 77).

Leach, S. M., Cardoso, J.-F., Baccigalupi, C., Barreiro, R., Betoule, M., Bobin, J., Bonaldi, A., Delabrouille, J., De Zotti, G., Dickinson, C., et al. (2008). Component separation methods for the planck mission. *Astronomy & Astrophysics*, *491*(2), 597–615 (cit. on p. 4).

Lehmann, E. L., & Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media. (Cit. on p. 71).

Lei, M., & Clark, S. (2023). Probing the cold neutral medium through hi emission morphology with the scattering transform. *The Astrophysical Journal*, *947*(2), 74 (cit. on pp. 61, 73).

Lesaffre, P. (2018). *Dynamics of the galactic matter cycle* [Habilitation à Diriger des Recherches]. Observatoire de Paris-PSL. (Cit. on p. 9).

Lesaffre, P., Durrive, J.-B., Goossaert, J., Poirier, S., Colombi, S., Richard, P., Allys, E., & Béthune, W. (2024). Multiscale turbulence synthesis in 2d hydrodynamics. *In prep.* (cit. on p. 126).

Lesaffre, P., Falgarone, E., & Hily-Blant, P. (2024). The intermittency of turbulence in magneto-hydrodynamical simulations and in the cosmos. *Atmosphere*, *15*(2), 211 (cit. on pp. 8, 12, 57).

Levrier, F., Neveu, J., Falgarone, E., Boulanger, F., Bracco, A., Ghosh, T., & Vansyngel, F. (2018). Statistics of the polarized submillimetre emission maps from thermal dust in the turbulent, magnetized, diffuse ISM. *Astronomy & Astrophysics*, *614*, A124 (cit. on pp. 91, 119).

Levrier, F., Falgarone, E., & Viallefond, F. (2006). Fourier phase analysis in radio-interferometry. *Astronomy & Astrophysics*, *456*(1), 205–214 (cit. on p. 88).

Licquia, T. C., & Newman, J. A. (2015). Improved estimates of the milky way's stellar mass and star formation rate from hierarchical bayesian meta-analysis. *The Astrophysical Journal*, *806*(1), 96 (cit. on p. 2).

Liese, F., & Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, *52*(10), 4394–4412 (cit. on p. 78).

LiteBIRD Collaboration et al. (2023). Probing cosmic inflation with the LiteBIRD cosmic microwave background polarization survey. *Progress of Theoretical and Experimental Physics*, *2023*(4), Article 042F01, 042F01 (cit. on p. 5).

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, *35*(1), 857–876 (cit. on p. 72).

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of atmospheric sciences*, *20*(2), 130–141 (cit. on p. 23).

Lyth, D. H. (1997). What would we learn by detecting a gravitational wave signal in the cosmic microwave background anisotropy? *Physical Review Letters*, *78*(10), 1861 (cit. on p. 5).

Mac Low, M.-M., & Klessen, R. S. (2004). Control of star formation by supersonic turbulence. *Rev. Mod. Phys.*, *76*, 125–194 (cit. on p. 9).

Mallat, S., Zhang, S., & Rochette, G. (2020). Phase harmonic correlations and convolutional neural networks. *Information and Inference: A Journal of the IMA*, *9*(3), 721–747 (cit. on p. 61).

Mamajek, E. E. (2008). On the distance to the Ophiuchus star-forming region. *Astronomische Nachrichten*, *329*(1), 10 (cit. on p. 95).

Marchal, A., & Martin, P. G. (2023). On the origin of the north celestial pole loop. *The Astrophysical Journal*, *942*(2), 70 (cit. on p. 61).

McKee, C. F., & Ostriker, E. C. (2007). Theory of star formation. *Annu. Rev. Astron. Astrophys.*, *45*(1), 565–687 (cit. on pp. 2, 8, 87).

Miller, J. (1990). Statistical mechanics of euler equations in two dimensions. *Phys. Rev. Lett.*, *65*, 2137–2140 (cit. on p. 30).

Miville-Deschênes, M. .-., Martin, P. G., Abergel, A., Bernard, J. .-., Boulanger, F., Lagache, G., Anderson, L. D., André, P., Arab, H., Baluteau, J. .-., Blagrave, K., Bontemps, S., Cohen, M., Compiegne, M., Cox, P., Dartois, E., Davis, G., Emery, R., Fulton, T., ... Zavagno, A. (2010). Herschel-SPIRE observations of the Polaris flare: Structure of the diffuse interstellar medium at the sub-parsec scale. *Astronomy & Astrophysics*, *518*, Article L104, L104 (cit. on p. 91).

Miville-Deschênes, M.-A., Duc, P.-A., Marleau, F., Cuillandre, J.-C., Didelon, P., Gwyn, S., & Karabal, E. (2016). Probing interstellar turbulence in cirrus with deep optical imaging: No sign of energy dissipation at 0.01 pc scale. *Astronomy & Astrophysics*, *593*, A4 (cit. on p. 12).

Miville-Deschênes, M.-A., Lagache, G., Boulanger, F., & Puget, J.-L. (2007). Statistical properties of dust far-infrared emission. *Astronomy & Astrophysics*, *469*(2), 595–605 (cit. on pp. 87, 91, 119).

Miville-Deschênes, M.-A., Murray, N., & Lee, E. J. (2017). Physical properties of molecular clouds for the entire milky way disk. *The Astrophysical Journal*, *834*(1), 57 (cit. on p. 2).

Mousset, L., Allys, E., Price, M. A., Aumont, J., Delouis, J.-M., Montier, L., & McEwen, J. D. (2024). Generative models of astrophysical fields with scattering transforms on the sphere. *arXiv preprint arXiv:2407.07007* (cit. on p. 61).

Myers, P. C., Dame, T. M., Thaddeus, P., Cohen, R. S., Silverberg, R. F., Dwek, E., & Hauser, M. G. (1986). Molecular Clouds and Star Formation in the Inner Galaxy: A Comparison of CO, H ii, and Far-Infrared Surveys. *The Astrophysical Journal*, *301*, 398 (cit. on p. 2).

Neyman, J. (1936). *Su un teorema concernente le cosiddette statistiche sufficienti*. Istituto Italiano degli Attuari. (Cit. on p. 69).

Nguyen, M. Q., Shadloo, M. S., Hadjadj, A., Lebon, B., & Peixinho, J. (2019). Perturbation threshold and hysteresis associated with the transition to turbulence in sudden expansion pipe flow. *International Journal of Heat and Fluid Flow*, *76*, 187–196 (cit. on p. 31).

Nielsen, F. (2013). Cramér-rao lower bound and information geometry. *Connected at Infinity II: A Selection of Mathematics by Indians*, 18–37 (cit. on p. 72).

Norman, C., & Silk, J. (1980). Clumpy molecular clouds-a dynamic model self-consistently regulated by t tauri star formation. *Astrophysical Journal, Part 1, vol. 238, May 15, 1980, p. 158-174. NATO-supported research; 238*, 158–174 (cit. on p. 9).

Ntormousi, E., & Hennebelle, P. (2019). Core and stellar mass functions in massive collapsing filaments. *Astronomy & Astrophysics*, *625*, A82 (cit. on pp. 20, 117).

Nunes, M. A., & Balding, D. J. (2010). On optimal selection of summary statistics for approximate bayesian computation. *Statistical applications in genetics and molecular biology*, *9*(1) (cit. on p. 74).

Ortiz-León, G. N., Loinard, L., Dzib, S. A., Kounkel, M., Galli, P. A. B., Tobin, J. J., Evans, I., Neal J., Hartmann, L., Rodríguez, L. F., Briceño, C., Torres, R. M., & Mioduszewski, A. J. (2018). Gaia-DR2 Confirms VLBA Parallaxes in Ophiuchus, Serpens, and Aquila. *The Astrophysical Journal Letters*, *869*(2), Article L33, L33 (cit. on p. 95).

Ossenkopf-Okada, V., Csengeri, T., Schneider, N., Federrath, C., & Klessen, R. S. (2016). The reliability of observational measurements of column density probability distribution functions. *Astronomy & Astrophysics*, *590*, A104 (cit. on p. 99).

Padoan, P., Federrath, C., Chabrier, G., Evans, N., Johnstone, D., Jørgensen, J. K., McKee, C. F., Nordlund, Å., Beuther, H., Klessen, R., et al. (2014). The star formation rate of molecular clouds. *Protostars and Planets VI*, *77* (cit. on pp. 31, 53).

Padoan, P., & Nordlund, Å. (2002). The stellar initial mass function from turbulent fragmentation. *The Astrophysical Journal*, *576*(2), 870 (cit. on p. 18).

Padoan, P., & Nordlund, Å. (2011). The Star Formation Rate of Supersonic Magnetohydrodynamic Turbulence. *The Astrophysical Journal*, *730*(1), Article 40, 40 (cit. on p. 53).

Palmeirim, P. a., André, P., Kirk, J., Ward-Thompson, D., Arzoumanian, D. a., Könyves, V., Didelon, P., Schneider, N., Benedettini, M., Bontemps, S., et al. (2013). Herschel view of the taurus b211/3 filament and striations: Evidence of filamentary growth? *Astronomy & Astrophysics*, *550*, A38 (cit. on p. 93).

Palmer, T., & Hagedorn, R. (2006). *Predictability of weather and climate*. Cambridge University Press. (Cit. on p. 25).

Pardo, M., & Vajda, I. (1997). About distances of discrete distributions satisfying the data processing theorem of information theory. *IEEE transactions on information theory*, *43*(4), 1288–1293 (cit. on p. 78).

Park, C. F., Allys, E., Villaescusa-Navarro, F., & Finkbeiner, D. (2023). Quantification of high-dimensional non-gaussianities and its implication to fisher analysis in cosmology. *The Astrophysical Journal*, *946*(2), 107 (cit. on pp. 71, 121).

Peek, J., & White, R. (2021). Search By Image: Citizen Science and Deep Learning for next-generation archives. *Bulletin of the AAS*, *53*(6) (cit. on pp. 73, 113).

Peek, J., & Burkhart, B. (2019). Do androids dream of magnetic fields? using neural networks to interpret the turbulent interstellar medium. *The Astrophysical Journal Letters*, *882*(1), L12 (cit. on pp. 35, 37, 72, 88, 116).

Penzias, A. A., & Wilson, R. W. (1965). A Measurement of Excess Antenna Temperature at 4080 Mc/s. *The Astrophysical Journal*, *142*, 419–421 (cit. on p. 4).

Pilbratt, G. L., Riedinger, J. R., Passvogel, T., Crone, G., Doyle, D., Gageur, U., Heras, A. M., Jewell, C., Metcalfe, L., Ott, S., & Schmidt, M. (2010). Herschel Space Observatory. An ESA facility for far-infrared and submillimetre astronomy. *Astronomy & Astrophysics*, *518*, Article L1, L1 (cit. on p. 91).

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society*, *32*(4), 567–579 (cit. on p. 70).

Planck Collaboration et al. (2014a). Planck 2013 results. xi. all-sky model of thermal dust emission. *Astronomy & Astrophysics*, *571*, A11 (cit. on p. 19).

Planck Collaboration et al. (2014b). Planck 2013 results. xxx. cosmic infrared background measurements and implications for star formation. *A&A*, *571*, A30 (cit. on p. 3).

Planck Collaboration et al. (2015). Planck intermediate results. xix. an overview of the polarized thermal emission from galactic dust. *Astronomy & Astrophysics*, *576*, A104 (cit. on p. 5).

Planck Collaboration et al. (2017). Planck intermediate results-l. evidence of spatial variation of the polarized thermal dust spectral energy distribution and implications for cmb b-mode analysis. *Astronomy & Astrophysics*, *599*, A51 (cit. on p. 5).

Planck Collaboration et al. (2020a). Planck 2018 results-i. overview and the cosmological legacy of planck. *Astronomy & Astrophysics*, *641*, A1 (cit. on pp. 4, 5, 38).

Planck Collaboration et al. (2020b). Planck 2018 results-iv. diffuse component separation. *Astronomy & Astrophysics*, *641*, A4 (cit. on p. 4).

Planck Collaboration et al. (2020c). Planck 2018 results. XII. Galactic astrophysics using polarized dust emission. *Astronomy & Astrophysics*, *641*, Article A12, A12 (cit. on p. 92).

Politano, H., & Pouquet, A. (1995). Model of intermittency in magnetohydrodynamic turbulence. *Physical Review E*, *52*(1), 636 (cit. on p. 57).

Pollard, D. (2013). A note on insufficiency and the preservation of fisher information. In *From probability to statistics and back: High-dimensional models and processes–a festschrift in honor of jon a. wellner* (pp. 266–276, Vol. 9). Institute of Mathematical Statistics. (Cit. on p. 72).

Polyanskiy, Y., & Verdú, S. (2010). Arimoto channel coding converse and rényi divergence. *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1327–1333 (cit. on p. 78).

Pouteau, Y., Motte, F., Nony, T., González, M., Joncour, I., Robitaille, J. .-., Busquet, G., Galván-Madrid, R., Gusdorf, A., Hennebelle, P., Ginsburg, A., Csengeri, T., Sanhueza, P., Dell'Ova, P., Stutz, A. M., Towner, A. P. M., Cunningham, N., Louvet, F., Men'shchikov, A., ... Valeille-Manet, M. (2023). ALMA-IMF. VI. Investigating the origin of stellar masses: Core mass function evolution in the W43-MM2&MM3 mini-starburst. *Astronomy & Astrophysics*, *674*, Article A76, A76 (cit. on p. 99).

Regaldo-Saint Blancard, B., Allys, E., Boulanger, F., Levrier, F., & Jeffrey, N. (2021). A new approach for the statistical denoising of planck interstellar dust polarization data. *Astronomy & Astrophysics*, *649*, L18 (cit. on p. 61).

Regaldo-Saint Blancard, B., Levrier, F., Allys, E., Bellomi, E., & Boulanger, F. (2020). Statistical description of dust polarized emission from the diffuse interstellar medium-a rwst approach. *Astronomy & Astrophysics*, *642*, A217 (cit. on pp. 88, 101).

Remazeilles, M., Delabrouille, J., & Cardoso, J.-F. (2011). Foreground component separation with generalized Internal Linear Combination. *Monthly Notices of the Royal Astronomical Society*, *418*(1), 467–476 (cit. on p. 92).

Rényi, A. (1961). On measures of entropy and information. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, *4*, 547–562 (cit. on p. 77).

Richard, T. (2022). *Caractéristiques de la dissipation turbulente dans le milieu interstellaire* [Doctoral dissertation, Université Paris sciences et lettres]. (Cit. on p. 11).

Richard, T., Lesaffre, P., Falgarone, E., & Lehmann, A. (2022). Probing the nature of dissipation in compressible mhd turbulence. *Astronomy & Astrophysics*, *664*, A193 (cit. on p. 12).

Robitaille, J.-F., Motte, F., Schneider, N., Elia, D., & Bontemps, S. (2019). Exposing the plural nature of molecular clouds-extracting filaments and the cosmic infrared background against the true scale-free interstellar medium. *Astronomy & Astrophysics*, *628*, A33 (cit. on pp. 58, 88).

Ruelle, D. (1989). *Chaotic evolution and strange attractors* (Vol. 1). Cambridge University Press. (Cit. on p. 31).

Ruelle, D. (1991). The turbulent fluid as a dynamical system. *New perspectives in turbulence*, 123–138 (cit. on p. 31).

Sason, I. (2019). On data-processing and majorization inequalities for f-divergences with applications. *Entropy*, *21*(10), 1022 (cit. on p. 78).

Saydjari, A. K., Portillo, S. K., Slepian, Z., Kahraman, S., Burkhart, B., & Finkbeiner, D. P. (2021). Classification of magnetohydrodynamic simulations using wavelet scattering transforms. *The Astrophysical Journal*, *910*(2), 122 (cit. on pp. 35, 37, 61, 72, 88, 100).

Schekochihin, A. A. (2022). Mhd turbulence: A biased review. *Journal of Plasma Physics*, *88*(5), 155880501 (cit. on p. 100).

Schlafly, E. F., Green, G., Finkbeiner, D. P., Rix, H. .-., Bell, E. F., Burgett, W. S., Chambers, K. C., Draper, P. W., Hodapp, K. W., Kaiser, N., Magnier, E. A., Martin, N. F., Metcalfe, N., Price, P. A., & Tonry, J. L. (2014). A Large Catalog of Accurate Distances to Molecular Clouds from PS1 Photometry. *The Astrophysical Journal*, *786*(1), Article 29, 29 (cit. on p. 95).

Schmüdgen, K., et al. (2017). *The moment problem* (Vol. 9). Springer. (Cit. on p. 44).

Schneider, N., André, P., Könyves, V., Bontemps, S., Motte, F., Federrath, C., Ward-Thompson, D., Arzoumanian, D., Benedettini, M., Bressert, E., Didelon, P., Di Francesco, J., Griffin, M., Hennemann, M., Hill, T., Palmeirim, P., Pezzuto, S., Peretto, N., Roy, A., ... White, G. (2013). What Determines the Density Structure of Molecular Clouds? A Case Study of Orion B with Herschel. *The Astrophysical journal letters*, *766*(2), Article L17, L17 (cit. on p. 91).

Schneider, N., Ossenkopf-Okada, V., Clarke, S., Klessen, R., Kabanovic, S., Veltchev, T., Bontemps, S., Dib, S., Csengeri, T., Federrath, C., et al. (2022). Understanding star formation in molecular clouds-iv. column density pdfs from quiescent to massive molecular clouds. *Astronomy & Astrophysics*, *666*, A165 (cit. on p. 87).

Scoville, N. Z., & Good, J. C. (1989). The far-infrared luminosity of molecular clouds in the galaxy. *The Astrophysical Journal*, *339*, 149 (cit. on p. 2).

Seljak, U., & Zaldarriaga, M. (1997). Signature of gravity waves in the polarization of the microwave background. *Physical Review Letters*, *78*(11), 2054 (cit. on p. 5).

Srivastava, M. S., & Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, *99*(3), 386–402 (cit. on pp. 97, 121).

Stoica, P., & Selen, Y. (2004). Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine*, *21*(4), 36–47 (cit. on p. 73).

Teyssier, R. (2002). Cosmological hydrodynamics with adaptive mesh refinement. A new high resolution code called RAMSES. *Astronomy & Astrophysics*, *385*, 337–364 (cit. on p. 117).

Thuerey, N., Holl, P., Mueller, M., Schnell, P., Trost, F., & Um, K. (2021). *Physics-based deep learning*. WWW. (Cit. on p. 42).

Tsinober, A. (2009). *An informal conceptual introduction to turbulence*. Springer. (Cit. on p. 25).

Vacher, L., Aumont, J., Boulanger, F., Montier, L., Guillet, V., Ritacco, A., & Chluba, J. (2023). Frequency dependence of the thermal dust e/b ratio and eb correlation: Insights from the spin-moment expansion. *Astronomy & Astrophysics*, *672*, A146 (cit. on p. 5).

Vallée, J. P. (2008). New velocimetry and revised cartography of the spiral arms in the milky way—a consistent symbiosis. *The Astronomical Journal*, *135*(4), 1301 (cit. on p. 2).

Vazquez-Semadeni, E. (1994). Hierarchical structure in nearly pressureless flows as a consequence of self-similar statistics. *Astrophysical Journal v. 423, p. 681*, *423*, 681 (cit. on p. 53).

Vázquez-Semadeni, E., Ballesteros-Paredes, J., & Rodríguez, L. F. (1997). A search for larson-type relations in numerical simulations of the ism: Evidence for nonconstant column densities. *The Astrophysical Journal*, *474*(1), 292 (cit. on p. 87).

Ward-Thompson, D., & Whitworth, A. P. (2015). *An Introduction to Star Formation*. (Cit. on p. 19).

Wiener, N. (1939). The use of statistical theory in the study of turbulence. *Proc. Fifth Int. Cong. Appl. Mech*, *356* (cit. on p. 22).

Yan, Q.-Z., Zhang, B., Xu, Y., Guo, S., Macquart, J.-P., Tang, Z.-H., & Walsh, A. J. (2019). Distances to molecular clouds at high galactic latitudes based on Gaia DR2. *Astronomy & Astrophysics*, *624*, Article A6, A6 (cit. on p. 95).

Yoshimatsu, K., Schneider, K., Okamoto, N., Kawahara, Y., & Farge, M. (2011). Intermittency and geometrical statistics of three-dimensional homogeneous magnetohydrodynamic turbulence: A wavelet viewpoint. *Physics of Plasmas*, *18*(9) (cit. on p. 58).

Ysard, N., Jones, A. P., Guillet, V., Demyk, K., Decleir, M., Verstraete, L., Choubani, I., Miville-Deschênes, M.-A., & Fanciullo, L. (2024). Themis 2.0: A self-consistent model for dust extinction, emission, and polarisation. *Astronomy & Astrophysics*, *684*, A34 (cit. on p. 19).

Zamir, R. (1998). A proof of the fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory*, *44*(3), 1246–1250 (cit. on p. 72).

Zavagno, A., Dupé, F.-X., Bensaid, S., Schisano, E., Causi, G. L., Gray, M., Molinari, S., Elia, D., Lambert, J.-C., Brescia, M., et al. (2023). Supervised machine learning on galactic filaments-revealing the filamentary structure of the galactic interstellar medium. *Astronomy & Astrophysics*, *669*, A120 (cit. on p. 88).

Zhu, S. C., Wu, Y., & Mumford, D. (1998). Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, *27*, 107–126 (cit. on p. 116).

Zucker, C., Speagle, J. S., Schlafly, E. F., Green, G. M., Finkbeiner, D. P., Goodman, A., & Alves, J. (2020). A compendium of distances to molecular clouds in the Star Formation Handbook. *Astronomy & Astrophysics*, *633*, Article A51, A51 (cit. on p. 95).

RÉSUMÉ

Ce travail cherche à caractériser les structures multi-échelles et non Gaussiennes omniprésentes au sein du milieu interstellaire turbulent (MIS), en mettant néanmoins l'emphase sur le caractère limité du nombre d'observations dont on dispose. Les approches traditionnelles de ce problème, fondées sur des statistiques récapitulatives, n'exploitent pas pleinement la richesse informationelle encodée dans les structures du MIS, tandis que les techniques récentes d'apprentissage automatique, bien qu'extrêmement puissantes, nécessitent d'être entraînées sur des jeux de données adéquats. Si les simulations numériques constituent un outil attrayant pour fournir de tels jeux, la nature multi-physique et complexe de la dynamique du MIS rend sa reproduction *in silico* extrêmement difficile. Ainsi, la modélisation des propriétés non Gaussiennes du MIS ne peut reposer sur des modèles intégralement fondés sur les simulations et doit intégrer une étape d'apprentissage centrée sur les observations. Cependant, cela pose la difficulté de travailler sans supervision et avec une quantité très limitée de données d'entraînement, ce qui favorise fortement l'utilisation de descriptions compressées et de basse variance. Néanmoins, cette étude montre qu'en exploitant les symétries et régularités des processus physiques au travers de statistiques non expansives comme la *scattering transform*, il est possible d'explorer les propriétés non Gaussiennes du MIS à partir d'observations. En particulier, nous complétons le diagnostic standard, fondé sur les statistiques à un point, de l'évolution des nuages moléculaires du stade quiescent au stade actif de formation d'étoiles, par un diagnostic morphologique caractérisant le couplage entre échelles. En outre, nous développons une méthodologie qui permet de comparer, sans supervision, le pouvoir informatif de plusieurs statistiques récapitulatives dans un sens que nous définissons. Cela nous permet d'établir, à partir des observations, une distance morphologique entre les cartes de densité de colonne des nuages moléculaires, reposant sur une description non Gaussienne compacte. Les résultats de ce travail ouvrent de nouvelles perspectives sur le rôle que peuvent jouer les observations dans la caractérisation des structures non Gaussiennes du MIS.

MOTS CLÉS

Milieu interstellaire, turbulence, nuages moléculaires, structures non Gaussiennes, géométrie de l'information

ABSTRACT

This work addresses the challenge of characterizing the multi-scale, non-Gaussian structures in the turbulent Interstellar Medium (ISM), especially with limited observational data. Traditional approaches of this problem, based on summary statistics, do not fully exploit the rich information encoded in ISM structures, while recent machine learning techniques, though extremely powerful, require to be trained on appropriate datasets. If numerical simulations constitute an appealing tool to provide such required datasets, the complex multi-physics nature of the ISM makes it extremely challenging to be reproduced *in silico*. Thus, modeling the non-Gaussian properties of the ISM cannot rely solely on simulation-based approaches and must incorporate a learning step grounded in observations. However, this brings the difficulty to work without supervision and with a very limited amount of training data, which strongly favors the use of low variance and compressed descriptions. Still, this study demonstrates that, by leveraging physical symmetries and nonexpansive statistics like the scattering transform, it is possible to explore, from observations, non-Gaussian properties of the ISM. Specifically, we complement the standard one-point based diagnostic of molecular clouds' evolution from quiescent to active star-forming stages with a morphological diagnostic based on scale coupling. Additionally, we develop a methodology that allows to compare, without supervision, the informative power of multiple summary statistics in a sense that we define. This allows us to tailor, from observations, a morphological distance between column density maps of molecular clouds, based on a compressed non-Gaussian description. The results of this work open new perspectives for the role of observations in the characterization of non-Gaussian structures of the ISM.

KEYWORDS

Interstellar medium, turbulence, molecular clouds, non-Gaussian structures, information geometry